# The Effects of Big Data on Commercial Banks

Feb, 2024

**Abstract**

Drawing on a quasi-experiment and a structural model of loan demand and default, this paper analyzes the effects of providing an extensive amount of firm information, a proxy of making big data available, on commercial banks. Big data enables banks with high information technology capacity to improve screening ability and reallocate supply to high-quality borrowers; it also increases demand from all types of borrowers by improving the convenience of the loan-origination process. As a result, bank profitability increases sizably, and the effects are nearly only concentrated to those with high technology capacity that can utilize the data effectively.

**Keywords**: Information Technology, Small Business Lending, Hard Information, Big Data.
**JEL codes**: G21, O12, O38, L13, L25

# I Introduction

Over the last decade, businesses have increasingly turned to vast quantities of data to inform their decision-making processes. A Forbes report from 2018 (Marr, 2018) highlighted that every day, 2.5 quintillion bytes of data are generated. Often, this amount of data is too overwhelming for manual analysis. However, advancements in data storage and processing technologies have enabled business leaders to utilize *big data* to uncover patterns in customer behavior that are not immediately apparent to humans. As a result, data has emerged as a critical asset for driving business growth.

In the banking sector, the use of big data is especially crucial due to its reliance on data analysis for many of its operations. Yet, there is a noticeable gap in research on how big data affects banks' lending activities, primarily due to data availability issues and limited identification opportunities. Although "big data" lacks a universally accepted definition, it typically refers to datasets that are too large or complex for traditional data processing methods. Existing research has examined the impact of data diversity on lending decisions (Berg et al., 2019; Di Maggio et al., 2022), but the specific effects of increasing data volume, while keeping data diversity constant, have been less explored.

This paper aims to address this research gap by examining a quasi-experiment in China, offering initial insights into how an increase in the volume of firm data affects commercial banks. In 2014, Chinese local government agencies began sharing administrative data with commercial banks to enhance lending efficiency[1]. Privacy concerns led many agencies to collaborate with third-party financial service providers, who manage data storage and organization for the banks, while also handling data security legally. The focus of this study is on the earliest and largest of these third-party data providers, which started its operations in 2014 and has since maintained a market share of over 90% during the study period. Before the implementation of data sharing, banks typically required firms who decided to borrow from the banks to submit auditing reports for loan applications, which included various subsets of firm-related information [2]. With the advent of data sharing, data providers would first help government agencies aggregate

[1]Section A in the Online Appendix gives an outline of the policy guidance.

[2]Depending on lenders' requirement, information could include any of balance sheets, financial statements, tax histories, ownership structures, and credit and legal histories of firms and their directors.

all firms' information and then share the data altogether with banks. On average, each bank is shared with the information of more than 200 thousand firms with hundreds of variables at the initial provision of the data, and this information is continuously updated.

This study uses the data provider's market entry as a quasi-experiment to investigate the impact of increased data volume on loan characteristics. Starting in 2017, the data provider began supplying data to an arguably random group of banks in each province, maintaining these partnerships for two years. The analysis reveals that banks with access to the data (treatment) issued loans that were, on average, 2% larger in volume and featured a 33 basis point higher interest rate and a 22 basis point lower default rate compared to banks without access to the data (control). The finding of a reduced default rate for the treatment banks suggests that access to more extensive data enhances banks' ability to screen borrowers effectively. Further analysis using banks' proprietary credit scores supports this, showing that treated banks could better predict borrower defaults, as evidenced by a significant increase in the predictive power of their credit scoring models.

Access to a larger dataset enhances banks' ability to perform more accurate statistical analysis, leading to a reduction in information asymmetry and, consequently, a decrease in the default rate. Naturally, one might expect that lower default risks should result in lower average interest rates. However, the observation of higher interest rates among banks with access to the data indicates that improved screening is not the sole impact of big data on the loan market. The study reveals that following the data-sharing initiative, treated banks were able to offer loans more quickly and conveniently, primarily through providing more borrowers with online applications. With access to online applications, the loan processing time can be reduced from an average of 14 days to just about two hours, likely boosting demand among borrowers who prioritize speed (Buchak et al., 2018; Fuster et al., 2019; Wiersch et al., 2019), and, as a result, pushing up interest rates.

This demand-side effect is evident in the pricing strategy: the results reveal that treated banks adjusted their interest rates post-data sharing. First, changes in interest rates move with changes in banks' proprietary credit scores. Those with lower proprietary credit scores have higher interest rates and those with higher proprietary credit scores have lower interest rates. Notably, for loans processed more quickly, interest rate hikes were more pronounced for riskier borrowers, while reductions were less significant for safer ones.

This trend indicates that the acceleration in loan processing affected interest rates across all borrower categories.

Big data enables lenders to extract more high-dimensional information through systematic statistical inferencing. However, to use the data efficiently, banks need to have advanced technology stock. Therefore, the availability of a larger amount of data due to the data-sharing events is expected to have more significant effects on banks with high information technology (IT) capacity. Based on this conjecture, I continue to examine the impact of the data-sharing event on banks differentiated by their pre-existing technological capabilities, specifically their investment in information technology (IT). Banks were categorized based on their IT spending relative to total non-interest expenses before the data-sharing event. The findings suggest that banks with higher IT capabilities experienced more significant benefits from data-sharing, including improved risk assessment and more substantial changes in loan characteristics like interest rates, processing times, and default rates.

Heterogeneous screening ability by IT intensity suggests that treated banks, especially those with high IT intensity, could engage more in risk-based pricing. Through decreasing interest rates for previously unidentifiable low-risk borrowers and increasing rates for those with high risks, high-IT banks are expected to cream-skim high-type borrowers from low-IT banks. Consistently, I find that more high-quality borrowers started to form relationships with high IT-intensity banks as compared with low IT-intensity banks. On the other hand, more low-quality borrowers start to borrow from low IT-intensity banks as compared with high IT-intensity banks. The findings suggest that increases in the size of bank databases enable banks with higher IT intensity to cream-skim high-quality borrowers from low IT-intensity borrowers.

A limitation of the quasi-experiment is that only a subset of the banks are affected. As a result, it does not allow for the study of the equilibrium results of when all banks are affected. In addition, if focusing on all borrowers, the control group will also be affected by the event because the extensive margin adjustments affect borrower compositions. To cope with this, in the reduce-form analysis, I control for bank×firm fixed effects to hold bank-firm match constant, but this strategy disables the study at the extensive margin effects. Given these concerns, I then develop a structural model of loan application and

default that builds on Crawford et al. (2018) to study the equilibrium effects of the data-sharing event when all banks are provided the data and also incorporate the changes in borrower composition

In this model, borrowers are presumed to have a preference for online loan applications, and it's posited that the marginal costs associated with initiating loans are influenced by the type of application and the bank's screening capability. The model incorporates the impact of data-sharing by suggesting it alters the likelihood of banks offering online applications and enhances their screening abilities, with these effects being modulated by the banks' IT capabilities. The findings indicate a strong preference among borrowers for the speed of online applications, which, interestingly, do not lead to a higher default risk. Moreover, data-sharing enables banks with advanced IT infrastructures to adjust interest rates more significantlyâ€"increasing for high-risk borrowers and decreasing for low-risk onesâ€"highlighting two main mechanisms by which data-sharing influences the loan market: a screening-ability channel and a convenience channel.

The study explores the equilibrium effects of increasing access to concrete information on bank profitability through three counterfactual scenarios: 1. data sharing affecting both screening ability and convenience, 2. data sharing impacting only screening abilities, and 3. data sharing influencing only convenience. The outcomes illustrate that these mechanisms have distinct impacts on loan attributes. Specifically, when data sharing enhances screening ability alone, there's a notable reduction in default rates and a slight decrease in interest rates. Conversely, when it solely speeds up loan origination, default rates rise, and interest rates see a significant increase. However, when both factors are active, their effects on interest and default rates almost neutralize each other, yet improved risk-based pricing lowers marginal costs by approximately 8.22% and boosts markups by 18.82%.

While the average effects provide an overview of the market outcome, they obscure the varied impacts on banks with differing levels of information processing capabilities, namely IT capacity. Investigating the equilibrium effects of data sharing segmented by IT intensity reveals that banks with high IT capabilities experience more pronounced benefits, such as increased interest rates, reduced default rates, and significantly decreased

marginal costs, leading to a 25.54% rise in markups. In contrast, banks with lower IT capabilities see minor changes in interest rates and default rates, with a modest reduction in marginal costs. In the end, the markups of low-IT banks only increase by 4.35%

This analysis underscores a strong synergy between technology and data availability, showing that an increase in data significantly reduces marginal costs for banks with high IT infrastructure. Despite this cost reduction, prices do not fall due to a surge in demand facilitated by the convenience of faster loan processing, ultimately enhancing the profitability of high-IT banks significantly more than their low-IT counterparts. This indicates that, in the context of constant information technology, the advent of big data is poised to widen the profitability gap within the banking sector.

**Related Literature** This paper mainly contributes to three strands of literature. First, it contributes to a growing literature on fintech and information technology in banking[3]. This study aligns with studies on how the emergence of fintech and IT is affecting the traditional banking sector[4]. On the theoretical side, Hauswald and Marquez (2003) and He et al. (2020) show that technological progress in the banking sector could worsen the problem of the winner's curse, thereby increasing the average interest rate in the whole credit market. With structural estimation, Babina et al. (2020) shows that customer-directed data sharing increases entry by improving entrant screening ability and product offerings but harms some customers and can reduce ex-ante information production. This paper adds to this literature by providing a first set of empirical evidence on the heterogeneous effects of big data on loan attributes and lender activities. In addition, while the existing studies focus on adopting new technology or new types of information, this study analyzes the context where only the amount of data increases extensively but not the technology. In this case, I can dissect the interactive effects of data and information technology in affecting bank profitability by keeping one factor unchanged in the short run.

[3]Examples include Athreya et al. (2012), Livshits et al. (2016), Drozd and Serrano-Padial (2017), Jagtiani and Lemieux (2017), Buchak et al. (2018), Fuster et al. (2019), Berg et al. (2019), Frost et al. (2019), Hughes et al. (2019), Stulz (2019) Tang (2019), Di Maggio and Yao (2020), Babina et al. (2022), He et al. (2022), Gopal and Schnabl (2022), and Liu et al. (2022), etc. See Vives (2019) and Berg et al. (2021) for a review in banking.

[4]See Lorente et al. (2018), Hornuf et al. (2018), Calebe de Roure and Thakor (2019), Erel and Liebersohn (2020), and Aiello et al. (2020) for some examples.

This paper also relates to the recent literature on the implication of data ownership rights on market competition and welfare. The effects documented in the previous literature are usually ambiguous depending on how the data is used. For example, Farboodi et al. (2019) show that customer-generated data is valuable in forecasting business conditions. With structural estimation, Babina et al. (2020) show that customer-directed data sharing increases entry by improving entrant screening ability and product offerings, but harms some customers and can reduce ex-ante information production. He et al. (2022) and Parlour et al. (2022) emphasize that data sharing can increase the quality of lending but have ambiguous effects on consumer welfare and bank profits. In this paper, combining a quasi-experiment with structural estimation, I show that voluntary data sharing could simultaneously increase interest rates and decrease default rates. With detailed loan attributes, I can assess the relative importance of improved screening ability and improved convenience in determining the findings on interest rates and default rates.

Lastly, the structural estimation in this paper connects to the literature that employs structural techniques to quantitatively study the industrial organization of the financial markets. Recent literature has studied the retail deposits markets (Egan et al., 2017; Xiao, 2019; Egan et al., 2021), credit cards (Cuesta and Sepulveda, 2021; Nelson, 2022), mortgages (Buchak et al., 2018; Benetton, 2021; Guiso et al., 2022), and corporate loan Crawford et al. (2018); Ioannidou et al. (2022). This paper contributes to this literature by drawing from a quasi-experiment to quantitatively dissect the relative importance of screening ability and convenience through which financial technology and data-sharing affect interest rates and default rates.

# II Background

## *A.* Small Business Loan Market in China

In the early 2010s, small business credit origination in China primarily adhered to traditional relationship lending practices. Typically, small businesses established connections with loan officers at local bank branches, a practice that often included opening a business checking account for managing daily cash flows. For high-quality

businesses, this relationship extended further, with bank loan officers making visits to the company's headquarters to strengthen ties and gather soft information about the firm's quality, even when no borrowing occurred.

When seeking loans, companies usually approached the banks with which they had established relationships. The process involved visiting the bank branch and applying for a loan with the assistance of a loan officer. These officers would then request an auditing report, including a subset of information on balance sheets, financial statements, tax histories, ownership structures, and credit and legal histories of firms and their directors, from a third-party auditing firm. This auditing firm, in turn, would collect the necessary records from various government agencies with the company's authorization. Banks might request additional information as needed during this process. Once collected, this information, along with a credit score from banks' risk management department and a report summarizing any soft information gathered by the loan officer, would be used to finalize the loan terms offered to the company. This traditional loan origination process generally spanned approximately 14 calendar days.

In contrast, starting in 2012, many banks began offering an alternative through fast online applications. This modern approach allowed banks to directly gather firm information from government agencies, with the borrower's consent, to consolidate data from their records temporarily. Borrowers would then be promptly informed about the loan's approval status and, if approved, the terms of the loan. This streamlined process from starting the applications to receiving the funding could be completed in less than two days. However, due to various factors, including concerns over asymmetric information, a relatively small percentage of loans, typically less than 10% for most banks, were processed through these online applications.

### *B.* Data Sharing Policy

Since 2014, the local government agencies of many provinces in China have experimented with sharing administrative data with commercial banks. The policy aims to reduce the cumbersomeness of the auditing process and help banks reduce asymmetric information. For privacy concerns, many agencies contracted with third-party financial service companies to connect the banks and the governments. These companies have been helping

Figure 1. Types of Data Shared

This figure gives a list of the variables shared with the banks. The left panel is a screenshot of the provider's publicity material. The right panel is the English translation.

| 数据 | 数据内容 |
|---|---|
| 税务数据 | 1、税务登记信息<br>2、投资方信息<br>3、税务变更信息<br>4、申报信息<br>5、征收信息<br>6、利润表信息<br>7、资产负债表信息<br>8、供应商和客户信息<br>9、违法违章信息<br>10、稽查信息 |
| 工商数据 | 1、工商注册信息<br>2、股东信息<br>3、实际控制人信息<br>4、工商变更信息<br>5、管理层信息 |
| 司法数据 | 1、被执行人信息<br>2、法律诉讼信息 |
| 黑名单 | 1、银监会黑名单<br>2、小额贷款黑名单<br>3、P2P黑名单 |
| 反欺诈 | 1、反欺诈信息 |
| 征信数据 | 1、个人征信<br>2、企业征信 |

| Data | Data Content |
|---|---|
| **Tax Data** | 1. Tax Registration Information<br>2. Investors Information<br>3. Changes in Tax Category<br>4. Declaration Information<br>5. Taxation Administration Information<br>6. Cash Flow Statement<br>7. Balance Sheet<br>8. Information on Supplier and Customers<br>9 Law-Violation Information<br>10. Auditing and Inspection History |
| **Commercial Data** | 1. Business Registration Information<br>2. Share Holder Information<br>3. Information on Actual Controlling Shareholders<br>4. Changes in Business Registration<br>5. Information on Management Teams |
| **Judicial Data** | 1. Information on the Persons subject to Execution<br>2. Legal Action Information |
| **Blacklisting** | 1. CBRC Blacklisting<br>2. Petty Loan Blacklisting<br>3. P2P Blacklisting |
| **Anti-Fraud** | 1. Anti-Fraud Information |
| **Credit Registry Data** | 1. Individual Credit History<br>2. Business Credit History |

commercial banks store the data and claim legal responsibility for data security concerns. In this project, I use information from the earliest and largest third-party financial service company to examine how banks' lending decisions changed before and after the company stepped in. This provider's business started in 2014 and has had a market share of over 90% in the country throughout my sampling period from 2014 to 2018.

The data-sharing process takes two steps. First, a borrower must voluntarily participate in the program to allow government agencies to share their information. Specifically, the government agencies would first inform the firms about this program via different means of communication, including text messages, website notifications, WeChat official accounts, and in-person communication when the firms visit the agencies. The firms willing to participate in the program should then visit the agencies' websites to allow them to share the data. As official guidance from the central government, regional government agencies actively propagated this practice. Given the endeavor, government

agencies can receive permission from most of the registered corporations. In my sample, over 80% of all firms that have a record in the Credit Reference Registry of the People's Bank of China agreed to share their information prior to the time the provider had the initial sharing of the data. With firms' permission, government agencies aggregate their data to their local servers. After settling a contract with the provider, the agencies then share the data with the provider.

Second, the provider sets up a data interface with each partnered bank after settling a contract. The data was then stored on the provider's server, and the banks could retrieve the information from the interface using the intranet with the provider. At the same time, banks cannot approach the borrowers individually based on the data provided.

Figure 1 lists the types of variables shared with the banks altogether at once. It contains all information about firms' detailed balance sheet information, tax history, ownership structure, and firms' and the board of directors' credit history and history of legal activities. The shared information *does not* contain alternative data the banks could not get before the experiment, as all information is from government agencies and can be requested via auditing reports. Before the data-sharing, the banks could only request such information on a one-to-one basis when borrowers applied for a loan at these banks. After this event, with the borrower's permission, data is directly shared with all participating banks in bulk. Therefore, banks that the borrowers were not actively searching for could also get the data as long as they contract with the provider. Therefore, the impact of the event is the amount of information in the cross-section instead of new types of information. On average, each bank is shared with the information of more than 200,000 firms with more than one hundred variables for the initial provision of the data. Such information is then continuously updated. As for the amount of data shared, since more than 80% of all registered firms allowed sharing the data and the largest banks had a market share of around 10%, more than 85% of the shared data were from non-borrowers for all banks. The sudden increase in the data volume serves as a perfect laboratory to study the effects of the data revolution on lending activities.

The data-sharing program is similar to some previous studies of information sharing in the banking sector (Jappelli and Pagano, 2002; Liberti et al., 2022, 2019). However, the setting here provides a different channel by which more information changes banks'

lending decisions. For the setting here, a large amount of hard information, of which the specific *type* of information is previously known to the banks, is shared. That is, only data are shared, not technology. Usually, for other types of credit-registry expansion, both data and technology are shared to some extent. For example, in the US, banks can join PayNet to share their proprietary evaluation of their borrowers' riskiness with other members (Liberti et al., 2022). In addition, PayNet estimates and sells its proprietary credit scores using shared quantitative inputs. In this case, not only does PayNet increase the amount of information banks can access, but it also improves the technology to process the data for the banks that cannot utilize the data as efficiently as PayNet. In the case of Argentina (Liberti et al., 2019), banks also share their proprietary assessment of the borrowers to the credit registry. In both cases, the sharing of proprietary credit scores indirectly levels off discrepancies in information-processing abilities.

In addition, the data-sharing scheme in China is similar to a government-led open banking practice[5]. However, a difference is that, in open banking, customers choose to share their own *financial* data from their banks with all other banks or financial institutions. In the setting here, business owners choose to share information related to their economic activities from the governments instead of information only available from their financial accounts. In addition, information sharing relies on reciprocity for open banking and other credit registry expansion. That is, a bank can get information from other banks only if the other banks also join the credit registry. However, in the setting here, banks can retrieve data from all potential borrowers participating in the government-led program, regardless of whether these borrowers borrow from banks that are shared with the data.

# III Empirical Strategy

## *A.* Data

The loan-level data is a random 10% of that from the administrative agencies of a province where the loan information and the associated firm balance sheet information are available. The loan-level information includes the interest rate, maturity, loan volume,

[5]See Babina et al. (2022) for a discussion of open banking around the world.

loan application date, loan origination date, risk scores, the borrower's social identification number that uniquely identifies a firm in China, and a dummy indicating if the loan has defaulted. The associated firm-level information contains the borrowers' balance sheet information. The credit market in the province is relatively concentrated. Loan information is available from 22 registered commercial banks that contribute to over 90% of the total lending volume in this province[6].

### *B.* Identification

The sampling period consists of the early business years of the provider, or the "beta" stage as claimed by the provider. During this time, the provider had limited sales personnel and other resources to monitor a large number of banks about the data security issue. Because a sales team from the provider is usually in charge of one province, this effectively created a quota on the number of banks the provider could contract with within each market. As a result, the provider decided to only contract with a limited number of banks in each province. The provider first developed a potential partner list based on the banks' operating conditions. The list essentially excluded very small banks. After this step, the number of banks on the list is still larger than the provider's quota for most provinces. When deciding which banks on the potential partner list to contract with, the company informed the banks on the list about this opportunity at once through its sales department. Besides, during this beta stage, as guided by the government, the provider aimed to make sure that there was no data-security issue. Therefore, the provider had similar incentives to serve all registered banks. In the end, the provider contracted with the banks in a first-come-first-served manner until the predetermined number of partners is reached.

The firm's strategy gives the control and treatment groups to analyze the effects of enriched borrower information on commercial banks. Specifically, I first exclude banks that are not on the provider's potential partner list. I then use the banks that the provider contracted with as the treatment group and those that are on the potential partner list

[6]The concentration is similar to the small business loan markets in the US. From CRA, the total share of the top 16 banks in any state from 2011 to 2018 is on average 86% with an interquartile of 79% and 95%.

TABLE 1. Summary Statistics

This table gives the summary statistics of the loan-level data. Each panel except for Panel C gives the averages and associated standard deviations. In Panel C, the parentheses contain the $t$-statistics of the $t$-tests of differences in mean. log Volume is the log of the amount of each loan in 10-thousand CNY. Maturity is the loan maturity in months. Interest Rate is the interest rate (%) of the loan. Default is the percentage of the defaulted loan per year. log AT is the average of the borrowers' log total asset measure in 10-thousand CNY. Profit is the net profit over total assets (%). Leverage is the total debt outstanding over total assets. Origination time is the days it takes to originate the loan. The averages are weighted by loan volume. Response time is the time is in minutes it takes for the banks to respond to the contracting message sent by the provider. All variables are winsorized at 1% level by year-quarter. The sampling period is 2015Q2 to 2017Q1.

| log Volume | Maturity | Interest Rate | Defaulted | log AT | Profit | Leverage | Origination Time | Response Time | Nobs |
|---|---|---|---|---|---|---|---|---|---|
| Panel A: Treatment | | | | | | | | | |
| 5.18 | 27.08 | 6.83 | 0.08 | 7.51 | 0.06 | 0.48 | 13.32 | 12.35 | 174,173 |
| (1.08) | (6.91) | (1.47) | (0.27) | (1.22) | (1.69) | (0.41) | (21.33) | | |
| Panel B: Control | | | | | | | | | |
| 5.19 | 27.24 | 6.92 | 0.07 | 7.48 | 0.08 | 0.47 | 13.91 | 34.87 | 98,180 |
| (1.10) | (7.29) | (1.61) | (0.26) | (1.20) | (1.82) | (0.81) | (25.83) | | |
| Panel C: Difference in Mean | | | | | | | | | |
| 0.01 | 0.16 | 0.09 | -0.01 | -0.03 | 0.02 | -0.01 | 0.59 | | |
| (0.05) | (0.76) | (1.01) | (0.05) | (0.45) | (1.36) | (0.05) | (0.32) | | |

but not contracted with as the control group[7]. In the end, I have 272,353 observations from 2015Q2 to 2019Q2 from 12 of the 22 banks.

To explore the randomness of the treatment group assignment, I compare whether there is any statistical difference in the banks' observable information between the control and treatment groups before the experiment at the loan, firm, and bank levels. Table 1 compares the loan-level and firm-level variables between the treated and control banks before the experiment. Panels A and B give the averages and standard deviations of the variables for the treated and control banks. Averages are weighted by loan amount. Panel C gives the statistical significance of the differences using $t$-tests. The parentheses in Panel A and B contain the associated standard deviation, and the parentheses in Panel C contain the associated $t$-statistics. All variables are winsorized at 1% level within each year. From Panel B, before the experiment, the loans have an average volume of around 9.92 million CNY, with an average interest rate of 5.57% and a maturity of 26 months. In

[7]Note that the first-respond-first-serve strategy naturally gives rise to a regression discontinuity (RD) design. However, in the main analysis, the controlling of year-quarter and bank-firm fixed effects nests a sharp RD specification.

addition, 3.3% of the loans defaulted in the end. The average size of the firms is around 300 million CNY. The average profitability (gross profit over total assets) of the firms is 8.76%. The average leverage, calculated as total debt over total assets, is 0.37. The origination time gives the average days to originate the loans. This is around two weeks for both groups. From Panel C, the averages between the control and treatment groups were extremely close to each other. This further confirms the randomness of the firm's contracting strategy.

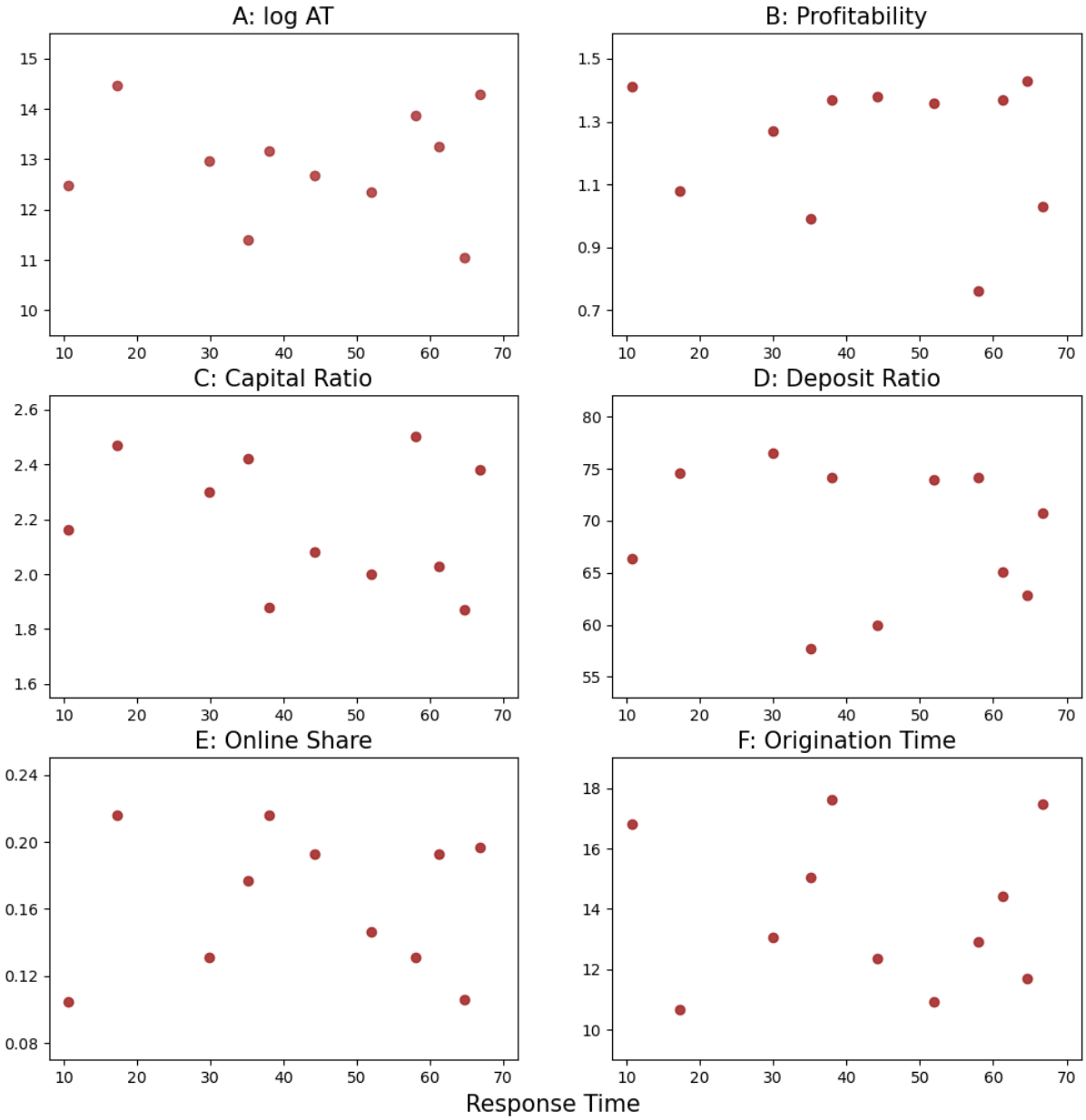
I continue to check if the time the banks take to reply systematically varies with their fundamentals. If the banks that would benefit the most from this program reply earlier, then the estimated average treatment effects (ATE) are confounded with selection biases. From Table 1, the average response times between the control and treatment groups are very close. Response time is the time for the banks to respond to the provider about getting access to the data in minutes. The average response time for the control group is 34.87 minutes, compared with 12.35 minutes for those in the treatment group. In addition, all banks on the list responded within 70 minutes. In addition, Figure 2 gives the scatter plots of bank characteristics with respect to response time. The plot shows that the response time is not correlated with banks' ex-ante size, profitability, capital ratio, or deposit ratio, share of online application, and origination time. Therefore, the heterogeneity in the response time is unlikely a result of selective participation.

A concern of the identification strategy is that the provider would charge a higher fee to the banks that would benefit more from the data. This is not a concern here as the fees charged during the sampling period are uniform to all banks and are very small. The fee only accounts for less than 1% of the total expenses for all banks. This is because, during this period, the provider is guided by government agencies to operate to ascertain the effectiveness of the policy instead of maximizing profits.

Finally, even if the treatment assignment is as good as random, directly comparing the loan outcome between the control group and treatment group suffers from a selection bias because of a change in the borrower composition. That is, after the policy changes, borrowers from the control group have incentives to change their lending relationships and decide to borrow from the treatment group. In this case, the control group will also be affected by the treatment. To avoid this issue, in all analyses, I only focus on

Figure 2. Bank Characteristics and Response Time

This figure assesses if response time is correlated with bank characteristics. Panel A gives the log of bank total assets. Panel B is the gross profitability (%), which equals to the ratio of gross profit and total assets. Panel C is the capital ratio (%) which is the ratio of total equity over total assets. Panel D is the deposit ratio (%) that is equal to the ratio of total deposits over total assets. Panel E is the share of loans originated through online applications. Panel F is the time for loan origination in days. The x-axes are the response time in minutes.




the borrowers that have borrowed both from the treatment group and control group in both the pre-experiment and post-experiment periods. Doing so allows me to control for firm×bank fixed effects, and only focus on the intensive margin about how loan terms

change conditional on the same bank-firm pairs. This strategy excludes any confounding factors that affect borrower composition. However, an unappealing feature of this strategy is that it ignores the extensive-margin dynamics. In section V, I use a structural model with loan demand and default to study the extensive margin dynamics.

# IV Results

## *A.* Loan Attributes

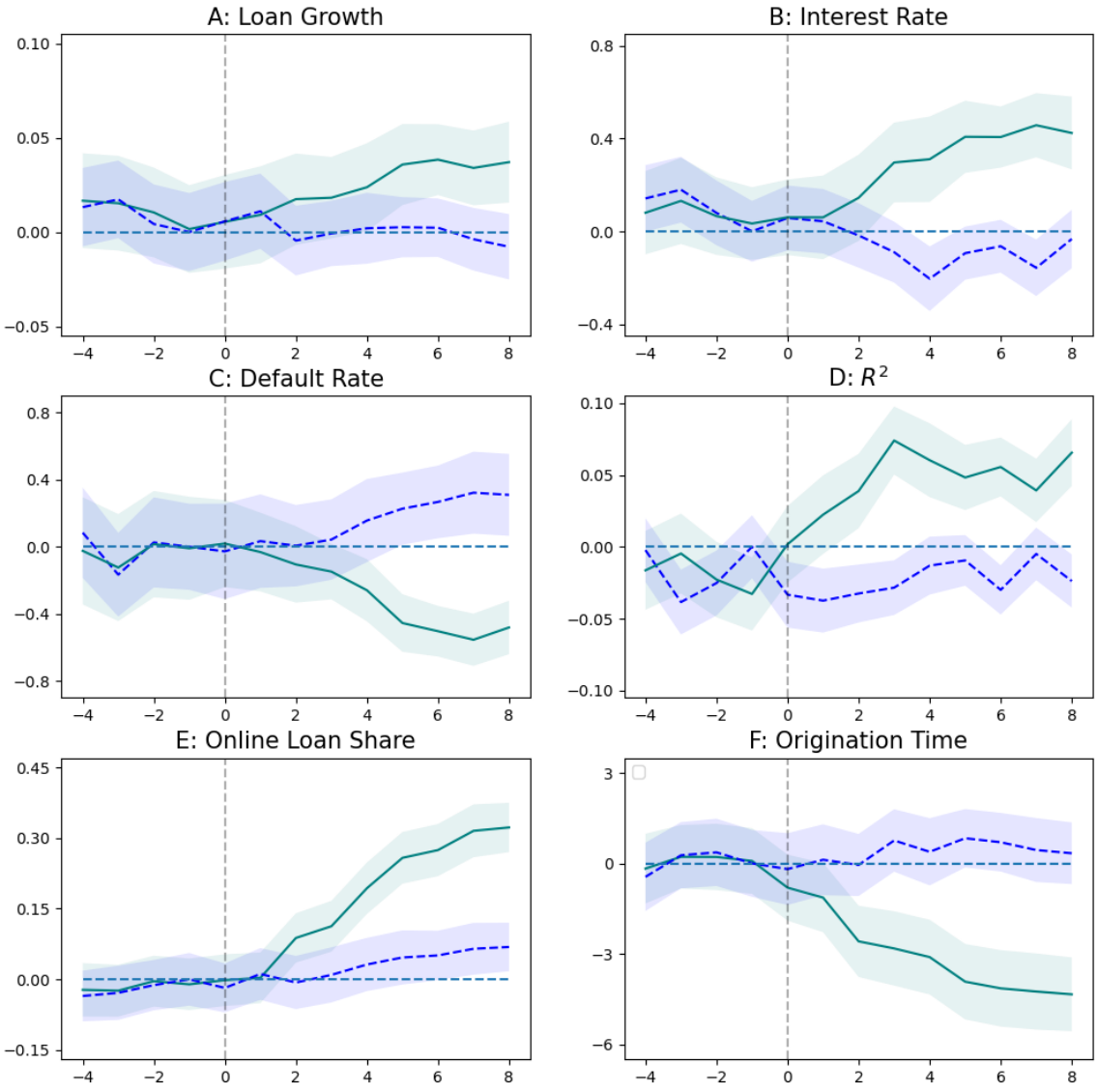
In this section, I study the ATE of the data-sharing policy. I start by looking at the effects on loan terms. Figure 3 plots the evolution of the various loan attributes between the control and treatment groups. Panels A, B, and C respectively give the evolution of loan growth, interest rates, and default rates. For each panel, the solid green line represents the treatment group, and the dashed blue line represents the control group. The $x$-axis is the number of quarters from the treatment quarter 2017Q2, labeled as time 0. For each panel, I vertically shift the plot by subtracting all values from the value of the control group at $t = -1$. Therefore, the $y$-axis is the change with respect to the control group at $t = -1$. Averages are weighted by the loan volume. The shaded region is the 95% confidence interval. The plot is based on borrowers that have borrowed both from the treatment group and control group in both the pre-experiment and post-experiment periods. All plots are residualized by firm-bank fixed effects and year-quarter fixed effects.

The figure shows some clear patterns. First, for all three variables, there is no distinguishable difference either in the pre-trends between the control and treatment groups. Second, there is a clear diverging pattern between the treatment and control groups after the data-sharing, indicating that the data-sharing event was not known previous to the shock. After the experiment, the loans from the treated banks see a weakly higher volume, a much higher interest rate, and a sizable decrease in default rate.

Table 2 gives the DID estimates. The odd columns and even columns respectively exclude and include the firm×bank fixed effects. All regressions are weighted by loan volumes. The results are qualitatively the same regardless of the fixed effects. Given controlling for bank-firm matches, the estimates in the even columns indicate that, after the data-sharing event, the loans from the treated banks to firms, as compared with

Figure 3. Changes in Loan Attributes

This figure gives the evolution of the loan attributes between the control and treatment groups. Panels A, B, and C, respectively, give the log loan volume, interest rate, and default rate. Panel D gives the screening ability, as measured by the pseudo-$R^2$ from predicting default using bank credit scores. Panel E gives the share of loans originated through online applications, and panel F gives the time for loan origination in days. For each panel, the green solid line captures the treatment group and the blue dashed line captures the control group. The $x$-axis is the number of quarters from the treatment quarter in 2017Q2. All values are subtracted by the value of the control group at $t = -1$. Averages are weighted by loan volume. The shaded region is the 95% confidence interval. For Panel D, the standard error is based on 500 bootstrap draws.




loans from the control banks to the same firms, have a 2% higher loan volume, 33 basis points higher interest rates, and 0.22 percentage points lower chances of defaulting. Quantitatively, the effects are quite different when the firm×bank fixed effects are not

TABLE 2. Loan Terms Outcomes

This table gives the heterogeneous treatment effects of the policy on the loan-level variables by bank IT intensity before the experiment. IT intensity is banks' average IT spending to total expenses before the experiment. log Volume is the log of the amount of each loan in 10-thousands CNY. log Time is the log loan origination time in days. Interest is the interest rate (%) of the loan. Default is an indicator that the loan is defaulted. Regressions are weighted by loan volume. The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | (1) log Volume | (2) log Volume | (3) Interest Rate | (4) Interest Rate | (5) Default | (6) Default |
|---|---|---|---|---|---|---|
| Treat | 0.04* | 0.02 | 0.22* | 0.33*** | -0.41*** | -0.22* |
| | (0.02) | (0.02) | (0.12) | (0.13) | (0.10) | (0.10) |
| Observations | 272,353 | 137,635 | 272,353 | 137,635 | 272,353 | 137,635 |
| R-squared | 0.053 | 0.197 | 0.042 | 0.244 | 0.047 | 0.162 |
| Year-Qtr FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm×Bank FE | No | Yes | No | Yes | No | Yes |

Standard Errors Clustered at Year-Quarter and Bank Level in Parentheses
*** $p$ <0.001, ** $p$ <0.01, * $p$ <0.05

included. Specifically, without conditional on the same bank-firm relationships, loans from the treated banks after the event have larger volumes, lower interest rates, and lower default rates. This is consistent with the mechanism that, at the extensive margin, the event enables treated banks to extend loans to higher-quality borrowers who borrow from the control banks before the experiment.

### *B.* Screening Ability

The results on loan terms indicate that borrowers from the treatment group are less likely to default after the experiment. To explore the reason for lower default rates, in this section, I study the effects of data-sharing on bank screening ability. Following Iyer et al. (2016), I study bank screening ability with a logistic regression that predicts borrowers' ex-post default decisions using the banks' standardized ex-ante proprietary risk score. I measure the screening ability by two statistics associated with the logistic regression: 1) the pseudo-$R^2$ and 2) the area under a receiver operating characteristic (ROC) curve. An ROC curve is a plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. As suggested by Iyer et al.

TABLE 3. Risk Score and Screening Performance

This table gives the predictive performance of banks' proprietary risk score (Score) separately for the control and treatment groups and before and after the experiment. Risk score is standardized by each bank. The analysis focuses on the borrowers that have borrowed from both before and the experiment and both from a control bank and from a treated bank. The parentheses in columns (1) to (4) contain the standard errors. The $p$-value of the DID estimates in panel A is based on 500 Bootstrapping draws, and is residualized by firm-bank fixed effects. The DID estimate in Panel B gives the difference-in-difference estimates between the changes in the AUC of the treated group and that of the control group, for which the $p$-value is calculated based on DeLong ER (1988). The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | Control | | Treatment | | |
|---|---|---|---|---|---|
| | (1) Before | (2) After | (3) Before | (4) After | (5) DID |
| | Panel A: Logistic Regression | | | | |
| Score | 1.10 | 1.10 | 1.11 | 1.16 | |
| | (0.01) | (0.01) | (0.01) | (0.01) | |
| Pseudo $R^2$ | 13.11% | 13.04% | 14.01% | 18.55% | 4.29% |
| | | | | | $p$-value = 0.00 |
| | Panel B: ROC | | | | |
| AUC | 0.7531 | 0.7521 | 0.7565 | 0.8087 | 0.0532 |
| | (0.0121) | (0.0098) | (0.0089) | (0.0098) | $p$-value = 0.00 |
| N | 42,554 | 45,025 | 24,137 | 25,919 | |

(2016), the ROC curve is a technique that is commonplace in the commercial financial banking markets. The area under the ROC curve (AUC) provides a more interpretable estimate of inference than the pseudo-$R^2$. The larger this number is, the higher the predictive power. The largest value AUC can get is 1, which indicates perfect forecast accuracy. The AUC of a random predictor is 0.5. [8]

Panel D of Figure 3 gives the evolution of the pseudo-$R^2$ from predicting the default probability using bank risk scores. A higher pseudo-$R^2$ indicates that the proprietary risk scores have better predictability of ex-post default. From the plot, while the pseudo-$R^2$ from the control group stays nearly constant across the sampling period, that from the treatment group increases sizably after the event. Therefore, treated banks have a much better screening ability after the experiment.

I continue to assess the effects of the event on bank screening ability quantitatively. The results are in Table 3. Panel A gives the logistic regression results, and Panel B gives the associated AUC. Columns (1) and (2) give the results for the control group, and

[8] See Iyer et al. (2016) for a detailed explanation and motivation.

columns (3) and (4) give the results for the treatment group. Panel A first confirms the risk score's strong predictive power of future default. The pseudo-$R^2$ is around 13% for both control and treated banks before the experiment. After the experiment, the pseudo-$R^2$ of the treated banks increases from 14.01% to 18.55%, while that for the control banks decreases marginally from 13.11% to 13.04%. The difference-in-difference (DID) estimate, which is residualized by firm-bank fixed effects, gives the average treatment effects (ATE) of the experiment on banks' screening ability. I calculate the DID estimates and the associated standard error through 500 bootstrapping draws. The estimate of 4.29% is both statistically and economically significant.

The ROC curves provide a more formal way to compare the predictive power between the control and treated banks before and after the experiment. Panel B gives the associated AUC of the logistic regression. I find that the AUC is around 0.75 for both control and treated banks before the experiment. After the experiment, the AUC of the treated banks increases from 0.7565 to 0.8087, while that for the control banks nearly remains unchanged. Following Iyer et al. (2016), I calculate the change in the performance of the treated banks' risk scores by $(0.8087 - 0.5)/(0.7565 - 0.5) \approx 1.20$. This is to say, treated banks' risk scores achieve 20% greater accuracy after the experiment. The DID estimate indicates that the increase in the treated banks' screening ability, as measured by AUC, is statistically significant.

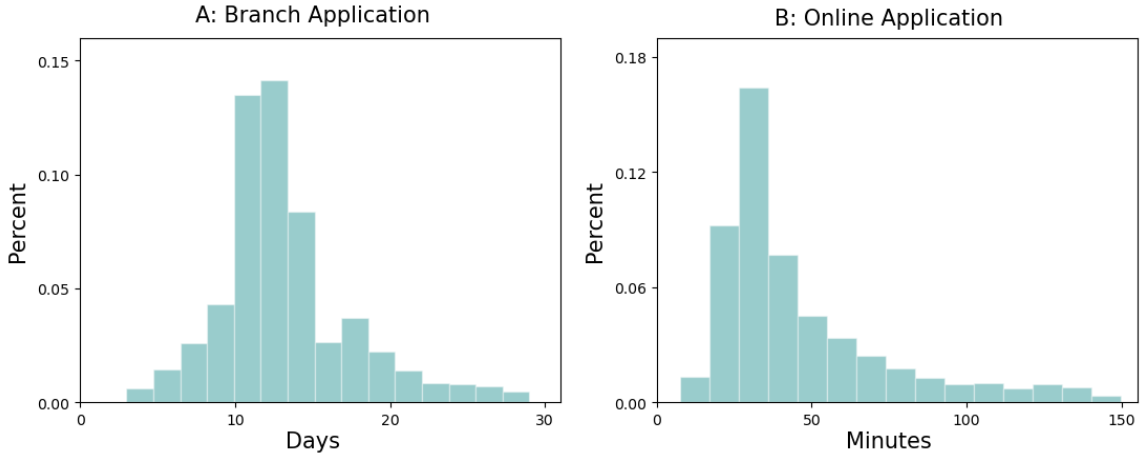
Since the results are conditional on the same bank-firm relationship, the results indicate that, after the data-sharing event, the treated banks' proprietary risk scores have much better predictability about borrowers' ex-post default probability than the control groups' to predict the default probability for the same borrowers. Therefore, the treatment group sees a much better screening ability than the control groups after the events, which implies lower realized default rates, as documented in the previous section.

### *C.* Convenience

The findings of an increase in the interest rate and a decrease in default rate are inconsistent with banks having better screening ability in a perfectly competitive market. Specifically, suppose banks break even on lending, and a better screening ability decreases the default rate, then interest rates should also decrease. On the other hand, in Athreya

Figure 4. Distribution of Origination Time

This figure plots the histogram of the loan origination time respectively for branch applications and online applications. Data is based on all banks in the pre-experiment period.




et al. (2012), Livshits et al. (2016), and Drozd and Serrano-Padial (2017), more advanced information technology reduces asymmetric information in the credit market. At the extensive margin, more contracts are offered to those previously denied borrowing. The entry of new lending contracts targeted at riskier borrowers gives rise to a higher default rate and interest rate. However, (Buchak et al., 2018; Fuster et al., 2019) show that fintech firms, apart from having differential screening abilities, could increase loan demand by offering a faster and more convenient loan origination process. In addition, in a recent survey by Wiersch et al. (2019) shows that the most frequently cited challenges with bank lenders were the application process and long wait times for credit decisions. Suppose borrowers value faster loan origination time. Then the data sharing could affect demand in addition to affecting screening ability. This will increase interest rates. In this section, I assess the event's effects on convenience by studying the outcomes on loan origination time and the proportion of online applications.

A key difference between branch applications and online applications is the time it takes to get the funding. Given a completely automatic loan origination process, online applications take a much shorter time to receive funding. Figure 4 plots the distribution of the time it takes to receive funding starting from the time of initiating the application. Panel A depicts branch applications and panel B plots online applications. As shown, in general, branch application usually takes two weeks to receive funding, and the process

could take as long as one month. In comparison, online applications mostly take less than three hours. This is a massive decrease in the time it takes to receive funding and is expected to improve the convenience of the origination process greatly[9].

However, the availability of online applications requires a better ability to use hard information at the cost of ignoring soft information. The availability of a large amount of hard information enables banks to spot hidden patterns in the cross-section through statistical analysis that are unable to be verified by humans. The ability of a finer recognition of borrower type increases with the amount of data, which reduces the standard errors in the inferencing process. The data-sharing event here increases the amount of hard data. The improved screening ability means that banks can supply more funds through online applications to reduce labor costs. Panels E and F respectively give the evolution of the share of online applications and the average loan origination time. Consistent with the conjecture that better screening ability enables banks to supply more loans online, treated banks have a much higher share of online applications, which results in a much shorter average origination time. Table 4 gives the ATE of the events on online application share and origination time. Condition on the firm-bank pairs, treated banks have 22% more loans originated through online applications. Accompanied by it, the treated banks take around 3.5 days less to extend the loans. Given an average of 13.5 days to extend the loan, this is equivalent to a 25% decrease.

To further assess how a faster loan origination process is accompanied by a higher interest rate, I separately study changes in interest rates for borrowers that, after the experiment, have a faster and slower loan origination time and lower and higher risk scores. The results are in Table 5. Regardless of the changes in origination time, the borrowers that are perceived as riskier (safer) by the banks have higher (lower) interest rates. This is consistent with treated banks increasing supply to high-quality borrowers. At the same time, for loans that have a faster origination time, increases in interest rates are larger for perceived riskier borrowers, but decreases in interest rates are smaller for

[9]The difference between traditional in-person loan granting style and the online style could be different from different countries. While it is difficult to know how long it takes to receive funding in the traditional practice, a useful comparison is to study the time it takes to get a loan from SBA and fintech platforms like LendingClub. The reports from here and here show that it usually takes more than a month to get loans from SBA, but a few hours from LendingClub from the US, which validates the significance about the convenience difference between traditional lending practice and the new online practice.

TABLE 4. Convenience Outcomes

This table gives the average treatment effects of the policy on shares of online applications and loan origination time in days. Regressions are weighted by loan volume. The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | (1) Online% | (2) Online% | (3) Times | (4) Times |
|---|---|---|---|---|
| Treat | 0.33** | 0.22** | -4.47** | -3.02*** |
| | (0.14) | (0.08) | (0.54) | (0.62) |
| Observations | 272,353 | 137,635 | 272,353 | 137,635 |
| R-squared | 0.051 | 0.211 | 0.037 | 0.169 |
| Year-Qtr FE | Yes | Yes | Yes | Yes |
| Firm×Bank FE | No | Yes | No | Yes |

Standard Errors Clustered at Year-Quarter and Bank Level in Parentheses
*** $p$ <0.001, ** $p$ <0.01, * $p$ <0.05

perceived safer borrowers. Specifically, faster origination comes with around 30 basis points higher interest rates than slower origination for both high-quality and low-quality borrowers. Therefore, the data-sharing event has two effects on the treated banks. First, it allows banks to reallocate funds to high-quality borrowers at lower interest rates. In addition, it enables the banks to supply funds faster, which increases demand from both high-quality and low-quality borrowers.

In sum, the results about screening ability and origination time indicate that, after acquiring a large amount of hard information, treated banks are able to have a better screening ability that enables banks to re-allocate credit supply to high-quality borrowers. At the same time, better screening ability enables treated banks to offer online loan applications, which improves the convenience of the loan origination process greatly. This is expected to increase loan demand for those who value a faster loan origination process.

### *D.* Heterogeneous Treatment Effects

The data-sharing event enables the treated banks to receive a large amount of hard data about firm information. As a characteristic of statistical inference over big data, the large volume often makes it impossible to process using traditional methods. Therefore, how effectively banks can exploit this great amount of information depends on the banks' information technology (IT) capacity. To test if banks with high IT spending can utilize

TABLE 5. Effects on Interest Rates by Quality and Convenience

This table gives the average treatment effects of the policy on interest rates. borrowers are split into four groups based on the changes in the loan origination time and changes in proprietary credit score. Regressions are weighted by loan volume. The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Slower Origination | | Faster Origination | |
| | Riskier | Safer | Riskier | Safer |
| Treat | 0.44* | -0.48*** | 0.78*** | -0.17* |
| | (0.20) | (0.16) | (0.19) | (0.10) |
| Observations | 10,867 | 29,381 | 31,046 | 45,623 |
| R-squared | 0.053 | 0.197 | 0.042 | 0.244 |
| Year-Qtr FE | Yes | Yes | Yes | Yes |
| Firm×Bank FE | Yes | Yes | Yes | Yes |

Standard Errors Clustered at Year-Quarter and Bank Level in Parentheses
*** $p$ <0.001, ** $p$ <0.01, * $p$ <0.05

big data more efficiently, I study the heterogeneous treatment effects of the experiment for banks with different levels of IT spending before the experiment. The data for IT spending at the bank level comes from a survey by the province's Banking and Insurance Regulatory Commission. I separate the banks into two groups based on their average IT intensity, which is the total IT spending over total non-interest expenses three years before the experiment, and study the heterogeneous treatment effects of the experiments for the two groups.

I first test if the banks with high ex-ante IT spending could use the shared data more effectively and, therefore, have a more accurate risk-scoring model. In Table 6, I study the changes in bank screening ability separately for those with high and low ex-ante IT intensity. Again, I focus on the borrowers that have borrowed at least once both before and after the experiment and both from the control and treatment groups to abstract from factors about borrower composition.

Columns (1) and (2) focus on the sample of banks with low IT spending. Columns (3) and (4) use the sample of banks with high IT spending. Column (5) gives the triple-difference (TD) estimates. Panel A shows that the screening ability hardly changes for the control group regardless of the ex-ante IT intensity. On the other hand, from Panel B, the screening ability increases greatly for those with high IT spending but only slightly

TABLE 6. Risk Score and Screening Performance by IT Intensity

This table gives the predictive performance of banks' proprietary risk score by bank IT-intensity group and before and after the experiment for banks in the control group only. IT-intensity group is split by the median of banks' IT spending to total expenses before the experiment. Risk score is standardized by each bank. Panel A focuses on the control group. Panel B focuses on the treatment group. Columns (1) and (2) present results for low IT-intensity banks. Columns (3) and (4) present results for high IT-intensity banks. The parentheses in columns (1) to (4) contain the standard errors. The $p$-value of the TD estimates in panels B1 is based on 500 Bootstrapping draws, and is residualized by firm-bank fixed effects. The $p$-values of the TD estimates in panel B2 is calculated based on DeLong ER (1988). The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | Low IT/Exp | | High IT/Exp | | |
|---|---|---|---|---|---|
| | (1) Before | (2) After | (3) Before | (4) After | (5) TD |
| | Panel A: Control | | | | |
| | Panel A1: Logistic Regression | | | | |
| Pseudo $R^2$ | 11.51% | 12.15% | 15.52% | 15.98% | |
| | Panel A2: ROC | | | | |
| AUC | 0.7314 (0.0108) | 0.7500 (0.0101) | 0.7587 (0.0098) | 0.7684 (0.0097) | |
| N | 18,036 | 19,585 | 24,518 | 25,440 | |
| | Panel B:Treatment | | | | |
| | Panel B1: Logistic Regression | | | | |
| Pseudo $R^2$ | 12.61% | 14.89% | 14.68% | 22.10% | 5.67% $p$-value = 0.00 |
| | Panel B2: ROC | | | | |
| AUC | 0.7218 (0.0219) | 0.7434 (0.0206) | 0.7649 (0.0214) | 0.8495 (0.0151) | 0.0674 $p$-value = 0.06 |
| N | 10,453 | 11,071 | 13,684 | 14,848 | |

for those with low IT spending. The pseudo-$R^2$ for the low-IT group in the treatment group is 12.61% before the experiment. After the experiment, the pseudo-$R^2$ increases to 14.89%. While for high-IT banks, the pseudo-$R^2$ increases from 14.68% to 22.10% after the experiment. Residualized by bank-firm fixed effects, the TD estimate for the changes in high-IT treated banks compared with the low-IT treated banks is 5.67% and is both statistically and economically significant.

Panel B confirms the results in Panel A using the ROC associated with the logistic regression. For the treated banks, the AUC increases from 0.7218 to 0.7434 for the low-IT banks. While for the high-IT banks, the AUC increases from 0.7649 to 0.8495. This increase is equivalent to a 31.94% improvement in the predictive accuracy, compared with a 9.74% higher predictive accuracy for the low-IT group.

TABLE 7. The Effects of the Event by IT Intensity

This table gives the heterogeneous treatment effects of the policy on the loan-level variables by bank IT intensity before the experiment. IT intensity is banks' average IT spending to total expenses before the experiment. log Volume is the log of the amount of each loan in 10-thousands CNY. log Time is the log loan origination time in days. Interest is the interest rate (%) of the loan. Default is an indicator that the loan is defaulted. Loan origination time is in days. Regressions are weighted by loan volume. The sampling period is from 2015Q2 to 2019Q2. All variables are winsorized at 1% level by year-quarter.

| | (1) log Volume | (2) log Volume | (3) Interest Rate | (4) Interest Rate | (5) Default | (6) Default | (7) Online% | (8) Online% | (9) Times | (10) Times |
|---|---|---|---|---|---|---|---|---|---|---|
| Treat | 0.02 | 0.01 | 0.11 | 0.06 | -0.09 | 0.17* | 0.16** | 0.05 | -0.29 | -0.09 |
| | (0.02) | (0.02) | (0.13) | (0.18) | (0.08) | (0.09) | (0.07) | (0.08) | (0.65) | (0.07) |
| Treat×High IT | 0.06*** | 0.03* | 0.23*** | 0.39*** | -0.53*** | -0.64*** | 0.23** | 0.26*** | -6.04*** | -4.68*** |
| | (0.02) | (0.02) | (0.09) | (0.10) | (0.09) | (0.09) | (0.08) | (0.08) | (0.97) | (0.09) |
| Observations | 272,353 | 137,635 | 272,353 | 137,635 | 272,353 | 137,635 | 272,353 | 137,635 | 272,353 | 137,635 |
| R-squared | 0.079 | 0.218 | 0.033 | 0.276 | 0.058 | 0.188 | 0.063 | 0.232 | 0.055 | 0.182 |
| Year-Qtr FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm×Bank FE | No | Yes | No | Yes | No | Yes | No | Yes | No | Yes |

Standard Errors Clustered at Year-Quarter and Bank Level in Parentheses
*** $p$ <0.001, ** $p$ <0.01, * $p$ <0.05

I continue to study the heterogeneous treatment effects of the data-sharing event on loan attributes by IT intensity. To do so, I fit the following DID specification:

$$Y_{j,k,t} = \lambda_{j,k} + \lambda_t + \beta_0 \times treat_{j,k,t} + \beta_1 \times treat_{j,k,t} \times HIT_k + \epsilon_{j,k,t}, \tag{1}$$

where $Y_{j,k,t}$ are various loan attributes. $\lambda_{j,k}$ is the firm-bank fixed effects, $\lambda_t$ is the year-quarter fixed effects, and $treat_{k,t} = 1$ if the bank $k$ has been shared with the data in year-quarter $t$. $HIT = 1$ if the bank's IT intensity is above the median. The inclusion of firm-bank fixed effects compares the effects of data-sharing within firm-bank pairs and abstracts from any impacts due to changes in borrower composition and bank-firm matching.

The results are in Table 7. Conditional on bank-firm pairs, after providing the data, banks with higher IT intensity see a 4% increase in average loan volume, 45 basis points increase in interest rate, 47 basis points lower default rates, 31% more loan origination from the online platforms, and 4.77 fewer days for loan origination. While in general, the effects of data-sharing have the same direction for low-IT banks, the effects are much smaller, with most of the effects being insignificant. Altogether, the results in tables 6 and 7 suggest that increasing the availability of hard information has a large positive impact

on bank screening ability, speed of originating the loan, and profitability. However, the effects are mainly concentrated in banks with high IT intensity.

### *E.* Cream-Skimming of High-IT Lenders

The effects of data-sharing are positive on screening ability, and that the effects are larger for high-IT banks suggests that treated banks, especially those with high IT intensity, could engage in better risk-based pricing. Through decreasing interest rates for previously unidentifiable low-risk borrowers and increasing rates for those with high risks, high-IT banks are expected to be able to cream-skim high-type borrowers from low-IT banks at the extensive margin.

To test this hypothesis, I continue to explore whether the experiment enables high-IT banks to attract more high-quality borrowers. Figure 5 plots two heat maps that show the flow of high-quality and low-quality borrowers after the experiment. I first split the borrowers into two groups by the sample median of their qualities. I define quality as one minus the predicted default rate using all banks' post-experiment proprietary credit scores, using a logistic model[10]. Then for each quality group. I split the borrowers into four groups based on their borrowing relationships: 1). low-IT banks in the control group; 2). high-IT banks in the control group; 3). low-IT banks in the treatment group; and 4). high-IT banks in the treatment group. If a borrower borrows from more than one bank, I assign it to the type of bank from which the borrower borrows the most from. Panels A and B of Figure 5 show the transitional matrix of the high-quality and low-quality borrowers. For each heat map, each cell shows the proportion of the borrowers that borrow from the type of banks shown by the row name before the experiment and then borrow from the types of banks shown by the column names after the experiment. The darker the color, the higher the proportion.

There are two clear patterns in the charts. First, the diagonals have darker colors. This indicates that a borrower that borrows from a certain type of bank before the experiment is more likely to borrow from the same type of bank after the experiment. Second, for high-quality borrowers, the color gets darker from the left to the right. While

[10]Similar results are obtained if I use a deep neural network with logistic activation function to predict default using only firm balance-sheet information.

Figure 5. Flow of Borrowers by Quality

This figure shows the proportion of the borrowers that borrowed from the type of banks shown by the row names before the experiment and then borrowed from the types of banks shown by the column names after the experiment. Panels A and B respectively show the flow proportion of high-quality and low-quality borrowers. Qualify is the one minus the predicted default rate using all banks' post-experiment credit scores. If a borrower borrowed from more than one bank, then the type of banks assigned to the borrowers is the one that the borrowers borrowed the most from.


for low-quality borrowers, the color gets darker from the right to the left. Given the order of the columns, this indicates that, after the experiment, the treated banks are more likely to make more loans to high-quality borrowers and fewer loans to low-quality borrowers. Among the treated banks, it is the high-IT banks that are more likely to make more loans to high-quality borrowers and fewer loans to low-quality borrowers. The results support the hypothesis that the data-sharing enables the treated banks to cream-skim high-quality borrowers from untreated banks and for high-IT treated banks to cream-skim high-quality borrowers from low-IT treated banks.

## V Structural Estimation

The empirical results in the previous sections help pin down some equilibrium data-sharing results when only some banks are affected. To study the equilibrium results of reducing the data-acquisition costs for all banks, in this section, I structurally estimate a model of loan application and default to explore the equilibrium effects of the data-sharing policy

when all banks are shared with the data, especially when banks have heterogeneous levels of IT intensity.

## *A.* Setup

### *1.* Demand and Default

The modeling of demand and default is similar to that in Crawford et al. (2018). There is one market in the economy[11]. Each year-quarter $t$, there are $J_t$ firms seeking credit to finance a project that requires an exogenous amount of $l_{j,k,t}$, where $k$ denotes bank $k$ among the $K_t$ banks active in the market. Firms select their main borrowing from one of the $K_t$ banks. Conditional on taking a loan, firms decide whether to default. Each bank $k$ chooses interest rate, $i_{j,k,t}$, to maximize expected profitability based on Bertrand-Nash competition.

Given these assumptions, let firms have the following indirect utility from their main borrowing:

$$\begin{aligned} U_{j,k,t} = \alpha_0 + \mathbf{X}_{k,t}\beta + \xi_{j,t} + \alpha_i i_{j,k,t} + \alpha_O O_{j,k,t} \\ + \alpha_Z Z_{j,k,t} + \alpha_{i,Z} i_{j,k,t} \times Z_{j,k,t} + \alpha_{O,Z} O_{j,k,t} \times Z_{j,k,t} \\ + \mathbf{Y}_{j,k,t}\eta + \epsilon_j + \nu_{j,k,t}, \end{aligned}$$

where $\mathbf{X}_{k,t}$ is a vector of bank-year determinants of demand, $i_{j,k,t}$ is the interest rate offered by bank $k$ to firm $j$ in year $t$, and $O_{j,k,t}$ is a dummy variable for online loan application. Therefore, $\alpha_O$ captures firm $j$'s preference for applying for the loan online. $Z_{j,k,t}$ is a dummy variable that equals to one if at year $t$, $j$ has borrowed before from $k$. It is a measure of the existence of lending relationships. $\mathbf{Y}_{j,k,t}$ is a vector of (non-interest) firm-bank-year determinants of demand, $\xi_{j,t}$ represents firm unobservable (to the econometrician) attributes in year $t$, and $\nu_{j,k,t}$ represents the unobserved shocks to $i$'s demand for bank $k$. $\epsilon_j$ represents firm $j$'s individual propensity to demand that is known to the firm but not the bank. It is modeled as a random coefficient on the constant $\alpha_0$, that is, $\alpha_j = \alpha_0 + \epsilon_j$. I let $U^0_{j,k,t} = \nu^0_{j,k,t}$ be the utility from the outside option, which is

[11] In China, business loan markets are usually defined at the province level. Given that I only have data from one province, I assume that there is only one market.

not borrowing from any of the banks active in the market at year $t$[12]. Firms choose their main banks to borrow from the bank that maximizes their utility, or else they choose not to borrow at all ($k = 0$).

Conditional on borrowing, each firm chooses to default if the indirect utility from doing so is larger than zero. The indirect utility is modeled as

$$\begin{aligned} U^D_{j,k,t} = \alpha^D_0 + \mathbf{X}_{k,t}\beta^D + \alpha^D_i i_{j,k,t} + \alpha^D_O O_{j,k,t} \\ + \alpha^D_Z Z_{j,k,t} + \alpha^D_{i,Z} i_{j,k,t} \times Z_{j,k,t} + \alpha^D_{O,Z} O_{j,k,t} \times Z_{j,k,t} \\ + \mathbf{Y}_{j,k,t}\eta^D + \epsilon^D_j, \end{aligned}$$

where $\epsilon^D_i$ represent firm $j$'s propensity to default.

Similar to Crawford et al. (2018), I allow the model to have asymmetric information, which is based on the correlation structure of the unobserved propensity to apply and default. That is, I assume that $\epsilon_j$ and $\epsilon^D_j$ are distributed following a bi-variate normal distribution:

$$\begin{pmatrix} \epsilon_j \\ \epsilon^D_j \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right).$$

A positive correlation between the firm-specific unobservables driving demand and default ($\rho$) is evidence of adverse selection: a positive correlation between $\epsilon_i$ and $\epsilon^D_i$ implies that firms with a higher unobservable propensity to demand credit are also more likely to default. At the same time, a positive $\alpha^D_i$ implies the existence of moral hazard: high repayment requirements on loans can reduce the incentives to exert effort, thus increasing the default probability. However, using $\alpha^D_i$ to imply moral hazard builds on the assumption that $\alpha_D$ is estimated by the component of price variation that is orthogonal to firms' unobservable characteristics, so that $\alpha_D$ doesn't mechanically capture the fact that observably riskier firms are offered higher interest rates. To do so, I follow Crawford et al. (2018) and estimate the indirect utility using a methodology that is similar to an instrumental variable (IV) regression (See Crawford et al. (2018) for details).

[12]The decision of not borrowing corresponds to the firms that are active but do not have any new loans in year $t$.

### *2.* Credit Supply

Bank $k$'s expected profits from offering borrower $j$ a loan with interest rate $i_{j,k,t}$ and amount $l_{j,k,t}$ is

$$\pi_{j,k,t} = (1 - \tilde{D}_{j,k,t}) i_{j,k,t} q_{j,k,t} l_{j,k,t} - c_{j,k,t} q_{j,k,t} l_{j,k,t}. \tag{2}$$

In (2), $\tilde{D}_{j,k,t} = \tilde{d}_{j,k,t}(1 - R_{j,k,t})$, where $\tilde{d}_{j,k,t}$ is firm $j$'s default probability and $R_{j,k,t}$ is the recovery rate in case of default. $q_{j,k,t}$ is the probability of application, and $c_{j,k,t}$ is the marginal costs of supplying the loan. Marginal cost is defined as

$$c_{j,k,t} = \kappa_1 \times O_{j,k,t} + \kappa_2 \times \tilde{s}_{j,k,t} + \psi_j + \psi_k + \psi_t + e_{j,k,t}, \tag{3}$$

where $\psi_j$, $\psi_k$, and $\psi_t$ are respectively the firm, bank, and year fixed effects. In addition, I allow marginal costs to depend on the time it takes to originate the loan. Similar to Einav et al. (2012), I assume that the banks can engage in risk-based pricing in addition to the expected default rates, as captured by the term $\kappa_2 \times \tilde{s}_{j,k,t}$. $\tilde{s}_{j,k,t}$ is bank $k$'s risk score of firm $j$ in year $t$. The inclusion of $\kappa_2 \times \tilde{s}_{j,k,t}$ indicates that per-loan cost is not constant but varies according to borrower risk.

The first-order condition of maximizing (2) yields

$$i_{j,k,t} = \underbrace{\frac{c_{j,k,t}}{1 - \tilde{D}_{j,k,t} + \tilde{D}'_{i_{j,k,t}} M_{j,k,t}}}_{\textit{Effective Marginal Cost}} + \underbrace{\frac{(1 - \tilde{D}_{j,k,t}) M_{j,k,t}}{1 - \tilde{D}_{j,k,t} + \tilde{D}'_{i_{j,k,t}} M_{j,k,t}}}_{\textit{Effective Markup}}, \tag{4}$$

where $\tilde{D}'_{i_{j,k,t}} = \tilde{d}'_{j,k,t}(1 - R_{j,k,t})$ is the marginal effects of setting a higher interest rate on default probability net of recovery. $M_{j,k,t} = -q'/q$ is bank $k$'s markup on a loan to firm $j$. The two terms on the right-hand side of (4) are respectively the effective marginal costs and effective markup. The decomposition of interest rates into a marginal cost term and a markup term is similar to any regular Bertrand-Nash pricing equation. The difference is that, in the existence of default, there is an additional term $\tilde{d}'_{i_{j,k,t}}$ in (4), which measures the effects of pricing on the sensitivity of default to interest rates.

## *B.* Modeling the Effects of the Experiment

### *1.* Screening Ability

Conventionally, when studying risk-based pricing, post-experiment $\tilde{s}_{j,k,t}$ is observed for all banks. Therefore, the estimation of (3) is directly based on the observed values of $\tilde{s}_{j,k,t}$. However, in the setting here, I only observe the post-experiment risk scores for the treated banks. To study the counterfactual scenario where all banks are shared the data, I construct a measure of the *optimal* post-experiment risk-scoring model, and model the heterogeneous screening ability as *how likely* different banks can use this technology. Specifically, I first fit a random-forest (RF) model with all bank's post-experiment risk scores to predict the five-year default probability for the loans originated after the experiment, and then construct the optimal post-experiment risk scores, $s_{j,k,t}$, as the standardized log default probability predicted by the RF model. $s_{j,k,t}$ can be thought as the *type* of the borrowers in describing the borrowers' default probability when borrowing from $k$ in year $t$.

I then construct bank $k$'s post-experiment risk score about borrower $j$ in time $t$ as

$$\begin{aligned} \tilde{s}_{j,k,t} &= (1 - p_{k,t}) \times \tilde{s}_{j,k,-1} + p_{k,t} \times s_{j,k,t} \\ p_{k,t} &= \kappa_3 \times treat_{k,t} \times I_k, \end{aligned} \tag{5}$$

where $\tilde{s}_{j,k,-1}$ is the standardized most recent risk scores available before data-sharing. $p_{k,t}$ captures the probability that bank $k$ is able to use the optimal technology. $I_k$ is bank $k$'s average IT intensity three years before the experiment. $treat_{j,t}$ is a dummy variable that is equal to one if the data is shared with the bank. $treat_{j,t} \times I_k$ captures the interaction between data-sharing and IT intensity. When $p_{k,t} = 0$, the bank cannot use the optimal technology, and the optimal risk score is the newest risk score before the experiment. When $\kappa_3 > 0$, banks with higher IT intensity can use the optimal credit scores with a higher probability.

### *2.* Convenience

Motivated by the empirical results, I model the experiment as affecting two dimensions. First, sharing the data changes the availability of online applications that increase the convenience of the loan origination process, including a much shorter origination time. This modeling choice is also supported by the findings in Fuster et al. (2019), who find that fintech lenders can originate mortgages at a faster time without incurring a higher default probability. The effect of data-sharing on the availability of online applications is modeled as

$$O_{j,k,t} = f(\boldsymbol{\kappa}_4, treat_{k,t} \times I_k, \tilde{s}_{j,k,t}),$$

where $\boldsymbol{\kappa}_4$ contains bank, firm, and year fixed effects. $treat_{j,t} \times I_k$ captures the interaction between data-sharing and IT intensity. $\tilde{s}_{j,k,t}$ captures the effects of the experiment on convenience indirectly through affecting screening ability. For simplicity, I let $f(\cdot)$ be a linear function. Therefore,

$$O_{j,k,t} = \boldsymbol{\kappa}_4 + \kappa_5 \times treat_{k,t} \times I_k + \kappa_6 \times \tilde{s}_{j,k,t}. \tag{6}$$

(6) estimates online application availability through a linear probability model. In this case, $O_{j,k,t}$ captures the availability of online applications through a continuous approximation. This modeling choice simplifies the estimation process. When $\kappa_5 < 0$, data-sharing decreases loan-origination time, and the effects increase with bank IT intensity. When $\kappa_6 > 0$, firms with higher credit scores, thus higher $\tilde{s}_{j,k,t}$, have a higher probability of having access to online applications.

## *C.* Estimation

### *1.* Demand and Default

The estimates of the structural model are presented in Table 8[13]. As shown, a significantly negative relationship exists between interest rate and loan demand. In addition, a positive

[13]A detailed description of the estimation process is in section B of the Online Appendix.

TABLE 8. Structural Estimates

This table gives the structural estimates. Standard errors are based on the inverse of the information matrix.

| | (1) Demand | (2) Default |
|---|---|---|
| Interest Rate | -0.36 | 0.44 |
| | (0.16) | (0.06) |
| Interest Rate × Relationship | -0.69 | 0.21 |
| | (0.22) | (0.04) |
| Online | 1.06 | 0.08 |
| | (0.06) | (0.12) |
| Online × Relationship | -0.26 | 0.03 |
| | (0.05) | (0.11) |
| log(Distance) | -0.23 | -0.41 |
| | (0.05) | (0.08) |
| log(AT) | -0.05 | -0.65 |
| | (0.12) | (0.14) |
| log(Volume) | 3.34 | -0.22 |
| | (0.12) | (0.07) |
| Age | 0.02 | 0.08 |
| | (0.41) | (0.31) |
| Profitability | 0.00 | -2.42 |
| | (0.37) | (0.57) |
| Leverage | 0.00 | -0.04 |
| | (0.01) | (0.01) |
| Maturity FE | Yes | Yes |
| Bank FE | Yes | Yes |
| Year FE | Yes | Yes |
| Relationship FE | Yes | Yes |
| N | 1,932,730 | 239,080 |
| Covariance Matrix | $\sigma = 0.29$ | |
| | (0.08) | |
| | $\rho = 0.39$ | $\sigma_P = 1$ |
| | (0.08) | |

number of $\rho$ and $\alpha_i^D$ indicates the existence of adverse selection and moral hazard. The coefficient in front of online applications in the demand equation is significantly negative. This implies that borrowers prefer an easier loan origination process. However, the coefficient in the default equation is insignificant. Therefore, similar to the findings in Fuster et al. (2019), faster origination is not at the cost of a higher default rate.

Previous lending relationships have a very strong effect on demand elasticity for interest rates and online applications. Similar to Ioannidou et al. (2022), demand is more sensitive to interest rates if there is a previous lending relationship. This is likely because borrowers with a prior relationship with the bank are more likely to be safer borrowers. Therefore, they are more price-sensitive as well. On the other hand, firms with a previous relationship are less sensitive to online applications. This is likely the case because, given the experience with this bank, firms are more certain about the final outcomes of the lending process and, therefore, less averse to waiting longer.

As for default, higher interest rates lead to a smaller increase in default probability or less moral hazard when there is a previous lending relationship. While regardless of the existence of lending relationships, origination time is not significantly related to default probability.

### *2.* The Effects of Data-Sharing

I first estimate the effects of data-sharing on origination time using the following DID specification

$$O_{j,k,t} = \lambda_j + \lambda_k + \lambda_t + \kappa_5 \times I_k \times treat_{k,t} + \kappa_6 \times \tilde{s}_{j,k,t} + e^T_{j,k,t}, \tag{7}$$

where $\lambda_j$, $\lambda_k$, and $\lambda_t$ are respectively the borrower, bank, and year fixed effects.

Directly estimating the marginal-cost equation (3) is difficult as it requires the separate identification of $\kappa_2$ and $\kappa_3$. Since I don't observe $p_{k,t}$, $\kappa_3$ cannot be identified easily. Instead, to estimate the marginal-cost equation, I first combine (3), (5), and (6), and express (3) as

$$c_{j,k,t} = \overline{c}_{j,k,0} + \tilde{\kappa}_1 \times treat_{j,t} \times I_j + \tilde{\kappa}_2 \times treat_{j,t} \times I_j \times \Delta s_{j,k,t} + \overline{\psi}_0 + e_{j,k,t} - \overline{e}_{j,k,0}, \tag{8}$$

where, $\overline{c}_{j,k,0}$ is the average marginal costs and log origination time before the experiment. $\Delta s_{j,k,t} = s_{j,k,t} - \tilde{s}_{j,k,-1}$ is the changes in the optimal risk scores. $\tilde{\kappa}_1 = \kappa_1 \times \kappa_5$, is the interaction effects of the experiment and IT intensity on marginal costs through affecting the availability of online applications. It captures the effects of convenience on marginal

TABLE 9. The Effects of Data-Sharing

This table presents the relationship between data-sharing and marginal costs (Cost) and that with loan origination time (log Time). $\Delta T_{j,k,t}$ is the changes in loan origination time. $I_j$ is bank average pre-experiment IT intensity. $\Delta s_{j,k,t}$ is the changes in the firm $j$'s optimal risk score as in (6).

| | (1) Cost | (2) Cost | (3) Online | (4) Online |
|---|---|---|---|---|
| $treat_{j,t} \times I_j$ | 0.01 | -0.01 | 0.18*** | 0.17*** |
| | (0.03) | (0.03) | (0.05) | (0.04) |
| $treat_{j,t} \times I_j \times \Delta s_{j,k,t}$ | -0.16*** | -0.14*** | | |
| | (0.02) | (0.02) | | |
| $s_{j,k,t}$ | | | 0.07*** | 0.08*** |
| | | | (0.02) | (0.02) |
| Firm FE | No | Yes | No | Yes |
| Bank FE | No | Yes | No | Yes |
| Year FE | No | Yes | No | Yes |
| N | 272,353 | 239,080 | 272,353 | 239,080 |

Standard Errors Clustered at Year-Quarter and Bank Level in Parentheses
*** $p$ <0.001, ** $p$ <0.01, * $p$ <0.05

costs holding screening ability constant. $\tilde{\kappa}_2 = (\kappa_1\kappa_6 + \kappa_2) \times \kappa_3$ is the effects of each unit adjustment of an optimal credit-scoring model on marginal costs. It captures the total effects of screening ability on marginal costs. Finally, $\overline{\psi}_0$ and $\overline{e}_{j,k,0}$ are respectively the averages of the year fixed effects and structural errors before the experiments. With (8), I directly estimate $\kappa_1$ and $\tilde{\kappa}_2$, can perform counterfactuals using (8). To estimate $\kappa_1$ and $\tilde{\kappa}_2$, I fit the following DID specification:

$$c_{j,k,t} - \overline{c}_{j,k,0} = \lambda_j^c + \lambda_k^c + \lambda_t^c + \tilde{\kappa}_1 \times treat_{j,t} \times I_j + \tilde{\kappa}_2 \times I_k \times treat_{k,t} \times \Delta s_{j,k,t} + e_{j,k,t}^c. \quad (9)$$

Table 9 gives the estimates of (7) and (9). Consistent with previous results, columns (4) through (6) show that there is a strong interaction effect of data-sharing and IT intensity on loan origination time and risk-based pricing. Specifically, focusing on column (5), given $\kappa_4 = -0.06$ and average IT intensity equals 3.3%, data sharing decreases the loan origination time for the bank with the average amount of IT intensity by around 20%.

Columns (1) to (3) gives the relationship between marginal costs and origination time and optimal risk scores. $\tilde{\kappa}_2 = 0.10$ implies that, for the bank with the average amount of

IT intensity, upon sharing the data, for each standard deviation increase in the risk score, marginal cost increases by around 33 basis points. However, the effect of loan origination time on marginal cost, $\kappa_1$, is slightly negative but insignificant[14].

An assumption made in (6) and (5) is that data-sharing is effective on origination time and screening ability only if banks' IT intensity is larger than zero. This is motivated by the empirical results that the effects of data-sharing is concentrated in banks with high ex ante IT intensity. In columns (3) and (6) of Table 9, I include the main effects of $treat_{j,t}$ in the estimation. The coefficient in front of $treat_{j,t}$ measures the effects of data-sharing on banks with zero IT intensity. Based on the estimates, the effects of data-sharing on marginal costs and origination time is both economically and statistically insignificant. Confirming the validity of the structural specification of (6) and (5).

### *D.* Model Fit

Panels A and B of Table 10 shows that the model is very effective in matching the equilibrium moments in the data. Before the experiment, the model generates an average default rate of 3.31% and an average interest rate of 5.56%, compared with 3.30% and 5.57% in the data. Effective marginal cost is on average 3.86%. This indicates an average effective markup of 1.70%.

Panel B gives the post-experiment fit of the model. Since the parameters in the demand and default equation are estimated using only the pre-experiment data, the moments in panel B can serve as an out-of-sample check of the model fits. In general, the model fits the data nicely, with an average default rate of 3.20% and an average interest rate of 5.66%, compared with 3.23% and 5.69% in the data.

### *E.* Counterfactual Analysis

#### *1.* Equilibrium Outcomes

I study three counterfactuals to assess the equilibrium outcomes of when a large amount of data is available to all the banks. For the first one, I set $treat_{k,t} = 1$ for all banks,

[14]When loan origination time is a result of screening, longer origination time is expected to be associated with higher costs. However, lower loan origination time could also be related to higher costs because of the need for more intense monitoring. So the net effect is ambiguous.

TABLE 10. Model Fit and Counterfactual Analysis

This table gives the summary statistics of the data and model outcomes. Panel A, B, and C respectively gives the pre-experiment, post-experiment, and counterfactual averages. Column (1) is the average default rate. Column (3) is the average interest rate. Column (5) is the average effective marginal costs. Column (7) is the average effective markup. The even columns are the percentage changes with respect to the pre-experiment level as estimated by the model.

| | | | (1) Default | (2) % Diff | (3) Interest Rate | (4) % Diff | (5) Effective MC | (6) % Diff | (7) Effective Markup | (8) % Diff |
|---|---|---|---|---|---|---|---|---|---|---|
| A: Pre-Experiment | Data | | 3.30 | | 5.57 | | | | | |
| | Model | All | 3.31 | | 5.56 | | 3.50 | | 2.06 | |
| | | High IT | 3.14 | | 5.47 | | 3.38 | | 2.10 | |
| | | Low IT | 3.57 | | 5.79 | | 3.79 | | 2.00 | |
| B: Post-Experiment | Data | All | 3.23 | | 5.69 | | | | | |
| | Model | All | 3.24 | | 5.71 | | 3.51 | | 2.20 | |
| C: Counter-Factual | Both | All | 3.26 | -1.51% | 5.66 | 1.80% | 3.21 | -8.22% | 2.45 | 18.82% |
| | Screening | All | 2.99 | -9.67% | 5.41 | -2.70% | 3.13 | -10.46% | 2.28 | 10.50% |
| | Convenience | All | 3.52 | 6.34% | 5.99 | 7.73% | 3.48 | -0.49% | 2.51 | 21.70% |
| D: Heterogeneity | | High IT | 2.96 | -5.73% | 5.63 | 2.93% | 3.00 | -11.11% | 2.63 | 25.54% |
| | | Low IT | 3.65 | 2.24% | 5.75 | -0.69% | 3.66 | -3.35% | 2.09 | 4.35% |

and regenerate average interest rates and default rates. This exercise is to study the equilibrium results when all banks are shared with the data. Then to dissect the effects of data-sharing on bank profitability, I study the case when only one of the convenience and screening-ability channels is at work. The results are in Panel C of Table 10. The odd columns respectively give the average default rate, the average interest rate, the average effective marginal costs, and the average effective markup. The even columns are the corresponding percentage changes with respect to the pre-experiment level as estimated by the model. As shown from the first rows of Panel C, data-sharing has an insignificant impact on the market average default rate and interest rate: the percentage changes are only -0.94% and 0.72% respectively. Looking at columns (6) and (8), despite an insignificant change in interest rate or default probability, there is a very big decrease in the average effective marginal cost: the market average effective marginal cost decreased by 7.53% from 3.86% before the experiment to 3.57%. Given a large drop in the effective marginal cost, effect markup increased by around 20% from 1.70% before the experiment to 2.03%.

The last two rows in Panel C dissect the effects of the experiment. As shown, the two channels mainly affect two different margins. When data-sharing only increases screening ability, the default rate decreases by more than 10% from 3.31% to 2.97%. This is accompanied by a 12 basis-point decrease in interest rates. However, when data-sharing only decreases origination time, demand increases, and interest rate increases by more than 7% from 5.56% to 5.96%. At the same time, a higher interest rate also comes with 20 basis points higher default rate. In sum, data-sharing allows banks to decrease the origination time for all types of borrowers. This leads to a higher demand, and thus higher interest rate from the average borrower. However, the increase in screening ability allows the bank to decrease interest rates only to borrowers with lower marginal costs, or lower ex-post default rates. Altogether, an increase in demand from the average borrowers and a reallocation in supply to low-risk borrowers generate a simultaneous increase in the interest rate and a decrease in the default. On net, for the whole market, the two effects nearly cancel out to yield an insignificant change in the average interest rate and average default rate. However, since better risk-based pricing ability directly decreases the marginal costs, average markup increases sizeably. This is to say, banks are able to have a lower marginal costs without decreasing the prices.

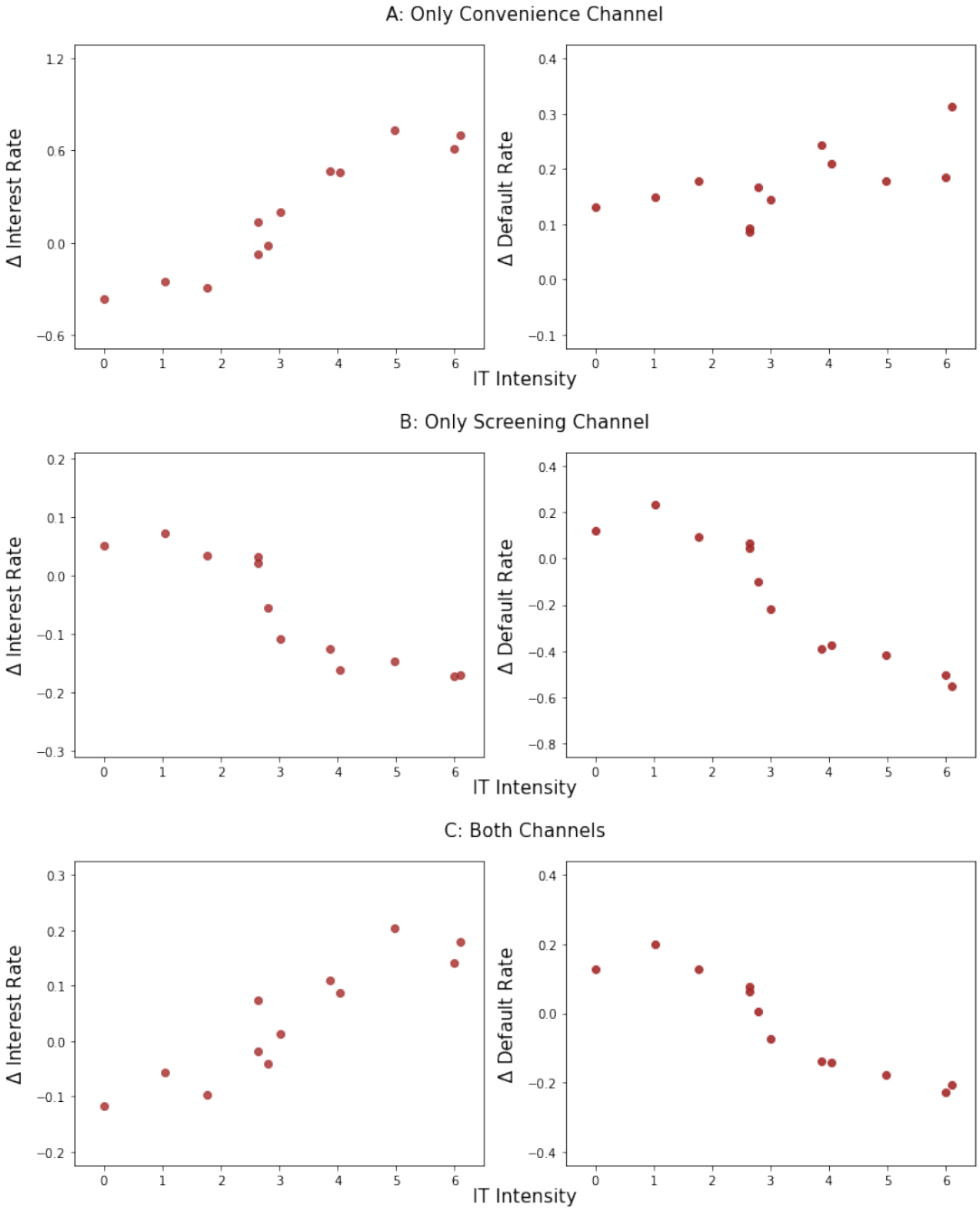
### *2.* Heterogeneity by IT Intensity

An important heterogeneity of the effects of data is bank's IT intensity. From the estimates in Table 9, data-sharing has insignificant effects on origination time or screening ability only for the banks with zero IT intensity. Therefore, the effects of data-sharing is expected to increase the profitability to banks with higher IT intensity.

To inspect this conjecture, I plot each bank's average change in interest rates and default rates when data is share with all banks by their IT intensity. The results are in Figure 6. Panel A gives the case when data-sharing only affects origination time; panel B gives the case when data-sharing only affects screening ability; and panel C gives the case when data-sharing affects both origination time and screening ability. From the plots, there is a strong positive interaction effect of data-sharing and IT intensity on interesting rate when data-sharing only affects origination time. Specifically, for high-IT banks, larger amount of firm data enables the bank to reduce origination time more, therefore facing

Figure 6. Counterfactuals

This figure gives the changes in interest rates and default rate under different counterfactual scenarios when all banks are shared with the data. Panel A gives that when the screening channel is shut down. Panel B gives that when the convenience channel is shut down. Panel C analyzes the case when both convenience and screening channels operates.




a higher demand and thus a higher interest rates. However, the interaction effects of data-sharing and IT intensity on default rate is only slightly positive.

From panel B, when data-sharing only affects screening ability, banks with higher IT spending extend more loans to borrowers with lower default rates. This results in a much steeper relationship between IT intensity and default rate. At the same time, looking at the left panel, the relationship between IT intensity and changes in interest rates, though also negative, is much flatter than that from Panel A. Therefore, based on the patterns from Panel A and Panel B, one can think of the screening-ability as a cost-reduction supply channel, which decreases both interest rates and default probability, and think of the convenience channel as a demand channel, which increases both interest rates and default probability.

Panels A and B of Figure 6 shed light on the relative strengths of the supply channel and demand channel respectively on default rates and interest rates. On the one hand, increases in screening ability decrease default rates much more than the increases in default rates caused by higher demand through moral hazard. On the other hand, increases in interest rates because of a higher demand for fast loan dominates the decrease in interest rate because of extended loan to safer borrowers. Altogether, data-sharing has a larger positive effect on interest rates, and a larger negative effect on default rates for high IT-intensity banks. This is confirmed by Panel C.

To assess the asymmetric effects of making big data available on banks with different IT intensity quantitatively, I give the summary statistics of bank profitability in Panel D of Table 10. Consistent with the patterns in Figure 6. The experiment has a much stronger effects on banks with higher IT intensity. Specifically, for banks with high IT intensity, the experiment decreases their default rate by 5.16% from 3.14% to 2.98%, and increases interest rate by 2.24% from 5.49% to 5.61%. At the same time, effective marginal cost decreases by more than 10% from 3.75% to 3.36%. In the end, banks with high IT intensity see a 29.5% increase in the effective markup. On the other hand, the experiment has little effects on banks with low IT intensity. Over all margins, the changes are economically insignificant.

To sum up, the counterfactual exercises confirm the conjecture in the experiment: big data is expected to increase loan demand through lower origination time, and decrease default rate through better screening ability. The effects are larger only for banks that have high IT capacity. The asymmetric effects on data on banks with different level of

data-processing abilities enable high IT banks to cream-skim good borrowers from low IT banks.

# VI Conclusion

In this paper, I combine a quasi-experiment that provides a large amount of hard data from local administrative agencies to commercial banks and structural estimation to shed light on the effects of big data on loan attributes and bank profitability. I first show that providing a great amount of hard data to banks extensively increases banks' screening ability. At the same time, the experiment increases interest rates and speed of originating the loan and decreases default rates. In addition, given the requirement of technology to process the large amount of data, the availability of a larger amount of data has more significant effects on banks with high information technology (IT) capacity.

The analysis here sheds light on several avenues for future research. First, I treat IT intensity as an exogenous variable, and study the heterogeneous effects of data-sharing by IT intensity. However, banks could adjust their IT spending when facing decreasing data-acquisition costs. For example, He et al. (2022) show that US commercial banks has been catching up on the investment of IT over the past decades. Future research could study the case when banks could optimally adjust their IT spending. In addition, I only focus on loan attributes but not borrower fundamentals. Future research could study how reduced data-acquisition costs to the banks spill over to the borrowers.

## References


Aiello, D., M. J. Garmaise, and G. Natividad (2020). Competing for deal flow in local mortgage markets. *Working Paper*.

Athreya, K., X. S. Tam, and E. R. Young (2012, July). A quantitative theory of information and unsecured credit. *American Economic Journal: Macroeconomics 4*(3), 153–83.

Babina, T., G. Buchak, and W. Gornall (2022). Customer data access and fintech entry: Early evidence from open banking. Available at SSRN: https://ssrn.com/abstract=4071214 or http://dx.doi.org/10.2139/ssrn.4071214.

Babina, T., A. Fedyk, A. X. He, and J. Hodson (2020). Artificial intelligence, firm growth, and industry concentration. Available at SSRN: https://www.ssrn.com/abstract=3651052.

Benetton, M. (2021). Leverage regulation and market structure: A structural model of the u.k. mortgage market. *The Journal of Finance 76*(6), 2997–3053.

Berg, T., V. Burg, A. Gombović, and M. Puri (2019, 09). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies 33*(7), 2845–2897.

Berg, T., A. Fuster, and M. Puri (2021, October). Fintech lending. Working Paper 29421, National Bureau of Economic Research.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–90.

Buchak, G., G. Matvos, T. Piskorski, and A. Seru (2018). Fintech, regulatory arbitrage, and the rise of shadow banks. *Journal of Financial Economics 130*(3), 453–483.

Calebe de Roure, L. P. and A. V. Thakor (2019). P2P Lenders versus Banks: Cream Skimming or Bottom Fishing? *SAFE Working Paper No. 206*.

Crawford, G. S., N. Pavanini, and F. Schivardi (2018, July). Asymmetric information and imperfect competition in lending markets. *American Economic Review 108*(7), 1659–1701.

Cuesta, J. I. and A. Sepulveda (2021). Price Regulation in Credit Markets: A Trade-Off between Consumer Protection and Credit Access. *Working Paper.*.

DeLong ER, DeLong DM, C.-P. D. (1988, Sep). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics 44*(3), 837–845.

Detragiache, E., P. Garella, and L. Guiso (2000). Multiple versus single banking relationships: Theory and evidence. *Journal of Finance 55*(3), 1133–1161.

Di Maggio, M., D. Ratnadiwakara, and D. Carmichael (2022, March). Invisible primes: Fintech lending with alternative data. Working Paper 29840, National Bureau of Economic Research.

Di Maggio, M. and V. Yao (2020, 12). Fintech borrowers: lax Screening or cream-skimming? *The Review of Financial Studies*.

Drozd, L. A. and R. Serrano-Padial (2017, March). Modeling the revolving revolution: The debt collection channel. *American Economic Review 107*(3), 897–930.

Egan, M., A. Hortacsu, and G. Matvos (2017, January). Deposit competition and financial fragility: Evidence from the us banking sector. *American Economic Review 107*(1), 169–216.

Egan, M., S. Lewellen, and A. Sunderam (2021, 08). The Cross-Section of Bank Value. *The Review of Financial Studies 35*(5), 2101–2143.

Einav, L., M. Jenkins, and J. Levin (2012). Contract pricing in consumer credit markets. *Econometrica 80*(4), 1387–1432.

Erel, I. and J. Liebersohn (2020). Does finTech substitute for banks? Evidence from the paycheck protection program. *Working Paper*.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp (2019, May). Big data and firm dynamics. *AEA Papers and Proceedings 109*, 38–42.

Flannery, M. and S. M. Sorescu (1996). Evidence of bank market discipline in subordinated debenture yields: 1983-1991. *Journal of Finance 51*(4), 1347–77.

Frost, J., L. Gambacorta, Y. Huang, H. S. Shin, and P. Zbinden (2019). Investment in ict, productivity, and labor demand : The case of argentina. *BIS Working Papers*.

Fuster, A., M. Plosser, P. Schnabl, and J. Vickery (2019, 04). The Role of Technology in Mortgage Lending. *The Review of Financial Studies 32*(5), 1854–1899.

Gopal, M. and P. Schnabl (2022, 06). The Rise of Finance Companies and FinTech Lenders in Small Business Lending. *The Review of Financial Studies*. hhac034.

Guiso, L., A. Pozzi, A. Tsoy, L. Gambacorta, and P. E. Mistrulli (2022). The cost of steering in financial markets: Evidence from the mortgage market. *Journal of Financial Economics 143*(3), 1209–1226.

Hauswald, R. and R. Marquez (2003, 07). Information Technology and Financial Services Competition. *The Review of Financial Studies 16*(3), 921–948.

He, Z., J. Huang, and J. Zhou (2020). Open banking: Credit market competition when borrowers own the data. Available at SSRN: https://ssrn.com/abstract=3736109.

He, Z., J. Huang, and J. Zhou (2022). Open banking: Credit market competition when borrowers own the data. *Working Paper, University of Chicago*.

He, Z., S. Jiang, D. Xu, and X. Yin (2022). Investing in lending technology: It spending in banking. Available at SSRN: https://ssrn.com/abstract=3936767 or http://dx.doi.org/10.2139/ssrn.3936767.

Hornuf, L., M. F. Klus, T. S. Lohwasser, and A. Schwienbacher (2018). How do banks interact with fintechs? forms of alliances and their impact on bank value. *CESifo Working Paper*.

Hughes, J., J. Jagtiani, and C.-G. Moon (2019). Consumer lending efficiency:commercial banks versus a fintech lender. *FRB of Philadelphia Working Paper No. 19-22*.

Ioannidou, V., N. Pavanini, and Y. Peng (2022). Collateral and asymmetric information in lending markets. *Journal of Financial Economics 144*(1), 93–121.

Ippolito, F., J.-L. Peydro, A. Polo, and E. Sette (2016). Double bank runs and liquidity risk management. *Journal of Financial Economics 122*(1), 135–154.

Iyer, R., A. I. Khwaja, E. F. P. Luttmer, and K. Shue (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science 62*(6), 1554–1577.

Jagtiani, J. and C. Lemieux (2017). Fintech lending: Financial inclusion, risk pricing, and alternative information. *FRB of Philadelphia Working Paper No. 17-17*.

Jappelli, T. and M. Pagano (2002). Information sharing, lending and defaults: Cross-country evidence. *Journal of Banking and Finance 26*(10), 2017 – 2045.

Liberti, J., J. Sturgess, and A. Sutherland (2022). How voluntary information sharing systems form: Evidence from a u.s. commercial credit bureau. *Journal of Financial Economics 145*(3), 827–849.

Liberti, J. M., A. Seru, and V. Vig (2019). Information, credit, and organization. Available at SSRN: http://ssrn.com/abstract=2798608.

Liu, L., G. Lu, and W. Xiong (2022, June). The big tech lending model. Working Paper 30160, National Bureau of Economic Research.

Livshits, I., J. C. Mac Gee, and M. Tertilt (2016, 03). The Democratization of Credit and the Rise in Consumer Bankruptcies. *The Review of Economic Studies 83*(4), 1673–1710.

Lorente, C., J. Jose, and S. L. Schmukler (2018, March). The fintech revolution: A threat to global banking? Research and Policy Briefs 125038, The World Bank.

Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *Forbes*. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=39d4a3fd60ba.

Martinez Peria, M. and S. Schmukler (2001). Do depositors punish banks for bad behavior? market discipline, deposit insurance, and banking crises. *Journal of Finance 56*(3), 1029–1051.

Nelson, S. T. (2022). Private information and price regulation in the us credit card market.

Parlour, C. A., U. Rajan, and H. Zhu (2022, 04). When FinTech Competes for Payment Flows. *The Review of Financial Studies 35*(11), 4985–5024.

Stulz, R. M. (2019). FinTech, BigTech, and the future of banks. *Journal of Applied Corporate Finance 31*(4), 86–97.

Tang, H. (2019, 04). Peer-to-Peer Lenders Versus Banks: Substitutes or Complements? *The Review of Financial Studies 32*(5), 1900–1938.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Vives, X. (2019). Digital disruption in banking. *Annual Review of Financial Economics 11*(1), 243–272.

Wiersch, A. M., S. Lieberman, and B. J. Lipman (2019). An update on online lender applicants from the small business credit survey.

Xiao, K. (2019, 10). Monetary Transmission through Shadow Banks. *The Review of Financial Studies 33*(6), 2379–2420.