The Effects of Big Data on Commercial Banks

Xiao Yin*

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 on commercial banks. Increasing the amount of data enables banks with high information technology (IT) capacity to improve screening ability and reallocate supply to high-quality borrowers; it also enables banks to attract demand through issuing loans that appear to be more convenient. However, the effects on banks with low IT capacity are negligible. Therefore, increasing the amount of data significantly raises the profitability of only banks with high information processing ability.

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

*Yin: UCL; xiao.yin@ucl.ac.uk. I deeply appreciate the valuable comments from Matteo Benetton, Enrico Sette (discussant), Yuriy Gorodnichenko, Vasso Ioannidou, Jordi Jaumandreu (discussant), and David Sraer. I am thankful for the comments from Bayes Business School, UC Berkeley, UCL, 4th Bank of Italy, Bocconi University and CEPR Conference on Financial Stability and Regulation, 13th MoFiR Workshop on Banking, and 22nd Annual International Industrial Organization Conference.

I Introduction

Over the last decade, businesses have increasingly turned to vast quantities of data to inform their decision-making processes. A recent report from Forbes (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 borrower information affects commercial banks. In 2015, Chinese local government authorities began sharing administrative data with commercial banks to enhance lending efficiency[1]. With the policy, local authorities share a large amount of borrower information with commercial banks. The information is listed in Figure 1. Overall, when first sharing the data, banks receive hundreds of characteristics of more than 200 thousand firms that have borrowed once over the past three years.

The data-sharing practice rolled out gradually, with different provinces selecting a fraction of commercial banks as experimental units to provide the data that local authorities process. This process is one example of the conventional policy experimentation in China. Specifically, policy experimentation, i.e., applying the policy innovation on selected experimental units, is a common reformation tool in China

[1]Section A in the Online Appendix outlines the policy guidance.

that aims to test the effectiveness of regulatory changes before massive policy roll-out. The selection of experimental units in this data-sharing practice is arguably exogenous conditional on banks with enough financial resilience (market share in the medium and small business loan market above 1%). Thereby, conditional on banks with medium and small business loan market share above 1%, I can study the partial equilibrium effects of data sharing by comparing the characteristics of banks selected as the experimental units (treatment group) and those not selected (control group).

The analysis reveals that treated banks issued loans that were, on average, 5% larger in volume and featured a 31 basis point higher interest rate and a 23 basis point lower default rate compared to banks in the control group. The finding of a reduced default rate for the treatment banks suggests that access to more extensive data potentially enhances banks' ability to screen borrowers. Further analysis using banks' proprietary credit scores supports this, showing that treated banks could better predict borrower defaults, as evident by a significant increase in the predictive power of their proprietary credit scores.

The results suggest that 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 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. For example, better information processing ability could increase loan demand by offering products with higher quality, potentially through improved convenience in the origination process (Buchak et al., 2018; Fuster et al., 2019). To shed light on big data shifting demand, I show that, following the data-sharing initiative, treated banks were able to offer loans more quickly and more conveniently, primarily by providing more borrowers with the opportunity to apply for loans online instead of visiting branches. With access to online applications, the loan processing time can be reduced from an average of 14 days to within one day, likely boosting demand among small business borrowers who prioritize speed Wiersch et al. (2019), and, as a result, pushing up interest rates.

While more hard information is expected to enable lenders to extract more high-dimensional information through systematic statistical inferencing, banks need to have advanced technology stock to use the data efficiently. 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 capacity 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 data available enable banks with higher IT intensity to cream-skim high-quality borrowers from low IT-intensity borrowers.

A limitation of the policy experimentation 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. To study the equilibrium effects of the data-sharing experimentation, I 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. The model allows for data-sharing to affect both the marginal costs of origination credit and the demand for credit. In addition, I assume that the effects are heterogeneous by bank IT capacity. Through this, I can separately quantify the impacts of data-sharing on supply and demand of credit.

I dissect the equilibrium effects of increasing access to hard information on bank profitability through three counterfactual scenarios: 1. data sharing affecting both screening ability and credit demand, 2. data sharing impacting only screening abilities, and 3. data sharing influencing only credit demand. 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 decrease in interest rates. Conversely, when it solely increases demand, default rates and interest rates increase. However, when both factors are active, their effects on interest rates almost neutralize, yet improved risk-based pricing lowers default rate by around 14 basis points and marginal costs by around 30 basis points. This results in an increase of markups by 25 basis points.

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 capacity experience more pronounced benefits, such as increased interest rates, reduced default rates, and significantly decreased marginal costs, leading to a 30 basis points 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 7 basis points. This analysis underscores a strong synergy between technology and data availability, showing that an increase in the amount of data significantly reduces marginal costs for banks with high IT capacity. Despite this cost reduction, prices do not fall due to a surge in demand, ultimately enhancing the profitability of high-IT banks significantly more than their low-IT counterparts.

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[2]. This study aligns with studies on how the emergence of fintech and IT is

[2]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. (2024), He et al. (2022), Gopal and Schnabl (2022), Liu et al. (2022), Blickle et al. (2024), etc. See Vives (2019) and Berg et al. (2021) for a review in banking.

affecting the traditional banking sector[3]. 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.

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.

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

[3]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.

production. He et al. (2020) 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.

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 officers, 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 in the early 2010s.

B. Data Sharing Policy

Since 2015, local government agencies in many provinces in China have experimented with sharing administrative data with commercial banks. The policy aims to reduce the cumbersomeness of collecting auditing reports and help banks reduce asymmetric information. The earliest province to enact this policy experimentation was Jiangsu in 2015. The provincial tax administration shared its data with a number of commercial banks. From 2016, the data-sharing practice started to proliferate in other provinces.

The data-sharing process takes two steps. The first step is to aggregate data locally at the government agencies. Beforehand, 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, etc. 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 companies that have borrowed once in the past three years. In my sample, over 80% of all firms that have a borrowing record in the Credit Reference Registry of the People's Bank of China agreed to share their information prior to the initial sharing of the data.

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

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

The second step is to share the data with the banks. With firms' permission, government agencies aggregate their data to their local servers. The data from each government agency in each province is aggregated into a central server that is usually managed by a third-party company. The company cleans the data and builds an interface for the banks to retrieve the data.

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 agreed to share the data. 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 updated every year. As for the amount of data shared, since more than 80% of all firms existing in the credit registry allowed sharing the data and the banks usually have smaller market shares, more than 85% of the information was from non-borrowers for all banks. The sudden increase in the data volume serves as a laboratory to study the effects of surges in the amount of data on lending activities.

The data-sharing program is similar to but not entirely the same as 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 hard information is shared, not soft information. Usually, for other types of credit-registry expansion, both hard information and soft information 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 shares banks' soft information as well as 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[4]. 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

[4]See Babina et al. (2024) for a discussion of open banking around the world.

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 the credit registry information from three provinces that have rolled out the policy experimentation. The loan-level information includes the interest rate, maturity, loan volume, 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. I also match the loan level information with firm information, mainly balance sheet information, from the administrative agencies.

B. Identification

1. Policy Experimentations

The identification strategy relies on policy experimentation concerning provincial government agencies' data-sharing practices. Policy experimentation has been a frequent means of facilitating institutional innovation in China since the 1980s. In general, policy experimentation comes as small-scale explorative pilot projects. In these projects, the local bureaucratic authorities roll out a potentially innovative policy on selected experimental units for a given period. The experimental units could be locations (districts, cities, provinces, etc.) or economic agents (firms, banks, etc.). The selected experimental units, conditional on the targeted population, generally should satisfy the following criteria:

1. Have institutional arrangements for bearing residual risks

2. Have differentiation to ensure that the outcomes are representative after widespread implementation
3. Have legally feasible plans for risk disposal and recovery

The first and third points, in most cases, exclude small and under-developed subjects that do not have the potential to enact the policy and, once failed, do not have enough resilience to recover. The second point ensures randomization so that the experimental (treated) units are representative of the targeted population, conditional on points one and three. Therefore, policy experimentations are akin to randomized control trials within selected populations.

Depending on the effectiveness of the experimented policy initiative, the local authorities then decide whether to scale it up to the whole city or province level. If the policy sees positive responses locally, the central authority would then roll out the policy nationally. The bottom-up gradualism in policy innovation has been deemed as effective for China's economic growth[5].

2. Pilot Projects of Data Sharing

The sample is based on three provinces that rolled out the data-sharing policy experimentation from 2015 to 2018. The experimentation lasted for two years. Afterwards, data were shared with all registered banks. The sampling period is a total of four years for each province, with two years respectively before and after the experimentation. When selecting the experimental units, local authorities first excluded banks that had a market share in the medium and small business loan market below 1%. This roughly kept the top 15 banks in each province ranked by total assets. In total, there were 43 banks left. I define these 43 banks as the participating subjects in the setting here. The authorities in each province then separated the banks into three or four groups based on total assets and arguably randomly selected one from each group as the experimental unit. I therefore define the selected banks as the treatment group, and the rest from the 43 participating subjects as the control group. In the end, there are 10 treated banks and 33 control banks. The credit markets in the three provinces are relatively concentrated.

[5]For examples discussing China's policy experimentation and gradualism, see Heilmann (2008) and Brunnermeier et al. (2017)

Loans from the 43 registered commercial banks contribute to over 80% of the total lending volume in these provinces[6]. However, since the selection of the experimental units was not through a random number generator, there could be intentional and unintentional human discretion across different dimensions. To explore this possibility, I compare the pre-experiment averages of borrower, bank, and loan characteristics and the pre-trends of the variables of interest between the two groups in later sections.

To explore the randomness of the treatment group assignment, I study whether there is any statistical difference in the observable information between the control and treatment groups before the experiment at the loan, firm, and bank levels, respectively in panels A, B, and C of Table 1. In each panel, sub-panels 1 and 2 give the averages for the treatment group and the control group, with standard deviations in parentheses; in sub-panel 3, I present the difference in the average between the two groups together with the t-statistics testing if the differences in the averages are statistically different from zero. All variables are winsorized at 1% level within each year.

The average loan has a volume of around 9.5 million CNY with an average interest rate of around 5.80% and a maturity of 27 months. In addition, 3.3% of the loans defaulted in the end. The average size of the firms is around 65 million CNY. The average profitability (gross profit over total assets) of the firms is 7%. The average leverage, calculated as total debt over total assets, is 0.47. The average gross profitability of the participating banks is around 1.2% and the average deposit-to-asset ratio is around 67%. For all characteristics, the differences in the averages between the control and the treatment groups are insignificant, validating the success of the randomization of the process.

IV Results

A. Loan Attributes

I first present some basic results about the effects of the data-sharing experimentation on loan attributes. Figure 2 plots the evolution of the various loan attributes between

[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

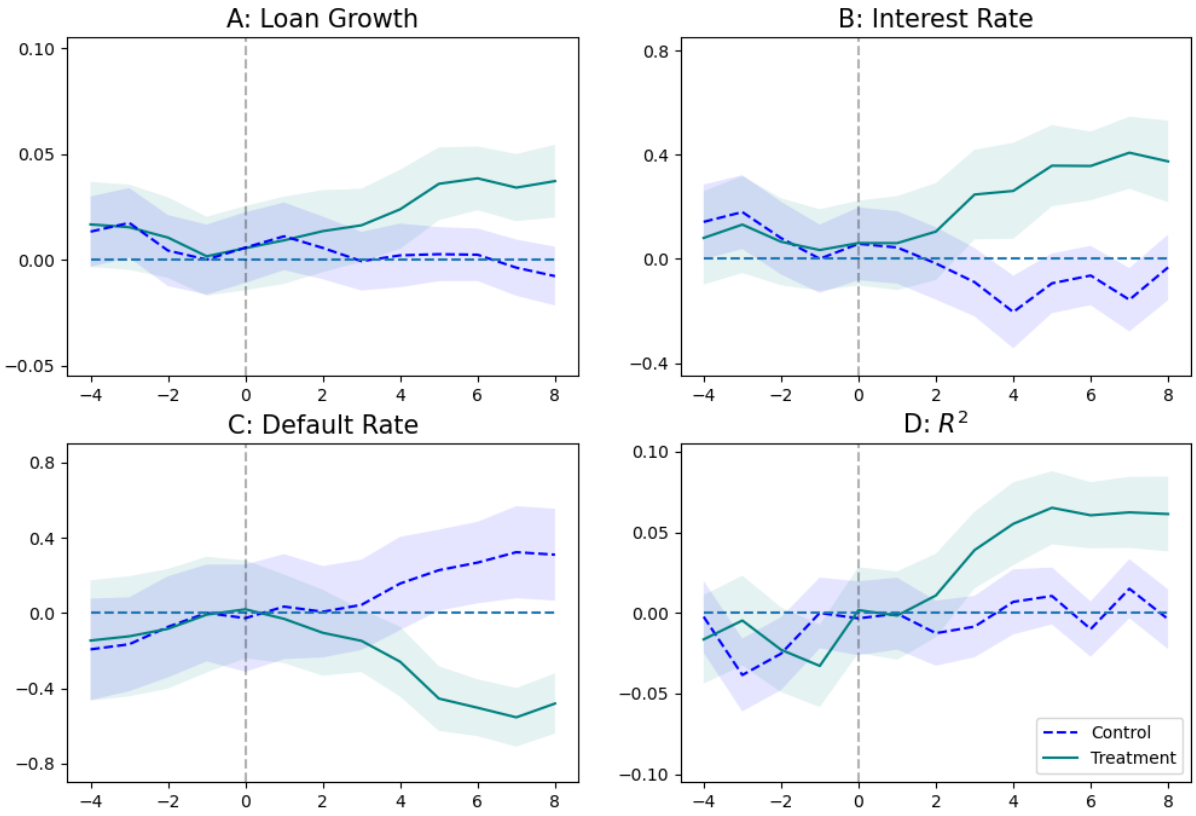
Each panel except for panels A3, B3, and C3 gives the averages and associated standard deviations. In panels A3, B3, and C3, the parentheses contain the t-statistics of the t-tests of differences in mean. In Panel A, Volume is the amount of each loan in 10-thousand CNY. Maturity is the loan maturity in months. Interest Rate is the interest rate (%) of the loan. Default is the percentage of the defaulted loan per year. In Panel B, AT is the average of the borrowers' total asset measure in 10-thousand CNY. Profit is the net profit over total assets (%). Leverage is the total debt outstanding over total assets. The averages are weighted by loan volume. All variables in Panel C are scaled by total asset (%). All variables are winsorized at 1% level by year-quarter. Sample is based on information before data-sharing.

	(1)	(2)	(3)	(4)	(5)
Panel A: Loan					
	Volume	Maturity	Interest Rate	Defaulted	Nobs
A1: Treatment	924.33	27.08	5.73	3.32	495,307
	(737.26)	(6.91)	(1.47)	(18.17)	
A2: Control	975.92	27.24	5.82	3.23	152,387
	(811.12)	(7.29)	(1.61)	(17.94)	
A3: Difference in Mean	-51.59	0.16	0.09	-0.09	
	(-0.21)	(0.76)	(1.01)	(-0.59)	
Panel B: Borrowers					
	AT	Profit	Leverage	Emp	Nfirms
B1: Treatment	6462.04	6.20	48.00	144.39	119,826
	(8901.22)	(29.00)	(41.00)	(176.92)	
B2: Control	7137.23	8.10	47.30	121.39	36,283
	(9312.10)	(34.00)	(81.08)	(153.86)	
B3: Difference in Mean	-675.19	1.90	-0.70	-23.00	
	(-1.58)	(0.60)	(-0.34)	(1.55)	
Panel C: Banks					
	Profit	Capital	Deposit	Cost	Nbanks
C1: Treatment	1.22	2.14	66.61	1.50	33
	(1.21)	(4.16)	(20.35)	(1.23)	
C2: Control	1.23	2.19	67.93	1.53	10
	(0.90)	(3.85)	(18.81)	(1.62)	
C3: Difference in Mean	0.01	0.05	1.32	0.03	
	(0.08)	(0.22)	(0.59)	(0.20)	

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 quarters, labeled as time 0. For each panel, I vertically shift the plot by subtracting all values from the value of the control

Figure 2. 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. 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 quarters. 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.



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. All plots are residualized by firm fixed effects, bank fixed effects, and year-quarter fixed effects.

The figure shows some clear patterns. First, there is no distinguishable difference in the pre-trends between the control and treatment groups for all three characteristics. Second, there is a clear diverging pattern between the treatment and control groups after the data-sharing event, starting from around six months after sharing the data. This confirms the exogeneity of the event. After the experiment, the loans from the treated

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. 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	0.05**	0.05**	0.24**	0.31***	-0.44***	-0.23**
	(0.02)	(0.02)	(0.11)	(0.11)	(0.13)	(0.11)
Observations	647,694	416,373	647,694	416,373	647,694	416,373
R^2	0.07	0.19	0.05	0.20	0.05	0.16
Year-Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes

Standard Errors Clustered at Firm Level in Parentheses
*** p <0.001, ** p <0.01, * p <0.05

banks see a higher average volume, a higher average interest rate, and a lower default rate.

Table 2 gives the difference-in-difference (DID) estimates. All columns include year-quarter fixed effects and bank fixed effects. In the even columns, I further include firm fixed effects. All regressions are weighted by loan volumes. The results are similar regardless of the firm fixed effects. After the data-sharing event, the loans from the treated banks to firms have a 5% higher loan volume, 31 basis points higher interest rates, and 23 basis points lower chances of defaulting. In Table C.1 in the online appendix, I present results with equal weights. The results are similar to those in Table 2.

One concern of Table 2 is that the borrowers might change their lending relationship between the control and the treatment groups. Therefore, the experimentation also affects the control groups through changing borrower composition. To study this possibility, in Table 3, I control for bank-firm fixed effects to hold the lending relationships constant. While the number of observations is reduced by around 35%. The results are similar. This means that for a firm that has borrowed both from bank A and bank B before the experiment. Suppose bank A is now treated, then the firm's loan term from bank A, as compared with bank B, has a higher volume, higher interest rate, and a lower likelihood to

TABLE 3. Loan Terms Outcomes with Firm-Bank FE

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. 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	0.06***	0.04*	0.28**	0.39**	-0.29**	-0.04
	(0.02)	(0.02)	(0.16)	(0.16)	(0.13)	(0.13)
R^2	0.37	0.27	0.25	0.38	0.26	0.14
Observations	276,832	276,832	276,832	276,832	276,832	276,832
Equally weighted	No	Yes	No	Yes	No	Yes
Year-Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm×Bank FE	Yes	Yes	Yes	Yes	Yes	Yes

Standard Errors Clustered at Firm Level in Parentheses
*** p <0.001, ** p <0.01, * p <0.05

default. While it seems puzzling that the same firm has different probabilities of defaulting to different banks, this is however not a result of firms' strategic behaviors. In the even columns of Table 3, I fit the regression with equal weights. While the effects on other characteristics are similar, there is no significant difference in the default rate anymore. Therefore, the lower default rate after controlling for bank-firm fixed effects comes from treated banks initiating loans with higher volumes to safer borrowers.

Another concern is that data-sharing will change the strategic behaviors of the borrowers. For example, when default information is shared with all banks, borrowers could choose to take less risk because default choices cannot be concealed from other lenders anymore. Therefore, data-sharing acts as a discipline device to keep borrowers from defaulting (Diamond, 1984; Padilla and Pagano, 2000). To test this possibility, I refit Table 2 while controlling for firm×period fixed effects where period is a dummy for post-experiment. The results, which are in Table C.2, show that the effects are similar if controlling for firm×period fixed effects. This indicates that changes in firm behavior after the experimentation are unlikely reasons for the changes in loan performance.

The decrease in default rate potential suggests that data-sharing reduces asymmetric information. However, 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 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. Therefore, perfect competition indicates a positive relationship between changes in interest rates and changes in default rates. This is inconsistent with the findings here.

However, the movement of interest rates and default rates can be opposite if one allows for imperfect competition. For example, Buchak et al. (2018) and 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. I continue to study how data-sharing affects supply and demand separately.

B. Screening Ability

I first study the effects of data-sharing on bank screening ability. Following Iyer et al. (2016), I study 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. (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. [7]

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

TABLE 4. 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). All variables are winsorized at 1% level by year-quarter.

	Control		Treament		
	(1) Before	(2) After	(3) Before	(4) After	(5) DID
	Panel A: Logistic Regression				
Score	1.27	1.29	1.27	1.38	
	(0.02)	(0.02)	(0.01)	(0.01)	
Pseudo R^2	11.33%	11.05%	11.27%	16.17%	5.18%
					p-value = 0.00
	Panel B: ROC				
AUC	0.7501	0.7487	0.7535	0.8097	0.0576
	(0.0091)	(0.0078)	(0.0079)	(0.0083)	p-value = 0.00
N	136,385	132,563	71,497	75,928	

Panel D of Figure 2 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 event's effects on bank screening ability quantitatively. The results are in Table 4. 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 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 11% for both control and treated banks before the experiment. After the experiment, the pseudo-R^2 of the treated banks increases from 11.27% to 16.17%, while that for the control banks stays roughly constant. The difference-in-difference (DID) estimate, which is residualized by firm, bank, and year-quarter 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 5.18% is both statistically and economically significant.

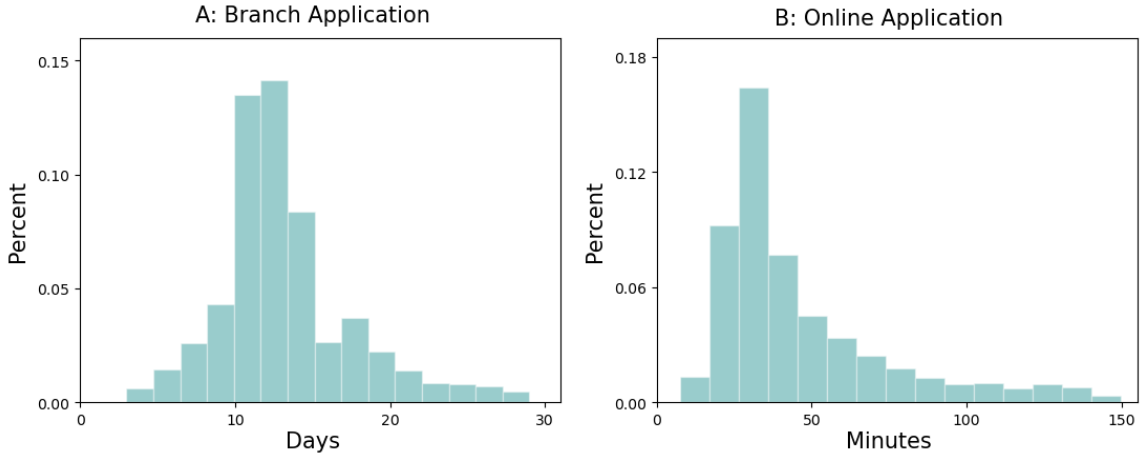
The ROC curves provide a more formal way to compare the predictive power between the control and treated banks before and after the experiment. Panel B gives the associated AUC of the logistic regression. I find that the AUC is around 0.75 for both control and treated banks before the experiment. After the experiment, the AUC of the treated banks increases from 0.7487 to 0.8097, 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 $(0.8097 - 0.5)/(0.7535 - 0.5) \approx 1.22$. This is to say, treated banks' risk scores achieve 22% 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. Therefore, the results show that data-sharing has greatly increased the screening ability of the treated banks.

C. Demand for Convenience

Apart from reduced asymmetric information, data-sharing enables banks to offer products with better quality or convenience. As shown by Buchak et al. (2018) and Fuster et al. (2019), fintech firms could increase loan demand by offering a faster and more convenient loan origination process. While Buchak et al. (2018) and Fuster et al. (2019) focus on the mortgage market, convenience, including loan speed, is especially valuable for small/medium business borrowers. For example, in a recent survey, Wiersch et al. (2019) shows that the most frequently cited challenges for businesses in the US to borrow from traditional bank lenders were the cumbersome application process and long wait times for credit decisions. Therefore, suppose borrowers value faster loan origination time. Then data sharing could affect demand in addition to affecting screening ability. The increased demand due to improved quality will increase interest rates. While there are potentially many factors through which better information processing ability affects loan quality, in this section, I assess the event's effects on quality through two specific margins: the proportion of online applications and loan origination time to shed light on the demand channel.

Figure 3. 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.



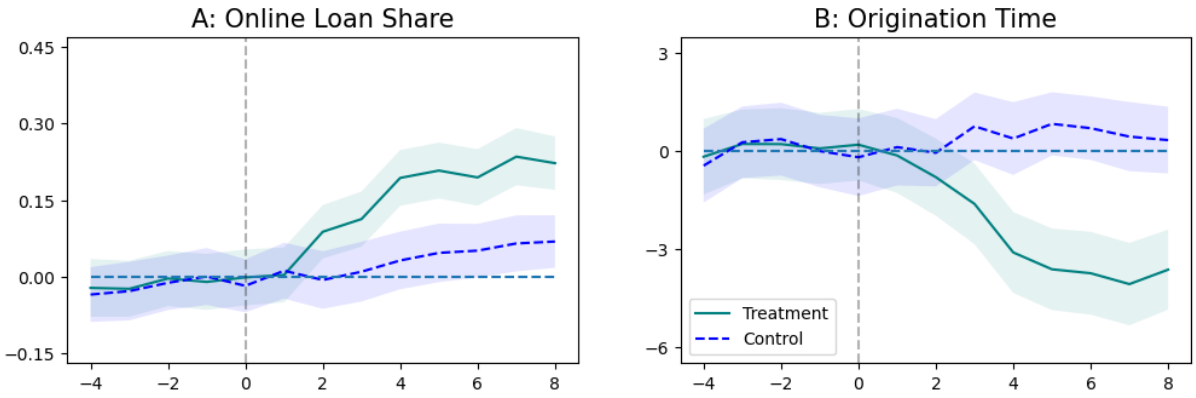
A key difference between branch applications and online applications is the time it takes to get the funding. Given a completely digital loan origination process, online applications take a much shorter time to receive funding. Figure 3 plots the distribution of the time it takes to receive funding starting 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[8].

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 to recognize borrower types more accurately increases with the amount of data, which is expected

[8]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 from all other countries, a useful comparison is to study the time it takes to get a loan from SBA and fintech platforms in the US like LendingClub. Recent reports from lendingClub (2023) and Bankrate (2024) 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 of the convenience difference between traditional lending practice and the new online practice.

Figure 4. Changes in Origination Time

Panel A gives the share of loans originating through online applications, and panel B 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 quarters. 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.



to reduce 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 branch labor costs. Panels A and B of Figure 5 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 5 gives the ATE of the events on online application share and origination time. In particular, treated banks have 20% more loans originating through online applications after the experimentation. Accompanied by it, the treated banks take around 4 days less to extend the loans. Given an average of 14 days to extend the loan, this is equivalent to a 29% 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 6. While sorting on characteristics after the experiment potentially induces selection issues, this exercise nonetheless offers insights about how changes in interest rates comove with changes in origination time before and after the

TABLE 5. 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. All variables are winsorized at 1% level by year-quarter.

	(1) Online%	(2) Online%	(3) Times	(4) Times
Treat	0.31**	0.20**	-4.17***	-3.96***
	(0.13)	(0.07)	(0.94)	(0.62)
Observations	647,694	416,373	647,694	416,373
R^2	0.03	0.18	0.07	0.19
Year-Qtr FE	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes

Standard Errors Clustered at Firm Level in Parentheses
*** p <0.001, ** p <0.01, * p <0.05

experiment. In particular, 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 perceived safer borrowers. Specifically, faster origination comes with 23 basis points higher interest rates than slower origination when including both high-quality and low-quality borrowers. Therefore, the data-sharing event could affect demand by allowing banks to supply high-quality funds, e.g., faster origination time, which increases demand from both high-quality and low-quality borrowers.

While origination speed could affect credit demand, it is not the only potential channel that shifts demand. From column (3) of Table 6, the average interest rate increases by 15 basis points even for borrowers with a slower loan origination speed. This indicates that data-sharing increases the monopolistic power of treated banks through other margins. Possible examples include that treated banks can offer more online applications. Even controlling for time, this could greatly reduce the pecuniary or non-pecuniary costs of applying for a loan from branches. In addition, better information could enable banks to have more differentiated products and increase price discrimination

TABLE 6. 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. All variables are winsorized at 1% level by year-quarter.

	(1)	(2)	(3)	(4)	(5)	(6)
	Slower Origination			Faster Origination		
	Riskier	Safer	All	Riskier	Safer	All
Treat	0.55**	-0.32***	0.15*	0.65***	-0.05	0.38***
	(0.22)	(0.11)	(0.08)	(0.22)	(0.09)	(0.08)
Observations	69,341	73,329	149,356	124,824	133,182	267,003
R^2	0.298	0.276	0.055	0.241	0.103	0.175
Year-Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	Yes	No	Yes	Yes
Bank FE	No	Yes	Yes	No	Yes	Yes

Standard Errors Clustered at Firm Level in Parentheses
*** p <0.001, ** p <0.01, * p <0.05

or steer those who demand non-interest rate related characteristics into more expensive products (Benetton et al., 2022). While it's hard to dissect all possible factors that shift demand, I further explore the effects of data-sharing on loan pricing through borrower preference in Section V.

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 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

TABLE 7. 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). 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	10.48%	10.93%	11.73%	11.44%	
	Panel A2: ROC				
AUC	0.7418	0.7527	0.7635	0.7635	
	(0.0113)	(0.0088)	(0.0108)	(0.0098)	
N	55,819	60,255	75,054	77,820	
	Panel B:Treatment				
	Panel B1: Logistic Regression				
Pseudo R^2	10.65%	13.48%	11.60%	19.06%	5.35%
					p-value = 0.00
	Panel B2: ROC				
AUC	0.7597	0.7802	0.7732	0.8285	0.0457
	(0.0230)	(0.0229)	(0.0126)	(0.0133)	p-value = 0.00
N	31,059	32,813	40,509	43,044	

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 7, 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 for those with low IT spending. The pseudo-R^2 for the low-IT group in the treatment group is around 3 percentage points higher after the experiment. While for high-IT banks, the pseudo-R^2 increases by around 7.5 percentage points. Residualized by fixed effects, the TD estimate for the changes in high-IT treated banks compared with the low-IT treated banks is 5.35% and is both statistically and economically significant. Similarly, for the treated banks, the AUC increases from 0.7597 to 0.7802 for the low-IT banks. Meanwhile, for high-IT banks, the AUC increases from 0.7732 to 0.8285. This increase is equivalent to a 20% improvement in the predictive accuracy, compared with an 8% higher predictive accuracy for the low-IT group.

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:

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

where $Y_{j,k,t}$ are various loan attributes. λ_j, λ_k, and λ_t are the firm fixed effects, bank fixed effects, and 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 before the experiment. 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 8. After providing the data, banks with higher IT intensity see a 6% increase in average loan volume, 40 basis points increase in interest rate, 44 basis points lower default rates, 26% more loan origination from the online platforms, and 4.37 fewer days for loan origination. While in general, the effects of data-sharing mostly 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 7 and 8 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.

TABLE 8. 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. 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	(8) Online%	(8) Online%	(9) Times	(10) Times
Treat	0.03	0.01	0.11	0.06	-0.08	0.16*	0.04	0.03	-0.31	-0.05
	(0.02)	(0.02)	(0.13)	(0.18)	(0.05)	(0.08)	(0.07)	(0.06)	(0.60)	(0.06)
Treat×High IT	0.06***	0.05***	0.21***	0.34***	-0.51***	-0.60***	0.26***	0.23***	-6.09***	-4.32***
	(0.02)	(0.02)	(0.07)	(0.09)	(0.08)	(0.07)	(0.07)	(0.05)	(0.90)	(0.08)
Observations	647,694	416,373	647,694	416,373	647,694	416,373	647,694	416,373	647,694	416,373
R^2	0.092	0.243	0.068	0.276	0.089	0.212	0.056	0.241	0.096	0.264
Year-Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Standard Errors Clustered at Firm Level in Parentheses
*** p <0.001, ** p <0.01, * p <0.05

E. Cream-Skimming of High-IT Lenders

The effects of data-sharing are positive on screening ability, and the larger effects for high-IT banks suggest that treated banks 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 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[9]. 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

[9]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.

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 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, high-IT banks are more likely to make more loans

to high-quality borrowers and fewer loans to low-quality borrowers. The results support the hypothesis that 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 the effects of data-sharing when only some banks are affected. However, when the data is shared with more banks, the competitive advantages will be attenuated. 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 are three markets in the economy, and each market represents one province[10]. 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:

$$\begin{aligned} U_{j,k,t} = \alpha_0 + \mathbf{X}_{k,t}\beta + \xi_{j,t} + \alpha_r r_{j,k,t} + \alpha_i IT_k \times I_{k,t} \\ + \alpha_Z Z_{j,k,t} + \alpha_{r,Z} r_{j,k,t} \times Z_{j,k,t} + \alpha_{i,Z} IT_k \times I_{k,t} \times Z_{j,k,t} \\ + \mathbf{Y}_{j,k,t}\eta + \epsilon_j + \nu_{j,k,t}, \end{aligned}$$

[10]In China, business loan markets are usually defined at the province level.

where $\mathbf{X}_{k,t}$ is a vector of bank-year determinants of demand, $r_{j,k,t}$ is the interest rate offered by bank k to firm j in year t, $Z_{j,k,t}$ 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. ϵ_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 α_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 not borrowing from any of the banks active in the market at year t[11]. Firms choose their main banks to borrow from the bank that maximizes their utility, or else they choose not to borrow at all ($k = 0$).

An important demand shifter is $IT_k \times I_{k,t}$, which captures the interaction between bank IT spending and the data-sharing event. Therefore, α_i captures firm j's preference for applying for a loan from high-IT banks that have been shared with the data. There are several possibilities for this preference. First, as shown in Section IV.C, borrowers could prefer banks that can utilize the data better as it increases convenience during the loan application process. However, as shown by column (3) of Table 6, demand is likely higher even for loans that do not originate faster. Therefore, I use $IT_k \times I_{k,t}$ to capture the total effects of the interaction between data sharing and IT capacity on demand.

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

$$\begin{aligned} U^D_{j,k,t} = \alpha^D_0 &+ \mathbf{X}_{k,t}\beta^D + \alpha^D_r r_{j,k,t} + \alpha^D_i IT_k \times I_{k,t} \\ &+ \alpha^D_Z Z_{j,k,t} + \alpha^D_{r,Z} r_{j,k,t} \times Z_{j,k,t} + \alpha^D_{i,Z} IT_k \times I_{k,t} \times Z_{j,k,t} \\ &+ \mathbf{Y}_{j,k,t}\eta^D + \epsilon^D_j, \end{aligned}$$

where ϵ^D_i represent firm j's unobserved 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

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

default. That is, I assume that ϵ_j and ϵ_j^D are distributed following a bi-variate normal distribution:

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

A positive correlation between the firm-specific unobservables driving demand and default (ρ) is evidence of adverse selection: when ρ is positive, firms with a higher unobservable propensity to demand credit are also more likely to default. At the same time, a positive α_r^D 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 α_r^D to imply moral hazard builds on the assumption that α_r^D is estimated by the component of price variation that is orthogonal to firms' unobservable characteristics, so that α_r^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).

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

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

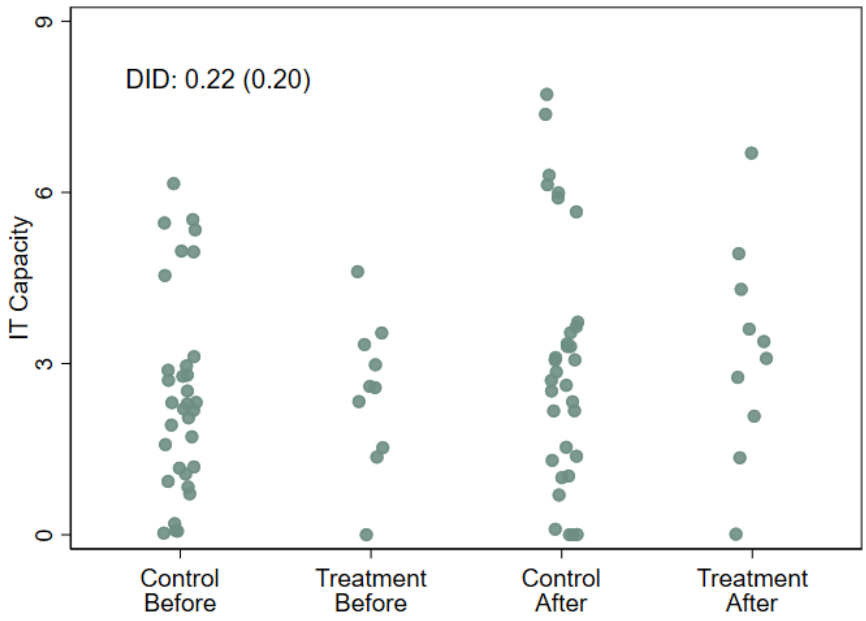
In (2), $\tilde{D}_{j,k,t} = \tilde{d}_{j,k,t}(1 - R_{j,k,t})$, where $\tilde{d}_{j,k,t}$ is firm j's default probability and $R_{j,k,t}$ is the recovery rate in case of default. $q_{j,k,t}$ is the probability of application, and $c_{j,k,t}$ is the marginal costs of supplying the loan. Marginal cost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

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

where ψ_j, ψ_k, and ψ_t are respectively the firm, bank, and year fixed effects. In addition, I allow marginal costs to depend on the interaction between IT capacity and data-sharing.

Figure 6. IT Capacity

This figure gives a scatter plot of the IT capacity, defined as the average of IT-related spending (including hardware, software, and labor) to non-interest expenses. The DID gives the diff-in-diff estimates. The parenthesis gives the standard error.



This captures the ability of data-sharing to reduce the cost of initiating a loan (e.g., potentially through the automatic lending system or reduced labor costs). Similar to Einav et al. (2012), I assume that 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 risks as observed by banks.

Note that I do not allow banks to change IT investments. That is, IT_k is fixed at the bank level. This assumption might be violated if banks could choose to invest more in IT after receiving the data. To verify this possibility, I plot the average IT capacity for the banks before and after the experiment in Figure 6. As shown, there is a 0.22 percentage point increase in IT capacity for the treated banks after the experiment relative to the control group. However, the change is insignificant. This is expected as investment is lumpy and is expected to have rigidity within a short period.

The first-order condition of maximizing (2) yields

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

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. Estimation

1. Demand and Default

The estimates of the structural model are presented in Table 9[12]. As shown, a significantly negative relationship exists between interest rate and loan demand. In addition, a positive number of ρ and α_r^D indicates the existence of adverse selection and moral hazard. The coefficient in front of $Treat \times I$ in the demand equation is significantly positive. This implies that borrowers prefer banks with higher IT capacity that are shared with the data. Therefore, the data-sharing increased demand significantly. However, the coefficient in the default equation is insignificant. When $Treat \times I$ captures the demand for increased convenience, this result is similar to Fuster et al. (2019) such that faster or more convenient 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

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

TABLE 9. 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	-0.42	0.40
	(0.14)	(0.07)
Interest Rate × Relationship	-0.67	0.24
	(0.25)	(0.03)
IT × Data	0.85	0.08
	(0.10)	(0.10)
IT × Data × Relationship	-0.30	0.03
	(0.06)	(0.13)
log(Distance)	-0.23	-0.46
	(0.05)	(0.07)
log(AT)	-0.05	-0.69
	(0.11)	(0.15)
log(Volume)	3.83	-0.21
	(0.12)	(0.08)
Age	0.02	0.08
	(0.41)	(0.29)
Profitability	0.00	-2.21
	(0.32)	(0.59)
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	2,419,668	403,278
Covariance Matrix	$\sigma = 0.27$	
	(0.09)	
	$\rho = 0.35$	$\sigma_P = 1$
	(0.07)	

relationship are less sensitive to data-sharing. 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, are less affected by the increased convenience or other demand shifters due to better lending technology to process larger amounts of data.

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, data-sharing does not significantly affect default probability.

2. The Effects of Data-Sharing on Credit Scores

In general, 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, while for those that are not treated, optimal credit scores after sharing the data are not observed. To study the counterfactual scenario where all banks are shared with the data, I construct a measure of the *optimal* post-experiment risk-scoring technology, 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 banks' post-experiment risk scores to predict the three-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 considered 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.

TABLE 10. The Effects of Data-Sharing

This table presents the relationship between data-sharing and changes in marginal costs, $c_{j,k,t} - \overline{c}_{j,k,0}$. $I_{k,t} = 1$ is bank k in time t has received the data. $\Delta s_{j,k,t}$ is the changes in the firm j's optimal risk score as in (6).

	(1) $c_{j,k,t} - \overline{c}_{j,k,0}$	(2) $c_{j,k,t} - \overline{c}_{j,k,0}$	(3) $c_{j,k,t} - \overline{c}_{j,k,0}$	(4) $c_{j,k,t} - \overline{c}_{j,k,0}$
$I_{k,t} \times IT_k$	-0.05*	-0.06*	-0.06*	-0.07*
	(0.03)	(0.04)	(0.03)	(0.04)
$I_{k,t} \times IT_k \times \Delta s_{j,k,t}$	0.12***	0.10***	0.14***	0.15***
	(0.02)	(0.03)	(0.03)	(0.04)
$\Delta s_{j,k,t}$			0.04*	0.04*
			(0.02)	(0.02)
Firm FE	No	Yes	No	Yes
Bank FE	No	Yes	No	Yes
Year FE	No	Yes	No	Yes
N	403,278	403,278	403,278	403,278

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

Directly estimating the marginal-cost equation (3) is difficult as it requires the separate identification of κ_2 and κ_3. Since I don't observe $p_{k,t}$, κ_3 cannot be identified directly. Instead, to estimate the marginal-cost equation, I combine (3) and (5), and express (3) as

$$c_{j,k,t} = \overline{c}_{j,k,0} + \kappa_1 \times treat_{j,t} \times I_j + \tilde{\kappa}_2 \times treat_{j,t} \times I_j \times \Delta s_{j,k,t} + \overline{\psi}_0 + e_{j,k,t} - \overline{e}_{j,k,0}, \quad (6)$$

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. $\tilde{\kappa}_2 = \kappa_2 \times \kappa_3$ 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-quarter fixed effects and structural errors before the experiments. With (6), I directly estimate κ_1 and $\tilde{\kappa}_2$. Then I can explore counterfactuals using (6). To estimate κ_1 and $\tilde{\kappa}_2$, I fit the following DID specification:

$$c_{j,k,t} - \overline{c}_{j,k,0} = \lambda_j^c + \lambda_k^c + \lambda_t^c + \tilde{\kappa}_1 \times treat_{j,t} \times I_j + \tilde{\kappa}_2 \times I_k \times treat_{k,t} \times \Delta s_{j,k,t} + e_{j,k,t}^c. \quad (7)$$

Table 10 gives the estimates (7). Consistent with previous results, columns (1) and (2) show that there is a strong negative interaction effect of data-sharing and IT intensity on loan origination time and risk-based pricing. Specifically, focusing on column (2), given $kappa_2 = -0.06$ and average IT intensity equals 3.3%, data sharing decreases the marginal cost of the bank with the average amount of IT intensity to lend to firms with no changes in credit score by around 20 basis points. $kappa_2 = 0.10$ 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. In columns (3) and (4), I control for the main effects of changes in credit score. The effects of data sharing on marginal costs are similar.

C. Model Fit

Panels A and B of Table 11 show that the model is effective in matching the equilibrium moments in the data. Before the experiment, the model generates an average default rate of 3.27% and an average interest rate of 5.67%, compared with 3.25% and 5.65% in the data. Effective marginal cost is on average 3.56%. This indicates an average effective markup of 2.11%. After the experiment, the average default rate and average interest rate from the model are respectively 5.74% and 3.08%, compared with 3.16% and 5.65% in the data.

D. 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, 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 screening ability and demand channels is at work. The results are in Panel C of Table 11. 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

TABLE 11. 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	(6) % Diff	(5) Effective MC	(6) % Diff	(7) Effective Markup	(8) % Diff
A: Pre-Experiment	Data		3.25		5.65					
	Model	All	3.27		5.67		3.56		2.11	
		High IT	3.12		5.47		3.21		2.26	
		Low IT	3.66		5.65		3.78		1.88	
B: Post-Experiment	Data	All	3.16		5.65					
	Model	All	3.08		5.74		3.57		2.17	
C: Counter-Factual	Both	All	3.11	-4.93%	5.63	-0.77%	3.27	-8.12%	2.36	11.61%
	Supply	All	2.93	-10.55%	5.32	-6.25%	3.08	-13.39%	2.24	5.77%
	Demand	All	3.36	2.71%	5.99	5.64%	3.56	-0.01%	2.43	15.16%
D: Heterogeneity		High IT	2.86	-8.38%	5.63	2.85%	3.07	-4.51%	2.56	13.33%
		Low IT	3.82	4.44%	5.64	-0.26%	3.69	-2.41%	1.95	4.08%

are the corresponding percentage changes with respect to the pre-experiment level as estimated by the model. Interestingly, while the results from Section IV show large effects of data-sharing on the default rate and interest rate of the treatment banks, the equilibrium outcomes when all banks are shared with the data are weaker. Specifically, the average default rate decreases by 16 basis points, and the average interest rate decreases by 4 basis points. The average reduction in marginal cost is larger and is 29 basis points as compared with the pre-experiment level. The results are expected because, when more banks are shared with the data, the increased competitive advantages for the treatment banks are weakened.

2. Heterogeneity by IT Intensity

An important heterogeneity of the effects of data availability is banks' IT intensity. From the estimates in Table 9 and Table 10, the effects of data-sharing are expected to increase the profitability of 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 shared with all banks

by their IT intensity. The results are in Figure 7. Panel A gives the case when data-sharing only affects borrower demand; Panel B gives the case when data-sharing only affects marginal costs; and Panel C gives the case when data-sharing affects both the supply and demand channels. The plots show a strong positive interaction effect of data-sharing and IT intensity on interesting rates when data-sharing only affects origination time. Specifically, for high-IT banks, a larger amount of firm data enables the bank to reduce origination time more, therefore facing a higher demand and, thus higher interest rates. However, the interaction effects of data-sharing and IT intensity on default rate are only slightly positive.

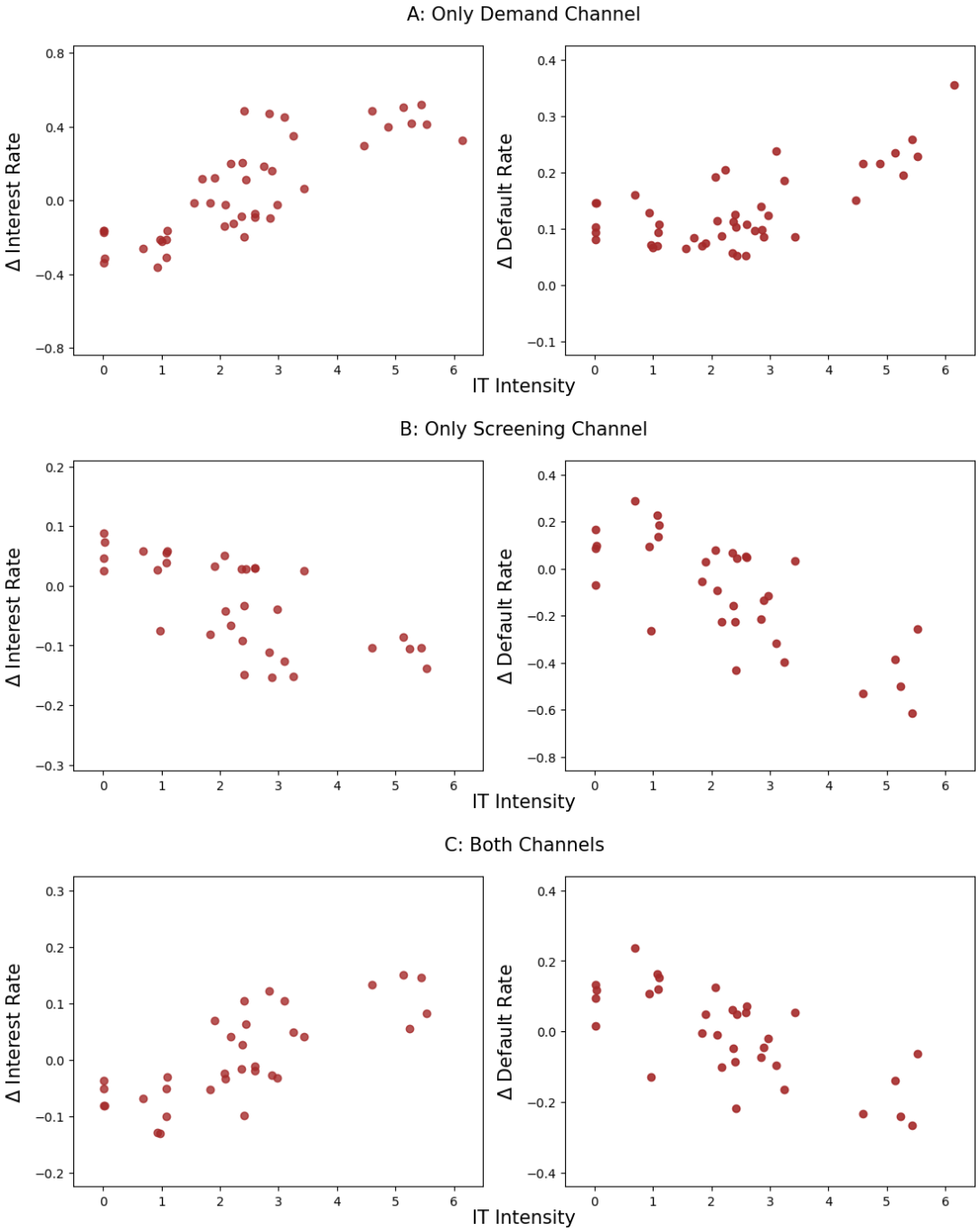
From panel B, when data-sharing only affects screening ability, banks with higher IT spending extend more loans to borrowers with lower default rates. This results in a much steeper relationship between IT intensity and default rate. At the same time, looking at the left panel, the relationship between IT intensity and changes in interest rates, though also negative, is much flatter than that from Panel A.

Panels A and B of Figure 7 shed light on the relative strengths of the supply channel and demand channel, respectively, on default rates and interest rates. For high IT banks, 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 dominate the decreases in interest rates because of extended loans 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 intensities quantitatively, I give the summary statistics of bank profitability in Panel D of Table 11. Consistent with the patterns in Figure 7. Data sharing has a much stronger effect on banks with higher IT intensity. Specifically, for banks with high IT intensity, data sharing decreases their default rate from 3.12% to 2.86% while increasing interest rates marginally from 5.47% to 5.63%. At the same time, effective marginal cost decreases from 3.21% to 3.07%. In the end, banks with high IT intensity see a 13.33% increase in the effective markup. On the other hand, data-sharing has little effect on banks with low

Figure 7. Counterfactuals

This figure gives the changes in interest rates and default rates 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 demand channel is shut down. Panel C analyzes the case when both demand and screening channels operate.



IT intensity. Across all dimensions, the changes are economically insignificant. In the end, banks with low IT capacity see a 4.08% increase in the effective markup.

To sum up, the counterfactual exercises confirm the conjecture in the experiment: big data is expected to increase loan demand through preferences for banks with better technology and decrease default rates through better screening ability. The effects are larger only for banks that have high IT capacity. The asymmetric effects of data on banks with different levels 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 in China and structural estimation to shed light on the effects of big data on loan attributes and bank profitability. I show that providing a great amount of hard data to banks extensively increases banks' screening ability. In particular, providing a large amount of hard data decreases default rates through reallocating funds to safer borrowers. Meanwhile, through improving the quality of the loans, the availability a large amount of hard data enables bank to increase demand from the average borrowers. In addition, given the requirement of technology to process a large amount of data, the availability of a larger amount of data has more significant effects on banks with high 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. This is motivated in the context here that, within a two-period period, there is limited evidence of changes in IT capacity. However, banks could adjust their IT spending in the longer run 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., S. A. Bahaj, G. Buchak, F. De Marco, A. K. Foulis, W. Gornall, F. Mazzola, and T. Yu (2024). Customer data access and fintech entry: Early evidence from open banking. Technical report, National Bureau of Economic Research.

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.

Bankrate (2024). How long do you have to wait for SBA loan approval? https://www.bankrate.com/loans/small-business/sba-loan-approval/.

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.

Benetton, M., G. Buchak, and C. R. Garcia (2022). Wide or narrow?: Competition and scope in financial intermediation.

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.

Blickle, K., Z. He, J. Huang, and C. Parlatore (2024). Information-based pricing in specialized lending. Technical report, National Bureau of Economic Research.

Brunnermeier, M. K., M. Sockin, and W. Xiong (2017). China's gradualistic economic approach and financial markets. *American Economic Review 107*(5), 608–613.

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*.

Diamond, D. W. (1984, 07). Financial Intermediation and Delegated Monitoring. *The Review of Economic Studies 51*(3), 393–414.

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.

Forbes (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.

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., 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.

Heilmann, S. (2008). Policy experimentation in chinaâ€™s economic rise. *Studies in comparative international development 43*(1), 1–26.

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.

lendingClub (2023). How Long Does It Take to Get Approved for a Loan? https://www.lendingclub.com/help/personal-loan-faq/how-long-does-it-take-to-get-approved-for-a-loan.

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.

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.

Padilla, A. and M. Pagano (2000). Sharing default information as a borrower discipline device. *European Economic Review 44*(10), 1951–1980.

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.