The Effects of Big Data on Commercial Banks

Xiao Yin

Apr 4, 2024

UCL

Surge of Data

- More than 2.5 mil TB of data created each day (Forbes 2018)
  - volume of data is too great for humans to handle
  - the use of *big data* to extract high-dimensional information (Brunnermeier et al. 2021)

Surge of Data

- More than 2.5 mil TB of data created each day (Forbes 2018)
    - volume of data is too great for humans to handle
    - the use of *big data* to extract high-dimensional information (Brunnermeier et al. 2021)
- Particularly interesting for banking
    - reliance on data collection/processing
- Little work studying the effects of big data on commercial banks

Surge of Data

- More than 2.5 mil TB of data created each day (Forbes 2018)
  - volume of data is too great for humans to handle
  - the use of *big data* to extract high-dimensional information (Brunnermeier et al. 2021)
- Particularly interesting for banking
  - reliance on data collection/processing
- Little work studying the effects of big data on commercial banks
- This paper:
  - a quasi-experiment in China
  - the effects of providing banks with a large amount of firm information

Background

- From 2014: local gov. experimented with sharing data with banks.

Background

- From 2014: local gov. experimented with sharing data with banks.
- Gov. agencies worried about data security issues.
- Some third-party firms were established:
  - gather, store, and clean data
  - share data for a fee
  - take legal responsibility for data security

Background

- From 2014: local gov. experimented with sharing data with banks.
- Gov. agencies worried about data security issues.
- Some third-party firms were established:
  - gather, store, and clean data
  - share data for a fee
  - take legal responsibility for data security
- Identification: the largest data provider's market entry strategy.
  - compare banks the provider contracted and not contracted
  - provider's market share: over 90% from 2014 to 2018

Information Provided

- Data shared:

Data	Data Content	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
		Blacklisting	1. CBRC Blacklisting 2. Petty Loan Blacklisting 3. P2P Blacklisting
		Anti-Fraud	1. Anti-Fraud Information
Judicial Data	1. Information on the Persons subject to Execution 2. Legal Action Information	Credit Registry Data	1. Individual Credit History 2. Business Credit History

- No new characteristics.
    - pre-experiment: auditing companies request information from admin under borrowers' permission.

Information Provided

- Data shared:

Data	Data Content	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
		Blacklisting	1. CBRC Blacklisting 2. Petty Loan Blacklisting 3. P2P Blacklisting
		Anti-Fraud	1. Anti-Fraud Information
Judicial Data	1. Information on the Persons subject to Execution 2. Legal Action Information	**Credit Registry Data**	1. Individual Credit History 2. Business Credit History

- No new characteristics.
  - pre-experiment: auditing companies request information from admin under borrowers' permission.
- Main effect: volume of information
  - $>$ **200 thousand** firms, average 125 characteristics at initial provision
  - information periodically updated

Information Provided

- Data shared:

Data	Data Content	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
		Blacklisting	1. CBRC Blacklisting 2. Petty Loan Blacklisting 3. P2P Blacklisting
		Anti-Fraud	1. Anti-Fraud Information
Judicial Data	1. Information on the Persons subject to Execution 2. Legal Action Information	Credit Registry Data	1. Individual Credit History 2. Business Credit History

- No new characteristics.
  - pre-experiment: auditing companies request information from admin under borrowers' permission.

- Main effect: volume of information
  - $>$ **200 thousand** firms, average 125 characteristics at initial provision
  - information periodically updated
  - *big data*: data with **massive** size, not new information type

Outline

Methodologies

Results

Structural Estimation

Conclusion

Data

- One province where granular data is available
- Sample period: 2014 - 2018
    - two years around data-sharing
- Loan-level data: random 10% from credit registry.
    - loan amount, interest rate, application date, proprietary credit scores, default, etc.
- Firm balance sheets: tax administrative
    - total asset, emp. size, age, etc.
- Data available for 22 banks
    - comprise of $>$ 90% market share

Identification

- Provider's market entry decisions from 2014 to 2018.
- Focusing on data security instead of profits $\Rightarrow$ uniforming pricing.
- Limited resources to monitor all banks
  - one sales team $\Leftrightarrow$ one or two provinces
    $\Rightarrow$ a quota on the N. banks/province.

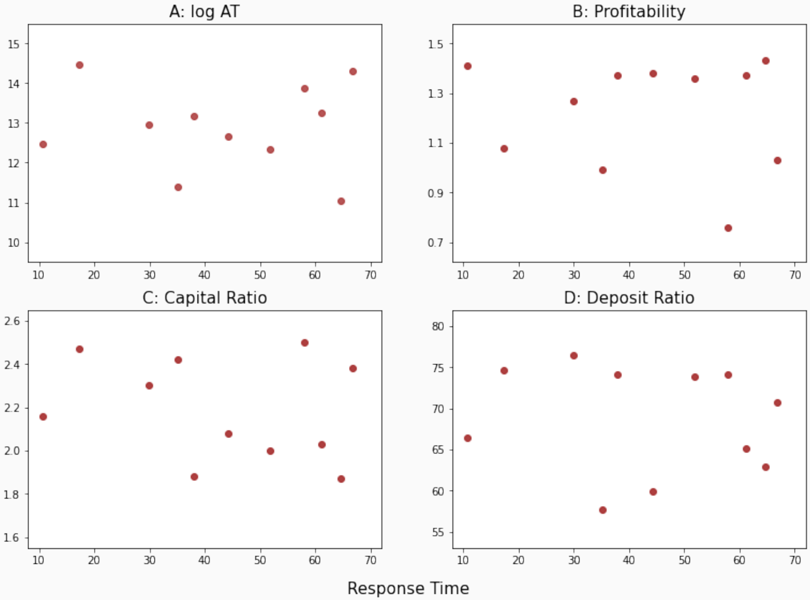
Identification

- Provider's market entry decisions from 2014 to 2018.
- Focusing on data security instead of profits $\Rightarrow$ uniforming pricing.
- Limited resources to monitor all banks
    - one sales team $\Leftrightarrow$ one or two provinces
      $\Rightarrow$ a quota on the N. banks/province.
- Only contracted with a limited number of banks in each province.
    1. excluding very small banks.
    2. the company informed the rest about this opportunity by provinces at once.
    3. made contracts in a first-come-first-serve manner.

Identification

- Provider's market entry decisions from 2014 to 2018.
- Focusing on data security instead of profits $\Rightarrow$ uniforming pricing.
- Limited resources to monitor all banks
  - one sales team $\Leftrightarrow$ one or two provinces
    $\Rightarrow$ a quota on the N. banks/province.
- Only contracted with a limited number of banks in each province.
  1. excluding very small banks.
  2. the company informed the rest about this opportunity by provinces at once.
  3. made contracts in a first-come-first-serve manner.
- Markets defined by provinces
  - excluding very small banks.
  - contracted as treatment, not contracted as control

Exclusion Restriction



Summary Statistics

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

- Parentheses
  - Panels A and B: standard deviations
  - Panels C: *t*-stats

Outline

Screening Ability

- Logistic regression of *ex post* default on *ex ante* proprietary risk scores.

Screening Ability

- Logistic regression of *ex post* default on *ex ante* proprietary risk scores.
- What could go wrong?
    - borrowers change lending relationship $\Rightarrow$ control groups are affected.

Screening Ability

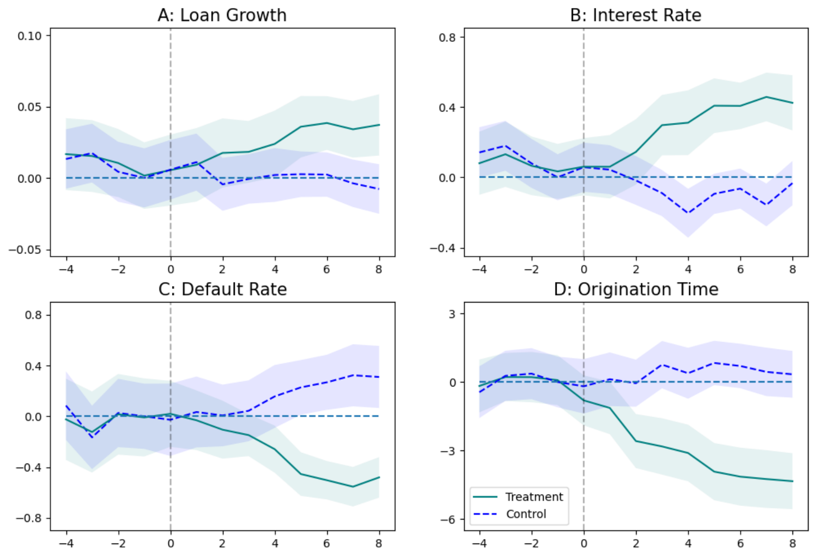
- Logistic regression of *ex post* default on *ex ante* proprietary risk scores.
- What could go wrong?
  - borrowers change lending relationship $\Rightarrow$ control groups are affected.
  - main analysis: control for firm$\times$bank fixed effects
  - holding borrower compositions fixed $\Rightarrow$ only focus on supply-side impact

Screening Ability

- Logistic regression of *ex post* default on *ex ante* proprietary risk scores.
- What could go wrong?
  - borrowers change lending relationship $\Rightarrow$ control groups are affected.
  - main analysis: control for firm$\times$bank fixed effects
  - holding borrower compositions fixed $\Rightarrow$ only focus on supply-side impact

	Control		Treatment		
	(1) Before	(2) After	(3) Before	(4) After	(5) DID
Score	1.10	1.10	1.11	1.16	
	(0.01)	(0.01)	(0.01)	(0.01)	
Pseudo R^2	13.11%	13.04%	14.01%	18.55%	4.29%
					p-value = 0.00
N	42,554	45,025	24,137	25,919	

Evolution of Loan Level Characteristics



Treatment Effects by Technology

- Big data: very large volume and complex variety
  - impossible to process using traditional methods.
  - surge of data $\Rightarrow$ asymmetric effects due to technology capacity
- Quasi-exp as lab for increases in data amount.
  - short-run: holding technology constant.
- Treatment effects by ex-ante technology capacity.

Loan Characteristics

$$Y_{i,j,t} = \alpha_{i,j} + \alpha_t + \beta_0 Treat_{i,j,t} + \beta_1 Treat_{i,j,t} \times High\ IT_j + \epsilon_{i,j,t}$$

	(1) log Volume	(2) Interest	(3) Org. Time (days)	(4) Default
Treat	0.01	0.06	-0.09	0.17*
	(0.02)	(0.18)	(0.07)	(0.09)
Treat × High IT	0.03*	0.39***	-4.68***	-0.64***
	(0.02)	(0.09)	(0.10)	(0.09)
N	137,639	137,639	137,639	137,639
Time FE & Firm × Bank FE	Yes	Yes	Yes	Yes

Standard Errors Clustered at Bank × Year-Quarter Level in Parentheses

- $Y_{i,j,t}$: aggregated firm-level variables; $\alpha_{i,j}$ bank×firm FE; α_t: year-qtr FE.
- $Treat_{i,j,t}$: dummy for firm *i* borrowing from treated bank *j* at *t*.
- *High* IT_j: *j*'s IT exp/non-int exp before exp above median.

Screening Ability

	Low IT/Exp		High IT/Exp		
	(1) Before	(2) After	(3) Before	(4) After	(5) TD
			Panel A: Control		
Pseudo R^2	11.51%	12.15%	15.52%	15.98%	
N	18,036	19,585	24,518	25,440	
			Panel B: Treatment		
Pseudo R^2	12.61%	14.89%	14.86%	22.10%	5.67% p-value = 0.00
N	10,453	11,071	13,684	14,848d	

Cream-Skimming of High-IT Banks

- Data improves accuracy in risk assessment
    - more so for high IT banks.
- Heterogeneous screening ability $\Rightarrow$ cream-skimming

Cream-Skimming of High-IT Banks

- Data improves accuracy in risk assessment
  - more so for high IT banks.
- Heterogeneous screening ability $\Rightarrow$ cream-skimming
- Focusing on extensive-margin dynamics
  - how borrowers with different types change relationships

Cream-Skimming of High-IT Banks

- Data improves accuracy in risk assessment
  - more so for high IT banks.
- Heterogeneous screening ability $\Rightarrow$ cream-skimming
- Focusing on extensive-margin dynamics
  - how borrowers with different types change relationships
- Use all post-exp proprietary scores to predict default.
  - high-quality if $p(def)$ above median

Cream-Skimming of High-IT Banks

Panel A: High Quality

	Low Control	High Control	Low Treated	High Treated
Low Control	0.51	0.07	0.16	0.26
High Control	0.04	0.48	0.19	0.29
Low Treated	0.03	0.05	0.55	0.37
High Treated	0.06	0.07	0.19	0.68

Panel B: Low Quality

	Low Control	High Control	Low Treated	High Treated
Low Control	0.69	0.21	0.07	0.03
High Control	0.23	0.59	0.14	0.04
Low Treated	0.34	0.21	0.41	0.04
High Treated	0.29	0.23	0.15	0.33

- Similar to a Markov transition matrix
    - row name: bank type before exp
    - col name: bank type after exp

Outline

Main Findings

- Main finding: interest rate $\nearrow$, default $\searrow$, loan origination time $\searrow$
    - more so for high IT banks.

Main Findings

- Main finding: interest rate $\nearrow$, default $\searrow$, loan origination time $\searrow$
  - more so for high IT banks.
- Data improves accuracy in risk assessment
  - supply shock given better risk pricing. Einav et al. (2012)
  - interest rate $\searrow$, default $\searrow$

Main Findings

- Main finding: interest rate $\nearrow$, default $\searrow$, loan origination time $\searrow$
    - more so for high IT banks.
- Data improves accuracy in risk assessment
    - supply shock given better risk pricing. Einav et al. (2012)
    - interest rate $\searrow$, default $\searrow$
- Less loan origination time
    - demand shock given more convenience. Buchak et al. (2018)
    - interest rate $\nearrow$, default ?.

Main Findings

- Main finding: interest rate $\nearrow$, default $\searrow$, loan origination time $\searrow$
  - more so for high IT banks.
- Data improves accuracy in risk assessment
  - supply shock given better risk pricing. Einav et al. (2012)
  - interest rate $\searrow$, default $\searrow$
- Less loan origination time
  - demand shock given more convenience. Buchak et al. (2018)
  - interest rate $\nearrow$, default ?.
- Identification only permits exploring PE effects.
  - what if all banks are shared with the data?

Main Findings

- Main finding: interest rate $\nearrow$, default $\searrow$, loan origination time $\searrow$
  - more so for high IT banks.
- Data improves accuracy in risk assessment
  - supply shock given better risk pricing. Einav et al. (2012)
  - interest rate $\searrow$, default $\searrow$
- Less loan origination time
  - demand shock given more convenience. Buchak et al. (2018)
  - interest rate $\nearrow$, default ?.
- Identification only permits exploring PE effects.
  - what if all banks are shared with the data?
- Standard discrete-choice model with credit demand and default

  Crawford et al. (2018), Ioannidou et al. (2022)
  - incorporate both channels to general the findings?
  - equilibrium effects when data shared to all banks?

Setup

- At yr-qtr t: one market, J_t firms, K_t banks.
    - loan data available for one province
    - credit markets usually broadly defined at province level

Setup

- At yr-qtr t: one market, J_t firms, K_t banks.
  - loan data available for one province
  - credit markets usually broadly defined at province level
- Borrower j:
  - takes loan volume $l_{j,k,t}$ as given.
  - choose one bank to borrow from.
  - conditional on borrowing: choose to default or not.

Setup

- At yr-qtr t: one market, J_t firms, K_t banks.
  - loan data available for one province
  - credit markets usually broadly defined at province level
- Borrower j:
  - takes loan volume $l_{j,k,t}$ as given.
  - choose one bank to borrow from.
  - conditional on borrowing: choose to default or not.
- Bank k:
  - chooses interest rate $i_{j,k,t}$
  - facing adverse selection
  - maximizes expected profitability *à la* Bertrand-Nash competition

Modeling the Experiment

- Convenience
  - data-sharing decreases time of originating loans Buchak et al. (2018)

Modeling the Experiment

- Convenience
  - data-sharing decreases time of originating loans Buchak et al. (2018)
  - demand increase due to preference for faster time

Modeling the Experiment

- Convenience
  - data-sharing decreases time of originating loans Buchak et al. (2018)
  - demand increase due to preference for faster time
- Screening ability
  - marginal cost depends on credit score Einav et al. (2012)
  - data-sharing narrows gaps between bank-perceived borrower types and borrowers' true types

Modeling the Experiment

- Convenience
  - data-sharing decreases time of originating loans Buchak et al. (2018)
  - demand increase due to preference for faster time
- Screening ability
  - marginal cost depends on credit score Einav et al. (2012)
  - data-sharing narrows gaps between bank-perceived borrower types and borrowers' true types
  - marginal cost decreases for higher-quality borrowers
  - reallocating supply due to finer type discovery

Modeling the Experiment

- Convenience
  - data-sharing decreases time of originating loans Buchak et al. (2018)
  - demand increase due to preference for faster time
- Screening ability
  - marginal cost depends on credit score Einav et al. (2012)
  - data-sharing narrows gaps between bank-perceived borrower types and borrowers' true types
  - marginal cost decreases for higher-quality borrowers
  - reallocating supply due to finer type discovery
- Heterogeneity: interaction effects between data-sharing and IT intensity

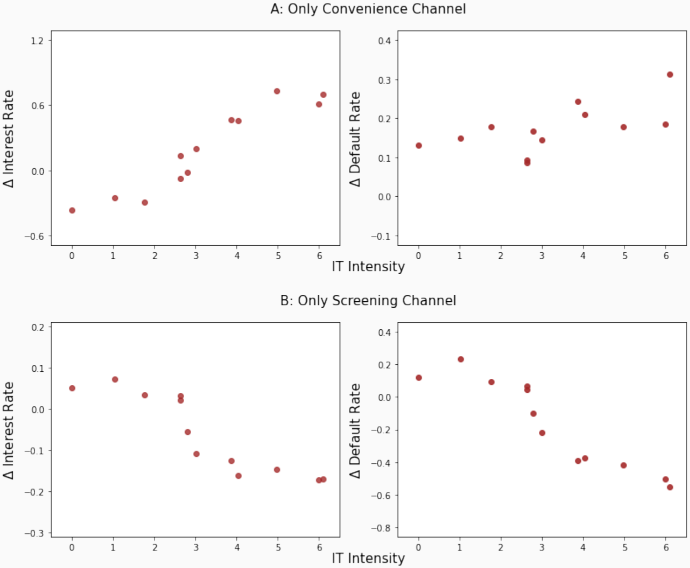
Model Fit

		(1) Default	(2) Interest Rate	(3) Effective MC	(4) Effective Markup
A: Pre-Experiment	Data	3.30	5.57		
	Model	3.31	5.56	3.50	2.06
B: Post-Experiment	Data	3.23	5.69		
	Model	3.24	5.66	3.51	2.20

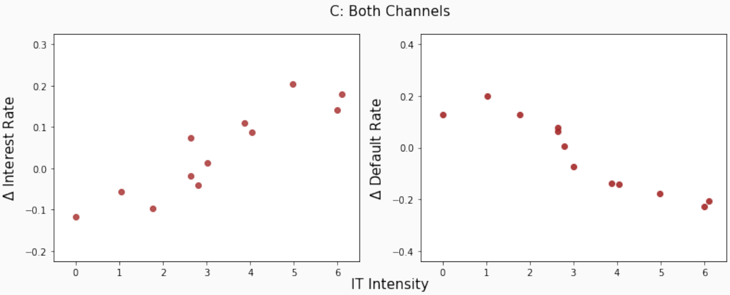
Estimates

	(1) Demand	(2) Default
Interest Rate	-0.39	0.44
	(0.14)	(0.06)
Interest Rate $\times$ Relationship	-0.73	0.24
	(0.21)	(0.05)
log(Days)	-1.66	0.08
	(0.23)	(0.12)
log(Days) $\times$ Relationship	-0.68	0.05
	(0.15)	(0.14)
FE: Maturity, Bank, Time, Relationship	Yes	Yes
N	1,932,730	239,080
Covariance Matrix	$\sigma = 0.30$	
	(0.07)	
	$\rho = 0.37$	$\sigma_P = 1$
	(0.04)	

Decomposition by IT Intensity – Decomposition



Incorporating both Channels



	(1) Default	(2) Interest Rate	(3) Effective MC	(4) Effective Markup	(5) % Diff
All	3.26	5.66	3.21	2.45	18.82%
High IT	2.96	5.63	3.00	2.25	25.54%
Low IT	3.65	5.75	3.66	2.09	4.35%

Outline

Methodologies

Results

Structural Estimation

Conclusion

Conclusion

- Effects of providing a large amount of data on banks.
- Surge of data increases profitability.
- Decomposition exercise: big data
  - simplified process of borrowing $\Rightarrow$ increase demand.
  - better risk-based pricing $\Rightarrow$ adjust supply by safer borrowers.
- Effects much larger for high IT banks
  - counterfactual markup: data shared to all
  - high IT: $\nearrow$ 25%; low IT: ~ 0
- Open question: what if banks can adjust technology?
  - might even amplify the heterogeneity
  - large banks invest more in IT He et al. (2023)