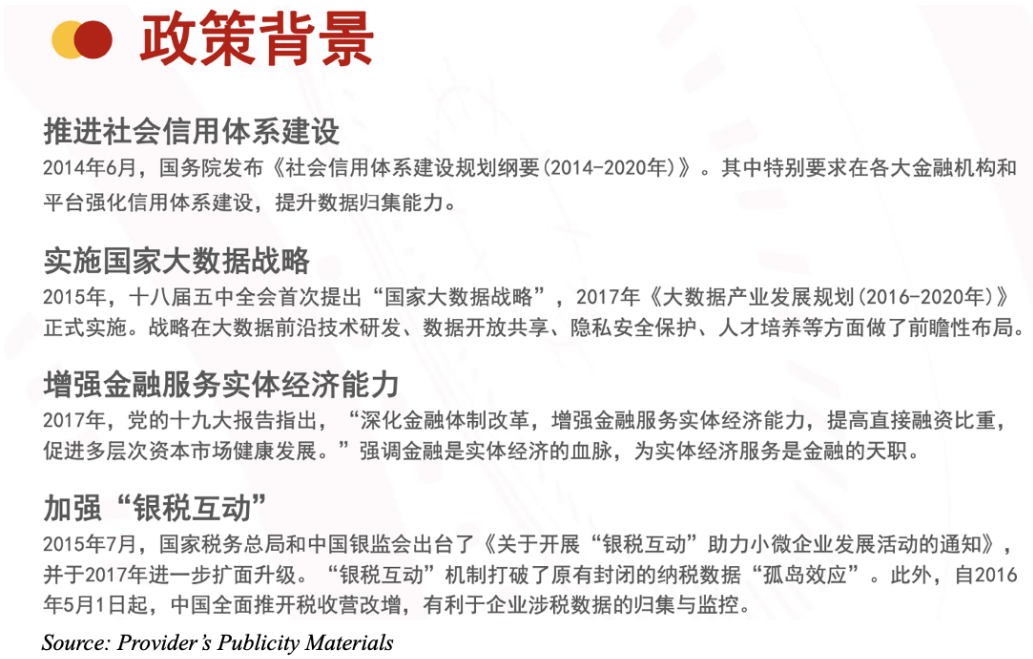
# Online Appendix

for "The Effects of Big Data on Commercial Banks" by Xiao Yin

## A. Policy Background




## Policy Background

**Promoting the Development of a Social Credit System**
In June 2014, The State Council issued the *Outline of the Plan for the Construction of the Social Credit System (2014-2020)*. In particular, it requires major financial institutions and platforms to strengthen the construction of the credit system and improve the ability of data collection.

**Implementing the national Big Data strategy**
In 2015, the Fifth Plenary Session of the 18th CPC Central Committee put forward the "National Big Data Strategy" for the first time, and in 2017, the *Big Data Industry Development Plan (2016-2020)* was formally implemented. The strategy has made a forward-looking layout in the research and development of cutting-edge big data technologies, data open sharing, privacy security protection, personnel training and other aspects.

**Strengthening the Ability of the Financial Sector to Serve the Real Economy**
In 2017, the report of the 19th CPC National Congress pointed out that "we will deepen the reform of the financial system, enhance the financial sector's ability to serve the real economy, increase the proportion of direct financing, and promote the healthy development of the multi-tiered capital market." He stressed that finance is the blood of the real economy and serving the real economy is the primary duty of finance.

**Strengthening "Bank-Tax Interaction"**
In July 2015, the State Administration of Taxation and the China Banking Regulatory Commission issued the *Notice on Carrying out the Activity of "Bank-Tax Interaction"* to help the development of small and micro enterprises, which was further expanded and upgraded in 2017. "Bank-Tax Interaction" mechanism breaks the original closed tax data "autarky effect". In addition, since May 1, 2016, China has comprehensively promoted the replacement of business tax with value-added tax, which is conducive to the collection and monitoring of tax-related data of enterprises.

## B. Estimation Details

The structural estimation follows Crawford et al. (2018) and Ioannidou et al. (2022). There are two steps. The first step is price prediction, and the second step is the joint estimation of demand and default.

### B.1. Price Prediction

Following Crawford et al. (2018) and Ioannidou et al. (2022), as the first stage of the estimation process, I need to predict the loan interest rates and the availability of online applications for each borrower from each bank. Given the assumption that there is only one market in a given year, I include all banks in a borrower's choice set. Afterward, I predict the interest rates and availability of online application of online applications not observed in the data. There are three steps. First, I use a random forest (RF) to predict online application availability across all loans that each borrower is offered by all banks it borrowed from. The predictors include loan volume, maturity, distance to the closest bank branch, bank-year fixed effects, borrower fixed effects, as well as each bank's proprietary credit scores. Since information of firms that have borrowed at least once is included in the credit registry, bank's proprietary credit scores are available for all these firms. The inclusion of bank-year fixed effects controls for systematic differences across banks in their reliance on soft information when setting interest rates. The inclusion of borrower fixed effects control for firm-level unobservables that determine borrowers' ability to get access to online applications. More importantly, the availability of each bank's proprietary credit scores controls for each bank's soft information on each borrower.

To estimate RF model, I first split the sample into a 50% of training sample and a 50% of test sample. The model is then fitted using the training sample with 10-fold cross-validation. Table B.1. gives the out-of-sample confusion matrix based on the test sample. Panels A and B respectively give the results excluding and including the proprietary credit scores. As shown, with proprietary credit scores, the RF model is very successful in predicting the availability of online applications, with an error rate of only around 10%. In addition, the proprietary credit scores are very effective in increasing the model's

Table B.1: Confusion Matrix for Online Application Availabiliy

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | A: Without Credit Score | | B: With Credit Score | |
| | | Online | Branch | Online | Branch |
| Observed | Online | 9012 | 4156 | 11457 | 1711 |
| | Branch | 18033 | 36440 | 6632 | 46841 |

forecastability, with error rates decreasing by more than 25% when including the credit scores.

As a second step, I use an OLS that includes the same set of controls to predict interest rates. The pricing model is as follows

$$i_{j,k,t} = X_{j,k,t}\gamma + \lambda_{j,k,t} + \tau_{j,k,t},$$

where $X_{j,k,t}$ includes loan volume, maturity, distance to the closest bank branch, and banks' proprietary credit scores. $\lambda_{j,k,t}$ includes the bank-year fixed effects, borrower fixed effects. The pricing model is similar to Crawford et al. (2018) and Ioannidou et al. (2022). However, one novel addition to previous literature is the inclusion of banks' proprietary credit scores, which controls for the effects of each bank's own soft information on each borrower in the process of determining prices. Similar to the RF model, the model is fitted using the training model.

Table B.2 assesses the effectiveness of the pricing model. Columns (1) to (3) present the out-of-sample adjusted $R^2$ fitted using the test sample. In column (1), the pricing model is fitted without borrower fixed effects and proprietary credit scores. Columns (2) and (3) add borrower fixed effects and proprietary credit scores sequentially. As shown, both borrower fixed effects and proprietary credit scores are crucial in increasing the goodness of fit. The adjusted $R^2$ increases from 0.37 in column (1) to 0.74 in column (3). In addition, columns (4) to (6) give the results of regressing default on pricing residuals. The results show that price residuals cannot forest borrower's default online after the inclusion of borrower fixed effects and proprietary credit scores. This indicates that both controls are important in controlling for banks' assessment of the firms' riskiness.

Table B.2: Price Prediction for Interest Rate

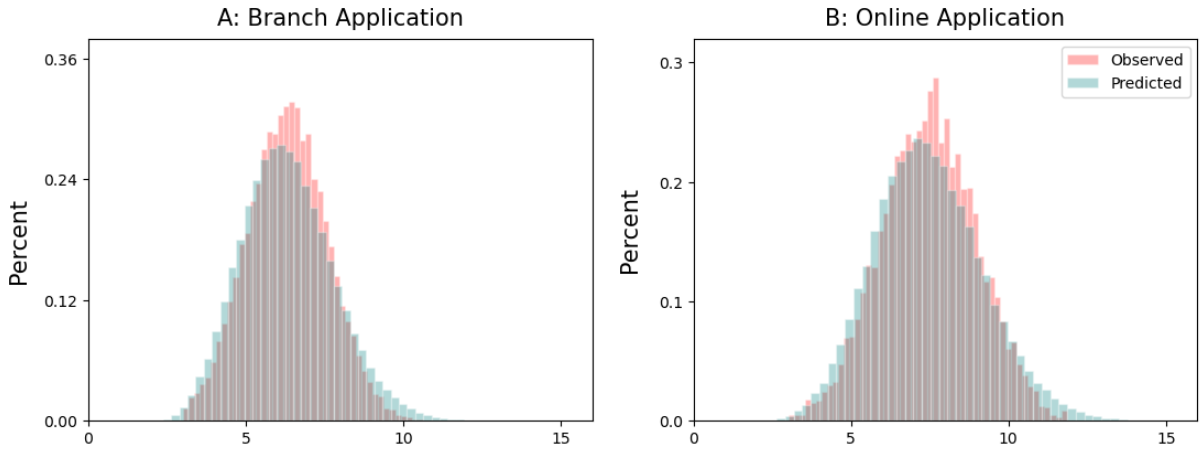
| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Observed Interest | | | Default | | |
| Interest Rate Residual | | | | 0.07*** | 0.04* | 0.01 |
| | | | | (0.02) | (0.02) | (0.03) |
| Firm FE | No | Yes | Yes | No | Yes | Yes |
| Credit Score | No | No | Yes | No | No | Yes |
| Bank FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Adjusted $R^2$ | 0.37 | 0.52 | 0.74 | 0.07 | 0.14 | 0.28 |
| N | 66,090 | 66,090 | 66,090 | 66,090 | 66,090 | 66,090 |

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

The final step is to use the fitted prediction model to predict interest rates on online application availability for the unmatched borrower-bank pairs. For borrowers that have borrowed more than once, prediction contract terms are straightforward. However, for those who only borrowed once, I cannot include the firm fixed effects in the prediction model. Following Crawford et al. (2018), I use propensity score matching. Specifically, I match those who have only borrowed once to those who borrowed more than once on the firms' age, log total asset, leverage, employment size, loan volume, loan maturity, city, and year. I then randomly assign one of the five borrowers that have the closest propensity scores and assign the firm fixed effect from the latter to the borrower. With the assigned fixed effects, I then use the prediction models to predict interest rates.

To assess the performance of the prediction model, Figure B.1 plots the histograms of the interest rates for both the interest rates of the observed sample and the predicted interest rates using the test sample. The plots therefore assess the out-of-sample prediction accuracy. The left panel gives the branch application sample, and the right panel gives the online application sample. In both panels, the distributions of the observed and predicted interest rates align with each other nicely, indicating that the prediction models are very successful. In addition, 20% are online applications in the observed sample, compared with 23% in the forecasted sample.

Figure B.1: Distribution of Observed and Predicted Interest Rates




## B.2. Simulated Maximum Likelihood Estimation

I estimate the demand and default system using a two-step method based on maximum simulated likelihood and instrumental variables estimation. In the first stage, using data on firms' choices of bank and default, I estimate the firm-level and bank-level parameters across the two equations, $\eta = \{\alpha^D, \eta, \eta^D, \beta, \beta^D\}$, and the variance-covariance matrix of the errors in the system $\sigma$ and $\rho$. I also recover the bank-market-year specific constants (mean utilities) in the demand model ($\delta_{k,t} = \alpha_0 + \mathbf{X}_{k,t}\beta + \xi_{j,t}$) following Berry et al. (1995).

The indirect utility from demand can be written as d

$$U_{j,k,t} = \delta_{k,t} + \alpha_i i_{j,k,t} + +\alpha_{i,Z} i_{j,k,t} \times Z_{j,k,t} + \underbrace{\alpha_T T_{j,k,t} + \alpha_Z Z_{j,k,t} + \alpha_{T,Z} T_{j,k,t} \times Z_{j,k,t} + \mathbf{Y}_{j,k,t}\eta}_{V_{j,k,t}} + \epsilon_j + \nu_{j,k,t},$$

where $\nu_{j,k,t}$ is assumed to follow a T1EV distribution. Then the probability that borrower $j$ in year $t$ chooses bank $k$ is given by

$$q_{j,k,t} = \int \frac{\exp\{\hat{\delta}_{k,t} + \alpha_i i_{j,k,t} + \alpha_{i,Z} i_{j,k,t} Z_{j,k,t} + V_{j,k,t}\}}{1 + \sum_l \exp\{\hat{\delta}_{l,t} + \alpha_i i_{j,l,t} + \alpha_{i,Z} i_{j,l,t} Z_{j,l,t} + V_{j,l,t}\}} f(\epsilon_j) d\epsilon_j.$$

The probability of default conditional on borrowing is

$$p_{j,k,t} = \int \Phi_{\epsilon_i^D|\epsilon_i} \left( \frac{\mathbf{z_p} - \tilde{\mu}_{\epsilon_i^D|\epsilon_i}}{\tilde{\sigma}_{\epsilon_i^D|\epsilon_i}} \right) f(\epsilon_i|D=1) d\epsilon_i,$$

$$\begin{aligned} \mathbf{z_p} = {} & \alpha_0^D + \mathbf{X}_{k,t}\beta^D + \alpha_i^D i_{j,k,t} + \alpha_T^D T_{j,k,t} + \alpha_Z^D Z_{j,k,t} \\ & + \alpha_{i,Z}^D i_{j,k,t} \times Z_{j,k,t} + \alpha_{T,Z}^D T_{j,k,t} \times Z_{j,k,t} + \mathbf{Y}_{j,k,t}\eta^D, \end{aligned}$$

where

$$\epsilon_i^D|\epsilon_i \sim N(\underbrace{\rho\epsilon_i\sigma_D}_{\tilde{\mu}_{\epsilon_i^D|\epsilon_i}}, \underbrace{\sigma_D^2(1-\rho^2)}_{\tilde{\sigma}_{\epsilon_i^D|\epsilon_i}}).$$

The two probabilities give the joint likelihood

$$\ln L = \sum_i \mathbf{q}_{j,k,t} \left[ \ln(\overline{q}_{j,k,t}) + \mathbf{p}_{j,k,t} \ln(\overline{q}_{j,k,t}^D) + (1 - \mathbf{p}_{j,k,t}) \ln(1 - \overline{q}_{j,k,t}^D) \right],$$

where $\mathbf{q}_{j,k,t} = 1$ if $j$ chooses $k$ at $t$, and $\mathbf{p}_{j,k,t} = 1$ if firm $j$ defaults.

### B.3. Identification

The availability of proprietary credit scores and a large set of fixed effects explains a majority of supply decisions. However, such characteristics do not perfectly explain the pricing strategies. This is evident from column (3) of Table B.2, which shows a large but not perfect out-of-sample $R^2$ of the price prediction equation in forecasting interest rates. Therefore, the loan interest rate may be endogenously related to unobservables that influence borrowers' demand and default. If this is the case, the estimates of the price sensitivities in both the demand and the default models will be biased. Following Crawford et al. (2018) and Ioannidou et al. (2022), I use the control function approach suggested by Train (2009) to address this potential endogeneity concern[15]. This approach involves an initial step where both predicted and actual interest rates are regressed against the same observable factors utilized in the demand and default analysis, augmented by

[15]In Crawford et al. (2018), the identification of price sensitivity in the demand model is based on 2SLS. However, due to the inclusion of the interaction of interest rates and lending relationship, 2SLS requires more IVs, which requires a stronger assumption on the exclusion restrictions. Therefore, I use control functions for the estimation of both demand models and default models.

Table B.3: First Stage Results

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Observed Interest Rates | | Predicted Interest Rates | |
| Deposit Interest Rates | 1.17*** | 0.84*** | 0.66*** | 0.53*** |
| | (0.09) | (0.08) | (0.11) | (0.10) |
| Other-Market Interest Rates | | 0.44*** | | 0.28*** |
| | | (0.04) | | (0.05) |
| Controls | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes |
| Bank FE | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes |
| Adjusted $R^2$ | 0.36 | 0.49 | 0.18 | 0.25 |
| N | 239,080 | 239,080 | 1,693,650 | 1,693,650 |

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

a set of instrumental variables. In the subsequent step, the residuals from the pricing regression are included as control variables within the demand and default models to mitigate the impact of unobserved factors associated with pricing, thus ensuring that the residual pricing variance is independent of the unobserved variables affecting demand and default.

For the identification of interest rate sensitivity in both models, I employ two instrumental variables: interest rates from household savings and the standard Hausman instrument, which draws from interest rates in different provinces. These instruments act as indirect measures for the cost of bank funding. They meet the exclusion criterion as the household savings market operates distinctly from corporate lending, and loan markets across various provinces are subject to separate regulatory bodies. Consequently, fluctuations in these distinct markets do not correlate with the hidden factors determining a firm's banking preferences or its default risk. Table B.3 outlines the initial findings for both observed and predicted loan interest rates, indicating the relevance of these instruments with coefficients aligning with expectations.

Concerns about the exclusion restriction's potential breach when incorporating deposit market rates into the pricing model arise due to the intertwined nature of bank risk and its funding sources and costs (Flannery and Sorescu, 1996; Detragiache et al., 2000; Martinez Peria and Schmukler, 2001; Ippolito et al., 2016). Following Ioannidou

et al. (2022), addressing these worries, I focus on minor household deposits, protected by deposit insurance and implicit state backing, hence unaffected by the bank's risk profile (Egan et al., 2017). Moreover, the similarity in results obtained using either of the instruments, or none, suggests that any potential bias from endogeneity is minimal.

### B.4. Estimation

With the predicted interest rates and control functions, the model is then estimated with simulated maximum likelihood estimation. The specific procedure is as follows

1. given initial value of $\theta$ and calculate homogeneous $\tilde{\delta}_{kt}$.
2. for each firm, simulate $NS = 100$ $\epsilon_i$'s following Halton simulation. Then get

$$\overline{Pr}_{jkt} = \frac{1}{NS} \sum_{n=1}^{NS} \frac{\exp\{\hat{\tilde{\delta}}_{kt} + V_{jkt}\}}{1 + \sum_l \exp\{\hat{\tilde{\delta}}_{lt} + V_{jlt}\}}$$

3. get $\hat{\sigma}_{kt} = \sum_i \overline{Pr}_{jkt}/I$, where $I$ is the total number of firms.
4. update $\tilde{\delta}_{kt}$ as

$$\tilde{\delta}_{kt}^{r+1} = \tilde{\delta}_{kt}^{r} + \ln(s_{kt}) - \ln(\hat{\sigma}_{kt}),$$

   until all $\tilde{\delta}_{kt}^{r+1} = \tilde{\delta}_{kt}^{r}$. Let this be $\hat{\tilde{\delta}}_{kt}$.
5. get the probability of default

$$\begin{aligned}
\overline{Pr}_{jkt}^{D} &= \frac{1}{NS} \sum_{n=1}^{NS} \Phi\left(\frac{X - \sigma_D \rho \epsilon_i}{\sigma_D^2(1-\rho^2)}\right) \\
X &= \alpha_0^D + \mathbf{X}_{k,t}\beta^D + \alpha_i^D i_{j,k,t} + \alpha_O^D O_{j,k,t} + \alpha_Z^D Z_{j,k,t} \\
&\quad + \alpha_{i,Z}^D i_{j,k,t} \times Z_{j,k,t} + \alpha_{O,Z}^D O_{j,k,t} \times Z_{j,k,t} + \mathbf{Y}_{j,k,t}\eta^D
\end{aligned}$$

6. form the log-likelihood function

$$\ln L = \sum_j \mathbf{d}_{jkt} \left[\ln(\overline{Pr}_{jkt}) + f_{jkt} \ln(\overline{Pr}_{jkt}^{D}) + (1 + f_{jkt}) \ln(1 - \overline{Pr}_{jkt}^{D})\right],$$

   where $d_{jkt} = 1$ if $j$ chooses $k$ at $t$, and $f_{jkt} = 1$ if firm $j$ defaults.
7. MLE to get estimates of the $\tilde{\eta}$s.

8. get vcov matrix by

$$vcov\left(\hat{\eta}\right) = \left[-\frac{\partial^2}{\partial \eta^2} \ln L\left(\hat{\eta}\right)\right]^{-1}$$