Journal of Behavioral and Experimental Economics 96 (2022) 101806

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

The role of social and psychological related soft information in credit analysis: Evidence from a Fintech Company

Yao Wang [*,a], Zdenek Drabek [a,b], Zhengwei Wang [c]

[a] *Charles University in Prague, IES FSV, UK*
[b] *CERGE-EI: The Center for Economic Research and Graduate Education Economics Institute*
[c] *Tsinghua University, PBC (People's Bank of China) School of Finance*

ARTICLE INFO

Keywords:
Soft information
Fintech
Microfinance
Credit analysis
Behavioral economics

ABSTRACT

Improvements in the quality of information in credit appraisal are paramount to the greater efficiency of credit markets. The existing research to assess the role of soft information in credit markets has so far been very limited and inconclusive due to differences in approaches and methodological limitations. The aim of this paper is to discuss the role of social and psychological related soft information in predicting defaults in the P2P lending market and to assess the importance of such information in Fintech credit analysis. Using a unique dataset from the pioneer P2P lending platform RRDai.com and alternative models of testing, we compared the predictive performance of soft information, hard information and combined hard and soft information on defaults. The results show that soft information can provide valuable input into credit appraisals. Soft information shows high predictive power in our test, and combined with hard information, it increases the power of our model to predict defaults.

1. Introduction

Perhaps one of the most interesting new features of the financial industry in the past decade is the development of new technologies for data generation and management. New technologies and better information reduce uncertainties and increase efficiencies in lending. They offer opportunities to improve access to credit and build better default predicting models. Traditionally, the financial sector has relied primarily on financial statements, denoted in the literature as 'hard information', as the predictor of creditworthiness. However, 'hard information' together with collateral may not always fully secure repayment of loans, and loans based on collateral actually sometimes have higher default rates. Pari passu, credit scoring systems, while contributing to increasing credit availability for small businesses, have also not been as effective as expected.

To address the drawbacks of traditional (hard) information-based credit rationing systems, soft information derived from social and psychological factors has become a complementary approach. With the development of data management and drawing on ideas from "identity economics", originating in the work of Akerlof & Kranton (2000), the availability of social and psychological information, (i.e. *soft information*) is increasing, and the costs of collecting such information are decreasing (Liberti & Petersen, 2018). This provides us with the motivation and opportunity to explore the role of "identity" in credit appraisal.

The importance of soft information has dramatically increased with the emergence of Peer to Peer (P2P) lending markets.[1] In contrast to bank lending to small and medium-sized enterprises (SMEs), P2P lending does not require the presence of branches and loan offices in local communities.[2] Borrowers fill in online loan application forms and

* Corresponding author.
E-mail address: wangyaoflora@gmail.com (Y. Wang).

[1] New technologies have spectacularly transformed the industry by reaching out to market segments which have not been well served in the past. The first P2P platform, Zopa, started in the UK in 2005, and was followed by Prosper and Lending Club in 2006. In 2007, P2P platforms emerged in other European countries (e.g., Smava in Germany, TrustBuddy in Sweden, Prestiamoci in Italy), China (e.g., PPDai, RenrenDai), and Japan (e.g., Maneo). Since 2009, P2P platforms have been booming on a global scale. For an earlier review of website-based lending see, for example, Ashta & Assadi (2009).

[2] In China, the SME sector was serving 10 million clients in 1995, the early days of SME lending; the number today is around 300 million. Microfinance institutions have been commercialized over time and, today, around 100 specialized funds have invested and loaned about US$ 12.5 billion. The growth of P2P lending has been equally spectacular. For more information on Chinese P2P platforms, see Appendix A. More information also appears in Section 3.

https://doi.org/10.1016/j.socec.2021.101806
Received 8 December 2020; Received in revised form 23 September 2021; Accepted 30 November 2021
Available online 1 December 2021
2214-8043/© 2021 Elsevier Inc. All rights reserved.

choose what information they want to disclose which is then posted online. Typically, there are no restrictions on the amount borrowed, and the funding process comes to an end when the full amount of the loan request is reached. During the entire loan process, there is no financial intermediary serving as a credit rationing mechanism. Thus, the quality of information available to lenders and borrowers has become a major issue.

However, research exploring the role of soft information in credit appraisal for P2P markets is very limited and inconclusive. Most of the existing research covers banks and their credit appraisal systems. These studies typically look at the role of hard or soft information but rarely at the role of both hard and soft information together. What is particularly missing is strong evidence of how these different appraisal systems perform. The existing research is also heavily oriented towards an assessment of loan applications rather than assessments of defaults, and that can lead to serious misidentification of borrowers. Moreover, most of the research is typically based on a specific factor in lending and even less on exploring the role of social and psychological factors.

The aim of this paper is to answer the question of whether risk assessment can be improved by the incorporation of social and psychological related soft information into appraisals of credit risk in the presence of imperfect hard information. We build a model to analyze the determinants of loan defaults. It looks at the importance of soft and hard information in different scenarios. We compare the predictive performance of soft information, hard information, and combined hard and soft information on loan defaults. Our results show that soft information can provide valuable input for credit appraisal. The predictive power of soft information alone in our test was high, and together with hard information it improved the predicting power of loan appraisal. These results hold firmly after the application of a number of robustness tests and analyses.

The paper is divided into five sections. Section 2 reviews the relevant empirical literature. Its purpose is to identify the important advances in the debate on the quality of information and key gaps and limitations of the literature, which drive our approach and methodology. Section 3 describes our methodology: the data used in the study, and the econometric method we used. The results of our empirical tests are presented in Section 4. The results of sensitivity tests are reported in Appendix D and Appendix E. Our conclusions are summarized in Section 5.

2. Treatment of hard and soft information in the literature

The literature dealing with the role of information in credit appraisal in P2P platforms is fairly recent and draws heavily on the literature covering the same issue for the rest of the financial sector. It can be grouped into three streams, distinguished by three different approaches.

Hard Information-Based Approach and Its Limitations. Assessments of loan performance have traditionally been related to the use of various financial indicators (Horrigan, 1966). Indicators such as income level, ownership of property and other collateral, and debt serve to generate credit scoring in risk-based pricing, in which the terms of a loan offered to borrowers, including the interest rate, are based on the probability of repayment. These financial indicators, known in the literature as *hard information*, are also used in creditworthiness analysis and to assess the probability of the success of a loan in P2P markets. Following this practice, traditional models of loan determinants, which emphasize the key role played by financial (hard) information, show how the credit scoring system impacts the lending behavior of banks (e.g., Berger, Frame, & Miller, 2005a; Berger, Miller, Petersen, Rajan, & Stein, 2005b) and how it predicts the likelihood of loan defaults (Deyoung, Glennon, & Nigro, 2008). Verified bank account information and credit ratings were the key determinants of loan approvals and interest rates in Klafft (2009). Similarly, Iyer, Khwaja, Luttmer, & Shue (2009), Uchida (2011), and others have found that large lenders base loan judgments mostly on hard information (e.g., the debt-to-income ratio), even when other information is available. Xu & Zou (2010) found that only hard information is conveyed to bank headquarter's credit office despite the availability and transferability of both hard and soft information. Serrano-Cinca, Gutierrez-Nieto, & López-Palacios (2015) and, previously, Deyoung et al. (2008) also argue that the probability of default is significantly related to an applicant's annual income, housing situation, credit record, and indebtedness. In brief, collateral and other hard information are widely viewed as the most informative factors in credit approval.

However, the research also shows that the usefulness of hard information in the assessment of credit risk is limited. For one thing, sufficient hard information is sometimes not available. In addition, while credit scoring systems can provide an ordinal risk assessment, they do not provide an estimate of the borrower's default probability. For example, Iyer et al. (2009) showed that lenders can differentiate the creditworthiness of borrowers with different credit scores, but only within the same credit categories. Collateral, too, cannot always secure repayment behavior. As shown, for example, by Jiménez & Saurina (2004), loans with collateral may actually have higher default rates. Clearly, defaults cannot be entirely avoided using hard information. Other approaches, including various techniques based on soft information, should be taken into account in order to improve loan performance.

Soft Information-Based Approach. The second stream of literature originates in information theory from the perspective of asymmetric information under imperfect contracts. Following studies on credit rationing and information signaling (Stiglitz & Weiss, 1981; Spence, 1973 and Akerlof, 1970), attention has increasingly been paid to information other than financial indicators that may signal the ability and willingness of borrowers to repay loans. In these studies, soft information variables represent an important new element of information about borrowers by addressing the asymmetric information problem. The most commonly accepted distinction between soft and hard information can be traced back to Diamond (1984)'s theory of financial intermediaries and his distinction between banks and public bond markets or theories under the principal-agent framework which explored relationship lending (e.g. Godbillon-Camus & Godlewski, 2005; Stein, 2002).

Akerlof & Kranton (2000)'s identity economics has been particularly helpful in explaining various puzzles in standard economic literature. By emphasizing the role of the identity of agents in their economic choices, they make the point that economic decisions are not exclusively dependent on monetary incentives. In the context of lending in financial markets, the introduction of borrower's identity in credit appraisal must be considered as a factor determining loan applications or loan performance together with traditional financial indicators.

Soft information has been variously defined as including social characteristics of borrowers such as gender and age (e.g. Bertrand, Karlin, Mullainathan, Shafir, & Zinman, 2005), education (Liao, Lin, & Zhang, 2015), beauty (Ravina, 2012; Gonzalez & Loureiro, 2014; Duarte, Siegel, & Young, 2012), and culture (Bourdieu, 1986). Alternatively, soft information has included indicators such as social capital (e.g. Greiner & Wang, 2009; Liu, Brass, Lu, & Chen, 2015; Cao, 2013; Miu & Chen, 2014) or psychological factors such as responses to texts (e.g. Lea, Webley, & Walker, 1995; Dorfleitner et al., 2016). Another definition was used by García-Appendini (2007), who defines soft information as any kind of data other than transparent public information. In the relationship lending literature on SME finance, some researchers also used the physical distance between the lender and borrower as the proxy for soft information (Dell'Ariccia & Marquez, 2004; Berger et al., 2005a; Deyoung et al., 2008; Agarwal & Hauswald, 2010).

As a factor in understanding loan determinants, soft information has been increasingly used both by researchers in their empirical work and in actual lending practices by financial institutions. As Berger & Udell (2002) and others have shown, small business loans already rely more on relationship lending due to the paucity of hard information relating to small businesses. Recent empirical work has exclusively focused on soft information, including studies by Cornée (2017) and Ge, Feng, Gu, & Zhang (2017). However, the results of studies that rely exclusively on soft information are fragmented and inconclusive.[3] In addition, most of the research refers to the impact of soft indicators on the funding success rate. The results are far less clear about the value of soft information in predicting a borrower's repayment performance. Some studies have shown that online friendships are a sign of a lower probability of default, but other studies have found that membership in social networks does not signal more successful loan repayment.[4] Similarly, contradictory results occured with regard to the roles of appearance, language, and gender in repayment performance.

Combined Hard and Soft Information-Based Approach. The third stream of literature that has recently received attention is the joint use of hard and soft information. Some empirical research has indicated that a combination of hard and soft information can achieve a better predictive power than exclusive reliance on hard or soft variables (Grunert, Norden, & Weber, 2005; Godbillon-Camus & Godlewski, 2005; Dorfleitner et al., 2016 in addition to the study of Agarwal et al., 2011 noted above). However, the evidence in this field is even more limited, as these studies only look at banks and their lending practices. In addition, none of these studies examined the standalone role of social and psychological factors or in combination with hard factors. One exception was Ge et al. (2017) in their P2P study, but they only look at the role of soft indicators and completely disregarded the assessment of hard indicators. Another exception is Dorfleitner et al. (2016), they covered a broad range of soft and hard indicators, but their study is limited to only banks. Moreover, by concentrating on the analysis of texts and keywords, their methodology was too specific and not always applicable to different linguistic environments. Finally, the literature suffers from the same limitation noted in the other two streams the absence of any appraisal of the scope for misidentification in estimated models.[5]

The limited emphasis to date on the determinants of defaults is unfortunate, as defaults are ultimately important for both lenders and borrowers. Should the determinants of loan approvals differ from those of defaults, the loan approval process could lead to the provision of loans to the wrong applicants (i.e., to a Type II error in the estimating procedures).[6]

3. Methodology

This paper uses a binary classification model to assess the value of soft information in credit appraisals. We began with a brief description of our approach, the data, the scope of the analysis, and the definitions used. We then provided a description of the model. Since the model is tested using different variants, the description also includes an explanation of our analytical treatment of model discrimination.

3.1. Approach, data, scope, and definitions

Approach. We examine the determinants of loan defaults with a special interest in the role of soft information. Due to the poor quality of hard information data, especially with regard to lending to SMEs and to individuals for business purposes, the Chinese P2P market is currently critically dependent on soft information. The administration and management of hard credit information in China have been severely criticized and the country's credit bureaus are undergoing major reforms.[7] Moreover, the explosion of P2P lending in China has been accompanied by growing credit risk and a rising likelihood of defaults.[8] Several P2P platforms have recently been closed due to poor management of credit information. As we suspect that the traditional methods of risk appraisal may have led to the misidentification of borrowers (Type II error), we therefore concentrated on analyzing ”soft” determinants of defaults in order to better identify credit risk in the industry and to lower the cost of credit appraisal.

Definition. Following Akerlof & Kranton (2000), we define soft information as information transmitted by a selected social or psychological characteristic that captures the identity of the borrowers. It contains information about borrowers including age, education, gender, and race. In addition, even softer information like borrower's social networks, video interviews, profile pictures, and descriptions of prior borrowing stories are also included. This broad definition allows us to capture links between the relevant characteristics of the borrower and defaults as, for example, in Grunert et al. (2005). Needless to say, the definitions of soft information have evolved over time and different definitions have been adopted in the literature (Liberti & Petersen, 2018).

Choice of Determinants. Our specific choice of soft variables is driven by the theory and empirical literature. According to Piliavin & Charng (1990), for example, gender matters because women are more likely to be altruistic than men and women can, therefore, be expected to be less likely to default on their loans. Franke, Crown, & Spake (1997) provided a different angle on the gender issue with the same conclusion when, in their empirical study, they showed a difference between men and women in their perceptions of unethical behavior. Men and women also show differences in sympathy and empathy (Lennon & Eisenberg, 1987). In terms of marital status, Chaulk, Johnson, & Bulcroft (2003) argue that it has a significant negative relationship with risk tolerance. Theories of family development suggest that people's behavioral expectations and decision-making contingencies change after marriage. Potential losses from risky investment loom larger than potential gains for married people. Brown (2000), for example, suggests that marriage can add stability to life and results in lower rates of depression and alcohol abuse (Horwitz & White, 1998).

Age and education also very likely affect borrower behavior. We treat age as a soft variable, following writings including Gonzalez & Loureiro (2014) and Ge et al. (2017). Age matters, as people's thoughts, feelings, and behaviors are known to change throughout their lives. Their moral understanding, emotional development, self-confidence, and identity formation evolve and their self-control and emotional stability generally increase with age (e.g. Roberts & Mroczek, 2008). Education, in turn, has arguably been the most covered soft variable in different streams of literature. For example, the level of educational attainment can play a role in the perception of financial risk. Psychology in cognitive development theory, as a branch of educational psychology,

[3] The use of soft information is also known in the other arm of the Fintech industry - in non-bank financial institutions. Those institutions rely on proprietary models and use a combination of hard and soft information to evaluate credit risk. Their activities have increased, but they remain relatively high-cost since they are paying commissioned agents to bring in potential clients. See, for example, Agarwal, Ambrose, Chomsisengphet, & Liu (2011).

[4] One explanation is that social networks often involve social pressures which build up within the groups. It seems that this kind of pressure is less likely in online lending. We are grateful to Professor Raffer for this point.

[5] Until recent attempts by mostly Chinese scholars and a paper by Santoso, Trinugroho, & Risfandy (2020), studies of determinants of loans have typically been focused on applications rather than on defaults in P2P markets. However, none of these studies makes any attempt to discuss the issue of misidentification. See also, for example, Jiang, Wang, Wang, & Ding (2018) or Wang, Yu, & Zhang (2019).

[6] See, for example, Gonzalez & Loureiro (2014) and Ge et al. (2017) with regard to age, Dorfleitner et al. (2016) with regard to language, and Liao et al. (2015) with regard to education. Social capital has been found to be positively related to loan terms (e.g. Lin, Prabhala, & Viswanathan, 2013; Herrero-Lopez, 2009; Cao, 2013) but negatively related to defaults (e.g., Ge et al., 2017; Freedman & Jin, 2011; Miu & Chen, 2014; Lin et al., 2013; Cao, 2013).

[7] See, for example, Botsman (2017) and Chorzempa (2018) and footnote 14 and 16.

[8] The emphasis on loan appraisal could be justified in the past by the relatively successful performance of microfinance lending. However, since the explosion of P2P lending, credit risk is rising. For more info, see Lieberman, Paul, Watkins, & Anna (2018). The rise of defaults in P2P markets is also well documented in Cornée (2017).

Table 1
Distribution of listings.

Overdue	84
To be opened for bids	11
In default	590
Failing auctions	181,043
Completed repayment	13,901
In the application process	5,439
In the repayment process	50,819
Total	251,887

emphasizes, inter alia, the point that people's understanding of morality changes with the development of education (Slavin & Davis, 2006).

We assume that by communicating their personal information, borrowers aim to generate a *positive and trustworthy* overall perception about themselves,[9] Such information is signaled through various personal characteristics of borrowers and their social networks. The scenarios reflect three different theoretical and practical considerations which have been adopted in the empirical literature and described in the previous section. We assume that each of the scenarios is formally independent and, in the absence of a robust and generally accepted theory, the choice must be made with the help of econometric techniques. This assumption is key in the estimations of all three models.

Thus, our treatment of soft information includes social indicators: education level (Liao et al., 2015), age (e.g. Gonzalez & Loureiro, 2014), and gender (e.g. Barasinska & Schäfer, 2010; Ravina, 2012; Pope & Sydnor, 2011), which can be identified and verified. We also add other types of soft information including variables that refer to personal characteristics and social networks of borrowers which, in turn, represent other proxies for social capital and networks. Due to limitations of data, we were unable to use other soft indicators, but we believe that we have captured a sufficiently broad range of those variables which have been most frequently used in the literature.[10]

Data. We examine the role of soft information with a case study of the Chinese P2P market, using the P2P platform RenrenDai.com. The Chinese P2P market is compelling because of its size and rapid growth as shown in Fig. A.4.[11] Moreover, the market has developed hand-in-hand with the development of a rich database which is a valuable source of soft information.

RenrenDai was established in 2010. By October 2016, the total amount of its transactions exceeded 21.2 billion yuan. The platform targets microloans; the average loan amount was 71,000 yuan. The platform consisted of 251,887 listings from 2010 to 2014. Borrowers fill in loan application and publish all the information online, peer investors do the credit analysis by themselves and choose the loans to invest. When the full loan amount has been filled by the investors, the funding process ended. One loan can have several investors. Until the maturity of the loan, the borrower can repay the loan by full or on monthly installment. The platform has collection teams to enforce loan terms and minimize losses. The number of defaults during the period examined was 674 of 14,575 total listings, representing a relatively modest default rate of about 4.2 percent.[12] 'Failing auctions' are the loans that failed to get the fund. 'Loans in repayment process' means until the time we collect the dataset, the loans are still not reaching their first repayment date or haven't finished the repayment. We do not have the data of the percentage of payment completed. And the repayment can be paid by monthly installment, so we don't know whether they will repay fully. Thus, they are not included in our dataset for default repayment behavior analysis. A summary of listings appears in Table 1. The description of our dataset of hard and soft information is below.

Loans provided on the platform are used both for personal and business purposes. Unfortunately, the platform does not provide direct access to the loan purpose. According to an interview with the CEO of RenrenDai.com CEO interview, 70–80 percent of loans granted are for free lancer or micro business operational cash flow purposes.[13] Other common purposes include car loans, home renovation, and consumption.

When loans on RenrenDai are overdue, borrowers receive reminders by SMS followed by phone calls if necessary. The P2P platform will then hire loan recovery companies to recover the loans. If the recovery company cannot recover the loans, the platform will cover the loss by their margin account.[14] When loans are in default, those borrowers will be added to the credit bureau's black list and to the P2P industry black list.

As a product of financial innovation (Ding, Fung, & Jia, 2020), the shadow banking industry has reached $114 trillion according to the Financial Stability Board's annual report on non-bank financial intermediation. As a representative of shadow banking, P2P lending has rapidly developed in the past decade. The credit appraisal in the Fintech industry highly relies on alternative information compare to traditional banks. Understanding the benefit and risk of employing soft information in credit appraisal can provide experience to the traditional bank reform and contribute policy implications to regulators.

Contractual arrangements in China are heavily influenced by Chinese culture, which favors information derived from human relationships.[15] Soft information has, therefore, become a special requirement for contracts and for P2P markets in China, and very rich information is provided on the RenrenDai platform. We believe that the RenrenDai data represents a considerable improvement on data used in most comparable studies: it is more comprehensive, more specific, detailed, and classifiable. The Chinese market is interesting also because of its institutional specifics. The system of oversight allows verification of

[9] See, for example, Pötzsch & Böhme (2010) who show that trust can lead to better credit conditions. More recently, Thakor & Merton (2018) analyze competitive interactions between banks and non-bank lenders and, distinguishing between trust and reputation, they show that trust enables lenders in Fintech firms to have assured access to financing, while a loss of investor trust makes access conditional on market conditions and lender reputation. They further show that banks have stronger incentives to maintain trust. When borrowers' defaults erode trust in lenders, banks are able to survive the erosion of trust (and bail-outs by taxpayers) when Fintech lenders do not. More corroborative evidence on the importance of trust enhanced by soft information has been provided by Miu & Chen (2014); Ravina (2012); Barasinska & Schäfer (2010) and Serrano-Cinca et al. (2015). However, it should be noted that while trust is likely to be important in establishing better loan terms, the effect of trust on defaults, as required in our model, is more ambiguous.

[10] For example, one could use geographic distance or the length of the relationship between the lender and the borrower as proxies for soft information, but those data are, alas, not available on the platform. Unfortunately, we are also unable to see whether borrowers were able to draw on multiple loans from the same lender(s) due to the absence of data.

[11] The Chinese P2P markets are the largest in the world. According to data reported by the *Financial Times* (6 August, 2018) loans outstanding at the end of 2017 amounted to Rmb 1.2 tn ($180 bn). According to the National Bureau of Statistics of PRC, the transaction volume of P2P markets in China has actually been even higher - reaching 2.8 trillion yuan at the end of 2017, when the market contained 1931 platforms. Some platforms have recently faced major problems, including Money Cat, Money Pig, and Ezubao, and 168 platforms ended operations in July 2018 alone. For more information, see Appendix A.

[12] We have added 'overdue loans' to 'defaults' for practical reasons. Strictly speaking, this is not the correct procedure since some overdue loans may not end in default, but this procedure does not affect the main argument. The total number of listings is 14,575 (84+590+13,901).

[13] See https://www.renrendai.com/about/ma/6/593e589b0083b60f212288ac

[14] The procedures are fragile and lead to a systemic crisis such as in August 2018 when a collapse of a large P2P Group resulted in a panic of default spread.

[15] We have extracted from the data set as much information as is available. "Verified Weibo account" is all that the data offers. Unfortunately, no additional information about the number of contacts, likes etc. was available to us.

Table 2
Description of independent variables.

Variables	Description
Hard Information	
Income level	Category variable: Monthly income (yuan) of the borrower (1~7) Group 1: <1000 Group 2: 1001~2000 Group 3: 2000~5000 Group 4: 5000~10000 Group 5: 10000~20000 Group 6: 20000~50000 Group 7: >50000
Income verification	Dummy variable: income is verified-1; is not verified-0
Home ownership verification	Dummy variable: ownership is verified-1; is not verified-0
Car ownership verification	Dummy variable: ownership is verified-1; is not verified-0
Mortgage loans	Dummy variable: the borrower has a mortgage loan-1; doesn't have a mortgage loan-0
Soft Information	
Loan description	Length of the loan description
Age	Age of the borrower
Gender	Dummy variable: female-1; male-0
Marital status	Dummy variable: married-1; otherwise-0
Educational level	Years of education
Weibo verification	Dummy variable: the social network is verified-1; is not verified-0
Mobile verification	Dummy variable: the mobile number is verified-1; is not verified-0
Video verification	Dummy variable: finished the video verification-1; otherwise-0
Loan features	
Interest	Interest rate of the loan
Term	Length of the loan
Amount	Amount of the loan

mobile phone users, which enables lenders to trace and verify the real users of cellphones. This increases borrower transparency and enhances trust in the information provided by borrowers. In addition, like many other emerging markets, the Chinese financial markets have a short investment history and relatively low public financial literacy, so credit analyses based on a broad range of indicators are of utmost importance in this market. Furthermore, the Chinese P2P sector is regulated by monetary authorities. Though the regulatory system is probably relatively light, it is highly sensitive to systemic instability and operates on both the formal and informal levels.[16]

3.2. Model

As noted above, in prior literature, determinants of default have been studied from three different perspectives. Our model starts from the traditional approach to credit appraisal, which emphasizes the key role played by financial (hard) information. Variant 1 of the model contains, therefore, only hard information variables together with other control variables. Our second variant is entirely focused on soft information as a determinant of defaults, to which we add the same control variables. Finally, we explore the joint effects of hard and soft variables together with our control variables in variant III of our model.

We test two hypotheses:

Hypothesis 1. Credit appraisal based on appropriately selected soft information can have strong predictive power: i.e., soft information coefficients are significantly non-zero;

Hypothesis 2. The credit predicting model can be strengthened by soft information. Soft information can capture useful information that is not included in hard information for credit analysis.

In order to estimate the probability of default, we chose a binary regression estimation model – logit regression. A receiver operating characteristic (ROC) curve is used to compare the performance of soft and hard information models as one way of discriminating among different estimates.

$$\textbf{Model I}: Y_i = \alpha \text{ Hard Information} + \propto \text{ Control Variables} + \varepsilon \tag{1}$$

$$\textbf{Model II}: Y_i = \alpha \text{ Soft Information} + \propto \text{ Control Variables} + \varepsilon \tag{2}$$

$$\textbf{Model III}: Y_i = \alpha \text{ Hard Information} + \beta \text{ Soft Information} + \propto \text{ Control Variables } + \varepsilon \tag{3}$$

Y is the dependent variable which represents whether the loan has been repaid completely without delay. 1 represents 'default'; 0 represents 'repaid'. The control variables are loan features, including the interest rate, the length of the loan, and the amount of the loan.

The proxies for the hard information in our model are the key financial determinants that indicate the wealth and solvency of the borrower. They are the four key fundamental financial indicators that are available in our dataset: monthly income, home ownership, car ownership, and existing mortgage loans. The car and home ownership are dummy variables with the value of 1 for "ownership" and 0 for "none'. Following Order & Zorn (2000), we have also chosen monthly income as an independent variable. We include verification of income in the model to certify accuracy.

As soft information is difficult to measure, it is necessary to use proxies. The proxies in our model are summarized in Table 2. Our treatment of soft information is similar as in the literature: we use duration of education (e.g. Liao et al. (2015)), age (e.g., Gonzalez & Loureiro (2014)) and gender (e.g. Barasinska & Schäfer, 2010; Ravina, 2012; Pope & Sydnor, 2011). Following Lin et al. (2013), we also use the length of the loan purpose description as a linguistic indicator.

Due to limitations of the dataset, we cannot obtain data on the discussion groups on the RenrenDai.com platform. Thus, we use verification data from the largest Chinese social network, Weibo, as the second-best option and as our indicator of social impact. According to the "Weibo 2016 Development Report", there were 297 million active users of Weibo at the end of September 2016. This guarantees that Weibo verification is useful as a social image proxy. If an applicant's social network was verified, it is represented as "1", otherwise "0".

The Chinese P2P lending platform does not usually provide real photos as the profile pictures of the members. We have, therefore, chosen video verification as the proxy for the image indicator. This can also be regarded as a social indicator. During a verification process, borrowers are required to video themselves holding their ID cards and reading a statement accepting the general rules and conditions from of Renrendai.com, and then to upload the video with their loan application. If the applicant agreed to have video verification, it is represented as "1", otherwise "0". The explosion of mobile services provides the key element of Fintech 2.0, and mobile data is the preferred instrument of verification by Fintech companies, especially for big data companies. It is the essential source for anti-fraud measures since mobile numbers have been added to the real-name system in China, allowing tracking and verification of the real users of cellphones. In addition, mobile usage behavior is recognized as one of the most effective indicator of default in the industry. Thus, we add the mobile verification variable to our model. It is also a dummy variable: "1" equals verified, otherwise "0".

[16] Oversight of Chinese banks continues to be closely linked to the government's regulation and financial policy. See, for example, Zha (2011). Nevertheless, the P2P market is seen as lightly regulated and subject to imperfect regulations, but with implicit state support. The number of platforms was down to 1504 by June 2018, a reduction of 42 percent from its peak in 2015. The reduction reflected consolidation within the sector, but was also due to regulatory interventions by the oversight authorities.

Table 3
Four cases for binary classification.

		Predicted Class	
		Class 1	Class 0
Actual Class	Class 1	True Positives	False Negatives
	Class 0	False Positives	True Negatives

Table 4
Logit regression results for model I.

VARIABLES	(1) default	(2) default	(3) default
Income verified	0.765***	0.775***	–0.263
	(0.210)	(0.219)	(0.226)
1.Income	–0.795	–0.629	–0.739
	(1.015)	(1.021)	(1.043)
2.Income	–0.0458	–0.905***	–0.493
	(0.310)	(0.332)	(0.343)
3.Income	–0.320**	–0.355***	–0.360***
	(0.129)	(0.131)	(0.135)
5.Income	–0.256	–0.265	–0.360**
	(0.161)	(0.166)	(0.173)
6.Income	0.370***	0.431***	0.354**
	(0.132)	(0.136)	(0.139)
7.Income	0.444***	0.523***	0.382***
	(0.125)	(0.133)	(0.138)
Incomeverified 1.Income	0	0	0
	(0)	(0)	(0)
Incomeverified 2.Income	1.311	2.341**	2.384***
	(1.211)	(1.104)	(0.879)
Incomeverified 3.Income	0.471	0.487	0.513
	(0.295)	(0.310)	(0.320)
Incomeverified 5.Income	–1.117**	–1.256**	–1.178**
	(0.561)	(0.569)	(0.555)
Incomeverified 6.Income	–1.879***	–1.766***	–1.515***
	(0.558)	(0.566)	(0.574)
Incomeverified 7.Income	–2.518***	–2.342***	–1.913***
	(0.557)	(0.563)	(0.578)
Car verified	–0.0832	–0.201*	–0.0941
	(0.112)	(0.109)	(0.118)
Home verified	0.601***	0.491***	0.627***
	(0.124)	(0.119)	(0.126)
Mortgage loan	–0.482**	–0.394*	–0.525**
	(0.208)	(0.218)	(0.225)
Homeverified#Mortgage loan	–0.378	–0.523*	–0.384
	(0.267)	(0.280)	(0.290)
Interest		0.216***	0.274***
		(0.0118)	(0.0139)
Term		–0.0168***	–0.0403***
		(0.00456)	(0.00516)
Amount		–4.39e - 07	–1.91e - 07
		(4.10e - 07)	(3.79e - 07)
2011.year			0.417
			(0.726)
2012.year			1.248*
			(0.724)
2013.year			1.876***
			(0.725)
2014.year			3.187***
			(0.734)
Constant	–3.129***	–5.895***	–7.929***
	(0.0975)	(0.221)	(0.772)
Pseudo R2	0.0294	0.0852	0.1226
Observations	14,569	14,569	14,569

Heteroscedasticity-Robust, standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ *Note.* The numbers associated with the variable “income” refer to income groups. The sample included 7 income groups. Columns 1–3 represent different model specifications defined by differences in control variables.

3.3. Receiver operating characteristics (ROC) curves

Since our model is estimated in three different versions, we need to determine whether the model estimates can be discriminated purely on econometric, as opposed to theoretical, grounds. A receiver operating characteristic graph is a technique for visualizing and selecting classifiers based on their performance (Fawcett, 2006).

As shown in Table 3, there are four cases for the binary classification model:

True Positives: The predicted class is 1, and the actual class is 1;
True Negatives: The predicted class is 0, and the actual class is 0;
False Positives: The predicted class is 1, and the actual class is 0;
False Negatives: The predicted class is 0, and the actual class is 1.

The ROC curve is the graphical plot that shows the performance of a binary classifier by diagrammatizing the true positive rate (TPR) against the false positive rate (FPR) at different thresholds. The TPR and FPR are known as sensitivity and specificity classification functions in statistics which represent the proportion of positives and negatives of the detection accordingly. The formula for TPR and FPR is as below:

$$TPR = TP/(TP + FN) \tag{4}$$

where TP stands for “true positive” and FN stands for “false negative’. Equation (4) represents the rate of correctly diagnosed numbers among all positive numbers in the sample. Similarly,

$$FPR = FP/(FP + TN) \tag{5}$$

where FP stands for “false positive” and TN for “true negatives’. Equation (5) represents the rate of wrongly diagnosed numbers among all negative numbers in the sample.

The ROC curve can be plotted by the TPR and FPR ratios against their different thresholds. TPR (sensitivity) data are plotted on the vertical axis and FPR (specificity) data on the horizontal axis. An important parameter of the ROC curves is the AUC - the area under the curve. AUC acts as a measure of the accuracy of the classifier, and it represents the probability of the classifier ranking a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). The closer the ROC curve is to the upper left-hand or the closer the AUC is to the value of 1, the truer are the positives defined, indicating a better classifier.

The area under the ROC curve is derived as:

$$ROC(AUC) = \int_1^0 TPR(x)FPR'(x)dx \tag{6}$$

3.4. Robustness tests

In order to verify the solidity of our models we carried out a number of robustness tests of our estimates and results. With Kernel density estimates, we analyzed the structure of interest rates. Since loan characteristics might also influence the loan performance, we analyzed our data in terms of maturity, loan amounts and default rates. We also carried out a test of independence for the chosen variables. This is done partly with the help of correlation matrix and partly through the analysis of the relevant frequency table.

4. Results

Results are presented for the three versions of our model. The predictive power of the hard information on default is tested first. We then compare the results with those in version II of the model, utilizing solely soft information as the key determinant. Finally, we combine hard and soft information in model III. The logit regression results are presented in the following section, and a comparison of the ROC curves for the three models is discussed in Section 4.2. The summary statistics for all variables in the three models are provided in Appendix C.

4.1. The logit regression results

Model I investigates the relationship between the probability of default and traditional hard financial indicators. The results are reported

Table 5
Income distribution.

Group	Monthly Income (yuan)	Freq.	Percent
1	$\leq$ 1000	51	0.35
2	1001–2000	312	2.14
3	2000–5000	4,464	30.6
4	5000–10000	3,235	22.20
5	10000–20000	2,013	13.82
6	20000–50000	2,116	14.52
7	$>$ 50,000	2,378	16.32
	Total	14,569	100.00

Table 6
Logit regression results for Model II.

VARIABLES	(1) default	(2) default	(3) default
Loan description	–0.00647***	–0.00641***	–0.00562***
	(0.000532)	(0.000551)	(0.000546)
Age	0.00174	–0.000483	0.00480
	(0.00558)	(0.00601)	(0.00596)
Gender	–0.317**	–0.262**	–0.231*
	(0.126)	(0.128)	(0.129)
Marriage	–0.353***	–0.266***	–0.202**
	(0.0972)	(0.0999)	(0.101)
Education	–0.122***	–0.117***	–0.122***
	(0.0155)	(0.0160)	(0.0165)
Mobile verified	–0.555***	–0.523***	–0.639***
	(0.126)	(0.129)	(0.132)
Weibo verified	–0.802***	–0.701***	–0.453***
	(0.147)	(0.150)	(0.154)
Video verified	0.908***	0.936***	0.976***
	(0.113)	(0.120)	(0.123)
Interest		0.191***	0.242***
		(0.0132)	(0.0144)
Amount		0.0687*	0.0609
		(0.0400)	(0.0444)
Term		0.0102*	–0.00653
		(0.00536)	(0.00595)
2011.year			0.423
			(0.740)
2012.year			0.929
			(0.739)
2013.year			1.403*
			(0.737)
2014.year			2.257***
			(0.746)
Constant	0.0863	–3.535***	–5.545***
	(0.351)	(0.574)	(0.943)
Pseudo R2	0.1127	0.1483	0.1694
Observations	14,571	14,571	14,571

Heteroscedasticity-Robust, standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ *Note.* Columns 1–3 represent different model specifications defined by differences in control variables.

in Table 4.

Table 4 presents our logit regression results for model I. The model investigates the relationship between traditional hard credit information and default behavior. The interest rate, amount and term are used to control for the omitted variable bias. Since we are using a panel dataset, year dummy variables are added to control for heterogeneity in the adjusted model (last column).

Variable income represents borrowers" monthly income; the seven income categories are shown in Table 5. The median income group (5000∼10000 yuan) is the reference group for the variable income category. The interaction effect of income and verified income is significant except in Goup 3 and Group 1. None of the borrowers in income Group 1 has verified their income thus been omitted. The coefficient proves that borrowers who earn 1001∼2000 yuan are more likely to default than those who are in the reference group. This is consistent with Order & Zorn (2000), who found that defaults and losses were higher in low-income groups. Borrowers who have higher than 10,000 yuan monthly income are less likely to default than the borrowers in the reference group. Car ownership as an indicator of stronger financial status is insignificant in the model and should not necessarily be regarded as a significant indicator of default behavior.

Some interesting results occurred in the case of the effect of home ownership. A home ownership certificate turns out to be significantly positively related to default behavior. This may indicate that traditional real estate collateral does not guarantee creditworthiness on online P2P lending platforms, or that there is an adverse selection problem in the online lending market. This finding is also consistent with results obtained by Jiménez & Saurina (2004). Moreover, the mortgage loan variable is significantly and negatively related to default behavior. In other words, if the applicant is in debt for a mortgage loan, he/she is less likely to default on the P2P lending platform. This, in turn, could indicate that borrowers with mortgage loans care more about their credit standing. The violation of the traditional use of home ownership as an indicator of default also hints at the need for other important information in the internet lending market. The goodness of fit indicator (Pseudo R2) is increasing along with the addition of control variables and year dummies. The same feature is consistent with the log-likelihood estimations. In general, the results show that hard financial factors representing the wealth and solvency of the borrower do not predict as well as expected; some even show opposite results to those expected in the P2P lending market.

Model II analyzes the relationship between the probability of default and soft credit information. The results are presented in Table 6.

Table 6 presents the logit regression results for model II which analyzes the relationship between soft information and the probability of default. As shown above, the length of the loan purpose description is negatively related to the probability of default, and the results remain consistent after adding the control variables and the fixed effects of the year. This means that the more words the applicant wrote on the loan purpose description, the less likely it is that such an applicant will default. Results for the effects of gender are consistent with the literature and show that women are less likely to default than men. In addition, marital status and educational level are also significant variables. Since the length of education is used to express the educational level, the results show that the longer the applicant spent in training or schooling, the less likely it is that he/she will default on P2P loans. We also discovered that borrowers with a higher educational level tend to borrow higher amounts over shorter terms. Borrowers with a master's degree or above have a higher average borrowing amount (67927.25 yuan) than the total average loan amount (47547.51 yuan), and they also tend to borrow for a shorter period of time (average 9.5 months) than the general population (12.4 months). This could be due to the higher income levels and higher demand for funds among borrowers with higher levels of education. The shorter terms may indicate that they tend to borrow safer loans and have the ability to repay them in a shorter period of time. Marital status is a significant factor both before and after the robustness treatment, and illustrates that people with a spouse are less likely to default.

The three social capital variables are all significantly related to the probability of default. Mobile verification and social network verification have negative correlation with the probability of default. We also found that borrowers with Weibo and mobile verification tend to borrower safer loans. Borrowers with Weibo verification have a much lower average borrowing amount (12027.19 yuan) than the overall average (47547.51 yuan). They also have quite short average terms, of 6.45 months. Similar results have been obtained for mobile verified borrowers; they also tend to borrower loans of lower amounts (average 19050.48 yuan) and over shorter terms (average 6.658986 months). This may be an indication of borrowers caring about their social image; thus, they tend to borrow safer loans and potentially have a lower risk of defaulting. However, for video verification, there is a positive correlation with defaults. Video verification is not a mandatory procedure; indeed only 37.56% of people are video-verified. This could suggest that

Table 7
Logit regression results for Model III.

VARIABLES	(1) default
Income verified	–0.184
	(0.231)
1.Income	–1.146
	(1.196)
2.Income	–0.268
	(0.351)
3.Income	–0.146
	(0.137)
5.Income	–0.389**
	(0.173)
6.Income	0.284*
	(0.150)
7.Income	0.283*
	(0.157)
Income verified1.Income	0
	(0)
Income verified2.Income	2.764***
	(0.803)
Income verified3.Income	0.409
	(0.336)
Income verified5.Income	–1.135*
	(0.583)
Income verified6.Income	–1.548***
	(0.594)
Income verified7.Income	–1.891***
	(0.578)
Car verified	–0.295**
	(0.116)
Home verified	0.455***
	(0.128)
Mortgage loan	–0.573**
	(0.225)
Homeverified#Mortgage loan	–0.0162
	(0.287)
Loan description	–0.00537***
	(0.000560)
Age	–0.00171
	(0.00623)
Gender	–0.254**
	(0.129)
Marriage	–0.130
	(0.106)
Educational	–0.121***
	(0.0171)
Mobile verified	–0.579***
	(0.136)
Weibo verified	–0.403**
	(0.157)
Video verified	1.006***
	(0.127)
Interest	0.243***
	(0.0151)
Amount	0.00969
	(0.0497)
Term	–0.00298
	(0.00637)
2011.year	0.343
	(0.743)
2012.year	0.831
	(0.743)
2013.year	1.386*
	(0.742)
2014.year	2.522***
	(0.754)
Constant	–4.903***
	(0.976)
Pseudo R2	0.189
Observations	14,566

Heteroscedasticity-Robust, standard errors in parentheses *** $p < .01$, ** $p < .05$, * $p < .1$ *Note.* The numbers associated with the variable "income" refer to income groups. The sample includes 7 income groups.

borrowers with a higher probability of default may have the incentive to disclose more information in order to make themselves seem more trustworthy.

The only variable that turns out to be insignificant is age. We also discovered that borrowers" age distribution for defaulted loans has a significant overlap with general loan distribution, thus providing robustness for this result. This is consistent with Santoso et al. (2020) but is not consistent with Pope & Sydnor (2011), whose findings reveal that the default rate is usually high within both the extremely young and extremely old age groups. We didn't observe this pattern in our dataset. This is probably becasue the percentages of extremely young and extremely old people are quite limited. Only 0.38% of borrowers are younger than 23, and only 0.27% are above 60. This may also be due to the fact that an especially young person does not usually have a high demand for funds, and especially old people are often unfamiliar with online lending.

Table 7 presents the logit regression results with the combined effect of soft and hard independent variables. The significance and the direction of all variables remained consistent with the previous models I and II except for the effect of car ownership, which turns from insignificant to significant. The pseudo R2 is increasing from the 0.123 (model I) and 0.169 (model II) to 0.189 (model III). The results for our control variables showed that the higher the interest rate the higher the probability of default. The amount and the term are insignificant – possibly, because most of the loans in the P2P platform are relatively small and short.[17] This suggests that the combination of hard and soft information can better predict loan performance. It should also be note that the improvement is unlikely to come from different loan terms given for loans based on hard and soft information. Using the Kernel density technique, we found that the terms of loans related to hard and soft information are normally distributed with means that were around similar values.

As can be seen in Table 7, the probability of default is increasing for the top two income groups. However, borrowers with "verified income" are shown to be less likely to default. Since only a small fraction of incomes was verified, we suspect that the hard information on income may have been misrepresented. We believe that combining hard and soft information can provide valuable input into load approvals by identifying possible sources of misrepresentation stemming from hard data as in this case.

In order to increase the confidence level in our findings, we take additional steps and tests in the following section. We use the ROC curve technique to help in discriminating among the three models. In addition, we carried out various tests and data examinations to check for the robustness of our results.

4.2. Model discrimination and tests of robustness

All three versions of our model generated significant results for most of the variables tested. We wanted to see if it is possible to identify which of the models performs best. Before addressing this from a theoretical point of view, we turned to the ROC statistical technique described in the methodology section. The ROC curves were used to measure the performance of the default prediction model. Visually, the more the curve approaches the upper left-hand corner (0,1), the better the performance of the model. An alternative way to assess the performance of the estimations is to look at AUC, as it is increasing with the addition of "better" information.

We have generated three ROC graphs corresponding to our three models and they are presented in Figs. 1–3. ROCs derived from model I (hard information) and model II (soft information) are shown in Fig. 1 and Fig. 2, respectively. ROC in blue represents the curve from the basic model (hard information and soft information respectively) without

[17] For details, see Appendix E.

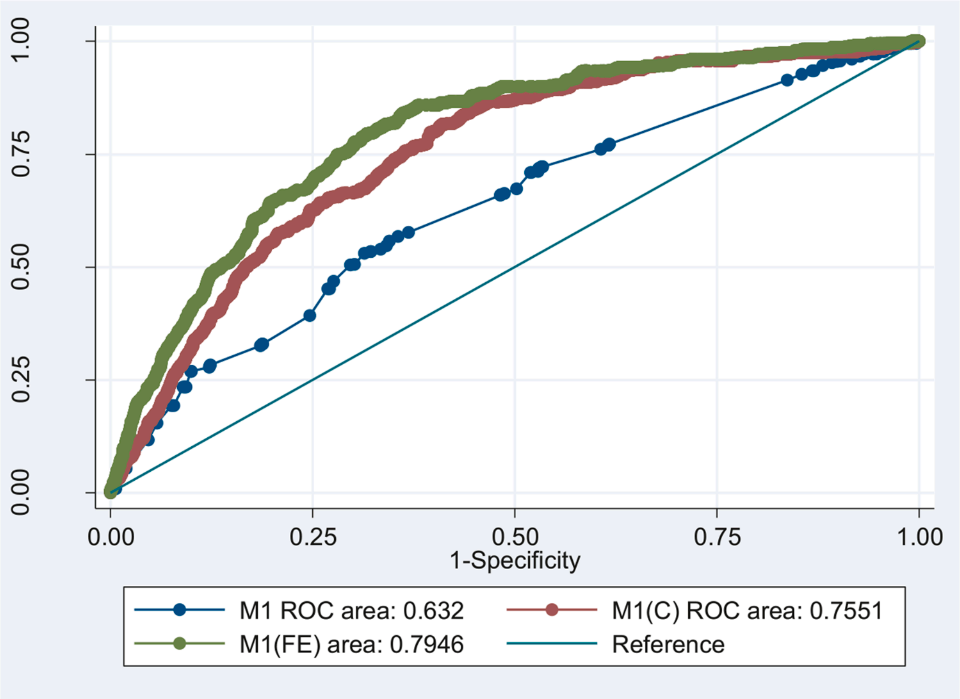


Fig. 1. ROC curves for model I.

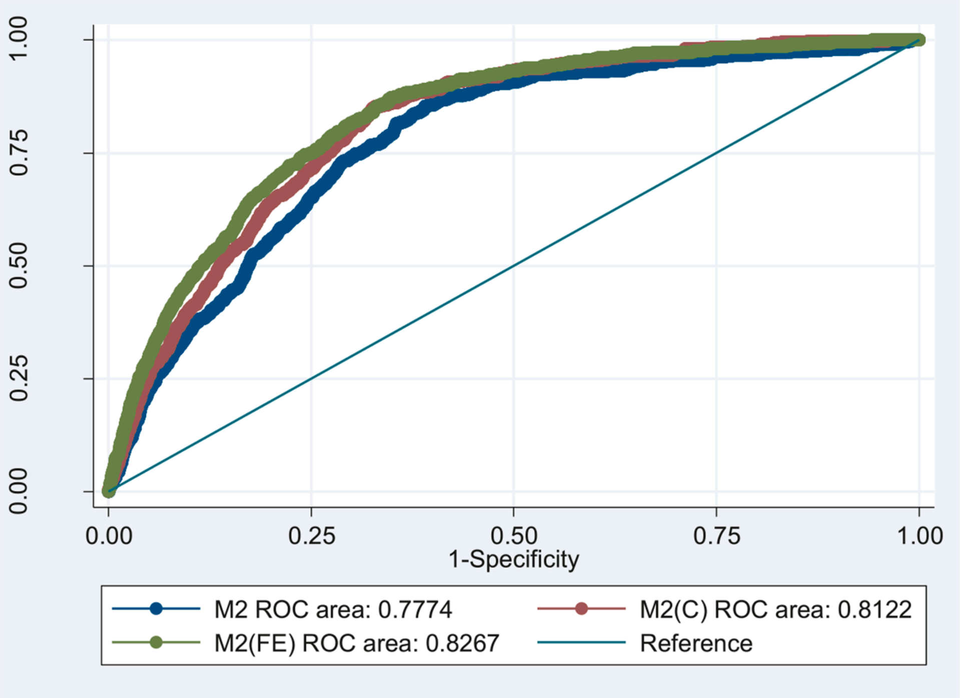


Fig. 2. ROC curves for model II.

control variables and a dummy for years. ROC in red represents the basic model plus control variables, and ROC in green represents the basic model plus control variables and a dummy variable for years. Fig. 3 presents the robustness model with control variables and year dummies for model I (blue), model II (red), and model III (green).

Starting with Fig. 1, the AUC in model I is increasing with the addition of the control variables and increasing even more with the addition of the year dummy. This is also in accordance with our results from the pseudo R square of model I.

As in model I, the AUC for model II (in Fig. 2) is increasing by adding the robustness treatment variables. However, the growth interval is not as large as in model I.

The AUC computations are summarized in Table 8. Recalling equation (6) above, we calculated and compared the AUC in model III (curve related to hard and soft information combined) with that of model I (hard information) and model II (soft information). The ROC in model III has the largest AUC; it is 0.0473 larger than the AUC in model I and 0.0151 larger than in model II. This indicates model III has the highest

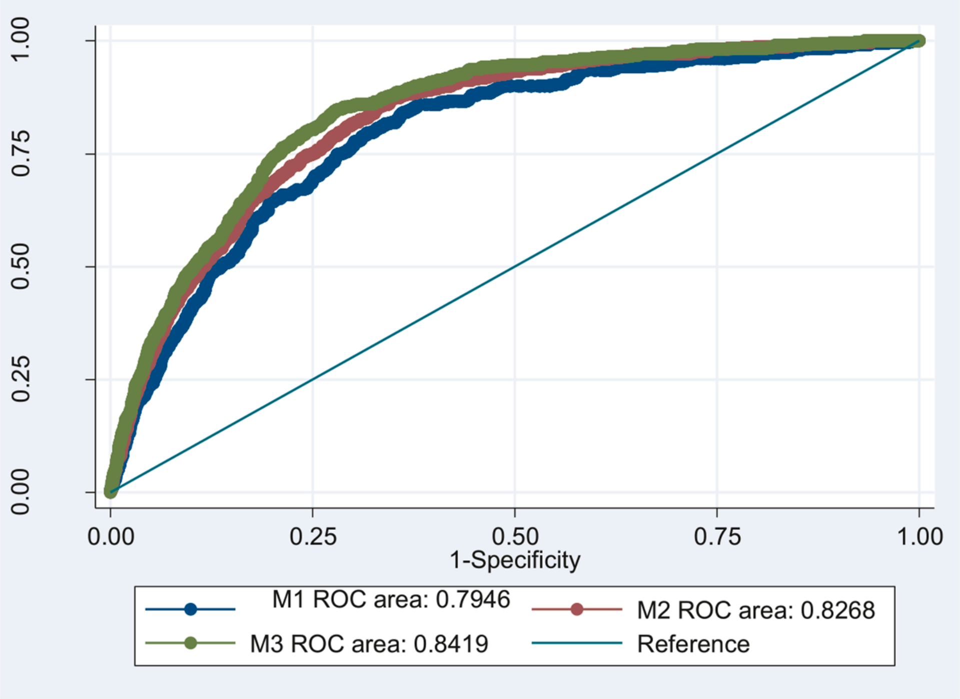


Fig. 3. ROC curves for model comparisons.

Table 8
ROC results of hard and soft information models.

	Obs	Area	Std. Err.	[95% Conf. Interval]	
Hard	14,566	0.7946	0.0084	0.77819	0.81098
Soft	14,566	0.8268	0.0073	0.81249	0.84107
Combined	14,566	0.8419	0.0069	0.82829	0.85553

Ho: area (Hard) = area (Soft) = area (Combine) chi2(1) = 133.48 Prob >chi2 = 0.0000

accuracy as a default screening classifier. Model III, which includes the soft information, has 4.73% higher probability of correctly distinguishing default and non-default borrowers than model which doesn't include soft information. Other interesting results were obtained from these tests when we compared the ROC curves of model I and model II. The closer the ROC curve is to the upper left-hand or the closer the AUC is to the value of 1, the truer are the positives defined, indicating a better classifier. As shown in Fig. 3, the curvature of the ROC for model II is bigger than model I, in other words, red curve is closer towards the upper left corner than blue curve. This indicates that soft information variables have a stronger effect on classifying default borrowers than hard information variables.

Additional analyses were conducted to check for the robustness of our results. Our sample of borrowers has a few characteristic features that could raise questions about the possibility of a bias generated by the aggregated values of defaults. Specifically, we have different groups of borrowers identified by income levels and borrowers identified by gender. More than 80 percent of the borrowers in our sample are male and more than 50 percent belong to only two income groups of the seven total groups.[18] After more detailed examinations of the structure of defaults, we did not find any abnormalities concerning the default pattern, neither among different income groups nor between males and females. In addition, since our data for 2014 only covers the first six months of the year, we also carried out a sensitivity test involving a comparison of data for comparable periods in the preceding years and, again, we did not find any irregularities. Finally, in order to test for changes in the regulatory environment that were introduced in 2015, we also analyze the structure of defaults before and after that date and obtained similar results.

As noted in the previous section, we have assumed that hard and soft information are independent of each other and do not lead to biased estimates. In the absence of perfect guidance from theory to identify a complete set of proxies for hard and soft information and due to limitations of data, we have to rely on further robustness tests. Using collinearity diagnostics based on the analysis of variance inflation factors, we did not find any evidence of multicollinearity. As shown in Appendix B, variance inflation factors (VIF) of the independent variables (shown in column 1 in the table) are in the range of 1.03 to 2.21 and with a mean VIF of 1.4. In other words, the variance of the estimated coefficients is inflated with very low factors and within the reasonable “rules of thumb” of 10.

In order to control for borrower misrepresentation, an overlap check of our hard and soft information variables has been conducted. The analysis confirmed that at least the key soft information variables (verified video, verified mobile and verified Weibo) do not overlap with the key hard variable – income – and that whatever overlap exists is small. In other words, the borrowers who verified their incomes were not the ones who verified their mobile and Weibo information. In addition, an analysis of interest rates charged to borrowers showed that applicants with Weibo or mobile verified information did not receive better terms than those without the soft information indicators.[19]Furthermore, as we have also noted above, analyses of determinants of loan approvals and defaults are subject to imperfect information, which raises the question of missing variables. We have, therefore, carried out

[18] See Appendix C

[19] Unfortunately, we were not able to examine the extent to which borrowers obtained funds from different lenders or whether a particular lender had bids on multiple loans. This information is not available on the Chinese platform as it is in the US dataset (Prosper).

additional tests using modified instrumental variables.[20] The results of these second stage estimations were similar to the results obtained in the first stage – all our estimators are statistically significant and the best results are obtained from the hybrid hard and soft information model.

We also tested the soft information explanation power in the screening process. We ran the regression with the same hard and soft variables for the successfully funded dummy (loan successfully funded - 1; loan not funded - 0). As shown in Table F.17, the pseudo R square is 0.4459 for the hard information model and 0.5519 for the soft information model. The area under the ROC curve (AUC) shows the same results. The AUC for the hard information model is 0.9126, while the AUC for the soft information model is 0.9395. The difference between the AUCs for the hard and soft information models for the successfully funded dummy is 0.0269 (0.9395-0.9126). This is quite similar to the difference between the hard and soft information models for the default dummy, which is 0.0322 (0.8268-0.7946). This indicates that the screening procedure does not bias the dataset used to test the default behavior, because investors employed both soft and hard information during the screening process.

As an additional test of robustness, we have carried a detailed analysis of the term structure of the loans, interest rates and other conditions of loans including, in particular, the use of soft indicators, for all the different classes of loans. Using different techniques of analysis, we have found that the interest rate structure was similar for all classes of loans. The term structure was also almost identical for all three classes. This is not surprising since the maturity was entirely short-term and determined by the conditions of the market. The default rates were similar on all three classes of loans., This suggests that the different purpose had small influence on loans default, if any.

These results lead to tentative conclusions. First, soft information provides valuable input into loan appraisals and predicting defaults. The results of the comparison of the hard and soft information models (Table 8) indicate that soft information may even be of equal importance to hard information in credit analyses performed by online lending systems. As the combined model with soft and hard information has the highest predictive value, this would suggest that soft information can strengthen the default predicting model.

5. Conclusion and discussion

This paper investigates the predictive power of soft and hard information on the loan performance in P2P lending. Our results of predictive power of the hard information is consistent with the existing literature. We also add evidence to the literature (e.g. Jiménez & Saurina (2004)) proving that collateral does not necessarily secure the non-default behavior. The estimates of the effects of gender, marital status, and educational level are all consistent with the literature, notwithstanding different views in the case of age (e.g. Ravina, 2012). The length of the loan purpose description performs very well in the estimation of the probability of default and is also consistent with Lin et al. (2013). All three social capital proxies – Weibo verification, video verification, and mobile verification – are statistically significant as determinants of defaults. However, there are some interesting findings in our results like the positive relationship between video verification and default possibility, and the opposite relationship with default for the verified and un-verified high-income group. This suggests the possibility of borrowers lying about the information they disclose online, in order to create a more trustworthy image. In practical terms, we need to take measures to control the possible lying behaviour of borrowers when using subjective social-related soft information. A possible solution could be building a deep learning algorithm to depict the social image of the borrower and detect the contradicting information in the pool of data, and then assigning penalty score of the unauthentic behavior.

It is quite likely that loan appraisals using better soft data could be further enhanced by other and, perhaps, better proxies for social and psychological factors. Clearly, this field is open and will undoubtedly develop over time. Better information already exists at various levels of business, such as more advanced social media data. With more comprehensive information technology and an enlarged dataset about repayment history, further research can be performed to analyze different repayment behaviors from different social identities. However, it is increasingly unlikely that such data will be accessible to financial markets due to rising concerns about data privacy as exemplified by the privacy protection laws adopted this year by the European Union and state of California.

Perhaps the most interesting and somewhat surprising result is that even on its own, soft information can play an important role in credit appraisal and in predicting defaults. We obtained even better results when we combined hard and soft information in our model III. These results are consistent with experiences in the Fintech industry from other countries, and they are also consistent with the findings of Cornée (2017).[21]

It could be said that our method of assessing the role of soft information may lead to biased estimates. Critics could argue that a bias could be generated by the absence of soft information in our hard information model and vice versa in our soft information model. However, if our model I and II are biased, it could also be argued that the results will be biased even in model III, since we are likely using imperfect information. Our model III may be the best and most accurate, but it may still not be optimal. Given the manner in which we use soft information, the proxies can only provide a lead as to which soft information should be used to predict defaults, but they cannot identify the intensity of that effect. The only perfect solution would have to come from a theory that would identify the complete set of proxies for hard and soft information and from the availability of such data. Without such a theory, the best that can be done are robustness tests and those, as we have seen, are quite encouraging.

Finally, we should also acknowledge that the incorporation of psychological and social factors into soft information could complicate international comparisons. Since psychological and social factors are influenced by the culture of a given country, it is quite likely that the relevant sets of psychological and social factors should vary from country to country. *Pari passu*, the value of identical models applied to different countries may be diminished, as would be our ability to generalize.

Some of the policy implications of this work are evident. As our results emphasize the importance of soft information, they provide empirical evidence in support of measures to encourage greater use of soft information in addition to hard information in credit analysis. The importance of soft information is considerably greater in situations when hard information is missing or has poor quality. The importance and availability of soft information will increase with the development of technology and information "hardening" tools. This is also in line with the expansion of credit in the age of big data. However, if implemented, this would considerably increase the challenges for regulators. Microfinance banks and non-bank financial institutions are already regulated by local or regional banking supervisors. Moreover, regulatory agencies would have to pay far more attention to lending based on the use of soft information, its quality, its dissemination, and data privacy, which will

[20] The re-specifications included squaring some of the independent variables, introducing interaction terms between "amount" and "interest", "term" and "amount", and in a few cases dropping some of the variables. The relevant specifications of the models are, therefore, slightly different in the two stages but the models retain the fundamental features. The results are reported in Appendix D.

[21] While credit scores continue to be important both in the US Community Banking sector and for the US Fintech firms, the value of soft information in credit appraisal is increasingly recognized by both of these industry segments. We are grateful to I. Lieberman for sharing his findings on this with us.

require a considerably different range of skills than in traditional lending. Legislative steps are very likely to be needed in order to fully reflect technological changes in the Fintech industry and in financial markets.

Acknowledgment

Financial assistance from GACR under project number 18-04630S is gratefully acknowledged. We have benefited from discussions with the staff and graduate students of the Institution for Economic Studies (IES) at Charles University and the staff of Tsinghua University. We are also grateful to I. Lieberman and N. Hicks, former lead economists at the World Bank and professors K. Raffer of the University of Vienna, and H. Poenisch of Zhejiang International Business School and colleagues at US and UK for their insights and detailed comments. We have also benefited from the helpful comments of Professor J.A.C. Brada, who helped us improve the presentation and the arguments. Comments by anonymous referees for the Working Paper series of IES of Charles University and Journal of Behavioral and Experimental Economics are gratefully acknowledged.

Appendix A. Chinese P2P Key Market Indicators

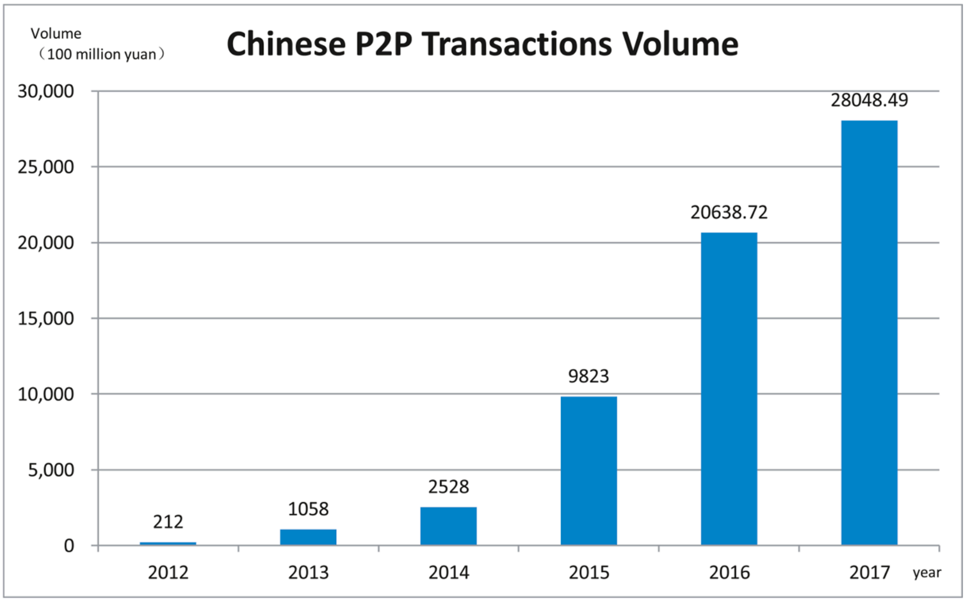


Fig. A.4. Chinese P2P key market indicators. Source: Annual P2P Industrial Report, https://www.wdzj.com/.

Appendix B. Collinearity Diagnostics Table B.9 provides the results of variance inflation factors of the independent variables.

Table B.9
Collinearity diagnostics.

Variable	VIF	SQRT VIF	Tolerance	R-Squared
Income verified	1.04	1.02	0.9648	0.0352
Income	1.44	1.20	0.6932	0.3068
Car verified	1.54	1.24	0.6480	0.3520
Home verified	1.65	1.28	0.6071	0.3929
Mortgage loan	1.26	1.12	0.7937	0.2063
Loan description	1.52	1.23	0.6562	0.3438
Age	1.31	1.15	0.7620	0.2380
Gender	1.03	1.02	0.9687	0.0313
Marriage	1.18	1.09	0.8478	0.1522
Education	1.11	1.05	0.9020	0.0980
Mobile verified	1.42	1.19	0.7028	0.2972
Weibo verified	1.41	1.19	0.7068	0.2932
Video verified	1.53	1.24	0.6520	0.3480
Interest	1.10	1.05	0.9085	0.0915
Amount	2.21	1.49	0.4529	0.5471
Term	1.69	1.30	0.5928	0.4072
Mean VIF	1.40			

Note. Column 1 includes the independent variables of the model. Figures in column 2 show variance inflation factors (VIF), figures in column 3 provide corresponding figures for squared root VIF. The tolerance indicators computed as 1- R squared are in column 4 and R squared figures for correlation between the given independent variable and the rest of independent variables are shown in column 5. Since the tolerance is just the reciprocal of the VIF, they essentially provide the same information and are included for the convenience of readers.

Appendix C. Statistical Summary of Variables Table C.10 provides the statistical summary of the independent variables.

Table C.10
Statistical summary of variables.

Variable	Obs	Mean	Std. Dev.	Min	Max
Length of Loan Description	14,575	259.7457	96.12812	3	367
Age	14,575	35.76123	7.967914	21	72
Interest (APR%)	14,575	13.31848	2.607268	3	24.4
Term (months)	14,575	12.40254	9.528335	1	36
Amount (yuan)	14,575	47547.51	128784.2	3000	3000000
Educational Level	Freq.	Percent			
High School	4,806	32.98			
Technical College	5,594	38.39			
University	3,837	26.33			
Master or Higher	334	2.29			
Total	14,571	100.00			
Income (yuan)	Freq.	Percent			
≤1000	51	0.35			
1001∼2000	312	2.14			
2001∼5000	4,464	30.64			
5001∼10000	3,235	22.20			
10001∼20000	2,013	13.82			
20000∼50000	2,116	14.52			
>50000	2,378	16.32			
Total	14,569	100.00			
Home Ownership	Freq.	Percent			
No	8,084	55.46			
Yes	6491	44.54			
Total	14,575	100.00			
Gender	Freq.	Percent			
Female	2,636	18.09			
Male	11,939	81.91			
Total	14,575	100.00			
Income Verification	Freq.	Percent			
Unverified	13,228	90.76			
Verified	1,347	9.24			
Total	14,575	100.00			
Mortgage loans	Freq.	Percent			
Don’t have	12,084	82.91			
Have	2,491	17.09			
Total	14,575	100.00			
Home Ownership Verification	Freq.	Percent			
No	10,838	74.36			
Yes	3,737	25.64			
Total	14,575	100.00			
Car Ownership	Freq.	Percent			
No	8,439	57.90			
Yes	6,136	42.10			
Total	14,575	100.00			
Car Ownership Verification	Freq.	Percent			
No	10,489	71.97			
Yes	4,086	28.03			
Total	14,575	100.00			
Marriage Status	Freq.	Percent			
Single	3,611	24.78			
Married	10,964	75.22			
Total	14,575	100.00			
Weibo Verification	Freq.	Percent			
No	12,100	83.02			
Yes	2,475	16.98			
Total	14,575	100.00			
Video Verification	Freq.	Percent			
No	9,101	62.44			
Yes	5,474	37.56			
Total	14,575	100.00			
Mobile Verification	Freq.	Percent			
No	11,971	82.13			
Yes	2,604	17.87			
Total	14,575	100.00			

Appendix D. Sensitivity Tests The sensitivity tests results for Model I, II, III are presented in Table D.11, Table D.12, Table D.13accordingly.

Table D.11
Sensitivity tests results for Model I.

VARIABLES	(1) default	Test Results
Income verified	–0.263	–0.0157
	(0.226)	(0.241)
1.Income	–0.739	–0.798
	(1.043)	(1.033)
2.Income	–0.493	–0.309
	(0.343)	(0.322)
3.Income	–0.360***	–0.190
	(0.135)	(0.132)
5.Income	–0.360**	–0.290*
	(0.173)	(0.165)
6.Income	0.354**	0.437***
	(0.139)	(0.140)
7.Income	0.382***	0.509***
	(0.138)	(0.140)
Incomeverified#1.Income	0	0
	(0)	(0)
Incomeverified#2.Income	2.384***	2.147
	(0.879)	(1.524)
Incomeverified#3.Income	0.513	0.347
	(0.320)	(0.318)
Incomeverified#5.Income	–1.178**	–1.185**
	(0.555)	(0.582)
Incomeverified#6.Income	–1.515***	–1.679***
	(0.574)	(0.571)
Incomeverified#7.Income	–1.913***	–2.074***
	(0.578)	(0.572)
Car verified	–0.0941	0.0536
	(0.118)	(0.104)
Home verified	0.627***	0.658***
	(0.126)	(0.112)
Mortgage loan	–0.525**	–.913***
	(0.225)	(0.209)
Homeverified#Mortgage loan	–0.384	0.0841
	(0.290)	(0.271)
Interest	0.274***	1.415***
	(0.0139)	(0.136)
Term	–0.0403***	
	(0.00516)	
Amount	(–1.91e-07)	–2.54e-06***
	(3.79e-07)	(9.27e-07)
Interest square		–0.0342***
		(0.00406)
Amount square		9.21e - 13*
		–4.92E - 13
2011.year	0.417	0.570
	(0.726)	(0.733)
2012.year	1.248*	1.264*
	(0.724)	(0.732)
2013.year	1.876***	1.799**
	(0.725)	(0.733)
2014.year	3.187***	3.122***
	(0.734)	(0.746)
Constant	–7.929***	–17.54***
	(0.772)	(1.350)
Pseudo R2	0.1226	0.1288
Observations	14,569	14,569

Heteroscedasticity-Robust, standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table D.12
Sensitivity tests results for model II.

VARIABLES	(1) default	Test Results
Loan description	–0.00562***	–0.00645***
	(0.000546)	(0.000493)
Age	0.00480	–0.000916
	(0.00596)	(0.00608)
Gender	–0.231*	–0.311**
	(0.129)	(0.127)
Marriage	–0.202**	–0.311***
	(0.101)	(0.100)
Educational	–0.122***	–0.114***
	(0.0165)	(0.0169)
Mobile verified	–0.639***	–0.460***
	(0.132)	(0.122)
Weibo verified	–0.453***	–0.593***
	(0.154)	(0.149)
Video verified	0.976***	1.092***
	(0.123)	(0.106)
Interest	0.242***	
	(0.0144)	
Amount	0.0609	–2.17e - 06**
	(0.0444)	(8.78e - 07)
Term	–0.00653	0.335***
	(0.00595)	(0.0268)
Amount square		7.08e - 13
		(5.66e - 13)
Term square		–0.0105***
		(0.000964)
2011.year	0.423	–0.0389
	(0.740)	(0.735)
2012.year	0.929	–0.175
	(0.739)	(0.733)
2013.year	1.403*	0.0167
	(0.737)	(0.731)
2014.year	2.257***	1.222*
	(0.746)	(0.736)
Constant	–5.545***	–1.970**
	(0.943)	(0.812)
Pseudo R2	0.1694	0.1642
Observations	14,571	14571

Heteroscedasticity-Robust, standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table D.13
Sensitivity tests results for Model III.

VARIABLES	(1) default	Test Results
1.Income verified	–0.184	–0.190
	(0.231)	(0.244)
1.Income	–1.146	–1.215
	(1.196)	(1.082)
2.Income	–0.268	0.0557
	(0.351)	(0.326)
3.Income	–0.146	–0.359***
	(0.137)	(0.136)
5.Income	–0.389**	–0.524***
	(0.173)	(0.169)
6.Income	0.284*	0.0144
	(0.150)	(0.146)
7.Income	0.283*	0.0352
	(0.157)	(0.150)
1.Income verified#1.Income	0	0
	(0)	(0)
1.Income verified#2.Income	2.764***	1.623
	(0.803)	(1.731)
1.Income verified#3.Income	0.409	0.459
	(0.336)	(0.320)
1.Income verified#5.Income	–1.135*	–0.825
	(0.583)	(0.581)
1.Income verified#6.Income	–1.548***	–1.407**
	(0.594)	(0.573)
1.Income verified#7.Income	–1.891***	–1.894***
	(0.578)	(0.568)
Car verified	–0.295**	–0.311***
	(0.116)	(0.107)
1.House verified	0.455***	0.502***
	(0.128)	(0.114)
1.Mortgage Loan	–0.573**	–0.141
	(0.225)	(0.214)
1.Houseverified#1Mortgage loan	–0.0162	–0.512*
	(0.287)	(0.274)
Loan description	–0.00537***	–0.00626***
	(0.000560)	(0.000510)
Age	–0.00171	0.116**
	(0.00623)	(0.0471)
Gender	–0.254**	–0.332***
	(0.129)	(0.128)
Marriage	–0.130	–0.314***
	(0.106)	(0.105)
Educational	–0.121***	–0.117***
	(0.0171)	(0.0174)
Mobile verified	–0.579***	–0.407***
	(0.136)	(0.125)
Weibo verified	–0.403**	–0.524***
	(0.157)	(0.151)
Video verified	1.006***	1.115***
	(0.127)	(0.110)
Interest	0.243***	
	(0.0151)	
Amount	0.00969	–2.72e - 06***
	(0.0497)	(9.76e - 07)
Term	–0.00298	0.332***
	(0.00637)	(0.0274)
Amount square		8.68e - 13
		(5.80e - 13)
Term square		–0.0105***
		(0.000998)
Age square		–0.0015759**
		(0.0006425)
2011.year	0.343	–0.104
	(0.743)	(0.737)
2012.year	0.831	–0.219
	(0.743)	(0.736)
2013.year	1.386*	0.0488
	(0.742)	(0.735)
2014.year	2.522***	1.420*
	(0.754)	(0.744)
Constant	–4.903***	–3.780***
	(0.976)	(1.171)
Pseudo R2	0.189	0.1831
Observations	14,566	14,566

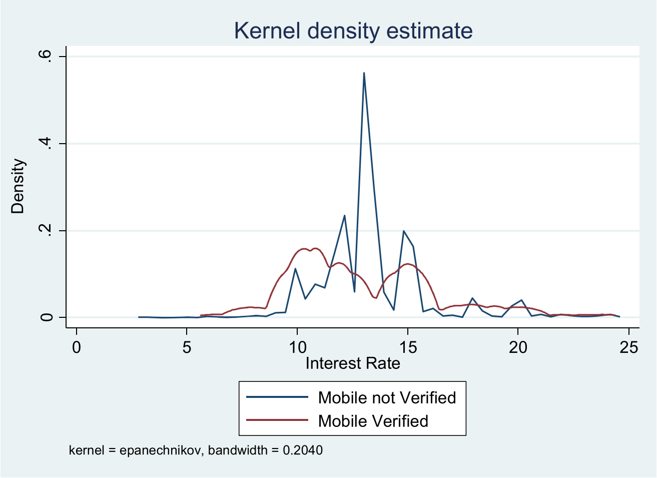
Heteroscedasticity-Robust, standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

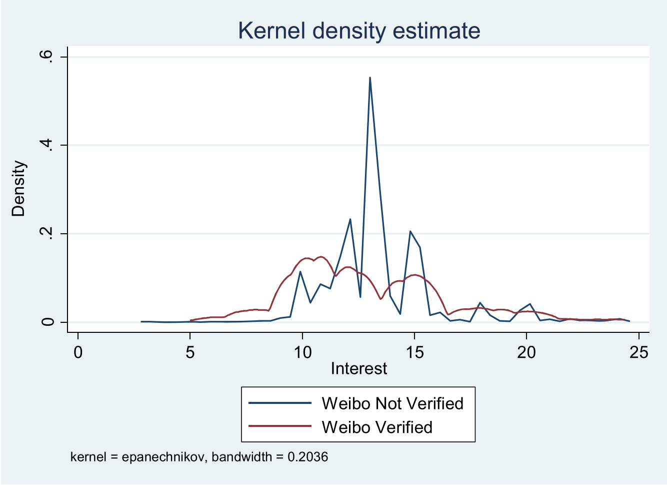
Appendix E. Loan Classes: Amounts, Interest rates, Maturity and Defaults

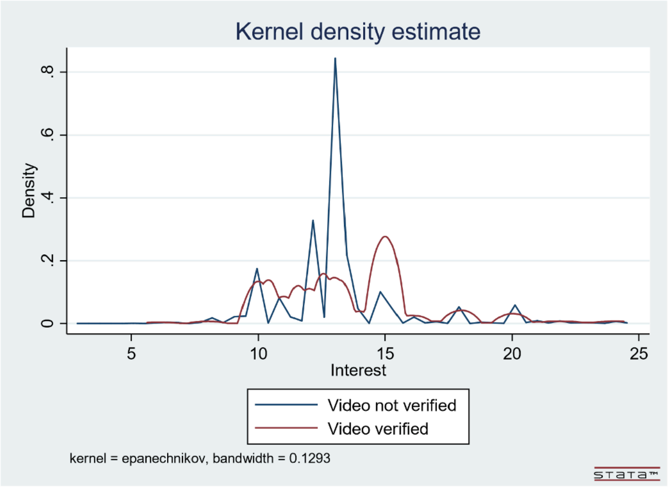
Purpose of loans. The following analysis of the purposes of loans and their properties was carried out as another test of robustness. Given the large size of our data, we have carried out the analysis using a random sample of 687 selected from our large population of 14 575. We read each loan description and divided them manually into three classes of loans – loans for personal consumption (25.47% of the total), loans for business (37.26% of the total), and loans without any clear indication of the purpose (37.26% of the total). The analysis focused on the structure of interest rates, the maturity of loans, the distribution of social capital indicators, and the soft information predictive power across the three groups. The results are reported below.

Structure of interest rates. Using Kernel density estimates, it can be seen that the distribution of interest rates is very similar both for Weibo verified and non-verified loans, and for mobile – verified and for non-verified loans. A vast majority of loans are in the range of 10–15%. In other words, we cannot observe any significant difference between interest rates on loans granted based on soft information and those that were not.

The interest rate structure was similar for all three classes of loans – loans for personal consumption, loans for business purposes, and loans for which it was impossible to identify the actual purpose. The average rates of interest were: 13.46% for personal loans, 14.73% on loans for business purposes, 13.69% on undefined loans, and 13.99% for the total sample (total= 687). The average rate for the entire sample was 13.99% (compared to 13.31% for the whole population of 14,575).







Term structure. The term structure was also almost identical for all three classes. This is not surprising, since the term was uniquely short-term (that is, all loans had a maturity period of less than 3 years).

Amounts of loans. The bulk of loans were for small amounts. The loan amounts were highly skewed to the lowest range starting from 10,000 yuan to 30,000 yuan (38%). Almost 75% of loans were below 50,000 yuan as shown in Table E.14.

Default rates. In the three classes of loans of our random sample, default rates were similar and as follows (in percent of the total sample of 687 loan applications): loans for consumption = 6.29%, loans for business purposes=7.42% and loans without clear indications of purpose = 5.47%. The distribution of defaults suggests that the different classes had small influence on loan defaults, if any.

Soft information predictive power. We tested the predictive power of soft information on the business purpose loans; the results show that the pseudo R square of the soft information model based on business loans is 0.1497, while that for all other groups is 0.2614. The predictive power on busienss loans is lower, which indicates soft information is truly representing the borrowers" willingness to repay since business loans' credit risk also depends on the business operational status.

Social capital variables. All of the chosen variables are mandatory fields in the application form. The only soft factors that the applicant can choose whether to disclose are the following three social capital factors: Weibo verification, video verification, and mobile verification. If they chose not to disclose or did not go through the verification process, then the field is marked as "0". If they disclosed, then the field is marked as "1". After dividing the data into with and without disclosure for these three factors, we found that the default rate of the dataset with this soft information disclosed is

Table E.14
Distribution of loan amounts.

Amount (yuan)	Freq.	Percent
3000–10000	5551	38.09%
10000–20000	1462	10.03%
20000–30000	1411	9.68%
30000–40000	1051	7.21%
40000–50000	1168	8.01%
50000–60000	889	6.10%
60000–70000	401	2.75%
70000–80000	710	4.87%
80000–90000	83	0.57%
90000–100000	699	4.80%
100000–200000	849	5.83%
200000–300000	301	2.07%
Total	14575	100.00%

Table E.15
Distribution of verified variables.

Verified Info	Count
Weiboverified	2475
VideoVerified	5474
Mobileverified	2604
Incomeverified	1347
Weiboverified & VideoVerified & Mobileverified	845
Incomeverified & Weiboverified	230
Incomeverified & Videoverified	647
Incomeverified & Mobileverified	373
Incomeverified & Mobileverified & WeiboVerified	135
Incomeverified & Mobileverified & VideoVerified	274
Incomeverified & Weiboverified & VideoVerified	158
Incomeverified & Mobileverified & WeiboVerified & Videoverified	116

Table E.16
Correlation matrix of variables.

	description	age	gender	married	education	mobileverified	weiboverified	videoverified	incomeverified
description	1.000								
age	0.1865	1.000							
gender	0.0970	0.0656	1.000						
married	−0.2935	−0.0943	−0.0806	1.000					
education	−0.0943	−0.1645	−0.0037	0.1242	1.000				
mobileverified	−0.2543	−0.1408	−0.0857	0.4431	0.1100	1.000			
weiboverified	−0.2267	−0.1919	−0.0711	0.4071	0.1680	0.3972	1.000		
videoverified	−0.4471	−0.1047	−0.1085	0.2960	0.0288	0.3364	0.1709	1.000	
incomeverified	−0.1353	−0.0436	−0.0133	0.0375	0.0450	0.0818	0.0008	0.0692	1.000

1.3%, while without disclosure is 4.83%. This result indicates that people who choose to disclose their softer social factors are actually those who have less default probability. Thus, there is no risk of adverse selection as a result of applicants' willingness to disclose soft information.

The following tables provide cross-tabulations of data and the indications of correlations among different variables. Table E.15 illustrates the extent to which social capital variables were used in processing the loan applications and the extent of the overlap. As shown by the data below, the extent of overlap was very small. For example, all three social capital variables were equal to one in only 845 cases out of our sample of 14 775, i.e. 5.7%. This overlap is small and unlikely to lead to the conclusion that the overlap affects a particular class of loans, and, by extension, that it significantly affects our findings. The overlap is even smaller for only two of our social capital variables. As a further test of the independence of our independent variables, the data in Table E.16 show a relatively small level of correlation between different hard and soft variables as well as between all soft variables.

Furthermore, using another random sample of loan applications selected from a crawling date in our data between 24 July 2014 and 11 August 2014, a sample of 67 applications was identified, in which the applications comprised all three social capital indicators. Among those, 37 applications were without a clearly identified purpose, 14 applications were for personal consumption and 16 were for business purposes. Clearly, all three social capital variables seem to be "normally distributed" across all three loan classes. Moreover, as in our larger sample, the final purpose of the loans could not be identified for the majority of the loans. The share of loans for personal consumption was relatively small. Thus, the differences in default rates were unlikely to be due to different purposes of the loans.

Appendix F. Comparison of Explanation Power of Success and Default Models

Table F.17
Model explanation power of success and default models.

Model	Pseudo R square	AUC
Default (Hard)	0.1226	0.7946
Default (Soft)	0.1694	0.8268
Default (Combined)	0.1890	0.8419
Success (Hard)	0.4459	0.9126
Success (Soft)	0.5519	0.9395
Success (Combined)	0.5953	0.9508

References

Agarwal, S., Ambrose, B. W., Chomsisengphet, S., & Liu, C. (2011). The role of soft information in a dynamic contract setting: Evidence from the home equity credit market. *Journal of Money Credit & Banking, 43*(4), 633–655. doi: j.1538–4616.2011.00390.x.

Agarwal, S., & Hauswald, R. (2010). Distance and private information in lending. *Review of Financial Studies, 23*(7), 2757–2788. https://doi.org/10.1093/rfs/hhq001

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488–500. https://doi.org/10.2307/1879431

Akerlof, G. A., & Kranton, R. E. (2000). Identity economics. *The Quarterly Journal of Economics, 115*(3), 715–753. https://doi.org/10.1162/003355300554881

Ashta, A., & Assadi, D. (2009). An analysis of European online micro-lending websites. *Working Papers Ceb*.doi:citeulike-article-id:12156499

Barasinska, N., & Schäfer, D. (2010). Does gender affect funding success at the peer-to-peer credit markets? evidence from the largest german lending platform. *Social Science Electronic Publishing*. https://doi.org/10.1111/geer.12052

Berger, A. N., Frame, W. S., & Miller, N. H. (2005a). Credit scoring and the availability, price, and risk of small business credit. *Journal of Money, Credit and Banking*, 191–222.doi:stable/3838924

Berger, A. N., Miller, N. H., Petersen, M. A., Rajan, R. G., & Stein, J. C. (2005b). Does function follow organizational form? evidence from the lending practices of large and small banks. *Journal of Financial Economics, 76*(2), 237–269. https://doi.org/10.1016/j.jfineco.2004.06.003

Berger, A. N., & Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *The economic journal, 112* (477), F32–F53. https://doi.org/10.1111/1468-0297.00682

Bertrand, M., Karlin, D., Mullainathan, S., Shafir, E., & Zinman, J. (2005). What's psychology worth? A field experiment in the consumer credit market. *Technical Report*. National Bureau of Economic Research. https://doi.org/10.3386/w11892

Botsman, R. (2017). Big data meets big brother as China moves to rate its citizens. *Wired UK*.

Bourdieu, P. (1986). The forms of capital. cultural theory: An anthology. In J. Richardson (Ed.), *Handbook of theory and research for the sociology of education* (pp. 241–258). New York: Greenwood.

Brown, S. L. (2000). The effect of union type on psychological well-being: depression among cohabiters versus marrieds. *Journal of health and social behavior*, 241–255. https://doi.org/10.2307/2676319

Cao, X. (2013). Measurement and the role of social capital in online p2p lending markets.

Chaulk, B., Johnson, P. J., & Bulcroft, R. (2003). Effects of marriage and children on financial risk tolerance: A synthesis of family development and prospect theory. *Journal of Family and Economic Issues, 24*(3), 257–279. https://doi.org/10.1023/A:1025495221519

Chorzempa, M. (2018). China needs better credit data to help consumers (no. PB18-1). https://www.piie.com/system/files/documents/pb18-1.pdf.

Cornée, S. (2017). The relevance of soft information for predicting small business credit default: Evidence from a social bank. *Journal of Small Business Management*. https://doi.org/10.1111/jsbm.12318

Dell'Ariccia, G., & Marquez, R. (2004). Information and bank credit allocation. *Journal of Financial Economics, 72*(1), 185–214. https://doi.org/10.1016/S0304-405X(03)00210-1

Deyoung, R., Glennon, D., & Nigro, P. (2008). Borrower-lender distance, credit scoring, and loan performance: Evidence from informational-opaque small business borrowers. *Journal of Financial Intermediation, 17*(1), 113–143. https://doi.org/10.1016/j.jfi.2007.07.002

Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *Review of Economic Studies, 51*(3), 393–414. https://doi.org/10.2307/2297430

Ding, N., Fung, H.-G., & Jia, J. (2020). Shadow banking, bank ownership, and bank efficiency in china. *Emerging Markets Finance and Trade, 56*(15), 3785–3804.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., Castro, I. D., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending - evidence from two leading european platforms. *Journal of Banking & Finance, 64*, 169–187. https://doi.org/10.1016/j.jbankfin.2015.11.009

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies, 25*(8), 2455–2483. https://doi.org/10.1093/rfs/hhs071

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Franke, G. R., Crown, D. F., & Spake, D. F. (1997). Gender differences in ethical perceptions of business practices: A social role theory perspective. *Journal of applied psychology, 82*(6), 920. https://doi.org/10.1037/0021-9010.82.6.920

Freedman, S. M., & Jin, G. Z. (2011). Learning by Doing with Asymmetric Information: evidence from Prosper. com. *Technical Report*. National Bureau of Economic Research. https://doi.org/10.3386/w16855

García-Appendini, E. (2007). Soft information in small business lending.

Ge, R., Feng, J., Gu, B., & Zhang, P. (2017). Predicting and deterring default with social media information in peer-to-peer lending. *Journal of Management Information Systems, 34*(2), 401–424. https://doi.org/10.2139/ssrn.968178

Godbillon-Camus, B., & Godlewski, C. J. (2005). Credit risk management in banks: Hard information, soft information and manipulation. *Mpra Paper, 55*(1–6), 114–125. https://doi.org/10.1080/07421222.2017.1334472

Gonzalez, L., & Loureiro, Y. K. (2014). When can a photo increase credit? the impact of lender and borrower profiles on online peer-to-peer loans. *Social Science Electronic Publishing, 2*, 44–58. https://doi.org/10.2139/ssrn.882027

Greiner, M. E., & Wang, H. (2009). The role of social capital in people-to-people lending marketplaces. *ICIS 2009 proceedings*, 29. https://doi.org/10.1093/oxfordhb/9780199282944.003.0022

Grunert, J., Norden, L., & Weber, M. (2005). The role of non-financial factors in internal credit ratings. *Journal of Banking & Finance, 29*(2), 509–531. https://doi.org/10.1016/j.jbankfin.2004.05.017

Herrero-Lopez, S. (2009). Social interactions in p2p lending. *The Workshop on Ssocial Network Mining & analysis* (pp. 1–8). https://doi.org/10.1145/1731011.1731014

Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios. *Journal of Accounting Research, 4*(3), 44–62. https://doi.org/10.2307/2490168

Horwitz, A. V., & White, H. R. (1998). The relationship of cohabitation and mental health: A study of a young adult cohort. *Journal of Marriage and the Family*, 505–514. https://doi.org/10.2307/353865

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?. 10.2139/ssrn.1570115.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research, 266*(1–2), 511–529. https://doi.org/10.1007/s10479-017-2668-z

Jiménez, G., & Saurina, J. (2004). Collateral, type of lender and relationship banking as determinants of credit risk. *Journal of Banking & Finance, 28*(9), 2191–2212. https://doi.org/10.1016/j.jbankfin.2003.09.002

Klafft, M. (2009). Peer to peer lending: Auctioning microcredits over the internet. *Social Science Electronic Publishing*.

Lea, S. E. G., Webley, P., & Walker, C. M. (1995). Psychological factors in consumer debt: Money management, economic socialization, and credit use. *Journal of Economic Psychology, 16*(4), 681–701. https://doi.org/10.1016/0167-4870(95)00013-4

Lennon, R., & Eisenberg, N. (1987). Gender and age differences in empathy and sympathy. *Empathy and its development*, 195–217.

Liao, L., Lin, J. I., & Zhang, W. (2015). Education and credit:evidence from p2p lending platform. *Journal of Financial Research*.

Liberti, J. M., & Petersen, M. A. (2018). Information: Hard and Soft. *Working Paper*. Northwestern University: Kellogg School of Management.

Lieberman, I., Paul, D., Watkins, T. A., & Anna, K. (2018). Microfinance: Revolution or footnote: The future of microfinance over the next 10 years. *Technical Report*. Lehigh University.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending. *Management science, 59*(1), 17–35. https://doi.org/10.1287/mnsc.1120.1560

Liu, D., Brass, D. J., Lu, Y., & Chen, D. (2015). Friendships in online peer-to-peer lending: pipes, prisms, and relational herding. *Mis Quarterly, 39*(3), 729–742. https://doi.org/10.2139/ssrn.2251155

Miu, L. Y., & Chen, J. L. (2014). The influence of social capitals on borrower's default risk in p2p network lending—a case study of the prosper. *Finance Forum*. https://doi.org/10.16529/j.cnki.11-4613/f.2014.03.002

Order, R. V., & Zorn, P. (2000). Income, location and default: Some implications for community lending. *Real Estate Economics, 28*(3), 385–404. https://doi.org/10.1111/1540-6229.00806

Piliavin, J. A., & Charng, H.-W. (1990). Altruism: A review of recent theory and research. *Annual Review of Sociology, 16*(1), 27–65. https://doi.org/10.1146/annurev.so.16.080190.000331

Pope, D. G., & Sydnor, J. R. (2011). What's in a picture? evidence of discrimination from prosper.com. *Journal of Human resources, 46*(1), 53–92. https://doi.org/10.1353/jhr.2011.0025

Pötzsch, S., & Böhme, R. (2010). The role of soft information in trust building: Evidence from online social lending. *International conference on trust and trustworthy computing* (pp. 381–395). Springer. https://doi.org/10.1007/978-3-642-13869-0_28

Ravina, E. (2012). Love & loans: The effect of beauty and personal characteristics in credit markets. 10.2139/ssrn.1101647.

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science, 17*(1), 31–35. https://doi.org/10.1111/j.1467-8721.2008.00543.x

Santoso, W., Trinugroho, I., & Risfandy, T. (2020). What determine loan rate and default status in financial technology online direct lending? evidence from indonesia. *Emerging Markets Finance and Trade, 56*(2), 351–369. https://doi.org/10.1080/1540496X.2019.1605595

Serrano-Cinca, C., Gutierrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in p2p lending. *PloS one, 10*(10), e0139427. https://doi.org/10.1371/journal.pone.0139427

Slavin, R. E., & Davis, N. (2006). *Educational psychology: Theory and practice* (8th). Pearson/Allyn & Bacon.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics, 87*(3), 355–374. https://doi.org/10.2307/1882010

Stein, J. C. (2002). Information production and capital allocation: decentralized versus hierarchical firms. *The Journal of Finance, 57*(5), 1891–1921. https://doi.org/10.1111/0022-1082.00483

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review, 71*(3), 393–410. https://doi.org/10.1126/science.151.3712.867-a

Thakor, R. T., & Merton, R. C. (2018). Trust in Lending. *Technical Report*. National Bureau of Economic Research.

Uchida, H. (2011). What do banks evaluate when they screen borrowers? soft information, hard information and collateral. *Journal of Financial Services Research, 40*(1–2), 29–48. https://doi.org/10.1007/s10693-010-0100-9

Wang, H., Yu, M., & Zhang, L. (2019). Seeing is important: The usefulness of video information in p2p. *Accounting & Finance, 59*, 2073–2103. https://doi.org/10.1111/acfi.12530

Xu, Z., & Zou, C. (2010). Banks' Loan approval right allocation and incentive mechanism design under the framework of hard and soft information:implications for SME finance. *Journal of Financial Research, 8*, 1–15.

Zha, J. (2011). *Tide players: The movers and shakers of a rising china*. New Press.