CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

INSTITUTE OF ECONOMIC STUDIES

Dissertation Thesis

2021/2022 Yao Wang

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

INSTITUTE OF ECONOMIC STUDIES

Dissertation Thesis

Three Essays on Asymmetric Information in SME Finance and Microfinance

Prague 2022

Author: **Yao Wang**

Supervisor: **Prof. Zdenek Drabek**

Academic Year: **2021/2022**

References

WANG, Yao. *Three Essays on Asymmetric Information in SME finance and Microfinance.* Praha, 2022. 153 pages. Dissertation thesis (PhD.). Charles University, Faculty of Social Sciences, Institute of Economic Studies. Supervisor Dr. Zdenek Drabek.

Abstract

This dissertation thesis consists of three essays on asymmetric information problem in small and medium sized enterprises (SMEs) finance and Microfinance. The aim of the thesis is to address the key problem in the credit rationing in the SME finance and microfinance and strive to improve the credit analyzing model with the help of soft information. The first essay investigates the factors that hinder the growth of SMEs using a World Bank dataset, and access to finance is found to be their biggest constrain to growth. Asymmetric information between small business owners and banks generates high interest rates, complex application procedures and high collateral requirements, which are found to be the biggest obstacles business owners face when they seek external financing. Small business owners who cannot get loans from banks will turn to microfinance as an alternative source of funds. In the second essay, a new dataset from disintermediated Peer to Peer (P2P) lending market is used to investigate credit rationing efficiency when there is no financial intermediary. The results show the existence of adverse selection where investors are predisposed to making inaccurate diagnoses of signals and gravitate to borrowers with low creditworthiness, while inadvertently screening out those with high creditworthiness. This implies that although disintermediation can decrease transaction costs, it increases credit risk because the peer lenders lack professional credit rationing experience. We also find that this misdiagnosis is particularly evident with finance oriented (hard) signals, while lenders can distinguish better the social and psychological related (soft) signals. Given that developing countries commonly lack a solid financial credit bureau and that financial information is hard to verify, in the third essay we examined whether the soft social and psychological information can be used to improve the credit analyzing model. The results show that soft social and psychological related information can improve the predictive power of the credit model and serve as a substitution when hard financial information is difficult to verify and under weak credit bureau conditions.

Keywords

Asymmetric Information; SMEs; Microfinance; Behavioral Economics; Soft Information; Fintech

Declaration

1 I hereby declare that I have compiled this thesis using the listed literature and resources only.

2 I hereby declare that my thesis has not been used to gain any other academic title.

3 I fully agree to my work being used for study and scientific purposes.

In Prague on 27.12.2021 Yao Wang

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Prof. Zdenek Drabek, for his patient guidance and advice. I am also grateful to all my colleagues from Charles University and Tsinghua University for their helpful suggestions on my papers. I am thankful to the full support and understanding from our Center of Doctoral Studies' colleagues and the Doctoral Committee throughout my PhD journey in Charles University. Finally, I would like to thank my husband and my family for their endless support and encouragement.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Staff Exchange program under the Marie Sklodowska-Curie grant agreement No.681228 and No.870245, and GACR grant No.18-04630S.

Table of Contents

List of Tables

Chapter 1 What are the biggest obstacles to growth of SMEs in developing countries? - An empirical evidence from an enterprise survey

Chapter 2 Adverse Selection in P2P Lending: Does Peer Screening Work Efficiently? —Empirical Evidence from a P2P Platform

Chapter 3 The role of social and psychological related soft information in credit analysis: Evidence from a Peer to Peer lending platform

List of Figures

Introduction

"Contract is a specification of actions that named parties are supposed to take at various times, generally as a function of the conditions that hold." (Shavell, 2004). Contract theory studies the form of the contract under a specific transaction environment with hidden information, hidden actions and contractual incompleteness. According to Borusseau & Glachant (2002), modern contract theory can be divided into three streams: incentive theory (Akerlof, 1970; Stiglitz, 1981; Spence, 1973; Hölmstrom, 1979), incomplete-contract theory (Hart and Grossman, 1986), and new institutional transaction costs theory (Williamson, 1979). Incentive theory investigates incentives from both sides of the contract before and after the contract is signed. My research examines the asymmetric information problem under incentive theory. The three essays in this dissertation contribute to the empirical literature on asymmetric information in SME finance and microfinance.

Synopsis We started by analyzing the factors that hinder the growth of the engine of the economy - small and medium-sized enterprises (SMEs) and found that access to finance was one of their biggest obstacles. Furthermore, asymmetric information between business owners and banks generated high interest rates, complex application procedures and high collateral requirements, which were reported to be the biggest hurdles business owners faced when they sought external financing. Small business owners who could not get loans from banks would turn to microfinance as an alternative source of funds. Then in the second essay, using a new dataset from disintermediated Peer to Peer (P2P) lending market, we investigated credit rationing efficiency when there was no financial intermediary. The results showed the existence of adverse selection where investors were predisposed to making inaccurate diagnoses of signals and gravitate to borrowers with low creditworthiness while inadvertently screening out those with high creditworthiness. This implied that although disintermediation could decrease transaction costs, it increased credit risk because the peer lenders lack professional credit rationing experience. We also found that this misdiagnose particularly happens on hard financial based signals while lenders were able to distinguish better the soft social and psychological related signals. Given that developing countries commonly lack a solid financial credit bureau and that financial information is hard to verify, we examined whether the soft social and psychological information can be used to improve the credit analyzing model in the third essay. The results show that soft social and psychological-related information can improve the predictive power of the credit model and serve as a substitution when hard financial information is difficult to verify and under weak credit bureau conditions.

Theoretical Context In the late 1960s and 1970s, George Akerlof, Joseph Stiglitz and Michael Spence developed asymmetric information theory and brought in the concept of hidden information and hidden actions. Hidden information occurs when one party to the contract has private information over the other. When the contract is drafted by the party which lacks private information, the uninformed party needs to screen the information the informed party possessed. This is the so-called screening problem. If the contract is offered by the informed party, this constitutes a signaling problem since the informed party can signal the private information through the type of contract it offers. Spence (1973, 1974) investigated education as a signal of intrinsic skill and found that employees with better skills had lower disutility of education and thus were more willing to educate themselves. such that the employer would like to pay educated employees more, even if education itself may not have any no additional value to him.

Akerlof (1970) used the automobile market as an example to explain the situation when one party had private information and regarded the second hand automobile market as the market for “lemons” since the seller had private information about the condition of the car and thus had the incentive to sell below-average quality cars. This lowered the quality of the whole market, but due to the asymmetric information, the buyer can only bargain according to the average price and preferred to buy the lower-quality cars, which caused the above-average quality cars to exit the market. This adverse selection problem in the loan market refers to a situation in which high-risk borrowers are those who are most eagerly looking for money and most likely to obtain the loan, with low-risk borrowers crowded out as a result. In financial markets, efficient credit rationing as a screening method can alleviate adverse selection in the lending process. The efficiency of credit rationing matters, especially in an environment without financial intermediaries. Hence in essay 2, we examined whether online P2P investors can accurately and effectively diagnose signals of creditworthiness during their decision-making. Our findings indicated that adverse selection exists in the investors' decision-making process, meaning that investors were predisposed to making inaccurate diagnoses of signals and to gravitate to borrowers with low creditworthiness, while inadvertently screening out those with high creditworthiness. This was especially true with hard financial-based signals. Specifically, signals such as income and property ownership were insignificant or provided contradictory guidance in terms of default. However, investors allocated disproportionate weights to this in the decision-making process of loan funding. Rather than hard financial signals, investors were more adept at diagnosing soft social signals. This aroused our interest in analyzing the role of soft information in loan default prediction, which is the aim of essay 3.

Hölmstrom (1979) studied moral hazard under a principal-agent relationship and derived assumptions for imperfect information to improve contracts' payoff. He states that any imperfect information about actions and state of nature can be used to improve contracts, additional information is of value because it allows a more accurate judgment of the performance of the agent. This provided the theoretical foundation for my assumption that a certain amount of soft information can compensate for the missing hard information in predicting ex-post moral hazard default behavior. Essay 3 compared the performance of default predicting with soft information, hard information, and combined soft and hard information. The results showed that the combination achieved the best predictive power. Moreover, the soft information model performed nearly the same as the hard information model, indicating that soft information can substitute for hard information when hard information is not available or is difficult to verify.

Contribution The vast majority of research (e.g., Levy, 1993; Pissarides et al., 2003; Gree and Thurnik, 2003; Lee, 2014) that analyzed the barriers to growth of SMEs was limited to one or two specific countries where the survey was carried out or to some small regions (e.g., Pissarides, 1999; Yin, 2012) due to data limitations. We used a unique dataset of 119 developing countries and strived to find common obstacles that SMEs in developing countries confronted. We attained to test whether some key findings in the literature can be generalized for developing countries as a group. Moreover, our research further examined the determinants of the barriers, and also explored the reasons for the financing barrier specifically. Our research provided a reference for research that analyze the financial difficulties in developing countries in general.

A large literature (e.g., Serrano-Cinca et al., 2015; Deyoung et al., 2008; Jimenez & Saurina, 2004; Berger, Frame, & Miller, 2005a; Berger, Miller, Petersen, Rajan, & Stein, 2005b; Pötzsch and Böhme, 2010; Dorfleitner et al. 2016; Wang et al. 2019; Zhang et al. 2017) has analyzed trust building between lenders and borrowers in SME finance or in microfinance. However, the bulk of research has focused on the role of hard financial information. Only in recent years, after the emergence of Fintech, has soft information started to become a topic. In the meantime, various researchers in the field have identified the inefficiency of the credit analyzing method, which depends solely on hard information. For example, Jiménez and Saurina (2004) found that collateral could not secure the repayment of loans and that loans with collateral sometimes have higher default rates. Berger et al. (2005) showed that a credit scoring system helped boost credit availability for small businesses, but its function for credit risk analysis was not as effective as expected. A similar result has been obtained by DeYoung et al. (2008), who found a positive relationship between the use of a credit scoring model and loan default rates. Therefore, based on the identity economics of Akerlof and Kranton (2000), we explored the role of soft information in credit analysis using a P2P lending dataset. This study is among the first to argue for the importance of soft information in credit analysis in microfinance.

Moreover, within the fraction of research that analyzed trust-building and credit analysis in the P2P market, only Iyer et al. (2016) targeted the efficiency of credit analysis. Their results indicated that misspecification of creditworthiness signals existed in two-thirds of the lenders. Their work also opened the first debate on whether the usage of soft information would compensate for the traditional credit analysis model and add more choice for credit model development after the 2008 financial crisis. However, Iyer et al. did not delve into the specific determinants which resulted in the misspecification. We extended their work and provided empirical evidence for the misspecification of the lenders’ screening mechanism in P2P lending and in a developing country. Research on this topic is slowly growing these days (e.g., Kim, 2021). Our research also put forward different opinions to the supportive voices of disintermediation and emphasized its potential risks due to the lack of public financial literacy and professional credit appraisers.

Policy Implications Reducing the external obstacles that impede the growth of SMEs requires government effort in building up a comprehensive financial infrastructure, with features such as a solid credit system for generating small business credit profiles, a user-friendly accounting and taxation system, and supportive lending and taxation policy for SMEs. In addtion, the government should commit to providing a small business-friendly commercial environment which includes: first, enhancing the basic infrastructures such as electricity, transportation and telecommunications; second, providing macroeconomic and political stabilities; third, perfecting the business law system; and fourth, establishing antitrust laws and encouraging healthy competition. Financial institutions also play an important role in helping SMEs' growth. Financial innovations on alternative lending methods, such as disintermediated lending platforms and, electronic applications, can provide convenience to small business borrowers. Traditional banks should also actively apply new technology to smooth the application process, reduce transaction costs, and build up big data based credit models. Alongside external changes, small business entrepreneurs should improve their financial literacy and understanding of the lending process and the credit rating system. Small business owners should note the importance of accounting and tax recording in getting funds from banks. A large group of literature (eg., Fagariba, 2016; Alkhatib et al. 2018; Vincent, 2021) confirms the common existence of tax evasion in SMEs in developing countries. Thus, understanding the logic of healthy business development is critical to SMEs in developing countries.

The growing size of the Fintech industry suggests that the misdiagnosis of borrowers’ credit signals in disintermediated financial institutions may pose systematic risk to financial systems, requiring regulators’ close attention. The misidentification of creditworthiness signals can be alleviated by a sophisticated and independent credit bureau and by increasing public financial literacy. Expanding the use of soft social information could also mitigate adverse selection in disintermediated financial institutions; this process must be accompanied by the establishment of transparent and effective oversight on the use of soft information in order to avoid abuse. To consolidate the system and prevent shadow banking from infiltrating the industry, the lending license regulations need to be tightened. In addition, in order to avoid the issue of lenders blindly pursuing profit without considering risks, the interest rate cap should be closely monitored in this field. P2P lending platforms could provide guidelines for credit analysis. Moreover, the verification process could be strengthened in the platform. This can be achieved by cooperation in data sharing between the private sector and the credit bureau.

The function of soft information in credit analysis is considerably greater in situations when hard information is missing or has poor quality. And the importance of soft information will increase with the development of technology and information "hardening" tools. Regulatory agencies would have to pay more attention to lending based on the use of soft information, its quality, its dissemination, and data privacy, which will 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.

References

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. The quarterly journal of economics, 488-500.

Akerlof, G. A., & Kranton, R. E. (2000). Identity economics. The Quarterly Journal of Economics, 115(3), 715–753.

Alkhatib, A. A., Abdul Jabbar, H., & Marimuthu, M. (2018). The effects of deterrence factors on income tax evasion among Palestinian SMEs. International Journal of Academic Research in Accounting, Finance and Management Sciences, 8(4), 144-152.

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.

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.

Brousseau, E., & Glachant, J. M. (2002). The economics of contracts: theories and applications. Cambridge University Press.

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.

Dorfleitner, Gregor, Christopher Priberny, Stephanie Schuster, Johannes Stoiber, Martina Weber, Ivan de Castro, and Julia Kammler. 2016. Description-text related soft information in peer-to-peer lending–Evidence from two leading European platforms. Journal of Banking and Finance 6: 169–87.

Fagariba, C. J. (2016). Perceptions of Causes of SMEs and Traders Tax Evasion: A Case of Accra Metropolis, Ghana. Journal of Business & Economic Management 4 (2), 017-039.

Gree, A. & Thurnik, C. (2003). Firm selection and industry evolution: the post country performance of new firm. Journal of Evolutionary Economics, 4 (4), 243-264.

Jim´enez, G., & Saurina, J. (2004). Collateral, type of lender and relationship banking as determinants of credit risk. Journal of Banking & Finance, 28(9), 2191–2212.

Hart, O. D., & Holmstrm, B. (1986). The theory of contracts.

Hölmstrom, B. (1979). Moral hazard and observability. The Bell journal of economics, 74-91.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. Management Science 62: 1554–77.

Kim, D. (2021). Can investors' collective decision-making evolve? Evidence from peer-to-peer lending markets. Electronic Commerce Research.

Lee, N. (2014). What holds back high-growth firms? Evidence from UK SMEs, Small Business Economics, 43(1), 183-195.

Levy, B. (1993). Obstacles to developing indigenous small and medium enterprises: an empirical assessment. The World Bank Economic Review, 7(1), 65-83.

Pissarides, F., Singer, M., & Svejnar, J. (2003). Objectives and constraints of entrepreneurs: Evidence from small and medium size enterprises in Russia and Bulgaria. Journal of Comparative Economics, 31(3), 503-531.

Pissarides, F. (1999). Is lack of funds the main obstacle to growth? EBRD's experience with small-and medium-sized businesses in Central and Eastern Europe. Journal of Business Venturing, 14(5), 519-539.

Potzsch, S., & Bohme, 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.

Serrano-Cinca, C., Gutierrez-Nieto, B., & Lo´pez-Palacios, L. (2015). Determinants of default in p2p lending. PloS one, 10(10), e0139427.

Shavell, S. (2009). Foundations of economic analysis of law. Harvard University Press.

Spence, M. (1973). Job market signaling. The quarterly journal of Economics, 87(3), 355-374.

Spence, M. (1974). Competitive and optimal responses to signals: An analysis of efficiency and distribution. Journal of Economic theory, 7(3), 296-332.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. The American economic review, 71(3), 393-410.

Vincent, O. (2021). Assessing SMEs tax non-compliance behavior in Sub-Saharan Africa (SSA): An insight from Nigeria. Cogent Business & Management , 8(1), 1938930.

Williamson, O. E. (1979). Transaction-cost economics: the governance of contractual relations. The journal of Law and Economics, 22(2), 233-261.

Wang, H., Yu, M., & Zhang, L. (2019). Seeing is important: The usefulness of video information in p2p. Accounting & Finance, 59, 2073–2103.

Yin Qiyue (2012), A study on the dilemma of China’s small business finance. (Doctoral Dissertation, Southwestern University of Finance and Economics).

Zhang, Yuejin, Haifeng Li, Mo Hai, Jiaxuan Li, and Aihua Li. 2017. Determinants of loan funded successful in online P2P Lending. Procedia Computer Science 122: 896–901.

1 What are the biggest obstacles to growth of SMEs in developing countries? - An empirical evidence from an enterprise survey

SMEs are drivers of economic growth and job creation in developing countries. It is paramount to determine the factors that hinder their growth. This paper uses the Enterprise Survey from the World Bank which covers data from 119 developing countries to investigate the biggest obstacles SMEs are confronting and the determinants that influence the obstacles as perceived by enterprise managers. The results show that SMEs perceive access to finance as the most significant obstacle which hinders their growth. The key determinants among firms' characteristics are size, age and growth rate of firms as well as the ownership of the firm. The latter - the role of the state in financing SME - is particularly intriguing. External reasons for the financing dilemma are also examined. It is shown that the main barriers to external financing are high costs of borrowing and a lack of consultant support.

1.1 Introduction

Small and medium sized enterprises (SMEs) potentially constitute the most dynamic firms in emerging economies (Pissarides, 1999). The empirical evidence from around the globe shows that the ubiquity of SMEs has grabbed the world’s attention. The original idea formed at the end of the 19th century that large firms are the greatest support for the economy has been challenged since the 1950s. Nowadays, the significant role SMEs play in the economy cannot be underestimated. For example, Ayyagari et al. (2011) investigated the role SMEs play in creating jobs and showed that SMEs with less than 250 employees were the engine of growth in many countries. Beck et al. (2005) added that SMEs constituted over 60% of total employment in manufacturing in most developing countries. According to the data from the Chinese National Bureau of Statistics, SMEs represented 99.4% of all enterprises in China in 2012, and they contributed to 59% of China’s GDP and accounted for 60% of total sales. All these figures reflect the importance of SMEs both in developed and developing economies.

However, their importance notwithstanding, SMEs are confronted with significant obstacles which impede their development. This paper aims at sorting out the biggest obstacles SMEs face in developing countries and determining the factors affecting the obstacles for firms to grow. Only in this way can we offer effective recommendations to policymakers in those countries in their quest for faster and healthier growth of their economies.

A considerable number of scholars have investigated the obstacles that affect the development of SMEs within specific areas. However, very little research has been directed toward developing economies as a group. By researching developing countries as a group, we believe that some common problems that they all face can be revealed. In this paper, the Enterprise Survey of the World Bank, which contains 119 developing countries, will be used to test the biggest obstacles to the growth of SMEs in developing countries. Firstly, the five most significant obstacles will be taken from the 18 obstacles which are described in the survey. Then a hypothesis will be made based on work done by other researchers. Econometric models will then be set up to examine the relationship between the obstacles and the chosen factors. Moreover, a specific variable "sme" will be generated to emphasize the significance of the problems SMEs face compared to larger firms. A further investigation will be carried out to identify the determinants of the main obstacles to growth and also the intensity of the barriers.

In the following section, a review of recent literature is presented in order to provide a brief summary of the relevant work that has been carried out so far. Section 3 provides a brief description of our approach and hypotheses and a description of the dataset used in our analysis. As we shall see, the data is unique in its coverage and richness. In addition, the section includes a description of the methodology used in choosing the variables for our model. Section 4 introduces the model. Section 5 presents the results of our tests with a relevant discussion. The final section concludes our presentation with a brief summary of the results and a brief discussion of the approach adopted in this study.

1.2 Literature Review and Hypotheses

The literature dealing with barriers to the growth of SMEs is relatively rich. Levy’s (1993) research on the leather industry in Sri Lanka and the construction and furniture industry in Tanzania is one of the interesting examples of papers from the 1990s. Levy has identified three major constraints - access to finance, access to non-financial inputs, and high cost. His results showed that financial constraints were the main obstacles for firms to grow. Moreover, a high tax constraint was also identified as an important obstacle for the smallest firms. Since then, the research has focused on specific sectors to give more detailed and specific information about the difficulties SMEs face in the chosen industries. However, due to a lack of updated data and the expense of conducting the required surveys, the results cannot be used more broadly. Pissarides (1999) investigated whether a lack of funds is the main obstacle to SMEs' growth using survey data from the EBRD (European Bank for Reconstruction and Development). He pointed out that lack of financing became an obstacle to SMEs' growth in transitional economies due to poorly developed capital markets and where credit was granted according to historical working practice. In other words, state banks were more likely to lend to state or larger enterprises. Later on, Pissarides *et al.* (2003) used survey data from 437 CEOs of SMEs in Russia and Bulgaria to detect the biggest obstacles to SMEs' growth. Variables were chosen by ranking the highest-rated constraints. The top four constraints were defined as: “suppliers are not ready to deliver”, “access to land”, “finance problems” and “other production constraints”. Their results showed that the constraint on external finance was most serious, while other factors such as licensing did not appear to be as significant a problem as expected. More generally, Gree and Thurnik (2003) divided the obstacles into two groups: external and internal. Of the 30 obstacles chosen, finance turned out to be the most important. Other significant factors are “management skills” “location” “technology” “corruption” and “regulations”; which are similar to what was listed in the World Bank Enterprise Survey of emerging economies. Literature from the developing world suggests that access to finance is a common problem for SMEs, thus we have the first hypothesis:

H1: SMEs are more likely to perceive access to finance as the most significant obstacle to their growth compared with big firms;

An important element of the debate is the relationship between the characteristics of firms and barriers to their growth. A particularly interesting part of the debate concerns the role of different types of ownership of firms as factors of growth. For example, Richter and Schaffer (1996) found that private firms developed faster than state-owned firms, the latter typically focussing their objectives on employment expansion and less on the efficient utilization of resources. However, comparisons of small public and private firms remain rather rare and the debate is typically linked to the performance of large public enterprises. Numerous scholars from China, such as Yin (2012) and Ji (2011), have drawn the conclusion that state-owned firms are "too big to fail" and thus face much fewer obstacles, not only in finance, but also in sales and have greater growth compared with smaller businesses. In brief, types of ownership need to be taken into account in the analysis of the business environment in which SME operate. Based on these researches, we assume the ownership is an important factor that influences the access to finance of SMEs in developing countries, thus the second hypothesis is:

H2: Privately-owned enterprises are more likely than state-owned enterprises to perceive access to finance as a significant obstacle to their growth.

Furthermore, Beck (2007) summarized the empirical evidence on SMEs' financing constraints and showed that SMEs are more likely than large enterprises to be constrained by finance and other institutional obstacles. This brings us to the third hypothesis about the influence of the size on the financing constraint in developing countries:

H3: The probability of perceiving access to finance as a significant barrier to SMEs' growth has a negative correlation with the size of the enterprise.

Moreover, the bigger the firm, the less severe the perception that finance is the binding constraint.

Using the World Bank Enterprise Survey 2006-2009, Chavis *et al*. (2010) found that 31 percent of examined firms regarded access to finance as the major constraint. Moreover, 40 percent were young firms with less than 3 years' experience in the industry. Further analysis addressed the relationship between the firm's age and its access to finance. The empirical results showed that younger firms were more reliant on informal financing rather than bank financing. Bank finance gradually increased with age, while informal finance gradually decreased with age. Young firms were found to be twice as likely as older firms to use personal assets as collateral, which is consistent with the results from a study of US small firms (Avery et al.,1998). However, young firms in countries with stronger legislation and better credit information have less reliance on informal financial resources. These researches suggest a significant relationship between the age of the firm and its access to finance. This gives us the fourth hypothesis:

H4: The probability of perceiving access to finance as a significant obstacle to SMEs has a negative correlation with the age of the enterprise. Moreover, the younger the firm, the more severe is likely to be the perception that the financial barrier will be an issue.

A wealth of relevant literature attaches importance to high-growth firms[1]. Results from some studies suggest the importance of finance to high-growth firms but the evidence is not clear-cut. For example, Brush et al. (2009) stratified the growth paths into rapid, incremental and episodic and then investigated the impact of access to finance, market conditions and management on the growth of firms. The results show that Rapid growth firms were cash hungry machines while incremental growth firms have to find the right employees. And advanced management skills play an important role during episodic growth of firms, while

[1] As Henrekson and Johansson (2010) shows, high growth firms occupy only a small proportion of the total number of firms but create the majority of the jobs.

marketing strategy is a way to turn a business around when firms reach a plateau. Zarook et al. (2013) especially emphasized the positive impact of management experience on access to finance for SMEs. Mason and Brown (2013) investigated the policy effect on high growth firms and how to promote high growth firms through policy approaches. The importance of management skills to the growth of SMEs and access to finance brings us to the fifth hypothesis:

H5: As the top manager's working experience increases, the probability of perceiving access to finance as a significant obstacle decreases.

Lee (2014) developed the study of Brush et al. (2009) and investigated the obstacles that were holding back high growth of small firms in the UK. Using the Small Business Survey in the UK, firms were divided into high growth firms and potential high growth firms. He analyzed the effects of six key barriers to high growth and potential high growth firms. The selected variables were "recruitment" "government", "premises", "market conditions", "management" and "finance". The results showed that actual high growth firms were no longer constrained by market conditions, but they were significantly affected by the other five barriers. On the other hand, potential high growth firms were less likely to perceive "government" as a significant problem. Similarly, "recruitment" which was expected to be important by the author, appears to have been less significant. The author explains that the difference between expectations and the results may have been due to the matching process of potential high growth firms and also to the diversity of the interviewees' experiences. These literatures suggest that high growth firms may have special funding needs. From this comes the sixth hypothesis for high growth firms:

H6: High growth firms are more likely to perceive access to finance as a significant barrier than firms with a slower growth rate.

What emerges from the literature is that SMEs face a range of different barriers. A common finding in most of the studies is that SMEs face a financing problem – a problem of access to funding. But the studies also show that there is a considerable range of barriers depending on the conditions of specific markets. Another important finding is that obstacles to the growth of SMEs are determined by a variety of factors and, once again, the specific conditions may vary from country to country. The determinants can be grouped as "internal" or "external". Internal factors typically include a variety of firm characteristics. External factors usually refer to barriers related to access to credit. Both of these issues – barriers and their determinants - will be addressed in the following section together with an explanation of how we propose to deal with them in this study.

1.3 Data and Methodology

1.3.1 Aim and Approach

Many of the findings noted in the previous section are specific to countries in which the research was carried out (such as UK SMEs in Lee 2014), and cannot be generalized to other regions. The aim is to see whether some of the key findings can be generalized for developing countries as a group.

Our approach will be to analyze the role of barriers to growth by using a survey based on interviews with firm managers and other officers. Their answers to questions provide rich data on their *perceptions* of barriers to growth, which is an approach commonly used in the literature.[2]

The constraints as identified in the literature vary a great deal, and our task had to be narrowed down. Our analysis will be concentrated on five key barriers. The five obstacles will be identified in Section 1.3.3 below. Repeated in most of the literature is that "finance" is one of the biggest obstacles. As we shall see later in the text, "finance" was also identified as one of the major obstacles to SMEs in the World Bank Survey which we shall use in this study. Even though we shall identify and target five major barriers and provide commentaries, our main attention will be focussed on "finance" as the main obstacle. This partly reflects

[2] Please see also discussion in the following section.

the importance of “finance” in the World Bank survey as well as the result of our reading of most traits of the literature.

An attempt will also be made to identify the major determinants of the barriers. We shall start by selecting a range of factors identified in the literature and the description of the selection is also provided in Section 1.3.3 below.

1.3.2 Data

Our study draws on cross-country data. This choice was determined by the task at hand – our attempt to study the obstacles to the growth of SMEs in developing countries as a group. Following the practices of many institutions, a distinction is made between developed and developing countries, reflecting their differences in terms of the level of economic development, the level of industrialization, and the development and sophistication of markets that affect the business environment. Considerable differences exist, particularly in the range and depth of the financial industry, but also in regard to many other factors and attributes. Taking into account the important function of SMEs in developing economies. The analysis of problems faced by SMEs in developing countries in general has high economic significance and it is the originality of this research. However, a word of caution is necessary at this point as the use of cross-section data has its limitations. Perhaps the most serious limitation is the heterogeneity of individual country conditions which could lead to “identification” problems in regression analysis. This limitation typically means that studies of SMEs' performance are carried out with the help of time series or panel data. Such an approach would clearly be impossible in our case – the task would be far too complex and expensive. In using cross-section data we assume, therefore, that heterogeneity of countries is minimal or with differences not generating biases in our estimations.

Our analysis will draw on data obtained from a survey. The survey focuses on perceptions and views of managers of SMEs of barriers to growth, and it is legitimate to ask whether those perceptions are the true reflections of real barriers to growth. By using the survey, it is assumed that there is a close relationship between the perception of barriers and real barriers. The assumption has been discussed and questioned in the literature, for example, by Doern (2009). We believe, together with many other researchers in the field, that such an analysis of barriers is revealing and useful. The main conclusions of the study are consistent with the theory as well as with findings from many individual country studies.

The data used in this paper comes from the Enterprise Survey (ES) which is an ongoing project from the World Bank. The main objective of the survey is to assist the World Bank in pursuing one of its strategic goals to build a climate for investment, job creation and sustainable growth. To be more specific, the survey aims at providing investment indicators and also the constraint to the growth of the private sector to achieve the final target of enhancing employment and economic growth.

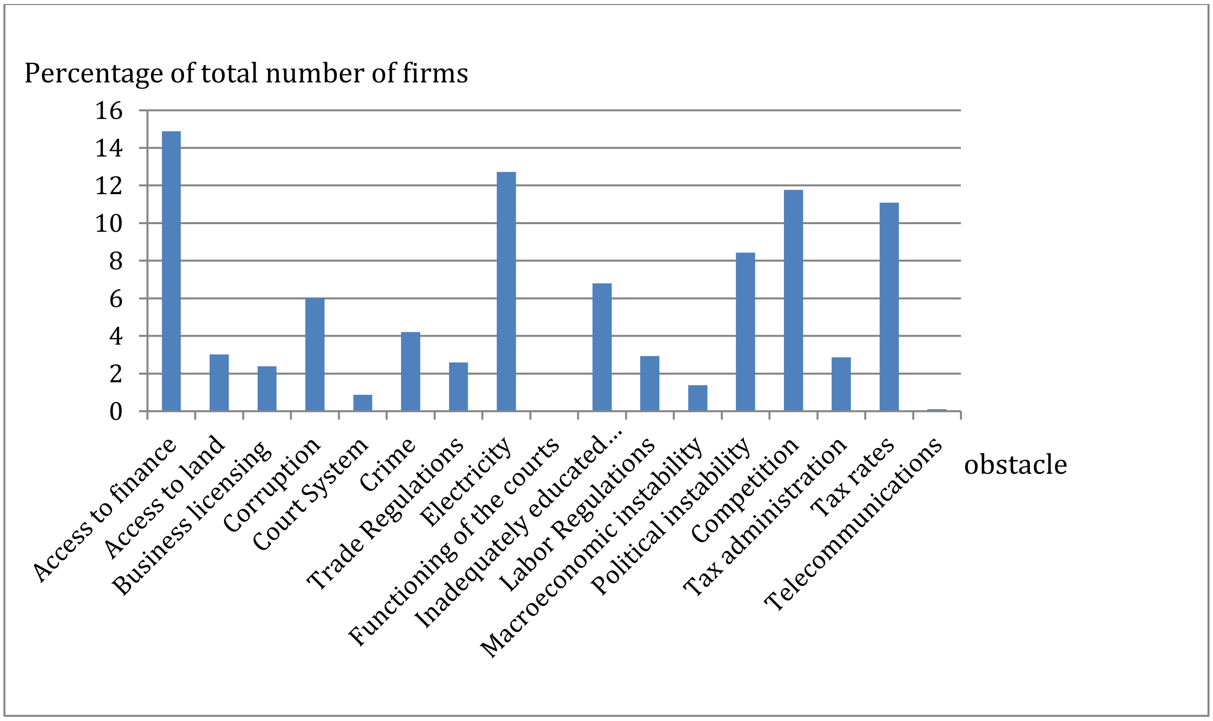
The survey is a firm-level survey conducted through 130,000 firms in 135 countries, of which 119 are conducted through the standard methodology. It includes 41 Sub-Saharan African countries, 29 from Eastern Europe and Central Asia, 31 are from Latin America and the Caribbean, 12 are in East Asia and Pacific, 4 are in South Asia, and only two are in the Middle East and North Africa. Thus the ES is a suitable dataset to investigate the economic environment and policies in developing countries.

The data is collected from face-to-face interviews with managing directors, accountants, human resource managers and other relevant firm staff by private contractors on behalf of the World Bank. Since 2002, over 73,000 interviewees have joined the survey. The survey contains responses from 2006 to 2014. In order to test the consistency check for the survey, there is a pilot questionnaire for each country which contains 20~25 interviews. If the regional differences are considerable, then an attempt is made to pilot the survey in all the major regions in that country.

1.3.3 Variables

In this section, we shall describe the selection of variables used in this paper and their features. The dependent variable is the obstacles firms are facing in their business. As the key barriers, we have selected the five most important obstacles which were identified in the World Bank survey. The choice was represented by the answers to the following survey question: “Which of the above obstacles is the biggest obstacle to the current operation of the firm?” The independent variables were chosen from the literature review. All the chosen variables are listed in Table 1.

Chart 1**:** The Main Barriers to Growth as Perceived by SMEs (In percent of the total number of firms)



The survey generated useful series of variables for investigating the perceived obstacles to firm growth in developing countries. The answer of the respondents from 119 developing countries for the period of 2006-2014 is shown in Chart 1. As shown by the chart, the five most severe problems were: Access to finance, Electricity, Political instability, Competition and Tax rate. These five variables are chosen as the dependent variables in the regression.[3] If the surveyed companies chose any of the listed obstacles as the most significant obstacles then the variable is set as “1”, otherwise “0”.

Table 1: Description of Variables (Barriers to Growth of SMEs)

Dependent (D) and Independent (I) Variables	Description
Finance (D)	Dummy variable: Access to finance is a major obstacles-1; is not a major obstacles-0
Tax (D)	Dummy variable: Tax rate is a major obstacle -1; is not a major obstacles-0
Competition (D)	Dummy variable: Competition is a major obstacles-1; is not a major obstacles-0
Electricity (D)	Dummy variable: Electricity is a major obstacles-1; is not a major obstacles-0
Political (D)	Dummy variable: Political is a major obstacles-1; is not a major obstacles-0
High growth firms (Hgf) (I)	Dummy variable: Firms with high growth rate enterprises number of employee bigger than 1.2^3=1.728 times as much as 3 years ago-1 number of employee less than 1.728 times as much as 3 years ago-0
SME (I)	Dummy variable: small and medium sized-1; large and very large-0
Ownership (I)	Have state ownership-0; Totally private-owned-1
Age (I)	Age of the firm
Experience (I)	Top manager’s years of working experience in the sector

[3] Please see also Table 3 further below and the accompanying discussion.

In order to proceed, we need to address other methodological issues related to definitions of concepts and characteristics of firms. First of all, as independent variables were selected for this paper “high growth firms”, ”employees”, “sme”, “age”, ”ownership” and “experience”. The choice was arbitrary but largely reflects again our reading of the most frequently discussed firm characteristics as determinants of SMEs’ performance. Turning now to “growth of SMEs”, Ayyagari *et al.* (2014) used the number of employees to measure the size-growth of SMEs and investigated the relationship between the size of the firm and the number of jobs it created. Organization for Economic Co-operation and Development (OECD) has provided the definition of high growth firms as those which achieved a 20% employment growth/annum for 3 consecutive years. This definition has been widely adopted in the literature. For example, Lee (2014) used the change in the growth of the number of employees as the definition in his study. In this paper, the variable “high growth firms” will also be defined by the number of employees. A comparison of the full-time employees over 3 years periods will give a clear indication of whether a firm is expanding. “High growth firms” will be a dummy variable; when a firm is expanding and it reaches a 20% growth rate then it can be defined as a high growth. enterprise. When the firm is growing fast it will be set as”1”, if it is not then it will be set as “0”. We define high growth firms based on the answers to two questions available in our dataset: “number of permanent full-time employees of this firm at end of last fiscal year” and “number of permanent full-time employees of this firm at end of 3 fiscal years ago”. We do not have the data for the second year. Thus, accumulated 20% growth rate for 3 years as the proxy for high-growth rate.

The variable “employees” comes from the survey question - “At the end of the fiscal year, how many permanent, full-time employees did this establishment employ?” An investigation made into the relationship of this variable can give us an idea of whether a firm's size will influence the obstacles it faces. As Lee’s (2014) research showed, the bigger the firm the fewer financial obstacles it will face and the more management obstacles it will have.

Since the research scope of this paper is SMEs, the variable “sme” is used to define whether the observations are SMEs. The World Bank Enterprise Survey (WBES) classifies enterprises with less than 20 employees as small size and those with 20-99 as medium size. In our dataset, there is a category variable size and it has three categories: small (number of employees <20), medium (number of employees 20~99), and large (number of employees 100 and over 100). So enterprises with less than 100 employees are grouped as SMEs. If the firm is an SME then it will be defined as 1, if it is not, then 0.

The variable “ownership” is a dummy variable. It comes from the survey question - “What percent of this firm is owned by the government?” The answer is the percentage of state ownership. A firm is defined as state-owned if the state has a share in the ownership – irrespective of the level. In such a case, the dummy variable for ”ownership” is set as “0”. When the firm is totally private (i.e. the answer to the above question is “0” percent of the firm is owned by the government”), the variable for “ownership” is set as “1”. Hypothesis 2 can thus be tested: whether state-owned enterprises will have any privilege in financing or in affecting other operations of the business (Yin, 2012).

The variable “age” comes from the survey question - “What was the established year of the enterprise?” We then use 2014 as the year of the survey and subtract it from the year of the establishment of the firm in order to get the age of the enterprises. This variable can address the question of whether young firms are experiencing more obstacles than older firms (Chavis *et al*., 2010).

Following Brush *et al*. (2009), the variable “experience” is used to describe how many years the top manager has been working in the industry. The question under investigation is “whether the company with experienced managers will be less likely to perceive access to finance as a significant obstacle than those with less experienced management.” (Hypothesis 5).

1.4 The Model

As noted above, the dependent variable in our analysis will be the firm's perception of the biggest obstacles to its current operations and it is a dummy variable. When the firm perceives a certain obstacle to be the obstacle to growth, then it is set as "1", otherwise "0". The independent variables are the characteristics of the firms which consist of both continuous and dummy variables. The estimation model of a specific obstacle can be constructed as

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 hgf + \beta_2 sme + \beta_3 age + \beta_4 employees + \beta_5 ownership \\ & + \beta_6 experience + \varepsilon \end{aligned} \tag{1}$$

where Y is the outcome variable which represents whether firm i perceives a specific obstacle to be the biggest obstacle to its current operation. The independent variables were described in Table 1 above.

The probit model is used for analysis since our outcome variables are discrete. Furthermore, since the outcome variable is ranked from 1 to 5, an ordered probit model is put into use to investigate the relationship between the severe level of the obstacles and the firm characteristics. The "severity" (the level) of the constraint is obtained from answers to the question - how severe the firm perceived a specific kind of obstacle to be the major constraint of its current operation. The answers were graded on a five-point scale: no obstacle at all (1), minor obstacle (2), moderate obstacle (3), severe obstacle (4) and very severe obstacle (5). [4]

1.5 Results

This section presents the results of the regression analysis. Using different outcome variables in our regressions equations, we shall first identify the most important barriers to the growth of SMEs. We shall then discuss the relationship between different firm characteristics and the probability of perceiving a given obstacle to play a significant role.

[4] This "order" variable is used in the literature as a proxy for the credit constraint when it comes to studying the obstacle "access to finance" (Kuntchev *et al.* 2013).

Finally, we shall present the results of our estimation of the relationship between the level of the financing constraint and the selected determinants.

Table 2: Marginal Effect of Probit Regression

	(1)	(2)	(3)	(4)	(5)
VARIABLES	finance	tax	competition	electricity	political
Sme	0.231***	0.0151	0.129***	0.00633	-0.125***
	(0.0151)	(0.0161)	(0.0157)	(0.0175)	(0.0175)
High growth firms	0.0664***	-0.0238	-0.0405**	-0.0476**	-0.0335
	(0.0181)	(0.0206)	(0.0199)	(0.0214)	(0.0236)
Age	-0.0000647**	0.00000668	-0.0000129	0.0000391	0.0000424
	(0.0000256)	(0.0000264)	(0.0000267)	(0.0000280)	(0.0000305)
Ownership	0.0103*	0.00776	-0.00472	0.00193	-0.000657
	(0.00546)	(0.00499)	(0.00574)	(0.00754)	(0.00725)
Experience	-0.000158	-0.00000796	-0.0000620	-0.00280***	-0.0000303
	(0.000208)	(0.0000597)	(9.91e-05)	(0.000607)	(0.000111)
Constant	-0.849***	-2.203***	-2.308***	-1.280***	-0.956***
	(0.134)	(0.168)	(0.215)	(0.364)	(0.185)
Pseudo R square	0.2645	0.1222	0.0850	0.2397	0.1648
Observations	85,018	86,376	86,835	86,752	84,071
Fixed effects	YES	YES	YES	YES	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In order to figure out the barriers faced by SMEs compared to big firms, the data used for Table 2 includes big firms. The estimates of the marginal effect (Table 2) reflect the extent to which SMEs are more likely to perceive finance, tax, competition, electricity and political factors as a significant constraint that impedes their growth. To be more specific, and most interestingly, SMEs are 23.1 percentage points more likely to perceive access to finance as the biggest obstacle to their growth than large firms. This confirms our Hypothesis 1. Moreover, the results also show that SMEs also have higher probability of perceiving competition as a significant obstacle than large firms. It also shows that SMEs worried less about political issues compared with large firms. Estimates of neither “tax” nor “electricity” turned out to be significant. This is probably because tax and electricity obstacles are a general problem for all enterprises in developing countries, thus it is not significant for SMEs. Dummy variables for country and industry were added to control the heterogeneity.

The focus will now be put on SMEs in the following analysis. Therefore, the data concerning large firms which have over 100 employees is eliminated from the dataset and 16322 big firms were deleted.

Table 3: Summary of the obstacles for SMEs

	#1Most		#2 Most		#3Most		Sum	
	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent
Non-response	3,712	5.21	905	5.25	1,394	8.12	6,011	6%
Access to finance	11,096	15.57	1,680	9.75	1,496	8.71	14,272	13.51%
Access to land	2,271	3.19	538	3.12	472	2.75	3,281	3.10%
Business licensing and Permits	1,681	2.36	505	2.93	595	3.47	2,781	2.63%
Corruption	4,385	6.15	1,265	7.34	1,230	7.16	6,880	6.51%
Court System	588	0.83	225	1.31	224	1.3	1,037	0.98%
Crime, theft and disorder	3,019	4.24	1,142	6.63	1,088	6.34	5,249	4.97%
Customs and Trade Regulations	1,724	2.42	448	2.6	435	2.53	2,607	2.47%
Electricity	9,469	13.29	1,288	7.47	1,003	5.84	11,760	11.13%
Functioning of the courts	9	0.01	18	0.1	30	0.17	57	0%
Inadequately educated workforce	4,344	6.09	910	5.28	1,084	6.31	6,338	6.00%
Labor Regulations	1,798	2.52	666	3.86	710	4.14	3,174	3.00%
Macroeconomic instability	999	1.4	1,200	6.96	1,198	6.98	3,397	3.21%
Political instability	5,798	8.14	1,084	6.29	1,056	6.15	7,938	7.51%
Practices of competitors	8,543	11.99	1,649	9.57	1,741	10.14	11,933	11.29%
Tax administration	1,983	2.78	832	4.83	847	4.93	3,662	3.47%
Tax rates	7,925	11.12	2,056	11.93	1,754	10.22	11,735	11.10%
Telecommunications	81	0.11	68	0.39	71	0.41	220	0.21%
Transportation	1,847	2.59	755	4.38	741	4.32	3,343	3%
Total	71,272	100	17,234	100	17,169		105,675	100%

Table 3 is a summary of the selected obstacles. The table has merged the top 3 obstacles from the survey. It can be clearly seen that access to finance has occupied the highest frequency of all the obstacles namely 14722 and it accounted for 13.51% of the total observations. This number exceeds the second most important obstacle “competition” with 2339. This provides further support for our selection of dependent variables and the emphasis on testing our Hypothesis 1.

Table 4: Marginal Effect of Probit Regression (SMEs)

	(1)	(2)	(3)	(4)	(5)
VARIABLES	finance	tax	Competition	Electricity	political
High growth firms	0.0845***	-0.0252	-0.0183	-0.0608***	-0.0306
	(0.0197)	(0.0229)	(0.0218)	(0.0232)	(0.0262)
Employees	-0.00364***	0.000539	-0.00140***	-0.000293	0.00107***
	(0.000312)	(0.000335)	(0.000324)	(0.000366)	(0.000380)
Age	-0.0000544*	-0.0000179	-0.00000917	0.0000316	0.0000339
	(0.0000287)	(0.0000313)	(0.0000304)	(0.0000318)	(0.0000354)
Ownership	0.0124**	0.00508	-0.00844	-0.0109	0.00137
	(0.00630)	(0.00564)	(0.00646)	(0.00829)	(0.00844)
Experience	-0.000130	0.00000373	-0.0000771	-0.00279***	-0.0000468
	(0.000169)	(0.0000631)	(0.000118)	(0.000679)	(0.000138)
Constant	-0.560***	-2.197***	-2.097***	-1.194***	-1.165***
	(0.143)	(0.183)	(0.220)	(0.371)	(0.206)
Pseudo R square	0.2633	0.1235	0.0858	0.2415	0.1643
Observations	68,795	70,158	70,575	70,578	67,778
Fixed effects	YES	YES	YES	YES	YES

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4 shows the new marginal effects of the selected independent variables for the five major obstacles as perceived by SMEs after the elimination of data on large firms. As noted above, “high growth firms” are represented by firms that achieved at least 20% growth every year times or about 72 percent over the three year period. The table clearly shows that when the firm in point is a high growth firm, then it will have a greater chance of perceiving access to finance to be an important obstacle than those firms which are not growing at a fast rate. This may be due to the fact that high growth firms are “cash hungry” machines as noted by Brush *et al.* (2009). Their rapid growth results in great demand for money since funds are a necessity for business expansion. The same conclusion has been reached by Brush *et al.* (2009) and it supports our Hypothesis 6. Moreover, the high growth firms appear to be less worried about tax, electricity, political stability as well as competitors from the informal sector.

The variable “employees” is used to define the number of employees of the enterprises at the time of completing the survey. This variable can be used to define the size of the enterprises. As shown in Table 4, when the size of the enterprises as measured by the number of employees is getting larger, the probability that the firm perceives access to finance as the greatest obstacle decreases. Shen *et al.* (2009) have indicated that small firms have to face more financing constraints and access to bank credit, at least based on the evidence from China. Moreover, for larger SMEs, the probability of perceiving informal competition decreases. On the other hand, larger SMEs will worry more about political stability than smaller ones.

Our tests concerning the role of the age of SMEs in determining access to finance tend to confirm our Hypothesis 4 as well as the main findings in the literature but the relationship tends to be weak. As Kuntchev *et al*. (2013) note, the interaction effect of firm size and age is significant and negatively correlated with the credit constraints of firms. Lee (2014), too, chooses age as a control variable in his research and explains the importance of age on the grounds that older firms may have a credit history and established relationships with banks – in contrast to younger firms..

The variable “ownership” is another important variable of our interest. As our estimates presented in Table 4 show, firms that have public ownership perceive fewer financial problems than those privately owned firms. This may due to the fact that state-owned enterprises have the government’s bail-out explicit or implicit guarantee which increases their creditworthiness. The effect of other determinants turns out to be insignificant.

The coefficient of the variable “experience” is significant only when it comes to “electricity” even though the signs of other estimated coefficients are correct. This suggests that as the working experience of top managers is increasing, the probability of the firm to perceive electricity as a significant obstacle is decreasing.

We shall now turn to the factors that influence the level of the relationship.

Table 5: Determinants of Financial Constraints: Results of Ordered Probit Regression

INDEPENDENT VARIABLES	(1) Finance (level)
High growth firms	0.0944***
	(0.0139)
Employees (Size)	-0.00223***
	(0.000209)
Age	-0.0000315
	(0.0000198)
Ownership	0.00692*
	(0.00410)
Experience	-0.00000663
	(0.0000492)
Observations	67,351
Country dummies	YES
Industry dummies	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5 shows the results of the ordered probit regression exploring the relationship between the level of the financing barrier and the selected variables. The dependent variable "level" comes from the survey question "How severe is access to finance as an obstacle to the current operation of the firm?" As the table shows, a significant negative correlation has been revealed between the level of financing constraint and the firms' size which implies that smaller firms experience more severe financing problems than larger firms.

Similarly, high growth firms will perceive financing problems to be more severe than those without high growth rates. Nevertheless, the result shows "ownership" is also negatively correlated with the level of the financing problem which is consistent with the finding of Yin (2012). Our findings imply that private firms will perceive access to finance as a more severe obstacle than firms with state ownership.

After determining the internal characteristics of firms that influence the importance of perceived financial constraints for SMEs, we shall now consider external factors which can act as constraints on the operations of SMEs. We shall do so by examining the role of conditions applied to bank loans.

Table 6: Reasons for Not Applying for a Loan

Main reason for not applying for new loans or new lines of credit	Freq.	Percent
Don't know	619	1.17
Refuse to answer	34	0.06
No response	7	0.01
Still in process	749	1.42
Skip	3	0.01
No need for a loan	28,742	54.53
Application procedures for loans are complex	5,064	9.61
Interest rates are not favorable	7,562	14.35
Collateral requirements are too high	3,666	6.95
Size of loan or maturity is insufficient	976	1.85
It is necessary to make informal payments	1,617	3.07
Did not think it would be approved	3,290	6.24
Other	384	0.73
Total	52,713	100

As the figures in Table 6 show, 54.53 percent of SMEs did not need a loan. This indicates that internal funds were the main source of financing for SMEs.[5] Among the SMEs which need external financing, it is evident that the financing difficulties usually result from the following reasons: (1) high interest rate; (2) complex application procedures; (3) high collateral requirements; (4) perception of SMEs that the application would not be approved; (5) informal payments. Those reasons can also be categorized into two groups: high expenses with loan processing and lack of consultant support. High interest rates, informal payments as well as the time demanding procedures all lead to high expenses related to obtaining funds from a bank. High requirements for collateral and lack

[5] As noted by Jiang et al (2014), Gert Wehinger (2014) and Abdulsaleh and Worthington (2013), internal financing is still a dominant form of financing for SMEs and prioritised compared to external financing.

of confidence imply a lack of credit guarantee institutions. The consequences are similar to the observations made by Beck and Demirguc-Kunt (2006) who concluded that asymmetric information between borrowers and lenders plus the high transaction costs are the two leading constraints that exacerbate the financing available for SMEs.

1.6 Conclusion

SMEs are drivers of economic growth and job creation. Moreover, SMEs are effective tools for poverty alleviation. As a result, the development of SMEs is vital to developing countries, and it is, therefore, paramount to determine the factors hindering their growth. This paper is an attempt to identify the main obstacles to growth and their determinants as perceived by SMEs. The five most significant obstacles perceived by SMEs managers were identified as - “access to finance”, “tax rate”, “competition”, “electricity” and “political factors”. Among those five obstacles, “access to finance” appears to be the biggest barrier, followed by “competition”.

The picture emerging from the evaluation of factors determining the managers’ perceptions of those obstacles is mixed. Among the selected variables, “experience” has been shown to be insignificant with one single exception while “high growth enterprises”, “age”, “employees” and “ownership” were all significantly correlated with access to finance. Nevertheless, the effect of “age” turned out to be relatively small. The results suggest, *inter alia,* that high growth firms perceive finance as the biggest obstacle to growth. This, in turn, confirms widely held beliefs that high growth firms have greater demand for funds than those slower-growing firms. SMEs with state ownership appears to have fewer financing problems than private SMEs. This, too, confirms the findings from the literature which have shown that firms with state participation had better access to bank financing due to implicit or explicit guarantees from the governments and due to other government interventions.

We have also made an attempt to evaluate the level of the financing problem. Perhaps the most interesting finding is that size and age were negatively correlated with a “severe” level of the financing constraint. This implies that, with increasing size and age, the bigger and older SMEs respectively will be less likely to perceive access to finance as a severe problem. This is a plausible conclusion which also provides more light on the finding noted above that age does not seem to be a strong driver of the financing problem.

Following the analysis of the internal factors affecting the access to finance of SMEs, we have also looked the role of the external factors. Those factors can be grouped under the heading of “terms of financing”. The role of external factors can be ascribed to imperfections of the financial system due to factors such as asymmetric information between banks and SMEs, financial market fragmentation and a lack of specialized banking or high transaction costs. Our results show that more than half of the SMEs did not need a loan which indicates that most of the SMEs preferred internal financing. For SMEs in need of external financing, the most serious constraints were high interest rates, complex application procedures, and high collateral requirements.

Reducing the external obstacles that impede the growth of SMEs requires government effort in building up a comprehensive financial infrastructure, with features such as a solid credit system for generating small business credit profiles, a user-friendly accounting and taxation system, and supportive lending and taxation policy for SMEs. In addition, the government should commit to providing a small business-friendly commercial environment which includes: first, enhancing the basic infrastructures such as electricity, transportation and telecommunications; second, providing macroeconomic and political stabilities; third, perfecting the business law system; and fourth, establishing antitrust laws and encouraging healthy competition. Financial institutions also play an important role in helping SMEs' growth. Financial innovations in alternative lending methods, such as disintermediated lending platforms and, electronic applications, can provide convenience to small business borrowers. Traditional banks should also actively apply new technology to smooth the application process, reduce transaction costs, and build up big data based credit models. Alongside external changes, small business entrepreneurs should improve their financial literacy and understanding of the lending process and the credit rating system. Small business owners should note the importance of accounting and tax recording in getting funds from banks. A large group of literature (eg., Fagariba, 2016; Vincent, 2021; Alkhatib et al. 2018) confirms the common existence of tax evasion in SMEs in developing countries. Thus, understanding the logic of healthy business development is critical to SMEs in developing countries.

It would be reasonable to ask to what extent is a perception of barriers to growth by managers of SMEs the true reflection of real barriers. As we have noted above, the assumption concerning the identity between the perception of barriers and real barriers is a common challenge in studies of this kind. While the analysis of real constraints on growth was not the subject of this paper we believe, together with many other researchers in the field, that an analysis of perceived barriers is revealing and useful, especially with the regard to the effects of firm characteristics. The main conclusions of the study are consistent with the theory as well as with findings from many individual country studies.

Nevertheless, as is the case with studies of a similar kind, we have faced limitations in data and methodology. Our aggregate approach of looking at all developing countries as a group may be intellectually interesting but, at the same time, our analysis may not be sensitive enough to country differences even though appropriate provisions have been made in our econometric analysis. Similar concerns could be raised about the absence in the analysis of the treatment of sectoral and regional differences. It would be, therefore, legitimate to ask whether the use of cross-country data in our analysis was optimal. Unfortunately, given the complexity of the task at hand, the use of panel data or time-series data had to be abandoned on practical and cost grounds. Nevertheless, we are encouraged that the main findings of this study are consistent with what is already known from the literature. They should provide additional evidence in the debate about enhancing the performance of SMEs in developing countries.

References

Abdulsaleh, A. M., & Worthington, A. C. (2013). Small and medium-sized enterprises financing: A review of literature. *International Journal of Business and Management*, *8*(14), 36.

Alkhatib, A. A., Abdul Jabbar, H., & Marimuthu, M. (2018). The effects of deterrence factors on income tax evasion among Palestinian SMEs. *International Journal of Academic Research in Accounting, Finance and Management Sciences,* 8(4), 144-152.

Ayyagari, M., Demirgüç-Kunt, A., & Maksimovic, V. (2011). Small vs. young firms across the world: contribution to employment, job creation, and growth. *World Bank Policy Research Working Paper*, (5631).

Ayyagari, M., Demirguc-Kunt, A., & Maksimovic, V. (2014). Who creates jobs in developing countries?. *Small Business Economics*, *43*(1), 75-99.

Avery, R. B., Bostic, R. W., & Samolyk, K. A. (1998). The role of personal wealth in small business finance. *Journal of Banking & Finance*, *22*(6), 1019-1061.

Beck, T., & Demirguc-Kunt, A. (2006). Small and medium-size enterprises: Access to finance as a growth constraint. *Journal of Banking & Finance*,*30*(11), 2931-2943.

Beck, T., Demirguc-Kunt, A., & Levine, R. (2005). SMEs, growth, and poverty: cross-country evidence. *Journal of Economic Growth*, *10*(3), 199-229.

Beck, T. (2007). Financing constraints of SMEs in developing countries: Evidence, determinants and solutions. *The World Bank. Washington DC*.

Brush, C. G., Ceru, D. J., & Blackburn, R. (2009). Pathways to entrepreneurial growth: The influence of management, marketing, and money. *Business Horizons*, *52*(5), 481-491.

Chavis, L., Klapper, L., & Love, I. (2010). International differences in entrepreneurial finance. *Age*, 1, p.2.

Doern, R. (2009). Investigating barriers to SME growth and development in transition environments a critique and suggestions for developing the methodology. *International Small Business Journal*, *27*(3), 275-305.

Fagariba, C. J. (2016). Perceptions of Causes of SMEs and Traders Tax Evasion: A Case of Accra Metropolis, Ghana. *Journal of Business & Economic Management* 4 (2), 017-039.

Gree, A. & Thurnik, C. (2003). Firm selection and industry evolution: the post country performance of new firm. *Journal of Evolutionary Economics,* 4 (4), 243-264.

Henrekson, M., & Johansson, D. (2010). Gazelles as job creators: a survey and interpretation of the evidence. *Small Business Economics*, *35*(2), 227-244.

Ji, H. (2011). The reasons and solutions for Chinese SMEs' financing dilemma. *Journal of special zone economy in China*, 2, 219-221.

Jiang, J., Li, Z., & Lin, C. (2014). Financing difficulties of SMEs from its financing sources in China. *Journal of Service Science and Management*, *2014*.

Kuntchev, V., Ramalho, R., Rodríguez-Meza, J., & Yang, J. S. (2013). What have we learned from the enterprise surveys regarding access to credit by SMEs?. *World Bank Policy Research Working Paper*, (6670).

Lee, N. (2014). What holds back high-growth firms? Evidence from UK SMEs, *Small Business Economics*, *43*(1), 183-195.

Levy, B. (1993). Obstacles to developing indigenous small and medium enterprises: an empirical assessment. *The World Bank Economic Review*, *7*(1), 65-83.

Mason, C., & Brown, R. (2013). Creating good public policy to support high-growth firms. *Small Business Economics*, *40*(2), 211-225.

Pissarides, F., Singer, M., & Svejnar, J. (2003). Objectives and constraints of entrepreneurs: Evidence from small and medium size enterprises in Russia and Bulgaria. *Journal of Comparative Economics*, *31*(3), 503-531.

Pissarides, F. (1999). Is lack of funds the main obstacle to growth? EBRD's experience with small-and medium-sized businesses in Central and Eastern Europe. *Journal of Business Venturing*, *14*(5), 519-539.

OECD. (2010). High-growth enterprises: What governments can do to make a difference?. *Paris: OECD*.

Qiyue Yin (2012), A study on the dilemma of China’s small business finance. (Doctoral Dissertation, Southwestern University of Finance and Economics).

Richter, A., & Schaffer, M. (1996). The performance of de novo private firms in Russian manufacturing. Enterprise Restructuring and Economic Policy in Russia, 253-74.

Shen, Y., Shen, M., Xu, Z., & Bai, Y. (2009). Bank size and small-and medium-sized enterprise (SME) lending: Evidence from China. *World Development*, *37*(4), 800-811.

Vincent, O. (2021). Assessing SMEs tax non-compliance behavior in Sub-Saharan Africa (SSA): An insight from Nigeria. *Cogent Business & Management, 8*(1), 1938930.

Wehinger, G. (2014). SMEs and the credit crunch. *OECD Journal: Financial Market Trends*, *2013*(2), 115-148.

Zarook, T., Rahman, M. M., & Khanam, R. (2013). Management skills and accessing to finance: evidence from Libya's SMEs. *International Journal of Business and Social Science*, *4*(7), 106-115.

Annex 1 World Bank Economic Survey - Summary Information

The World Bank survey is divided into two parts: the core questionnaire and the screener questionnaire. The core questionnaire is applied to all industries in all countries. The screener questionnaire is used to screen out the establishments that cannot meet the sampling requirement or will cause bias in the dataset. Two modules are created on the basis of the core instrument: the manufacturing module and the services module. The core instruments are implemented on two groups. One covers the business characteristics and the other covers the investment climate.

Annex 2: The summary of the survey sections

Section A Control Information	Basic information of the firm's properties like size, country, region, and etc.
Section B General Information	General Information including firm's legal status, ownership, year of registration, etc.
Section C Infrastructure Conditions	Covers firm's transportation methods, conditions of electrical and water connections, internet access, etc.
Section D Sales and Supplies	Covers firm's main products, annual total sales, raw materials, etc.
Section E Competition	Covers the firm's exposure to the market, number of competitors in the market, the price adjustment, etc.
Section F Capacity	Includes information about the firm's operations, hours per week, working capacity of the workers and machines, etc.

Section G Land Information	Covers issues of ownership of land, permission of using land, expense on security, etc.
Section I Crime issue	Includes information about security expenses, effects of crime on the business, etc.
Section J Business and Government relations	Covers issue related firm's licenses, tax rates and obstacles concerning the government, etc.
Section K Finance issue	Covers issues related to firm's sources of finance, loans availabilities, finance difficulties, etc.
Section L Labor Information	Covers firm's number of employees, education level, training of employees, etc.
Section N Productivity	Includes firm's total costs, total sales, net book value and all the indicators needed for calculating profitability.

The end of every section contains a question whether the obstacle in point is "No Obstacle, a Minor Obstacle, a Moderate Obstacle, a Severe Obstacle or a Very Severe Obstacle" to the current operations of the establishment.

Annex 2 Summary of Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Finance	87619	0.150	0.358	0	1
Tax	87619	0.111	0.314	0	1
Competition	87619	0.118	0.323	0	1
Electricity	87619	0.129	0.335	0	1
Political	87619	0.084	0.278	0	1
High Growth Firm	87619	0.102	0.302	0	1
SME	87619	0.815	0.388	0	1
Age	87602	22.186	16.243	1	124
Ownership	87619	0.985	0.120	0	1
Experience	87105	17.129	10.988	0	62

Annex 3 List of Countries

Country Code	Freq.	Percent	Cum.
Afghanistan	891	1.02	1.02
Albania	664	0.75	1.77
Angola	785	0.90	2.67
Antiguaandbarbuda	151	0.17	2.84
Argentina	2117	2.42	5.26
Armenia	734	0.84	6.10
Azerbaijan	770	0.88	6.98
Bahamas	150	0.17	7.15
Bangladesh	2946	3.36	10.51
Barbados	150	0.17	10.68
Belarus	633	0.72	11.40
Belize	150	0.17	11.57
Benin	150	0.17	11.75
Bhutan	250	0.29	12.03
Bolivia	975	1.11	13.14
Bosnia and Herzegovina	721	0.83	13.97
Botswana	610	0.69	14.66
Brazil	1802	2.06	16.72
Bulgaria	1596	1.82	18.54
BurkinaFaso	394	0.45	18.99
Burundi	270	0.31	19.30
Cameroon	363	0.41	19.71
CapeVerde	156	0.18	19.89
Centralafricanrepublic	150	0.17	20.06
Chad	150	0.17	20.23
Chile	2050	2.34	22.57
China	2700	3.08	25.65
Colombia	1942	2.22	27.87
Congo	151	0.17	28.04
Costarica	538	0.61	28.66
Croatia	993	1.13	29.79
Côte d'Ivoire	776	0.89	30.68

DRC	1228	1.40	32.08
Djibouti	266	0.30	32.38
Dominica	150	0.17	32.55
DominicanRepublic	360	0.41	32.96
Ecuador	1024	1.17	34.13
Elsalvador	1053	1.20	35.33
Eritrea	179	0.20	35.54
Estonia	273	0.31	35.85
Ethiopia	644	0.73	36.58
Fiji	164	0.19	36.77
Fyr Macedonia	726	0.83	37.60
Gabon	179	0.20	37.80
Gambia	174	0.20	38.00
Georgia	733	0.84	38.84
Ghana	494	0.56	39.40
Grenada	153	0.17	39.58
Guatemala	1112	1.27	40.85
Guinea	223	0.25	41.10
GuineaBissau	159	0.18	41.28
Guyana	165	0.19	41.47
Honduras	1087	1.24	42.71
Indonesia	1444	1.65	44.36
Iraq	756	0.86	45.22
Jamaica	376	0.43	45.65
Kazakhstan	1144	1.31	46.96
Kenya	1370	1.56	48.52
Kosovo	472	0.54	49.06
Kyrgyz Republic	505	0.58	49.64
LaoPDR	630	0.71	50.35
Latvia	271	0.31	50.66
Lesotho	151	0.17	50.84
Liberia	150	0.17	51.01
Lithuania	276	0.31	51.32
Madagascar	445	0.51	51.83
Malawi	150	0.17	52.00
Mali	850	0.97	52.97
Mauritania	237	0.27	53.24
Mauritius	398	0.45	53.70
Mexico	2960	3.37	57.07
Micronesia	68	0.08	57.15
Moldova	723	0.83	57.98
Mongolia	722	0.82	58.80
Montenegro	266	0.31	59.11
Mozambique	479	0.55	59.65
Myanmar	632	0.72	60.37
Namibia	329	0.38	60.75
Nepal	850	0.97	61.72
Nicaragua	814	0.93	62.65
Niger	150	0.17	62.82
Nigeria	1891	2.16	64.98
Pakistan	935	1.07	66.04

Panama	969	1.11	67.15
Paraguay	974	1.11	68.26
Peru	1632	1.86	70.12
Philippines	1326	1.51	71.64
Romania	1536	1.75	73.39
Russia	5224	5.97	79.35
Rwanda	453	0.52	79.87
Samoa	255	0.29	80.16
Senegal	606	0.69	80.85
Serbia	848	0.97	81.82
Sierra Leone	625	0.71	82.53
SouthAfrica	937	1.07	83.60
SriLanka	610	0.70	84.30
StKittsandNevis	150	0.17	84.47
StLucia	150	0.17	84.64
StVincentandGrenadines	154	0.18	84.82
Suriname	152	0.17	84.99
Swaziland	307	0.35	85.34
Tajikistan	360	0.41	85.75
Tanzania	1142	1.31	87.06
Timor Leste	150	0.17	87.23
Togo	155	0.18	87.40
Tonga	150	0.17	87.57
TrinidadandTobago	370	0.42	88.00
Turkey	1152	1.31	89.31
Uganda	1203	1.37	90.68
Ukraine	1853	2.12	92.80
Uruguay	1228	1.40	94.20
Uzbekistan	366	0.42	94.62
Vanuatu	128	0.15	94.76
Venezuela	820	0.94	95.70
Vietnam	1053	1.20	96.90
West Bank And Gaza	434	0.50	97.40
Yemen	477	0.54	97.94
Zambia	1204	1.38	99.32
Zimbabwe	599	0.68	100.00
Total	87620	100.00	

Annex 4 List of Industries

	Freq.	Percent	Cum.
Basic Metals & Metal Products	1062	1.21	1.21
Chemicals & Plastics & Rubber	3695	4.22	5.43
Construction	817	0.93	6.36
Electronics	894	1.02	7.38
Fabricated metal products	842	0.96	8.34
Food	7441	8.49	16.84
Hotels & Restaurants	345	0.39	17.23
IT & IT Services	1227	1.40	18.63
Leather Products	499	0.57	19.20

Machinery and equipment	1326	1.51	20.71
Manufacturing	11943	13.63	34.34
Motor Vehicles	419	0.48	34.82
Non metallic mineral products	1438	1.64	36.46
Other Manufacturing	7279	8.31	44.77
Other Services	19700	22.48	67.25
Post and telecommunications	3	0.00	67.26
Printing & Publishing	41	0.05	67.30
Recorded media	5	0.01	67.31
Rest of Universe	4523	5.16	72.47
Textiles, Garments, Leather & Paper	7099	8.10	80.57
Transport	233	0.27	80.84
Wholesale & Retail	16057	18.33	99.16
Wood & Furniture	732	0.84	100.00
Total	87620	100.00	

2 Adverse Selection in P2P Lending: Does Peer Screening Work Efficiently? —Empirical Evidence from a P2P Platform

The rapid development of online lending in the past decade, while providing convenience and efficiency, also generates large hidden credit risk for the financial system. Will removing financial intermediaries really provide more efficiency to the lending market? This paper used a large dataset with 251,887 loan listings from a pioneer P2P lending platform to investigate the efficiency of the credit screening mechanism on the P2P lending platform. Our results showed the existence of adverse selection in the investors' decision-making process, which indicated that the investors were predisposed to making inaccurate diagnoses of signals, and gravitated to borrowers with low creditworthiness while inadvertently screening out their counterparts with high creditworthiness. Due to the growing size of the Fintech industry, this may pose a systematic risk to the financial system, necessitating regulators' close attention. Since investors can better diagnose soft signals, an effective and transparent enlargement of socially related soft information together with a comprehensive and independent credit bureau could mitigate adverse selection in a disintermediation environment.

2.1 Introduction

Peer-to-peer (P2P) lending has passed the shakeout period and entered a steady growth period. Its development experience can provide valuable insight for current market players. The fast development of disintermediated online lending in the past decade, while providing convenience and efficiency, also generates significant concealed credit risk for the financial system (Huang 2018). For example, due to the fragile auditing process and high default rate, in August 2018 the Chinese P2P market ushered in its consolidation period and experienced a reduction of 42% in P2P platforms when 168 platforms ended operation. Even after the <Interim Administrative Measures for the Business Activities of P2P Lending> was established, the default rate in the P2P industry was still high (You 2018). According to Gao et al. (2021), Chinese P2P lending platforms have an astonishing default rate of 87.2%, based on data available in 2019. This raises questions. Does disintermediation really provide more efficiency to the lending market, or does it actually add unforeseen credit risk to the system? Does peer screening work efficiently? This paper used a large dataset with 251,887 loan listings from the pioneer P2P lending platform RenrenDai to investigate the efficiency of the credit screening mechanism under a disintermediated environment by comparing the performance of loan funding signals and repayment determinants.

A group of scholars (Dorfleitner et al. 2016; Santoso et al. 2020; Liao et al. 2015; Lin et al. 2013; Pötzsch and Böhme 2010; Khan and Xuan 2021) investigated the determinants of credit rationing in the field, but findings in the literature regarding the determinants of loan application success and repayment behavior were inconsistent. Moreover, due to data limitations, the analyses of the default determinants were insufficient. The purpose of this paper, therefore, is to contribute to the literature that explores the determinants of a loan application's performance and the default behavior of the online P2P lending platform. More importantly, a comparison of the results can provide evidence for our research question: Does the peer screening mechanism in the P2P platform efficiently diagnose the signals provided by borrowers in their loan applications? Due to limitations in the repayment history data, no similar study has been conducted using an emerging-market dataset. The only reference is Iyer et al. (2016), who explored the question by using a Prosper dataset and US credit bureau data. However, their paper did not explore the specific determinants which resulted in the misspecification. Our paper fills that gap and also enriches the literature on emerging markets. We used the dataset from P2P pioneer RenrenDai to test our hypothesis. We divided the information provided by the borrowers into two categories: hard (financial) information and soft (social) information. Our findings showed that the hard (financial) indicators were given great importance when lenders were deciding whether to lend money. However, hard information was either unimportant or even acted in the opposite direction when it came to predicting the repayment behavior of a borrower. Soft information had much less inconsistency in the two models. This proved the existence of a TYPE II error in the investors' decision-making process, which indicated that the investors were predisposed to making inaccurate diagnoses of signals and gravitated to borrowers with low creditworthiness while inadvertently screening out their counterparts with high creditworthiness. Due to the growing size of the fintech industry, this may pose a systematic risk to the financial system, necessitating regulators' close attention. Since, in contrast to hard financial-based signals, investors can better diagnose the soft signals, this implies enlarging socially related soft signals, and the buildup of a comprehensive credit bureau could mitigate the adverse selection in a disintermediation environment.

The paper is divided into five sections. The literature review provides an overview of the previous research concerning the determinants of loan application success and loan defaults in the P2P market. We compare inconsistencies to find the gaps, then we define our scope. In Section 3 we introduce general information about the dataset and present our model with a descriptive summary of the chosen variables. Section 4 analyzes the results of the model in detail. We conclude with a discussion of the policy implications in the last section.

2.2 Literature Review

In the 1950s and 1960s, (Debreu 1959; Arrow 1964) were the first to explore optimal contracts under uncertainty and laid the foundation for contract theory. In the late 1960s and 1970s, George Akerlof, Joseph Stiglitz, and Michael Spence formed incentive theory as a branch of contract theory and introduced the concepts of hidden information and hidden actions. The asymmetric information problem under incentive theory has been discussed at length in modern contract economies. Credit rationing (Stiglitz and Weiss 1981) and information signaling (Spence 1973) were the two major branches of the discussion.

One major class of contracting problems lies in hidden information, which is also regarded as adverse selection. It describes a situation in which one party to the contract has private information that the other does not. When the contract is drafted by the party that lacks private information, the uninformed party needs to screen the information possessed by the informed party. This is the screening problem. If the contract is offered by the informed party, this constitutes a signaling problem, since the informed party can signal the information they have through the type of contract offered. Akerlof (1970) used the automobile market as an example to explain the situation when one party had private information and regarded the second-hand automobile market as the market for "lemons" since the seller had private information about the condition of the car and thus had the incentive to sell below-average quality cars. This lowered the quality of the whole market, but due to the asymmetric information, the buyer can only bargain according to the average price and preferred to buy the lower-quality cars, which caused the above-average quality cars to exit the market. This situation, when low-quality products replace high-quality products, causing the entire market quality to decline, is called adverse selection. In the loan market, this refers to a situation in which high-risk borrowers are usually those who are most eagerly looking for money, and most likely to obtain the loan. How to mitigate adverse selection and how to efficiently use signals to screen borrowers has thus become a crucial and heated topic. Credit appraisal is the application of screening in the financial market; the borrower has private information about the quality of the business and the incentives of paying back. Our research investigated the efficiency of the screening mechanism in online lending and posits a possible approach for improvement.

Empirical research concerning credit analysis in peer-to-peer lending can be divided into two groups. One is targeted at analyzing the trust of lenders. This research area studies how lenders screen borrowers, or what the determinants are for the success of loan funding. The other trend investigates the borrower's repayment behavior, which indicates their creditworthiness; in other words, the potential factors that may signal the possibility of default.

From the perspective of lenders, the work of Pötzsch and Böhme (2010) is representative of the literature analyzing lenders' trust. Data was used from Germany's largest P2P platform, Smava, to analyze trust-building between borrowers and lenders. The interest rate was used as a proxy for trust level. The authors introduced the concept of soft information as the personal information the borrower was willing to disclose. The results showed that communicating personal information increased lenders' trust, but the impact was small and limited to educational and professional information. In addition, if the borrower used statements aimed at arousing pity, they were given a higher interest rate, indicating a loss of trust. Herzenstein et al. (2008), on the other hand, more comprehensively summarized the determinants of success in P2P lending into several groups: demographic characteristics, including gender, race, and marital status; financial strength, including credit ratings from credit bureaus, debt ratio, and house ownership; effort indicators, such as the effort to increase reputation, mainly through group activity and loan description; and loan decision variables, namely loan features such as amount, interest rate, and duration. Their results showed that all variables representing financial strength had a significant influence on funding success except house ownership, which was insignificant. Credit ratings from A to E were all positively related to success, except for high-risk grading, while the debt-to-income ratio was negatively related to success.

Results for demographic characteristics showed that women were more likely to receive funding, which thwarted expectations; marital status was not significant in the decision to grant a loan. African American racial identity had a negative effect on loan funding success. The effort to include a picture had no significant influence on success, but the effort to join in group activity and to give a loan description had a positive effect.

Besides these two representative works which summarized the determinants of success in funding applications, a large group of researchers examined the impact of a specific screening variable on the success of a loan application. Barasinska and Schäfer (2014) analyzed the impact of gender on the possibility of successful funding on the German P2P platform Smava; Gonzalez and Loureiro (2014) and Pope and Sydnor (2011) analyzed whether a profile picture would influence funding success. Similarly, Duarte et al. (2012) analyzed appearance and funding success, while Greiner and Wang (2009), Herrero-Lopez (2009), and Lin et al. (2013) focused on the impact of social capital on loan success. Wang et al. (2019) led an analysis of the impact of video information on loan success. Research in this field provided evidence of screening determinants from the lender's perspective but lacked a comparison with the borrower's repayment behavior. This may be due to data limitations, but without this comparison, we cannot diagnose the efficiency of these determinants. Looking from the lender's perspective can only provide information about the lender's preference but cannot show whether these preferences correctly recognize the borrower's creditworthiness. Our research is based on the determinants that previous studies provided, but in addition, we compared the results with the borrowers' repayment behavior to explore the real efficiency of the lenders' screening mechanism.

From the perspective of borrowers, Santoso et al. (2020) used data from three Indonesian P2P platforms to analyze the determinants of loan interest rates and default status. As an inconsistency in the existing literature, they also observed that factors such as age and gender have different results on three different platforms. The paper investigated the relationship of the chosen determinants with default probability and the loan interest rate. However, they did not link these two results or further investigate the phenomenon behind and the origin of the problem. Our paper aims to fill this gap and analyze whether borrower signals are correctly diagnosed by lenders. Dorfleitner et al. (2016) studied the effect of soft factors derived from the descriptive text on the probability of successful funding and probability of default on two European P2P lending platforms. Their results showed that typos, text length, and keywords evoking positive emotions are significantly related to funding success but have no impact on default probability. Their research provided the first evidence of linguistic factors in credit analysis; however, they focused solely on linguistic factors and did not examine the misdiagnosis of other soft factors when comparing lenders' judgment and borrowers' real behavior.

The first paper to touch on the efficiency of the lenders' diagnosis is that of Iyer et al. (2016). They used the advantage that they had acquired the true credit scores of the borrowers from the credit bureau, while the lenders on the American P2P lending platform Prosper only had information about the credit grading. As a predictor, they used the final interest rate collected by the borrower to assess whether the lenders on the platform would use the details available to assess the borrower's true credibility. The results showed that, within one credit category, the lenders were able to infer one-third of the variation in creditworthiness that was captured by credit scores. Their results also suggested that, on top of the traditional financial factors, non-standard "softer" information was also used in analyzing the borrower's credit risk, especially for lower credit rating borrowers. Although the paper concluded that lenders on the platform had one-third of the ability to infer the real creditworthiness of the borrower, it also pointed to misspecification, since only one-third had been captured – meaning that two-thirds hadn't. Iyer et al.'s paper opened the first debate on whether the usage of soft information would compensate for the traditional credit analysis model and add more choice for credit model development after the 2008 financial crisis. However, the authors did not delve into the specific determinants that resulted in the misspecification. Our paper is an extension of their work, in that we provide empirical evidence for the misspecification of the lenders' screening mechanism in P2P lending.

We further compared the literature on these two trends and found inconsistent results for the same variable in different models. For example, gender was insignificantly correlated with success in Pötzsch and Böhme (2010) but significantly correlated with success in Zhang et al. (2017), Herzenstein et al. (2008), and Pope and Sydnor (2011). At the same time, the female gender was shown to be positively related to default in Santoso et al. (2020) but negatively related to default in Ge et al. (2017) and insignificantly related in Pope and Sydnor (2011). Moreover, the results of Dorfleitner et al. (2016) showed that typos, text length, and keywords evoking positive emotions were significantly related to funding success but had no impact on default probability. People who mentioned education in their loan descriptions were more likely to obtain loans (results were significant), but mentioning education was shown to be insignificant in predicting default. Liao et al. (2015), found that people with higher degrees of education had a lower probability of default (significant) but were not more likely to get funding (insignificant). In Freedman and Jin (2008), the mention of education in loan descriptions had an insignificant influence on funding success, but people who did so were significantly less likely to default. Mentioning car ownership was not significantly related to success but was significantly and positively related to default. In addition, mentioning family was significantly and positively related to success but also significantly and positively related to default. Due to these inconsistencies, we doubt whether investors can truly diagnose the credit signals given by borrowers. If there are misdiagnoses, which factors resulted in these mismatches?

Thus, our hypotheses are as follows:

Hypothesis 1: Investors on the P2P platform can correctly diagnose the credit signals the borrowers provide and efficiently screen out low credit borrowers;

Hypothesis 2: Investors can more efficiently diagnose hard financially related signals than soft socially related signals.

2.3 Data, Model and Variables

The data we used is from one of the world's pioneer P2P platforms, RenrenDai, which was established in 2010. By October 2016, the total amount of its transactions exceeded 21.2 billion yuan. The platform targets microloans, with 71,000 yuan being the average loan amount. The platform consisted of 251,887 listings from 2010 to 2014. Borrowers fill out a loan application online to be published on the website. Peer investors conduct their own credit analyses and choose which loans to invest in. The funding process is completed when the entire loan amount has been filled by investors. Like crowdfunding, a single loan may have multiple investors. Thus, among the total listings, only 65,394 loans were funded. The borrowers can repay the loan in full or in monthly installments until it matures. Among the funded loans, 50,819 loans are still in the repayment process and 14,575 loans have reached maturity. In the finished loans, 13,901 loans completed the repayment process while the other 674 defaulted, representing a relatively modest default rate of about 4.2%. Detailed variable descriptions are presented below.

Since the dependent variable is binary, we use the logit model to test the determinants of loan funding and default in P2P lending. Our models are presented below:

Model I: $\text{Logit}\ (\textit{Funded}_i) = \beta_0 + \beta_1\, \textit{Hard Information}_i + \beta_2\ \textit{Soft Information}_i + \propto \textit{Control Variables}_i + \varepsilon \qquad (1)$

Model II: $\text{Logit}\ (\textit{Default}_i) = \beta_0 + \beta_1\, \textit{Hard Information}_i + \beta_2\ \textit{Soft Information}_i + \propto \textit{Control Variables}_i + \varepsilon \qquad (2)$

The dependent variable for Model I, the funding probability model, is a dummy variable which equals 1 when the loans have been successfully funded, otherwise 0. Model II is the default predicting model; the dependent variable default represents whether the loan has been repaid completely without delay. 1 represents ‘defaulted’; 0 represents ‘repaid’.

All the chosen hard and soft information variables are listed in Appendix A, Table A1. All the chosen variables are based on the references from the literature review. We use financially related information, income level and collateral as the hard information. Socially and psychologically related information such as age, gender, loan description, marital status, educational level and social media information are used as the soft information. Loan features are used as the control variables.

The hard information is represented by 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. Car and home ownership are dummy variables, with 1 indicating ‘ownership’ and 0 indicating ‘none’. We include verification of income in the model to certify accuracy.

As soft information is difficult to measure, proxies must be employed. Table 1 summarizes the proxies used in our model. Our approach to soft data is similar to that in the literature: we employ education duration (e.g., Liao et al. 2015), age (e.g., Gonzalez and Loureiro 2014), and gender (e.g., Gonzalez and Loureiro 2014; Barasinska and Schäfer 2014; Ravina 2019; Pope and Sydnor 2011). We also employ the length of the loan purpose statement as a linguistic indicator, as suggested by Lin et al. (2013) and Kim et al. (2020).

Since social impact has been proved to be a significant factor on loan success (Greiner and Wang 2009; Herrero-Lopez 2009; Lin et al. 2013), we use the verification data from Weibo (the largest Chinese social network) as our indicator of social impact. If an applicant’s social network was verified, it is represented as “1”, otherwise “0”.

Profile photos are shown to influence the funding success by Pope and Sydnor (2011). Since the profile photos on Renrendai.com are not always real pictures of the applicants, we choose video verification as the picture indicator's proxy. During the verification process, borrowers must record themselves holding their ID cards and reading a statement accepting general rules and conditions from Renrendai.com as part of the verification procedure, and then upload the video with their loan application. If the applicant accepts video verification, this is recorded as a "1," otherwise it is reported as a "0".

The expansion of mobile services is a fundamental component of Fintech 2.0, and mobile usage data is the preferred verification tool for fintech firms, particularly big data firms. Since mobile numbers were introduced to China's real-name system, allowing tracking and verifying of real cellphone users, mobile usage data has become a critical source for anti-fraud efforts. Furthermore, one of the most powerful indicators of default in the consumer finance market is mobile usage behavior. As a result, we included a variable for mobile verification in our model. This is also a dummy variable: "1" means verified, "0" means not verified.

Following Nigmonov et al. (2022) and Khan and Xuan (2021), we included the interest rate, the length of the loan, and the amount of the loan. The average interest rate is 14.9%, and the highest interest rate is 24.4%. The average amount is 60,637.93 yuan. Since the amount is quite large, we used the log of amount as the proxy to normalize the distribution. The loan term is from 1 month to 36 months. The average term is 16 months.

We summarize the descriptive statistics of all the independent variables for Model I and Model II in Table1 (a) and Table 1(b) accordingly below.

Table 1(a): Descriptive Summary of Independent Variables for Model I

Variable	Observation	Mean	Std.Dev.	Min	Max	Median	First Quartile	Third Quartile
Income	222,757	4	1.281	1	7	4	3	5
Car verified	251,842	0	0.200	0	1	0	0	0
House verified	251,842	0	0.206	0	1	0	0	0
Mortgage loan	251,842	0	0.341	0	1	0	0	0
Description	251,842	184	101.908	0	367	165	88	276

Age	251,842	31	7.688	1	86	29	26	35
Gender	251,842	0	0.370	0	1	0	0	0
Marriage	251,842	0	0.500	0	1	0	0	1
Education	236,656	14	1.755	12	19	15	12	15
Mobile verified	251,842	0	0.213	0	1	0	0	0
Weibo verified	251,842	0	0.174	0	1	0	0	0
Video verified	251,842	0	0.199	0	1	0	0	0
Interest	251,842	15	3.550	3	24.4	15	13	16
Amount	251,842	60641	100320	1000	3000,000	30,000	10,000	62,200
Log(Amount)	251,830	10.186	1.350	6.908	14.914	10.309	9.210	11.038
Term	251,842	16	10.676	1	36	12	6	24

Table 1(b): Descriptive Summary of Independent Variables for Model II

Variable	Observation	Mean	Std.Dev.	Min	Max	Median	First Quartile	Third Quartile
Income	14569	5	1.517	1	7	4	3	6
Car verified	14575	0	0.449	0	1	0	0	1
House verified	14575	0	0.437	0	1	0	0	1
Mortgage loan	14575	0	0.376	0	1	0	0	0
Description	14575	260	96.128	3	367	273	174	364
Age	14575	36	7.968	21	72	34	30	41
Gender	14575	0	0.385	0	1	0	0	0
Marriage	14575	1	0.432	0	1	1	1	1
Education	14571	14	1.788	12	19	15	12	16
Mobile verified	14575	0	0.383	0	1	0	0	0
Weibo verified	14575	0	0.375	0	1	0	0	0
Video verified	14575	0	0.484	0	1	0	0	1
Interest	14575	13	2.607	3	24	13.2	12	15
Amount	14575	47547	128784.21	3000	3000000	27100	6000	52900
Log (Amount)	14575	9.941	1.291	8.006	14.914	10.207	8.700	10.876
Term	14575	12	9.528	1	36	12	6	18

2.4 Results

Table 2 presents the logit regression results for the funding probability model and the default prediction model with coefficient and robust standard errors in brackets.

Table 1: Comparison of Logit Regression Results for Funding Probability and Default Predicting Model

	(1)	(1)
VARIABLES	Funded	Default
Hard Information Variables		
1.Income verified	2.832 ***	0.596 **
	(0.0629)	(0.232)
1.Income group 1	−0.668 ***	−0.874

	(0.105)	(1.086)
2.Income group 2	−1.660 ***	−0.604 *
	(0.0821)	(0.344)
3.Income group 3	−0.394 ***	−0.168
	(0.0191)	(0.134)
5.Income group 5	0.155 ***	−0.360 **
	(0.0232)	(0.168)
6.Income group 6	0.382 ***	0.233
	(0.0282)	(0.148)
7.Income group 7	0.475 ***	0.261 *
	(0.0323)	(0.156)
Income verified#1. Income group 1	0	0
	(0)	(0)
Income verified#2. Income group 2	1.136	2.803 ***
	(0.738)	(0.882)
Income verified#3. Income group 3	0.434 ***	0.471
	(0.0903)	(0.329)
Income verified#5. Income group 5	−0.308***	−1.156 **
	(0.106)	(0.580)
Income verified#6. Income group 6	−0.606 ***	−1.744 ***
	(0.116)	(0.584)
Income verified#7. Income group 7	−1.172 ***	−2.233 ***
	(0.117)	(0.577)
Car verified	0.448 ***	−0.394 ***
	(0.0440)	(0.110)
Home verified	0.0795	0.348 ***
	(0.0529)	(0.122)
Mortgage loan	−0.311 ***	−0.409 *
	(0.0231)	(0.216)
Homeverified#1Mortgage loan	0.240 ***	−0.179
Soft Information Variables	(0.0779)	(0.276)
Loan description	0.0130 ***	−0.00603 ***
	(9.02×10^{-5})	(0.000549)
Age	0.0653 ***	−0.00531
	(0.00103)	(0.00625)
Gender	0.274 ***	−0.274 **
	(0.0183)	(0.129)
Marriage	0.345 ***	−0.203 *
	(0.0167)	(0.104)
Education	0.0763 ***	−0.120 ***
	(0.00441)	(0.0167)
Mobile verified	−0.515 ***	−0.486 ***
	(0.0432)	(0.131)
Weibo verified	0.605 ***	−0.627 ***
	(0.0492)	(0.151)

Video verified	2.522 ***	1.007 ***
Control Variables	(0.0423)	(0.120)
Interest	−0.304 ***	0.195 ***
	(0.00352)	(0.0138)
Amount	−0.304 ***	0.0349
	(0.00817)	(0.0452)
Term	0.113 ***	0.0117 **
	(0.000935)	(0.00579)
Constant	−2.150 ***	−3.159 ***
	(0.110)	(0.603)
Pseudo R2	0.5883	0.1674
Observations	222,437	14,566

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

Table 2 shows the logit regression results for Model I and Model II. The results show that income has a positive relationship with success since we take the mean group 4 as the reference group. Income groups lower than 4 are less likely to receive loans, while groups higher than 4 are more likely than the average group to have loans funded. This reflects the common sense of peer investors, who believe higher income means better solvency and more trustworthiness. This is consistent with most of the research in the field such as Pötzsch and Böhme (2010). However, the default results suggest that this is not the case: the lower income group is negatively correlated to default, thus they actually have lower default possibility (e.g., income groups 2 and 3), while the high income group can default more – income groups 6 and 7 are more likely to default than income group 4, for example. This may be because borrowers intend to lie about their income to create a more trustworthy image to the lenders. However, the lenders did not recognize the risk of false information. Nor has the value of the income verification been recognized: the high verified income group has a lower default probability. Nevertheless, compared to income group 4, investors give more loans to income group 3 than to groups 5,6, and 7, which is a TYPE II error that provides loans to those with lower creditworthiness. This stems from the misdiagnosis signals from income. This also implies the necessity of key information verification on the P2P platform. Since there is no credit rationing process on the platform, the judgment is based solely on unprofessional lenders. The validity of the information provided on the platform becomes critical.

After comparing the logit regression results from both models, we can see that, except for car ownership, all other hard information variables have either opposite results when compared to each other or different significance levels.

The median income group 4 is used as the reference variable, revealing that lower- income groups (1,2,3) are less likely to receive loan funding than the median income group (4), whereas higher-income groups (5,6,7) are more likely to be funded. The funding probability model shows interesting results, in that the interaction effect of verified income and declared income elicit opposite results. Surprisingly, higher-income groups are less preferred by the investor. Combined with the results of the default predicting model, we find that verified higher-income groups show lower default probability. However, higher-income groups without income verification demonstrate a higher probability of default. The implication may be that people in higher-income groups are more inclined to be dishonest regarding their incomes. In Table 3, we further analyze the distribution of the income verification, the results show that the income verification percentage increases along with the increase of income levels. Applicants in income groups 1 and 2 are very unlikely to verify their income, with the verification percentage being only around 0.3%. On the other hand, the high-income groups all have a verification percentage above 14%. However, as we can see from the regression results, investors are less willing to lend to verified high-income groups than the average income group, despite the verified high-income groups having a lower probability of default. But investors are more willing to lend to unverified high-income groups, who actually have a higher probability of default. This induces TYPE II errors among the investors since they cannot diagnose the income verification in high-income groups as a positive signal of creditworthiness and hence lend more funds to those who have a higher probability of default.

Table 3 shows the distribution of the verified income group and the percentage it occupies of the total application according to income group.

Table 2: Verified Income Distributions

Income Group	Verified	Total	Percentage
1	4	1231	0.32%
2	20	7190	0.28%
3	5641	82,862	6.81%
4	8057	65,763	12.25%
5	4597	31,046	14.81%
6	3178	19,863	16.00%
7	2133	14,802	14.41%
Total	23,630	222,757	10.61%

Lenders tend to prefer borrowers with fixed assets such as houses or cars. However, only car ownership is seen to be a significant indicator of reduced probability of default. House ownership is positively related to default. This finding is consistent with Jiménez and Saurina (2004), which shows loans with collateral are often linked to higher default rates. This is probably because loans in the P2P market are usually small-sized and fixed assets ownership is only an indicator of solvency and is not served as the collateral when the borrower is in default, this makes a car easier to monetize, whereas the process of realizing a house for loan repayment is more time-consuming and complicated, compared to smaller assets. As far as the mortgage loan is concerned, investors prefer borrowers without any debt. However, the default model suggests that the probability of default is lower for people with mortgage loans. This could be attributed to the fact that people with mortgage loans are more concerned about their creditworthiness.

For soft information, mobile verification exhibits the opposite result in the logit regression. It is negatively correlated to funding probability, but also negatively correlated to default. This means that borrowers who have mobile verification are less likely to default but are also less likely to get the loan funded. From Table 4, we can see that the percentage of mobile verified in successful loans (4.77%) is much less than in defaulted loans (17.87%).

Additionally, the percentages of successful and non-default mobile and video verified loans differed substantially. Successful mobile verified loans represent 26.6% of all verified loans, among which only 3.9% defaulted. This is lower than the total default rate of 4.6%. This substantiates a positive relationship of the verified mobile with the high creditworthiness of the borrowers. However, lenders cannot effectively diagnose the signal and categorize borrowers by this feature.

Non-financial information can improve the prediction model and can sometimes even outperform financial information in predicting default, as shown by Fernando et al. (2020) and Bhimani et al. (2013) using business loans. Now we add further evidence from the microfinance dataset.

Table 4 shows the distribution of the mobile verification in funded and not funded loans, and in defaulted and not defaulted loans.

Table 3: Mobile Verification Distribution List.

Mobile Verification	**Funded**		
	0	**1**	**Total**
0	172,187	67,650	239,837(95.23%)
1	8815	3190	12,005(4.77%)
Total	181,002	70,840	251,842(100%)
Mobile Verification	**Default**		
	0	**1**	**Total**
0	11,398	573	11,971(82.13%)
1	2503	101	2604(17.87%)
Total	13,901	674	14,575(100%)

The video verification also showed opposite results in the logit regression comparison, which is consistent with Duarte et al. (2012), where borrowers' willingness to show their appearance does not indicate that they have higher creditworthiness. However, most lenders attach great trust to video verification since the indicator is significantly correlated to loan success. As shown in Table 5, in contrast to mobile verified, 61.29% of video verified loans succeed in funding, while 8.2% default, which is 3.6% higher than the total default rate of

4.6%. This may be because borrowers that bear higher risk are willing to offer more information, indicating a classic adverse selection case and a TYPE II error.

Table 5 shows the distribution of video verification in funded and not funded loans, and in defaulted and not defaulted loans.

Table 4: Video Verification Distribution List.

Video Verification	**Success**		
	0	**1**	**Total**
0	176,955	64,433	241,388(85.85%)
1	4047	6407	10,454(4.15%)
Total	181,002	70,840	251,842(100%)
Video Verification	**Default**		
	0	**1**	**Total**
0	8878	223	9101(62.44%)
1	5023	451	5474(37.56%)
Total	13901	674	14,575(100%)

We can also see from the significance level of the variables that all the hard information is significant in the funding probability model except house ownership, but it becomes less significant in the default predicting model. However, this is not the case for soft information variables, as the results of soft information are more consistent in both models. This suggests that lenders were less capable of diagnosing the signals from hard information compared to soft information.

From our regression results, we can see that investors were not able to effectively diagnose most of the useful information from the signals provided by borrowers, especially from hard financially-related signals. This indicates that investors on the P2P platform may have lacked financial literacy regarding credit appraisal. Their biased investment decisions may have created credit risk to the disintermediated financial system. On the other hand, the P2P investors react surprisingly well to soft signals. They correctly diagnosed the effect of age, gender, educational level, marital status, and social media on creditworthiness. This has important policy implications - in a financial environment with a weak credit bureau and limited financial literacy, soft information may perform even better on credit screening. Adding more socially-related soft information into the credit rationing model could mitigate adverse selection in disintermediated financial institutions.

2.5 Discussion and Conclusions

This paper examines whether online P2P investors can accurately and effectively diagnose signals of creditworthiness during their decision-making process. Our findings reveal that TYPE II errors exist in the investors' decision-making process. Comparisons of the signs used in determining both loan defaults and loan funding show that the investors were predisposed to making inaccurate diagnoses of signals and gravitate to borrowers with low creditworthiness, while inadvertently screening out their counterparts with high creditworthiness.

This happens most with hard finance-based signals. Specifically, signals such as income and property ownership were insignificant or typically provided contradictory guidance in terms of default. However, investors have allocated disproportionate weights to this in the decision-making process of loan funding. Surprisingly, investors were more adept at diagnosing soft social signals than hard financial signals. That is, all directions of soft signals in the loan funding process were found to be accurate reflections in the default prediction model with the exception of softer signals such as video and mobile verification. These results suggest that soft social information can be a compensatory solution when hard information is not solid enough. The absence of a solid credit bureau is typically the main problem for credit appraisal in developing countries, and as our results show, soft information can provide an alternative solution in credit analysis. Due to data limitations, our soft information is restricted to social identity information. However, with the development of artificial intelligence and machine learning, softer information relevant to social behavior such as social networks and mobile usage behavior can provide more comprehensive angles of credit analysis in microfinance and deserve further research.

Our paper clearly demonstrates the existence of TYPE II errors in the disintermediated lending market, indicating a high potential credit risk in financial markets. Disintermediation reduces transaction costs in the lending process, provides convenience to borrowers and offers alternative investments to the lenders. However, due to a lack of professional training and credit rationing skills, lenders in this industry may misdiagnose the credit signals sent by borrowers. Due to the growing size of the Fintech industry, this may pose a systemic risk to financial systems, warranting regulators' close attention. To consolidate the system and to prevent shadow banking from infiltrating the industry, the lending license regulations need to be tightened. In addition, in order to avoid the issue of lenders blindly pursuing profit without considering risks, the interest rate cap should be closely monitored in this field. P2P lending platforms could provide guidelines about credit analysis. Moreover, the verification process could be strengthened in the platform. This can be achieved by cooperation in data sharing between the private sector and the credit bureau.

In addition, we believe the misidentification of creditworthiness signals can be alleviated by a sophisticated and independent credit bureau, and by increasing public financial literacy. Expanding the use of soft social information could also mitigate adverse selection in disintermediated financial institutions. This process must be accompanied by the establishment of transparent and effective oversight on the use of soft information in order to avoid abuse.

Appendix A. List of Variables

Table A 1: Description of Independent Variables

Variables	Description
Hard Information	
Income level	Category variable: Monthly income of the borrower (1~7) Group 1: <1000 yuan Group 2: 1001~2000 yuan Group 3: 2000~5000 yuan Group 4: 5000~10000 yuan Group 5: 10,000~20,000 yuan Group 6: 20,000~50,000 yuan Group 7: >50,000 yuan
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 in percentage
Term	Length of the loan in months
Amount	Amount of the loan, used log of amount as the proxy

Appendix B. Robustness Check

To control for multicollinearity, we analyzed the variance inflation factors (VIF) of our chosen variables. As shown in Table B7, all the independent variables' VIFs are within 2, with an average of 1.27. In other words, the variance of the estimated coefficients is inflated with very low factors and within the reasonable rule-of-thumb of 10. For verification, we also calculated the square root of VIF, the R square for the correlation between the given independent variable and the rest of the independent variables, and the tolerance indicators, which are computed as 1- R square. The results prove the non-existence of multicollinearity.

Table B 1: Collinearity Diagnostics

Variable	VIF	SQRT VIF	Tolerance	R-Squared
Income verified	1.17	1.08	0.8515	0.1485
Income	1.4	1.18	0.7168	0.2832
Car verified	1.39	1.18	0.7174	0.2816
Home verified	1.34	1.16	0.7442	0.2558
Mortgage loan	1.14	1.07	0.8751	0.1249
Loan Description	1.23	1.11	0.8153	0.1847
Age	1.38	1.18	0.7229	0.2771
Gender	1.02	1.01	0.9771	0.0229
Marriage	1.26	1.12	0.7931	0.2069
Education	1.04	1.02	0.9591	0.0409
Mobile Verified	1.18	1.09	0.8465	0.1535

Weibo Verified	1.13	1.06	0.8851	0.1149
Video Verified	1.34	1.16	0.7447	0.2553
Interest	1.07	1.04	0.9328	0.0672
Amount	1.7	1.3	0.5898	0.4102
Term	1.58	1.26	0.6334	0.3666
Mean VIF	1.27			

References

Akerlof, George A. 1970. The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84: 488–500.

Arrow, Kenneth J. 1964. The role of securities in the optimal allocation of risk-bearing. The Review of Economic Studies 31: 91–96.

Barasinska, Nataliya, and Dorothea Schäfer. 2014. Is crowdfunding different? Evidence on the relation between gender and funding success from a German peer-to-peer lending platform. German Economic Review 15: 436–52.

Bhimani, Alnoor, Mohamed Azzim Gulamhussen, and Samuel da Rocha Lopes. 2013. The role of financial, macroeconomic, and non-financial information in bank loan default timing prediction. European Accounting Review 22: 739–63.

Debreu, Gerard. 1959. Theory of Value: An Axiomatic Analysis of Economic Equilibrium (No. 17). New Haven and London: Yale University Press.

Dorfleitner, Gregor, Christopher Priberny, Stephanie Schuster, Johannes Stoiber, Martina Weber, Ivan de Castro, and Julia Kammler. 2016. Description-text related soft information in peer-to-peer lending–Evidence from two leading European platforms. Journal of Banking and Finance 6: 169–87.

Duarte, Jefferson, Stephan Siegel, and Lance Young. 2012. Trust and credit: The role of appearance in peer-to-peer lending. Review of Financial Studies 25: 2455–84.

Fernando, Jayasuriya Mahapatabendige Ruwani, Leon Li, and Greg Hou. 2020. Financial versus non-financial information for default prediction: Evidence from Sri Lanka and the USA. Emerging Markets Finance and Trade 56: 673–92.

Freedman, Seth, and Ginger Zhe Jin. 2008. Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper.Com. NET Institute Working Paper No. 08-43. Available online: https://ssrn.com/abstract=1304138 (accessed on Dec 14.2021).

Gao, Ming., Jerome. Yen, and Matthew Tingchi. Liu. 2021. Determinants of defaults on P2P lending platforms in China. International Review of Economics & Finance 72: 334–48.

Ge, Ruyi, Juan Feng, Bin Gu, and Pengzhu Zhang. 2017. Predicting and deterring default with social media information in peer-to-peer lending. Journal of Management Information Systems 34: 401–24.

Gonzalez, Laura, and Yuliya Komarova Loureiro. 2014. When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. Journal of Behavioral and Experimental Finance 2: 44–58.

Greiner, Martina E., and Hui Wang. 2009. The Role of Social Capital in People-to-People Lending Marketplaces. ICIS 2009 Proceedings, 29. Available online: https://aisel.aisnet.org/icis2009/29 (accessed on Dec 14.2021).

Herrero-Lopez, Sergio. 2009. Social interactions in P2P lending. Paper presented at the 3rd Workshop on Social Network Mining and Analysis, Paris, France, June 28, pp. 1–8.

Herzenstein, Michal, Rick L. Andrews, Utpal M. Dholakia, and Evgeny Lyandres. 2008. The democratization of personal consumer loans? Determinants of success in online peer to-peer loan auctions. Bulletin of the University of Delaware 15: 274–77.

Huang, Robin Hui. 2018. Online P2P lending and regulatory responses in China: Opportunities and challenges. European Business Organization Law Review 19: 63–92.

Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo FP Luttmer, and Kelly Shue. 2016. Screening peers softly: Inferring the quality of small borrowers. Management Science 62: 1554–77.

Jiménez, Gabriel, and Jesús Saurina. 2004. Collateral, type of lender and relationship banking as determinants of credit risk. Journal of banking & Finance 28: 2191–212.

Khan, Mohammad Tariqul Islam, and Yong Yee Xuan. 2021. Drivers of lending decision in peer-to-peer lending in Malaysia. Review of Behavioral Finance Vol. ahead-of-print No. ahead-of-print.

Kim, Dongwoo, Kyuho Maeng, and Youngsang Cho. 2020. Study on the determinants of decision-making in peer-to-peer lending in South Korea. Asia-Pacific Journal of Accounting & Economics 27: 558–76.

Liao, Li, Lin Ji, and Weiqiang. Zhang. 2015. Education and Credit: Evidence from P2P lending platforms. Journal of Financial Research 3: 146–59.

Lin, Mingfeng, Nagpurnanand R. Prabhala, and Siva Viswanathan. 2013. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer to-peer lending. Management Science 59: 17–35.

Nigmonov, Asror, Syed Shams, and Khorshed Alam. 2022. Macroeconomic determinants of loan defaults: Evidence from the US peer-to-peer lending market. Research in International Business and Finance 59: 101516.

Pope, Devin G., and Justin R. Sydnor. 2011. What's in a picture? Evidence of discrimination from Prosper. com. Journal of Human Resources 46: 53–92.

Pötzsch, Stefanie, and Rainer Böhme. 2010. The role of soft information in trust building: Evidence from online social lending. In International Conference on Trust and Trustworthy Computing. Berlin/Heidelberg: Springer, pp. 381–95.

Ravina, Enrichetta. 2019. Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets. SSRN Electronic Journal. http://dx.doi.org/10.2139/ssrn.1107307.

Santoso, Wimboh, Irwan Trinugroho, and Tastaftiyan Risfandy. 2020. What determine loan rate and default status in financial technology online direct lending? Evidence from Indonesia. Emerging Markets Finance and Trade 56: 351–69.

Spence, Michael. 1973. Job market signaling. Quarterly Journal of Economics 87: 355–74.

Stiglitz, Joseph E., and Andrew Weiss. 1981. Credit rationing in markets with imperfect information. American Economic Review 71: 393–410.

Wang, Huijuan, Mengxia Yu, and Lu Zhang. 2019. Seeing is important: The usefulness of video information in p2p. Accounting & Finance 59: 2073–103.

You, Chuanman. 2018. Recent development of FinTech regulation in China: A focus on the new regulatory regime for the P2P lending (loan-based Crowdfunding) market. Capital Markets Law Journal 13: 85–115.

Zhang, Yuejin, Haifeng Li, Mo Hai, Jiaxuan Li, and Aihua Li. 2017. Determinants of loan funded successful in online P2P Lending. Procedia Computer Science 122: 896–901.

3 The role of social and psychological related soft information in credit analysis: Evidence from a Peer to Peer lending platform

Improvements in the quality of the 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.

3.1 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 increase 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.[6] 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.[7] Borrowers fill in online loan application forms and 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

[6] 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).

[7] 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.

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.

3.2 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 (2016), 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 headquarters' credit office despite the availability and transferability of both hard and soft information. Serrano-Cinca, Gutierrez-Nieto, & Lopez-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. (2016) 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 Jimenez & 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 the 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 Corn´ee (2017) and Ge, Feng, Gu, & Zhang (2017). However, the results of studies that rely exclusively on soft information are fragmented and inconclusive.[8] 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

[8] 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).

successful loan repayment.[9] Similarly, contradictory results occurred 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.[10]

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

[9] 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.

[10] 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).

approval process could lead to the provision of loans to the wrong applicants (i.e., to a Type II error in the estimating procedures).[11]

3.3 Method

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.

3.3.1 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.[12] Moreover, the explosion of P2P lending in China has been accompanied by growing credit risk and a rising likelihood of defaults.[13] 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

[11] 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).

[12] See, for example, Botsman (2017) and Chorzempa (2018) and footnote 19 and 21.

[13] 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 Corn´ee (2017).

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 borrowers’ 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, 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 investments 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, 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 of themselves.[14] 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 & Schafer, 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

[14] See, for example, Po¨tzsch & Bo¨hme (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 & Sch¨afer (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.

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.[15]

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.[16] 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 the 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 in full or in monthly installments. 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.[17] 'Failing

[15] 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.

[16] 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.

[17] 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). The borrowers which are currently paying the loans are excluded from the dataset. We only included the loans that have finished the repayment process. Only loans which are "repaid" or "defaulted" are included in the dataset.

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 on the completion percentage of the payments. 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.

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

Loans provided on the platform are used both for personal and business purposes. Unfortunately, the platform does not provide direct information about the loan's purpose. Some of the applicants disclose the purpose in the loan description. However, it is not mandatory, the information that can be used to define the purposes is inadequate. According to an interview with the CEO of RenrenDai.com, 70–80 percent of loans granted are for freelancer or micro business operational cash flow purposes.[18] Other common purposes include car loans, home renovation, and consumption. The borrowers use personal credit for the loans and make the application as an individual. Thus, we analyze the creditworthiness of the applicant based on individual credit information. We tried

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

to manually classify the loans by reading each loan description for a small random selected sample to check the impact of the loan purpose in Appendix E.

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.[19] When loans are in default, those borrowers will be added to the credit bureau's blacklist and to the P2P industry blacklist.

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 compared to traditional banks. Understanding the benefit and risks of employing soft information in credit appraisal can provide experience to the traditional bank reform and contribute to policy implications for regulators.

Contractual arrangements in China are heavily influenced by Chinese culture, which favors information derived from human relationships.[20] 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 mobile phone users, which enables lenders to trace and verify the real users of cell phones. This increases borrower transparency and enhances trust in the

[19] 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.

[20] 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.

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.[21]

3.3.2 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.

Following the empirical literature on the use of soft information in credit appraisal reviewed in section 3.2, such as Ravina (2012), Gonzalez & Loureiro (2014), and Dorfleitner et al. (2016), and the theoretical support and empirical evidence from psychology mentioned in section 3.3.1, under the choice of determinants subsection. We shall hypothesize that soft information can have predictive power in credit appraisal, thus we have the first hypothesis:

Hypothesis 1. Credit appraisal based on appropriately selected soft information can have predictive power in predicting default.

[21] 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.

Based on Faia & Paiella (2019)'s development of Theil (1967)'s Information Theory, as more precise information is available, projects' dispersion under partial information approaches the dispersion under full information. And empirical evidence from Grunert, Norden, & Weber(2005), Godbillon-Camus & Godlewski (2005), Dorfleitner et al. (2016) and Agarwal et al. (2011), the combination of soft and hard information can achieve better credit rationing results. Thus, we derived the second hypothesis:

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.

Model I: $Yi = \alpha$ Hard Information $+ \propto$ Control Variables $+ \varepsilon$ (1)

Model II: $Yi = \alpha$ Soft Information $+ \propto$ Control Variables $+ \varepsilon$ (2)

Model III: $Yi = \alpha$ Hard Information $+ \beta$ Soft Information $+ \propto$ Control Variables $+ \varepsilon$ (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 to the literature: we use duration of education (e.g., Liao et al. (2015)), age (e.g., Gonzalez & Loureiro (2014)) and gender (e.g., Barasinska & Schafer, 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.

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: 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

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 Renrendai.com, and then 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 cell phones. In addition, mobile usage behavior is recognized as one of the most effective indicators 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”.

3.3.3 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).

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

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:

$$TPR = TP/\,(TP + FN) \qquad (4)$$

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,

$$FPR = FP/ (FP + TN) \qquad (5)$$

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:

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

3.3.4 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 a correlation matrix and partly through the analysis of the relevant frequency table.

3.4 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.

3.4.1 The logit regression results

Model I investigates the relationship between the probability of default and traditional hard financial indicators. The results are reported in Table 4.

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)

VARIABLES	(1) default	(2) default	(3) default
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

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

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 Group 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.

Table 5: Income Distribution

Group	Monthly Income (yuan)	Freq.	Percent
1	⩽ 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	> 50000	2,378	16.32
	Total	14,569	100.00

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 Jimenez & 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: 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.

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 a negative correlation with the probability of default. We also found that borrowers with Weibo and mobile verification tend to borrow 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 borrow loans of lower amounts (average 19050.48 yuan) and over shorter terms (average 6.658986 months). This may be an indication of borrowers care about their social image; thus, they tend to borrow safer loans and potentially have a lower risk of default. 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 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 because 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 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.[22] This suggests that the combination of hard and soft information can better predict loan performance. It should also be noted 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.

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**

[22] For details, see Appendix E.

VARIABLES	(1) default
	(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)

VARIABLES	(1) default
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)

VARIABLES	(1) default
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

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Note. The numbers associated with the variable ‘income’ refer to income groups. The sample includes 7 income groups.

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.

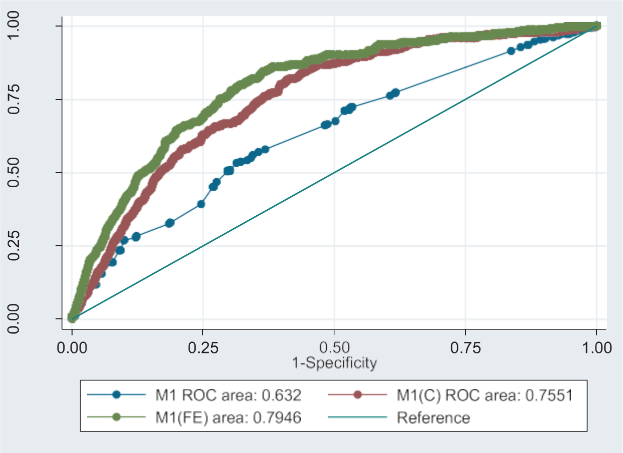
In order to increase the confidence level in our findings, we take additional steps and tests in the following section. We use the ROC curve technique to help in discriminating among the three models. In addition, we carried out various tests and data examinations to check for the robustness of our results.

3.4.2 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 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).

Figure 1: ROC Curves for Model I



Starting with Fig. 1, the AUC in model I is increasing with the addition of the control variables and increases 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.

Figure 2: ROC Curves for Model II

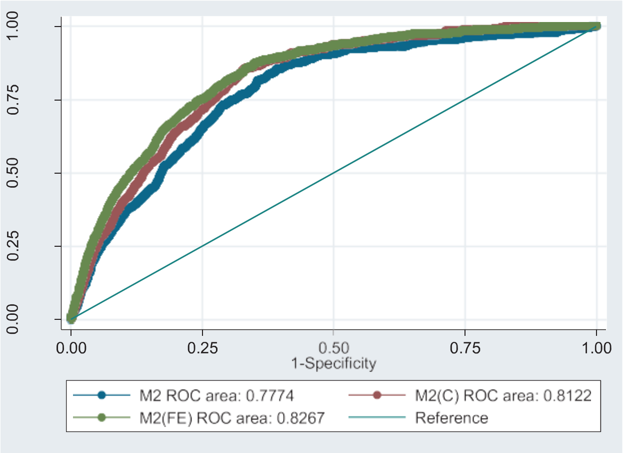
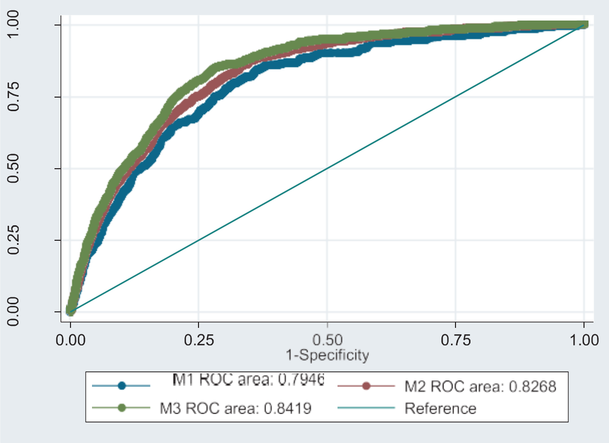


Figure 3: ROC Curves for Model Comparisons



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

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

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 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 the 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 the 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.[23] 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

[23] See Appendix C.

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.[24] 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 additional tests using modified instrumental variables.[25] 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

[24] 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).

[25] The respecifications 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.

(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 out 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 purposes had a small influence on loans default.

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.

3.5 Conclusion

This paper investigates the predictive power of soft and hard information on the loan performance in P2P lending. Our results of the predictive power of the hard information are consistent with the existing literature. We also add evidence to the literature (e.g., Jimenez & 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 unverified 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 behavior 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 assign penalty points for 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 the 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 Cornee (2017).[26]

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

[26] 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.

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. Banks and non-bank microfinance 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 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.

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.

Agarwal, S., & Hauswald, R. (2010). Distance and private information in lending. *Review of Financial Studies*, 23(7), 2757–2788.

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

Akerlof, G. A., & Kranton, R. E. (2000). Identity economics. *The Quarterly Journal of Economics,* 115(3), 715–753.

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

Barasinska, N., & Sch¨afer, 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*.

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.

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.

Berger, A. N., & Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organizational structure. *The economic journal*, 112 (477), F32–F53.

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.

Botsman, R. (2017). *Big data meets big brother as China moves to rate its citizens.* Wired UK. Available at: <https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion>

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.

Cao XB (2013) Measurement and the role of social capital in online P2P lending market," PhD Dissertation, June, 2014

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.

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

Corn´ee, S. (2017). The relevance of soft information for predicting small business credit default: Evidence from a social bank. *Journal of Small Business Management*.

Dell'Ariccia, G., & Marquez, R. (2004). Information and bank credit allocation. *Journal of Financial Economics*, 72(1), 185–214.

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.

Diamond, D. W. (1984). Financial intermediation and delegated monitoring. *Review of Economic Studies*, 51(3), 393–414.

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.

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.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Faia, E., & Paiella, M. (2019). Information and substitution in P2P markets. *CEPR DP*, *12235*.

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.

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

Garcia-Appendini, E. (2007, August). Soft information in small business lending. *In EFA 2007 Ljubljana Meetings Paper*.

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.

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.

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.

Greiner, M. E., & Wang, H. (2009). The role of social capital in people-to-people lending marketplaces. *ICIS 2009 proceedings*, 29.

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.

Herrero-Lopez, S. (2009). Social interactions in p2p lending. *The Workshop on Ssocial Network Mining & analysis* (pp. 1–8).

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

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.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. Management Science 62: 1554–77.

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.

Jim´enez, G., & Saurina, J. (2004). Collateral, type of lender and relationship banking as determinants of credit risk. *Journal of Banking & Finance*, 28(9), 2191–2212.

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.

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. DOI 10.3386/w25075

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.

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.

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*.

Order, R. V., & Zorn, P. (2000). Income, location and default: Some implications for community lending. *Real Estate Economics*, 28(3), 385–404.

Piliavin, J. A., & Charng, H.-W. (1990). Altruism: A review of recent theory and research. *Annual Review of Sociology*, 16(1), 27–65.

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.

Po¨tzsch, S., & Bo¨hme, 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.

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.

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.

Serrano-Cinca, C., Gutierrez-Nieto, B., & Lo´pez-Palacios, L. (2015). Determinants of default in p2p lending. *PloS one*, 10(10), e0139427.

Slavin, R. E., & Davis, N. (2006). *Educational psychology: Theory and practice (8th)*. Pearson/Allyn & Bacon. ISBN: 9781292020730

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

Stein, J. C. (2002). Information production and capital allocation: decentralized versus hierarchical firms. *The Journal of Finance*, 57(5), 1891–1921.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *American Economic Review*, 71(3), 393–410.

Theil, H. (1967). Economics and Information Theory, *Rand McNally and Company - Chicago*.

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.

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

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. 179p. ISBN 10: 1595586202

Appendix A. Chinese P2P Key Market Indicators

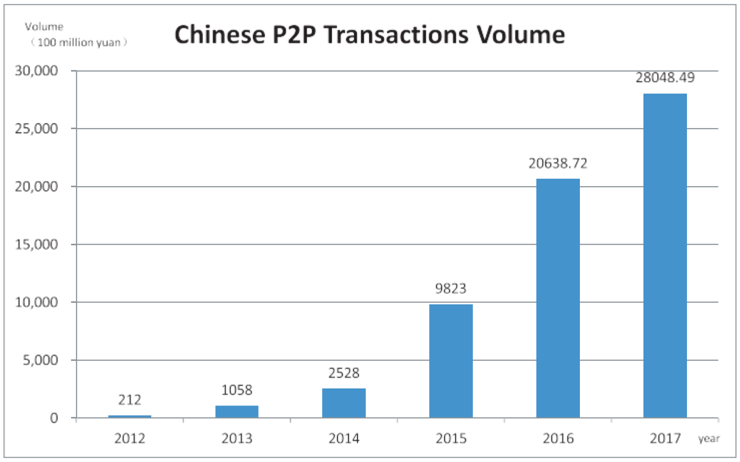


Figure A4: Chinese P2P Key Market Indicators

Source: Annual P2P Industrial Report , https://www.wdzj.com/

Appendix B. Collinearity Diagnostics

Table B9: 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- R2 are in column 4 and R2 figures for correlation between the given independent variable and the rest of the 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 C10: 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	6,491	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 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

Table D11: Sensitivity Tests Results for Model I

VARIABLES	(1) default	Test Results
Income verified	−0.263	−0.0157
	(0.226)	(0.241)

VARIABLES	(1) default	Test Results
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***

VARIABLES	(1) default	Test Results
	(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**	−0.913***
	(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 D12: 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***

VARIABLES	(1) default	Test Results
	(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)

VARIABLES	(1) default	Test Results
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 D13: 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

VARIABLES	(1) default	Test Results
	(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)

VARIABLES	(1) default	Test Results
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

VARIABLES	(1) default	Test Results
		(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. As another test of robustness, the following analysis of the purposes of loans and their properties was carried out. 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. The sample was divided into three classes of loans – loans for personal consumption (25.47% of the total), loans for business (37.26% of the total), and loans with no clear indication of the purpose of the loan. (37.26% of the total). The analysis focused on the structure of interest rates, maturity of loans, and the distribution of social capital indicators in the three groups of loans. 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 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 that were 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 end use. The average rates were: 13.46% for personal loans for consumption, 14.73% for loans for business purposes, 13.69% for undefined loans, and 13.99% for total sample (total sample= 687). The average rate for the whole sample was 13.99% (compared to 13.31% for the whole population of 14,575).

Figure E1: Kernel density Estimate for Mobile Verified and non-Verified

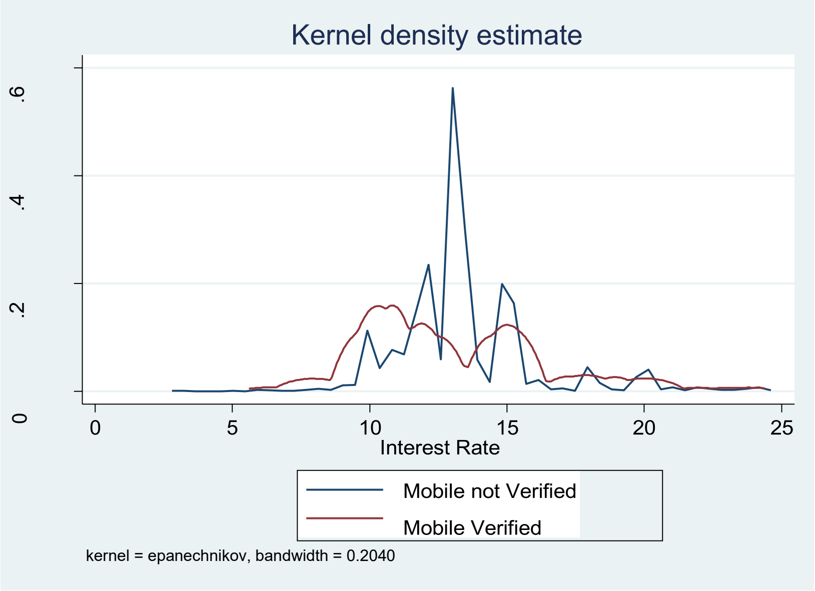


Figure E2: Kernel density Estimate for Weibo Verified and non-Verified

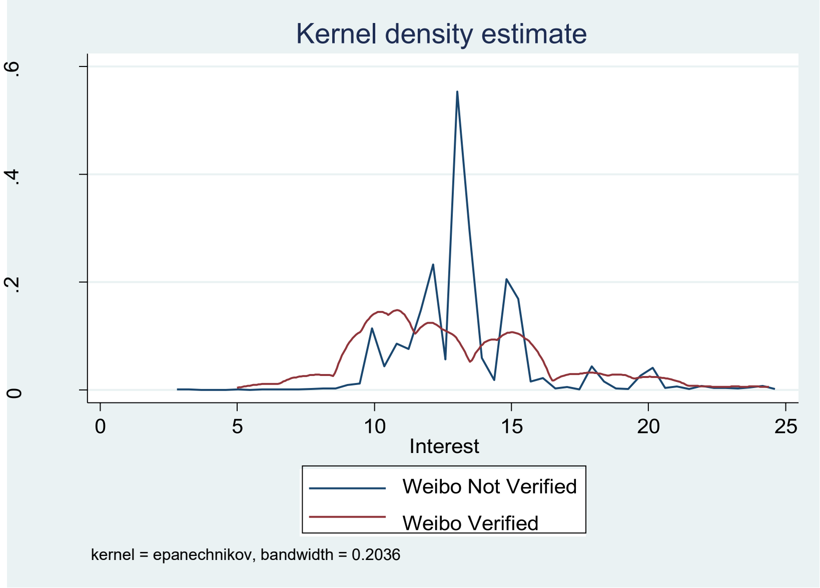
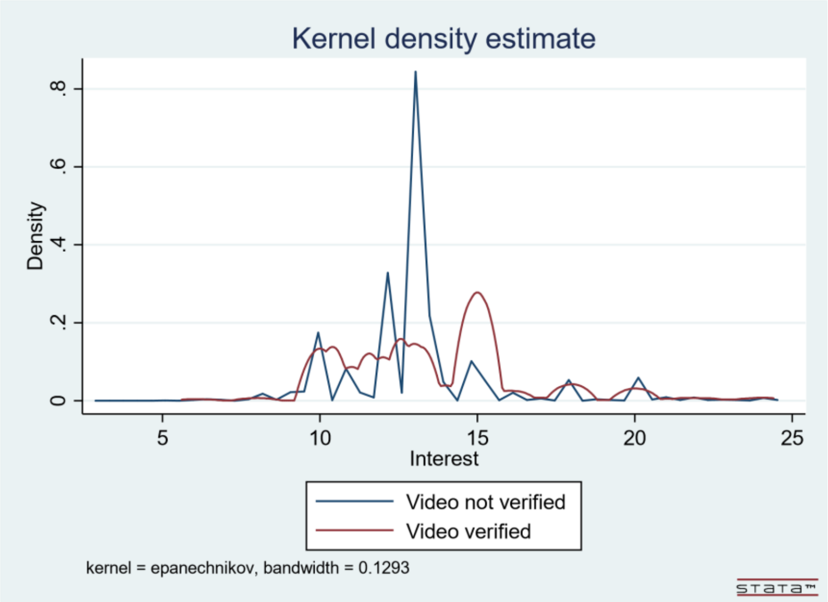


Figure E3: Kernel density Estimate for Video Verified and non-Verified



Term structure. The term structure was also almost identical for all three classes. This is not surprising since the term structure was uniquely short-term (all loans were with maturity of under 3 years).

Amounts of loans. The bulk of loans was for small amounts. The loan amounts were highly skewed to the lowest range starting from 10000 yuan to 30000 yuan (38%). Almost 75% of loans were below 50000 yuan.

Table E14: 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%

Amount (yuan)	Freq.	Percent
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%

Default rates. Default rates in the three classes of loans of our random sample were similar and as follows (in percent of the total sample of 683 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 asmall influence on loans, if any.

Social capital variables. The following tables provide cross-tabulations of data and the indications of correlations among different variables. Table E15 shows the extent to which social capital variables were used in processing the loan applications and the extent of the overlap. As the data below the line show, the extent of overlap was very small. For example, all three social capital variables equal to one in only 845 cases out of our sample of 14 575, i.e. 5.8%. The overlap is small and unlikely to lead to the conclusion that the overlap affects a particular class of loans, and, consequently, 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.

Table E15: Distribution of Social Capital 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 E16: Correlation Matrix of Variables

	Description	Age	Gender	Marriage	Educational	Mobileverified	Weiboverified	Videoverified	Incomeverified
Description	1.000								
Age	0.1865	1.000							
Gender	0.0970	0.0656	1.000						
Marriage	–0.2935	–0.0943	–0.0806	1.000					
Educational	–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	0.000	0.0692	1.000

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 contained 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. In sum, whatever differences in default rates existed, they were, therefore, unlikely to be due to different purposes of the loans.

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

Table F17: 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

4 Response to Opponents' Reports

I would like to thank all referees and members of the committee for their valuable comments and suggestions, which have helped to improve the quality of the thesis and inspired future research direction.

4.1 Connection to the Theoretical Model – Response to Prof. Karel Janda

Prof. Janda's comments concern the connection of our empirical research with theoretical models in the style of microeconomics of banking and information economics approach. It helped us define future research direction and provided insight for theoretical modeling in asymmetric information under the current big data age.

As noted by Freixas & Rochet (2008), a microeconomic theory of banking has evolved since the 1980s, primarily through a shift in emphasis from risk modeling to imperfect information modeling. This asymmetric information model is predicated on the assumption that different economic agents have different pieces of information on relevant economic variables and will use the information to their advantage. Based on Theil (1967)'s Information Theory and recent modelers' work from Ruckes (2004), Petriconi (2016), and Faia & Paiella (2019). We try to simulate the information premium model under the P2P lending environment and prove the importance of information precision in loan market efficiency.

We assume lenders are homogenous, risk-neutral investors. Borrowers are risk-neutral and seek funds for risky projects, whose success probabilities are heterogeneous and stochastically distributed.

Lender Side Model:

Random Return on P2P Loans:

$$R^i = \begin{cases} R^I & with\,probability\; p^i, \\ 0 & with\,probability\; 1 - p^i \end{cases}$$

where i denotes i^{th} projects, investors do not observe p^i, but form an expectation of such probability based on a signal, $\sigma_{i,q}$ that they may receive. We denote this estimated probability by

$$\pi^i = \varepsilon\left[p^i \middle| \sigma_{i,q}\right]$$

where ε_t denotes a Bayesian expectation.

Borrower Side Model:

Probability of funding a project for heterogenous risk-neutral borrowers according to a uniform density:

$$U\left[\bar{p} - \frac{\varepsilon}{2}, \bar{p} + \frac{\varepsilon}{2}\right]$$

where $\bar{p}$ denotes the unconditional mean.

Loan Pricing Model:

Based on Ruckes(2004) and Petriconi (2016), signal (soft and hard information) s_i can be summarized as:

$$\sigma_{i,q} = \begin{cases} s_i = p^i & with\, probability\ q, \\ s_i \sim U\left[\bar{p} - \frac{\varepsilon}{2}, \bar{p} + \frac{\varepsilon}{2}\right], & with\, probability(1-q) \end{cases}$$

with probability q, the signal conveys the project's true success probability. Probability q captures signals' precision. It follows that signals are distributed as a uniform:

$$\sigma_{i,q} \sim U\left[\bar{p} - \frac{q\varepsilon}{2}, \bar{p} + \frac{q\varepsilon}{2}\right]$$

Lenders received a signal and update their expectations:

$$\varepsilon\left\{p^i \middle| \sigma_{i,q}\right\} = \pi^i = q s_i + (1-q)\bar{p}$$

Thus, the expected return from each loan:

$$\varepsilon_t\left\{p^i \middle| \sigma_{i,q}\right\} R^I = \pi^i R^I$$

Information Premium Model:

The probability that a project will be funded under imperfect information:

$$\chi_q(\mu) = Pr\big(\varepsilon\big[p^i\big|\sigma_{i,q}\big] \leq \mu\big) = Pr\left(\sigma_{i,q} \leq \frac{\mu - (1-q)\bar{p}}{q}\right)$$

Since the distribution of the signals follows:

$$\sigma_{i,q} \sim U\left[\bar{p} - \frac{q\varepsilon}{2}, \bar{p} + \frac{q\varepsilon}{2}\right]$$

The mass of projects that won't be funded $X_q(\bar{\omega}) = 1 - \chi_q(\bar{\omega})$:

$$X_q(\mu)\begin{cases} 0 & if \mu \leq \bar{p} - \frac{q\varepsilon}{2} \\ \frac{1}{2} - \frac{\mu - \bar{p}}{q\varepsilon} & \bar{p} - \frac{q\varepsilon}{2} \leq \mu \leq \bar{p} + \frac{q\varepsilon}{2} \\ 1 & if \mu \geq \bar{p} + \frac{q\varepsilon}{2} \end{cases}$$

Based on Theil's (1967) index, information premium:

$$\Theta = \chi_q(\mu) - \chi_{q=1}(\mu)$$

As noted by Faia & Paiella (2019), $\chi_q(\mu)$ captures the amount of entropy among the funded loans under partial information for given signals. As signal precision increases, the dispersion or entropy widens, approaching the entropy under full information. Thus, we can conclude that with the increase of information precision q, information premium Θ decreases. Therefore, if more precise information is available, projects' dispersion under partial information approaches the dispersion under full information. This also matches the argument from a recent study by Yan et al (2015), which states that signaling costs can be reduced by the quality of the data and the quality of the analysis. The quality of the data includes four dimensions: Volume, Variety, Velocity, and Veracity. The quality of the analysis depends on the precision of the prediction model and the accuracy level of the machine learning algorithm.

Our research can serve as empirical evidence for these theoretical studies and emphasize the role of soft information as a new source of information and improve the precision of the information in credit analysis.

4.2 Policy Implications and Hypothesis Reference – Response to Prof. Ali M. Kutan

Prof. Kutan's comments helped us strengthen the policy implications of our findings and neaten the hypotheses' reference. I added the policy implications in the conclusion section in Chapter 1 and Chapter 2. I have repositioned the hypothesis in paper 1 and merged it with the literature review in Chapter 1.2 as suggested to help the readers better navigate the source of the hypothesis. I summarized the references for each hypothesis of paper 3 in Chapter 3.3.2. I also compiled the references for paper 3 below.

Empirical Evidence from Finance:

Based on Bertrand et al (2005), Liao et al (2015), Ravina (2012), Gonzalez & Loureiro (2014), Duarte et al. (2012), Greiner & Wang (2009), Liu et al. (2015), Cao (2013), Miu & Chen (2014), Lea et al (1995), Dorfleitner et al. (2016), Dell'Ariccia & Marquez (2004), Berger et al. (2005), Deyoung et al. (2008), Agarwal & Hauswald (2010), Berger & Udell (2002), Corn´ee (2017) and Ge et al. (2017), soft information can have predictive power in credit appraisal; thus we have the first hypothesis:

Hypothesis 1. Credit appraisal based on appropriately selected soft information can have predictive power in predicting default.

Based on Grunert, Norden, & Weber (2005), Godbillon-Camus & Godlewski (2005), Dorfleitner et al. (2016), and Agarwal et al. (2011), the combination of soft and hard information can achieve better credit rationing results, thus we derive the second hypothesis:

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.

Theoretical Support and Empirical Evidence from Psychology:

Gender: Piliavin & Charng (1990), Theory of Altruism, women are more likely to be altruistic than men and women can, therefore, be expected to be less likely to default on their loans.

Crown, & Spake (1997), empirical evidence for Social Role Theory, men and women have different perceptions of unethical behavior.

Lennon & Eisenberg (1987), men and women show differences in sympathy and empathy.

Marital Status: Chaulk, Johnson, & Bulcroft (2003), empirical evidence for Family Development and Prospect Theory, marriage has a significant negative relationship with risk tolerance.

Age: Roberts & Mroczek (2008), 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.

Education: Slavin & Davis (2006), Psychology in Cognitive Development Theory, as a branch of educational psychology, emphasizes, the point that people's understanding of morality changes with the development of education.

Theoretical reference from Behavioral Finance:

Akerlof & Kranton (2000), Identity Economics, by emphasizing the role of the identity of agents in their economic choices, make the point that economic decisions are not exclusively dependent on monetary incentives. Identity, a person's sense of self, also affects economic outcomes.

Theoretical reference from Information Theory:

Faia & Paiella (2019), development of Theil (1967)'s Information Theory, as more precise information is available, projects' dispersion under partial information approaches the dispersion under full information.

4.3 Definitions Clarification and Test Statistics – Response to Prof. Josef Brada

I am sincerely grateful for Prof. Brada's detailed and in-depth suggestions for further improvements. It helped greatly in improving the clarity and robustness of the papers.

4.3.1 Definition clarifications

Chapter 1 SMEs definition - Section 1.3.3. Countries may define SME in different ways, by sales, employment, etc. Does the survey account for this? What is the definition of SME — is it the same for all countries? The author chooses 100 workers; why? Also, there may be different approaches to surveying very small firms as well as medium sized firms. How does the survey address that? The reader should not have to go to the original survey to find this information.

Definitions of SMEs often vary in countries, but a unified methodology of the World Bank Enterprises Survey has been used in this study. The World Bank Enterprise Survey (WBES) classifies enterprises with less than 20 employees as small size and those with 20-99 as medium size. In the dataset, it has a category variable "size" and it divides the size into three categories: small (number of employees <20), medium (number of employees 20~99), and large (number of employees 100 and over 100). I added this explanation in section 1.3.3. While using a common definition is useful in international comparison, it is recognized that this may not take fully into account differences in labor intensity across sectors and segments of markets, as well as informal employment in some countries.

Chapter 1 - In the literature survey section, the author talks about firm "performance" or, equally, "growth". But it is unclear what is meant by this — is it growth of sales, of employment, of profits, profitability as measured by ROI? All of them? Any of them? Why is one more important or appropriate than the other? Clarifying this is key for the paper. Employment is the generally used measure of growth, but the author should discuss the potential problems with this measure.

Thank you very much for this comment. I added the explanation in section 1.3.3. It is the growth of employment that defines high-growth firms. It is based on Lee (2014) and OECD (2010)'s definition of growth. And also in our dataset, the size is defined by the number of employees, the size of the firm has been categorized into three groups: small (number of employees <20), medium (number of employees 20~99), and large (number of employees 100 and over 100). Organization for Economic Co-operation and Development (OECD) has provided the definition of high growth firms as those which achieved a 20% employment growth/annum for 3 consecutive years. Moreover, one of our main aims in the paper is to help developing countries in dealing with their acute problem of unemployment and under-employment. More importantly, we do not have the financial statistics which can be used to define the growth. Thus, we choose to use employees' numbers as the proxy of size, and also the growth in the number of employees as the proxy for the growth of the enterprises. However, we should be aware of different methodologies of measuring employment in different countries when using this method. Especially in developing countries, the existence of informal employment increases the statistical difficulty.

*Chapter 1 - How was the cutoff for high growth firms chosen? The text is confusing, perhaps because of language problems. The author writes "(OECD) has provided the definition of high growth firms as those which achieved a 20% employment growth within 3 years." I am not able to understand this — within is not the same as over. Within what period? Why only for the last 3 years of the survey period? In any case, the OECD does not seem to say that the firm has achieved 20% growth each year over a three-year period (1.2**3) = 1.78 —the sentence as written says it has increased employees by 20% over a three year period. It is a question of what the OECD says and how the author has put it into English. Moreover, it would be good to give the date of the survey and the years over which firm growth is calculated.*

Thank you very much for the comments. I have changed the wording in section 1.3.3 and also added the reference paper in the bibliography. The proxy method is chosen based on Lee (2014) and OECD's definition of high-growth firms. The OECD defines high-growth firms as those achieving 20 % employment growth/annum for 3 consecutive years (OECD 2010). The reason for using 120%^3 as the proxy is that we only have two variables that can be used to define the growth in our dataset. One is the "number of permanent full-time employees of this firm at end of last fiscal year", the other is "number of permanent full-time employees of this firm at end of 3 fiscal years ago". We do not have the data for the second year. Thus, I used 1.728 as the proxy to define high-growth firms. I added more explanation to this when defining high-growth firms in section 1.3.3.

Chapter 1 - Why not just use the growth rate of employment as a variable instead? If its coefficient is positive, then it should also be positive for the dummy and the continuous variable could yield better statistical results. What do small, large, very large mean in numbers? The categorization has an implicit assumption — namely that the effect of, say, growth rates for firms with a growth rate of more that 1.78 is the same no matter whether the firm grows at 20% a year of 50% per year. Also, it means that there is a significant difference between a firm that grows at 20% a year and one that grows at 19% that is as large as the difference between a firm that grows 20% and one that grows 1% per year.

Thank you for the suggestion of using the growth rate instead. The reason for using 120%^3 as the proxy is that we only have two variables that can be used to define the growth of firms in our dataset. One is the "number of permanent full-time employees of this firm at end of last fiscal year" and "number of permanent full-time employees of this firm at end of 3 fiscal years ago". We do not have the data for the second year. It is a survey data based on questionnaire answers from enterprise owners and managers. So there is no detailed financial statement available for checking the operational status of the company on yearly basis. And based on the literature, high-growth firms, this specific group of firms, have different needs for funds. We cannot ignore this variable and we want to analyze the problems faced by this group of firms instead of analyzing the relationship between growth rate and the difficulty in getting funds. Thus, we used 72.8% growth in three years as the proxy for high-growth firms.

Chapter 1 - Table 3. I assume "big firms" eliminated were those with over 100 employees? This needs to be stated before the Table is presented.

Yes, "big firms" eliminated were those with over 100 employees. I added the explanation in 1.5 Results Section.

Chapter 3 - What does "predicting a default" mean? If the data set uses information only on loans that have already defaulted, this is a problem because it means we are missing data from our sample of borrowers who have received a loan, may currently be paying on the loan but will default sometime in the future. So, we only have data on loans where the borrower defaulted "quickly". If I am wrong on this, please explain why. Footnote 17 does not really address this problem fully.

The borrowers which are currently paying the loans are excluded from the dataset. We only included the loans that have finished the repayment process, meaning the loans that already have a "result". So, there are only loans that are "repaid" or "defaulted". All other loans are excluded like: 1. Just got funded and didn't start the repayment process; 2. Loans that started the repayment process but haven't finished. I added more explanation in Footnote 17.

Chapter 2 - I am bit troubled by the interpretation of the main result that lenders are making type II errors. This is based on the marginal effects. The author points to some effects that are positive or negative — but the issue is not the difference between the effect and zero but between one effect and the other. So, say for one income level the effect is 0.005 and for another it is something else like -0.001. The real question is not whether one of these is significantly different from zero at 5% but whether the two coefficients are 2 standard deviations different from each other. If the SE of one variable is large, this may not be the case even if one coefficient is very significant. Here we are less interested in whether either coefficient is different from zero and *much more interested in whether one coefficient is different from the other at any real level of significance.*

We determined the type II error by the phenomenon that investors were predisposed to making inaccurate diagnoses of signals and gravitated to borrowers with low creditworthiness while inadvertently screening out their counterparts with high creditworthiness. We identify it by the wrong sign for the same variable in the funding model and the default predicting model. Taking the case of mortgage loans as an example, our results show that having a mortgage loan is a signal that has a significant negative relationship with default, meaning that borrowers with mortgage loans turn out to be less likely to default than those without mortgage loans. This suggests that borrowers with mortgage loans generally have higher creditworthiness. However, the lenders on the platform are unable to diagnose this signal correctly and prefer to lend to those without mortgage loans. There is a significant negative relationship between the mortgage loan and the probability of funding. This is probably due to the fact that borrowers with mortgage loans are creditworthy clients for banks which have passed the screening mechanism of professional creditors. Moreover, due to the mortgage loans in the banks, those borrowers care more about their credit history, and as a result, they tend to default less. However, the lenders on the platform misdiagnose this signal and are less willing to lend to them. The possible reason for this is that lenders on the platform treat mortgage loans as a signal of debt; consequently, they treat those borrowers as high-risk and tend to lend to borrowers without mortgage loans. Another explanation could be that these high creditworthy borrowers offer "lower" interest rates, which become less attractive to the lenders on the P2P platform. The misdiagnosis of the credit signal is what we defined as the TYPE II error in the credit appraisal process.

Another point that needs to be clarified is that our results are not marginal effects but the original coefficient of logit regression results. Since the coefficient represents changes in log odds, we only refer to the signs (positive/ negative), as we care more about whether the lenders can correctly diagnose the effect of the signals of the borrowers, e.g., whether they know that higher education indicates less risk of default. Hence we care more about the direction of the relationship between the dependent variable and the independent variable than to what extent the signals can increase/decrease the probability of default or getting funded.

Due to the fact that all the observations in the default model are successfully funded, we cannot run the multivariate logit regression. To test the robustness of the result and, following Menard (2011), I also tried to standardize the coefficient of the logit regression to be able to compare the two coefficients in two models. The fully standardized results are as below in Table 1 and Table 2. The sign of the results is not changing for both funding model and default predicting model, thus the comparison of the coefficient sign between the two models remains consistent to our original results.

Table 1 Standardized Coefficient for Funding Model

	b	z	P>z	bStdXY
1.Incomeverified	2.832	45.025	0.000	0.241
Income				
1	-0.668	-6.332	0.000	-0.014
2	-1.660	-20.223	0.000	-0.081
3	-0.394	-20.640	0.000	-0.053
5	0.155	6.687	0.000	0.015
6	0.382	13.535	0.000	0.030
7	0.475	14.703	0.000	0.033
1.Incomeverified# Income2	1.137	1.540	0.123	0.003
1.Incomeverified# Income3	0.434	4.806	0.000	0.019
1.Incomeverified# Income5	-0.308	-2.911	0.004	-0.012
1.Incomeverified# Income6	-0.606	-5.234	0.000	-0.020
1.Incomeverified# Income7	-1.172	-9.980	0.000	-0.032
Carverified	0.448	10.171	0.000	0.026
1.Houseverified	0.080	1.503	0.133	0.005
1.Mortgageloan	-0.311	-13.461	0.000	-0.031
1.Houseverified# Mortgageloan	0.240	3.081	0.002	0.010
Description	0.013	144.384	0.000	0.363
Age	0.065	63.422	0.000	0.140
Gender	0.275	14.976	0.000	0.028
Marriage	0.344	20.606	0.000	0.047
Education	0.076	17.308	0.000	0.037
Mobileverified	-0.515	-11.917	0.000	-0.032

Weiboverified	0.605	12.290	0.000	0.031
Videoverified	2.522	59.636	0.000	0.147
Interest	-0.304	-86.496	0.000	-0.299
Amount	-0.304	-37.186	0.000	-0.113
Term	0.113	121.342	0.000	0.339
Constant	-2.150	-19.526	0.000	.

b = raw coefficient
z = z-score for test of b=0
bStdXY = fully standardized coefficient

Table 2 Standardized Coefficient for Default Predicting Model

	b	z	P>z	bStdXY
1.Incomeverified	0.596	2.564	0.010	0.080
Income				
1	-0.874	-0.805	0.421	-0.024
2	-0.604	-1.754	0.079	-0.040
3	-0.168	-1.258	0.209	-0.036
5	-0.360	-2.145	0.032	-0.057
6	0.233	1.574	0.115	0.038
7	0.261	1.678	0.093	0.044
1.Incomeverified#Income2	2.803	3.177	0.001	0.021
1.Incomeverified#Income3	0.471	1.433	0.152	0.032
1.Incomeverified#Income5	-1.156	-1.992	0.046	-0.059
1.Incomeverified#Income6	-1.744	-2.987	0.003	-0.092
1.Incomeverified#Income7	-2.233	-3.871	0.000	-0.138
Carverified	-0.394	-3.580	0.000	-0.081
1.Houseverified	0.348	2.859	0.004	0.070
1.Mortgageloan	-0.409	-1.897	0.058	-0.071
1.Houseverified#Mortgage	-0.178	-0.647	0.518	-0.025
Description	-0.006	-10.992	0.000	-0.267
Age	-0.005	-0.850	0.395	-0.020
Gender	-0.274	-2.123	0.034	-0.049
Marriage	-0.203	-1.941	0.052	-0.040
Education	-0.120	-7.174	0.000	-0.140
Mobileverified	-0.486	-3.698	0.000	-0.086
Weiboverified	-0.627	-4.146	0.000	-0.109
Videoverified	1.007	8.363	0.000	0.225
Interest	0.195	14.076	0.000	0.234
Amount	0.035	0.773	0.440	0.021
Term	0.012	2.026	0.043	0.052
Constant	-3.159	-5.239	0.000	.

b = raw coefficient
z = z-score for test of b=0
bStdXY = fully standardized coefficient

Chapter 3 - The author writes: "appropriately selected soft information can have strong predictive power: i.e., soft information coefficients are significantly non-zero". Having significantly non-zero coefficients is no guarantee of "strong predictive power". The author should consult on how one might want to evaluate when "strong predictive power" exists.

The predictive power conclusion is coming from the ROC and the goodness fit of Model II and Model III, which is soft information solely model and the combination of hard and soft information respectively. I adjust the wording of the hypothesis.

Chapter 3 - Why does the author use dummies for different income levels rather than reported income? This just throws out information and weakens the regression.

Because the raw data of income is a category variable. We don't have the exact income of the borrower. Borrowers choose in which category of the income level they belong to. And category variable cannot be used as continuous. So, we used dummies to treat category variable income.

4.3.2 Summary of Statistics

Chapter 1 - It would be useful, one could almost say necessary, to see a Table of Summary Statistics. It is somewhat difficult to interpret Table 2 without seeing the summary statistics. Also, while usual measures like R-squared are not useful for probit regressions, it would help is the so-called McFadden (or pseudo) R2 were reported. The author might want to report the percentage of correct predictions by firm size for small, medium and large firms to see which group is driving the significant (or insignificant) marginal effects or if the results apply to all firms regardless of size. How should we interpret lack of significance of marginal effects for other "barriers"?

Tax and electricity probably are general problems all firms are facing in developing countries; thus, the test results are not significant for SMEs. I added the explanation to results section 1.5. I added the Pseudo R2 in the results. I also added the summary of variables in Annex 2.

Chapter 1 - Please clarify language on fixed effects/dummies.

The fixed effects are for country and sector. I added the sectors and countries list to Annex 3 and Annex 4.

Chapter 2 - Was there any effort to clean the data and eliminate outliers? For example, in the table of Summary Statistics, I see that the age of borrowers ranges from 1 to 86. I find it hard to believe that a 1-year-old is posting on the web for P2P loans. Also, please use the same number of decimal points for a variable in giving the mean and the max and min.

The task at hand is to test whether lenders on the platform can correctly diagnose the information provided by the borrowers. Since most of the information on the platform is not verified and cannot be verified now. The raw data provided by the borrowers can be used to test whether the lenders can distinguish the fake information. For example, by comparing the default model and funding model results, we can infer that some people probably lied about their income as verified income is more indicative in predicting default than unverified income. However, lenders on the platform seem to have ignored it and trusted the unverified income. If we intervene and eliminate the possible fake information, we can not spot this kind of misdiagnosis. This is also one of the problems we are worried about online lending because financial-related information is very hard to verify and lenders probably lack the skills to recognize fake information. This is also the reason for exploring the function of soft information in credit analysis because people are less likely to lie about soft information.

As shown in Table 1, 82 observations are below 18 years old. As a robustness check, I deleted the 82 observations and tested the results of the funding model. It shows almost no difference compared with the original results as shown in Table 2. As for the default model dataset, the age range is from 21 to 72 as shown in Chapter 2 summary statistics. It also reflects that lenders have the ability to correctly diagnose the credit signal age and didn’t lend to borrowers younger than 18 and older than 72. The decimal points of the mean and the max and min have been changed.

Table 1 Age Distribution for Funding Model

Age	Freq.	Percent	Cum.
1	2	0.00	0.00
2	1	0.00	0.00
13	1	0.00	0.00
14	1	0.00	0.00
15	10	0.00	0.01
16	11	0.00	0.01
17	56	0.02	0.03
18	186	0.07	0.11
19	636	0.25	0.36
20	1462	0.58	0.94
21	3356	1.33	2.27
22	7975	3.17	5.44
23	12170	4.83	10.27
24	16767	6.66	16.93
25	17543	6.97	23.89
26	18187	7.22	31.12
27	18867	7.49	38.61
28	16253	6.45	45.06
29	14384	5.71	50.77
30	12930	5.13	55.91
31	12822	5.09	61.00
32	12963	5.15	66.15
33	9263	3.68	69.82
34	8225	3.27	73.09
35	7747	3.08	76.17
36	6321	2.51	78.68
37	5346	2.12	80.80
38	4906	1.95	82.75
39	4330	1.72	84.47
40	4252	1.69	86.15
41	4089	1.62	87.78
42	3884	1.54	89.32
43	3561	1.41	90.73
44	3672	1.46	92.19
45	3164	1.26	93.45
46	2488	0.99	94.44
47	1827	0.73	95.16
48	1813	0.72	95.88
49	1910	0.76	96.64
50	1861	0.74	97.38
51	1953	0.78	98.15
52	997	0.40	98.55
53	706	0.28	98.83
54	748	0.30	99.13
55	533	0.21	99.34
56	498	0.20	99.54
57	398	0.16	99.70
58	319	0.13	99.82

59	190	0.08	99.90
60	117	0.05	99.94
61	46	0.02	99.96
62	43	0.02	99.98
63	18	0.01	99.99
64	7	0.00	99.99
65	12	0.00	99.99
66	1	0.00	99.99
67	1	0.00	99.99
68	4	0.00	100.0
69	2	0.00	100.0
70	1	0.00	100.0
72	1	0.00	100.0
73	4	0.00	100.0
86	1	0.00	100.0
Total	251842	100.00	

Table 4.2 Comparison of regression before and after dropping the age below 18

	(1)	(1)
VARIABLES	Funded	Funded
	(After)	(Before)
1.incomeverified		
	2.832***	2.832 ***
	(0.0629)	(0.0629)
1.income	-0.667***	−0.668 ***
	(0.106)	(0.105)
2.income	-1.659***	−1.660 ***
	(0.0821)	(0.0821)
3.income	-0.394***	−0.394 ***
	(0.0191)	(0.0191)
5.income	0.156***	0.155 ***
	(0.0232)	(0.0232)
6.income	0.382***	0.382 ***
	(0.0282)	(0.0282)
7.income	0.476***	0.475 ***
	(0.0323)	(0.0323)
1o.incomeverified#1o.income	0	0
	(0)	(0)
1.incomeverified#2.income	1.136	1.136
	(0.738)	(0.738)
1.incomeverified#3.income	0.434***	0.434 ***
	(0.0903)	(0.0903)
1.incomeverified#5.income	-0.308***	−0.308***
	(0.106)	(0.106)
1.incomeverified#6.income	-0.605***	−0.606 ***
	(0.116)	(0.116)
1.incomeverified#7.income	-1.173***	−1.172 ***
	(0.117)	(0.117)
carverified	0.448***	0.448 ***

	(0.0440)	(0.0440)
1.houseverified	0.0795	0.0795
	(0.0529)	(0.0529)
1.mortgageloan	-0.311***	−0.311 ***
	(0.0231)	(0.0231)
1.houseverified#1.mortgageloan	0.240***	0.240 ***
	(0.0779)	(0.0779)
description	0.0130***	0.0130 ***
	(9.02e-05)	(9.02×10^{-5})
age	0.0653***	0.0653 ***
	(0.00103)	(0.00103)
gender	0.275***	0.274 ***
	(0.0183)	(0.0183)
marriage	0.344***	0.345 ***
	(0.0167)	(0.0167)
educational	0.0763***	0.0763 ***
	(0.00441)	(0.00441)
mobileverified	-0.515***	−0.515 ***
	(0.0432)	(0.0432)
weiboverified	0.604***	0.605 ***
	(0.0492)	(0.0492)
videoverified	2.522***	2.522 ***
	(0.0423)	(0.0423)
interest	-0.304***	−0.304 ***
	(0.00352)	(0.00352)
amount	-0.304***	−0.304 ***
	(0.00818)	(0.00817)
term	0.113***	0.113 ***
	(0.000935)	(0.000935)
Constant	-2.148***	−2.150 ***
	(0.110)	(0.110)
	0.5883	0.5883
Observations	222,397	222,437

4.3.3 More Literature Support

. Chapter 1 - Perhaps it would help the reader if some mention were made of studies, both theoretical and empirical, of the particular problems facing SMEs and not just borrower firms in general in obtaining credit. That access to external finance is a key barrier to SME growth has been well established in the literature since the early 1950s. The author's claim to novelty for this study is that it covers many developing countries, in contrast to earlier studies that had more limited sectoral or country coverage. While that may be

true, since the hypothesis that SMEs are financially constrained is based on general theoretical principles, the author should explain more clearly why using more countries is necessary or useful. The author states, for example, that Lee's results for the UK "cannot be generalized to other regions". If the hypothesis is based on a robust theory and it is verified by studies of several regions, why should it have to be verified for every country in the world? After all, we have a robust theory that higher prices will reduce the quantity demanded. Is it necessary to prove that this holds for every country? Some effort should be made to buttress the novelty of the findings.

Developed and developing countries differ considerably in terms of their economic development and the level of industrialization. There are numerous reasons for these differences, but the emphasis should be placed on, in particular, differences in the effectiveness of the economic governance, corruption, the range and depth of the financial sector, large-scale unemployment or underemployment, poverty, child marriage, social and political disorders, illiteracy, to name just a few. Many of these factors became the elements of various theories of underdevelopment such as Prebisch's theory of dependency and theories elucidating colonialism, many of which were integrated with the writings of Paul Baran, Samir Amir, Andre Gunter Frank, Paul Streeten, Sanjaya Lall and others. The intellectual contributions of those writers have contributed to different approaches in policies toward developing countries. For example, the World Bank is providing loans to poor countries with different grant elements. Moreover, least developed countries may get grants instead of loans. Thus, as such, differentiating between developed and developing countries is evidently sensible in our attempts to generalize problems faced by SMEs. I added more explanations to address this issue in section 1.3.2.

Our aim of the paper is to analyze the biggest obstacles to the growth of SMEs with a particular focus on access to finance since it is the top listed obstacle according to the survey. I went through the literature and tried to expand the coverage of the specific problems SMEs faced. Besides the obstacles listed in our paper, literature concerning barriers generated by the technology adoption (Kapurubandara & Lawson, 2006; Kapurubandara, 2009; Kapurubandara & Lawson, 2008; Muriithi, 2017) among SMEs in developing countries, especially e-commerce related technology (D Al-Tayyar, et al., 2021 and Das et al., 2020) has formed a new trend of research direction. These could be interesting future research directions when data is available.

Chapter 2 - The specifications for getting a loan and defaulting on it are more or less the same. What are the implications of this? Is there a selection problem at work in the econometric sense because any borrower who defaults has already been selected to receive a loan (James J. Heckman, Econometrica, v ol. 47, No. 1 (Jan., 1979) on the basis of more or less the same criteria? But the estimates for getting a loan have greater predictive power. This suggests that predicting default is harder, perhaps not just for P2P lenders but also for other lenders (banks, buyers of corporate bonds, etc.) in differing circumstances. There is a huge literature on predicting bankruptcy and some of it could be used to answer the question (at least) of whether these predictive models are more accurate than the heuristic approach used by P2P lenders.

Literature on predicting default suggests that predicting default is harder due to the fact that the information that predicts default is hard to verify. The quality of information matters to the accuracy of the default predicting model. Rajan et al. (2015), in their paper, “The Failure of Models That Predict Failure: Distance, Incentives and Defaults”, using data on securitized subprime mortgages issued between 1997 and 2006, demonstrated that, as the degree of securitization increases, interest rates on new loans rely increasingly on hard information about borrowers. As a result, a statistical default model fitted during a low securitization period fails in a systematic way during a high securitization period: it underpredicts defaults among borrowers for whom soft information is more valuable in default prediction. These findings were rationalized in a theoretical model that highlights a decrease in lenders' incentives to collect soft information as securitization becomes more common, resulting in worse loans being issued to borrowers with similar hard information characteristics.

Since soft information about borrowers is unverifiable to a third party (as stated in Stein, 2002), lenders choose not to collect soft information about borrowers as distance increases. As a result, the set of borrowers who receive loans changes fundamentally as the securitization regime changes among borrowers with similar hard information characteristics. Consequently, the quality of predictions from default models that use parameters estimated using data from a period when a small proportion of loans are securitized breaks down. A more recent study from Al-Qerem et al. (2020) also emphasizes the importance of data quality and machine learning algorithms in the default prediction in their paper "Default Prediction Model: The Significant Role of Data Engineering in the Quality of Outcomes".

4.3.4 Purpose of the Loans – Response to Dr. Magda Pečená

I sincerely thank Dr. Pečená's effective comments and suggestions. They helped me increase the clarity of the paper and lead to interesting future research directions.

Loans provided on the platform are used both for personal and business purposes. According to an interview with the CEO of RenrenDai.com, 70–80 percent of loans granted are for freelancer or micro business operational cash flow purposes. Other common purposes include car loans, home renovation, and consumption. Unfortunately, the platform does not provide direct information about the loan's purpose. Some of the applicants disclose the purpose in the loan description. However, it is not mandatory, the information that can be used to define the purposes is inadequate. The borrowers make the application as an individual and use personal credit for the loans. Thus, we analyze the creditworthiness of the applicant based on individual credit information. We tried to manually classify the loans by reading the loan description for a small random selected sample to check the impact of the loan purpose in Appendix E and the results show a small influence on default. I added more explanation in Subchapter 3.3.1 under the data description section.

Microbusiness owners turn to P2P lending platforms as the last resort for funding reflects inefficiencies in business lending by traditional banks.

Financial innovation on new lending schemes and credit analyzing models for business loans deserves researchers' and entrepreneurs' additional effort.

As for the role of the collateral, fixed assets ownership is only an indicator of solvency and is not served as the collateral when the borrower is in default. Thus it cannot be used to repay the loans in case of default. The loans on the platform are pure credit lending. I added more explanation in chapter 2.4 when analyzing the result of fixed assets ownership.

4.3.5 Future Research Guide

Based on the comments from referees and the committee, we found interesting future research directions. The first possible direction is to derive a theoretical model for asymmetric information under the current technology environment with advanced big data engineering techniques and a mass of data from different segments. Secondly, the determinants of the financial constraints for SMEs are highly overlapped with the determinants used by banks to rate the creditworthiness of SMEs. This forms a dilemma for SMEs financing, as SMEs that need funds the most are those that are less favored by the banks. And lack of funds in turn results in obstacles to SMEs' growth. Thus, the question would be how to develop new credit models or find substitutional credit determinants for SMEs' credit rating in bank lending. The usage of soft information in the context of enterprise credit rating should be the subject of future research drawing on advances in psychology, sociology, and machine learning. Thirdly, as informally, SME owners turn to microfinance and borrow funds from the P2P lending market, lenders and P2P overseers will have to pay closer attention to distinguish between individual borrowers and enterprises borrowers. Financial innovation for business loans also deserves researchers' and entrepreneurs' additional effort. Fourthly, SMEs face a range of obstacles to growth as we have argued and shown. Many of these have been long-standing but new obstacles emerge as markets change and grow. For example, technology adoption, especially for e-commerce related technology, seems to have become a new obstacle for SMEs these days. New research exploring solutions for these problems will be needed. Finally, a more detailed division of geographic sections in the World Bank Enterprises Survey may lead to interesting findings concerning SMEs development in different groups of developing countries and is worth further research.

References

Al-Qerem, A., Al-Naymat, G., Alhasan, M., & Al-Debei, M. M. (2020). Default prediction model: the significant role of data engineering in the quality of outcomes. *Int. Arab J. Inf. Technol.*, *17*(4A), 635-644.

Chibelushi, C., & Costello, P. (2009). Challenges facing W. Midlands ICT-oriented SMEs. *Journal of Small Business and Enterprise Development, 16*(2), 210-239

Faia, E., & Paiella, M. (2019). Information and substitution in P2P markets. *CEPR DP*, *12235*.

Freixas, X., & Rochet, J. C. (2008). *Microeconomics of banking*. MIT press.

Das, S., Kundu, A., & Bhattacharya, A. (2020). Technology adaptation and survival of SMEs: A longitudinal study of developing countries. *Technology Innovation Management Review*, *10*(6).

D AL-TAYYAR, R. S., Abdullah, A. R. B., Abd Rahman, A., & Ali, M. H. (2021). Challenges and obstacles facing SMEs in the adoption of e-commerce in developing countries; A case of Saudi Arabia. *Studies of Applied Economics*, *39*(4).

Kapurubandara, M., & Lawson, R. (2006). Barriers to Adopting ICT and e-commerce with SMEs in developing countries: an Exploratory study in Sri Lanka. *University of Western Sydney, Australia*, 82(1), 2005-2016.

Kapurubandara, M., & Lawson, R. (2008). Availability of e-commerce support for SMEs in developing countries. *ICTer*, *1*(1).

Kapurubandara, M. (2009). A Framework to e‐Transform sMEs in developing countries. *The electronic Journal of Information Systems in developing countries, 39*(1), 1-24.

Menard, S. (2011). Standards for standardized logistic regression coefficients. *Social Forces*, *89*(4), 1409-1428.

Muriithi, S. M. (2017). African small and medium enterprises (SMEs) contributions, challenges and solutions. *European Journal of Research and Reflection in Management Sciences*, *5*(1), 36-48

Petriconi, S. (2016). Bank Competition, Information Choice and Inecient Lending Booms.“ *Mimeo, Bocconi university*.

Rajan, Uday, Amit Seru, and Vikrant Vig. "The failure of models that predict failure: Distance, incentives, and defaults." *Journal of financial economics* 115.2 (2015): 237-260.

Ruckes, M. (2004). Bank competition and credit standards." *The Review of Financial Studies*,*17*(4):1073-1102.

Theil, H. (1967). Economics and Information Theory, *Rand McNally and Company - Chicago*.

Yan, J., Yu, W., & Zhao, J. L. (2015). How signaling and search costs affect information asymmetry in P2P lending: the economics of big data. *Financial Innovation, 1*(1), 1-11.