Borrower's Self-Disclosure of Social Media Information in P2P Lending

Ruyi Ge
Shanghai Business School
gery@sbs.edu.cn

Bin Gu
Arizona State University
bin.gu@asu.edu

Juan Feng
City University of Hong Kong
juan.feng@cityu.edu.hk

Abstract

In peer-to-peer (P2P) lending, soft information, such as borrowers' facial features, textual descriptions of loan applications and so on, are regarded as potential signals to screen borrowers. In this study, we examine the signaling effect of a new category of soft information- social media information. Leveraging a unique dataset that combines loan data from a large P2P lending company with social media presence data from a popular social media site, and two natural experiments, we find two forms of social media information that act as signals of borrowers' creditworthiness. First, borrowers' choice to self-disclose their social media account is a predictor of their default probability. Second, borrowers' social media presence, such as their social network and social media engagement, are also predictors of default probability. This study proffers new insights for the screening process in P2P lending and novel usage of social media information.

1. Introduction

Peer-to-peer lending, also known as P2P lending, is the practice of lenders lending money to unrelated individuals without going through a traditional financial intermediary. Instead, the transactions are intermediated by P2P lending platforms, which provide online venues for lenders and borrowers to communicate and transact. The first P2P lending company, Zopa, was founded in UK in February 2005. Afterwards, dozens of imitators emerge across the world. By the year of 2015, there are thousands of P2P lending companies worldwide, and the loans funded on the biggest P2P lending platform, LendingClub, have reached $7 billion.

Peer-to-peer loans are unsecured personal loans. Lacking effective screening methods on small borrowers' creditworthiness, traditional financial institutions tend to do very little screening for small borrowers and rely excessively on collaterals [1-4]. However, in P2P lending markets, borrowers do not provide collateral as a protection to lenders against default. This practice, on the one hand, makes P2P lending particularly attractive for small borrowers who might otherwise turn to pay day lenders or credit card debt [5], and on the other hand, makes it very challenging for non-expert lenders who have to identify credible borrowers and assess default risk by themselves.

Essentially, the financial risk in loan markets is caused by information asymmetry between lenders and borrowers. In order to alleviate information asymmetry, P2P lending platforms encourage borrowers to submit as much relevant information as possible. The borrower information can be divided into two categories: standard "hard" information, which directly reflects borrowers' financial status or creditworthiness, such as credit score, debt-to-income ratio and annual income; and non-standard "soft" information, which has no direct relationship with borrowers' financial status or creditworthiness and usually posted by borrowers voluntarily, such as a borrower's picture or a textual description of his future plan [6].

It is found that lenders make use of both information categories to infer borrowers' creditworthiness. Soft information is a useful supplement to hard information in the loan underwriting process, especially for borrowers with poor credit, whose hard information is usually unattractive [7]. Prior studies has examined the signaling effect of a variety of soft information, such as borrowers' pictures, textual descriptions of the usage of loans, etc.[7], facial features [8-10] and social network characteristics on P2P platforms [11, 12]. Different from all these studies, we focus on a new and promising category of soft information - borrowers' self-disclosed social media information.

As one of the most transformative IT applications, social media changes people's life almost in every aspect. A recent report from Pew Research Center published in 2013 finds that 73% of online adults use a social media site of some kind (e.g., Facebook, Twitter,

URI: http://hdl.handle.net/10125/41834
ISBN: 978-0-9981331-0-2

LinkedIn), and 42% of them use two or more social media platforms. People use social media to communicate, collaborate, consume and even create, making social media a valuable information source about individuals. In P2P lending platforms, some borrowers voluntarily disclose their social media account, which makes their social media information accessible to lenders or P2P lending platforms. Are borrowers who choose to disclose their social media account more creditworthy than non-disclosing borrowers? Is the social media information they choose to disclose useful in assessing their default propensity? To our knowledge, no answers have been given for these questions yet.

Our study intends to answer the above questions by examining a combined data set obtained from both a P2P lending platform and a social media site. We first collected loan listing and borrower information from a P2P lending platform and marked the borrowers who disclosed their account with a certain social media site. Then, we collected these borrowers' social media information from the site. With these collected data, we model borrowers' default probability as a function of borrowers' choice to disclose their social media account or not, controlling for relevant factors such as borrowers' demographic characteristics and identity verifications. The result shows that borrowers who disclosed their social media information have a significant lower default probability compared to those who did not. In order to rule out the effect of self-selection, we leverage a natural experiment introduced by the P2P lending site that enabled borrowers to link to their social media sites. We further employ propensity score matching (PSM) technique to assess the relationship and the results are consistent. Furthermore we examine the relationship between borrowers' social media engagement and their default probability. We find social media engagement, such as the scope of the social network a borrower builds up and his activity level in a social media site, act as predictors of borrowers' default probability.

Our study makes three contributions to the P2P lending literature. First, we discover a predictive relationship between borrowers' choice to disclose their social media information and their default probability. Second, we found borrowers' social media engagement also predicts their default probability. These findings identify a new category of soft information that is useful for screening borrowers on P2P lending platforms. Finally, by examining a unique data set combining data from both a P2P lending site and a social media site, we have integrated borrowers' financial behavior with their social media characteristics for the first time in the literature.

2. Theoretical Background and Hypotheses

There is a stream of literature that focus on the information asymmetry in P2P lending markets. Since P2P lenders have less access to "hard" information such as borrower credit history, income, or employment status, they experience a higher degree of information asymmetry as compared to traditional lenders. To cope with the shortage of "hard" information, P2P lenders tend to make more use of "soft" information, such as borrowers' picture or a textual description of the purpose of a loan [8-10]. Moreover, friendship of borrowers exhibited on a P2P lending platform is examined, and certain types of friendships show signaling effects on default rate and others do not [11, 12].

Our study complements the recent literature by specifically examining a new type of soft information - social media information. Although it is related to social network, this category of information is different from the "friendship" studied by [11, 12] in two aspects. First, the "friendship" referred to in our studies is not located in a P2P lending platform, but located in a social media site instead. Second, social media information we examine here includes not only borrowers' friendship information but also borrowers' decision on disclosing their social media accounts, and borrowers' engagement in social media.

2.1 Disclosing Social Media Account as a Predictor of Default Probability

In our study, a borrower disclosing his social media account on a P2P lending platform means disclosing more information about himself, including the social network that he builds up in the social media site, which raises the possibility that his default behavior being known to his friends. Literature on social psychology shows that being honest, trustworthy and fair is important for a moral social image in the eyes of others [13, 14]. A moral failure damages one's social image, and consequently damages social bond to others [13-15] and lead to social punishment of being marginalized, ostracized, or excluded [14, 16]. Default on a loan is very likely to damage someone's social image and causes social punishment. Moreover, literature on social capital finds that social capital is a valuable resource [17, 18]. The sources of social capital lie not only in the structure and content of our social relations but also in trust [19-21]. Anything that makes someone less trustworthy, such as default on a loan, weakens his social capital. Finally, economic theories of social stigma points out that a default imposes a social stigma cost on a borrower if his friends know

about the default [22-24]. Therefore, borrowers who are at the risk of default are less inclined to share their social media accounts, in order to prevent their friends in the social media site from knowing their default in case it happens. We therefore propose:

HYPOTHESIS 1. *Borrowers who voluntarily disclose their social media accounts on a P2P lending platform are less likely to default*

2.2 Self-Disclosed Social Media Information as Predictors of Default Probability

For those borrowers who have disclosed their social media accounts, their social media information can be collected and used to predict their default probability. In this study, we focus on two social media metrics: the scope of borrowers' social network, and borrowers' engagement in the social media.

The scope of borrowers' social network refers to how many friends or acquaintances a borrower has in his social network. It has an effect on how much damage a default could cause to a borrower's social image and social capital, or how much stigma costs a default brings about. Specifically, the larger the social network, the more damage or costs a default can cause. Therefore, a borrower who has a larger social network should be more motivated to avoid a default. Accordingly, we propose Hypothesis 2a.

HYPOTHESIS 2a. *Borrowers who have a larger social network in the social media site are less likely to default on the P2P lending platform.*

Borrowers' engagement refers to how much a borrower is involved in the social media site, such as how many posts a borrower submit or how many dialogues a borrower hosts or joins. It is closely related to the time and efforts a borrower invests in the social media site. As these inputs are aimed to establish a good social image or good relationships with others, if a default behavior destroys the established social image and relationships, the loss to a borrower who has engaged substantially is more than that to a borrower who has engaged little. From another point of view, the more a borrower engages in building up his social image and relationships, the more he values the image and relationships. He should be more reluctant to default. Thus, we have

HYPOTHESIS 2b. *Borrowers who have more engagement in the social media site are less likely to default on the P2P lending platform.*

3. Data Collection

A key and notable contribution of our study is that we combine data related to borrowers' financial behavior in a P2P lending platform with their information in a social media site. In other words, our data consists of two parts: P2P lending data and social media data.

3.1 P2P lending data

The P2P lending data come from one of the largest online P2P lending platform in China. It was launched in June 2007. By the end of 2013, it has had over 600,000 members and nearly $173 million in funded loans. The company, as a platform providing matching between borrowers and lenders, requires borrowers to provide both loan and personal information for initial screening. Loan information includes loan amount, duration, and objectives of the loan. Personal information includes demographic information, education background, income status and any other information that the borrower is willing to provide. For a loan that passes initial screening, the company posts the relevant information on the website. Lenders examine such information and decide whether to invest, and if yes, how much to invest. The company does not provide any guarantee of loan payment. All the risks are borne by the lenders.

Our data sample covers all peer-to-peer lending listings on this company between January 2011 and August 2013. It consists of 35,457 loan records and 11,047 borrower records in total. Variables related to listed loans include loan amount, interest rate, opening and closing dates, credit grade from A (high quality) to HR (low quality) and the outcome of loan repayment. Variables related to borrowers contain borrower's demographic characteristics, including age, gender, education level and marital status, and verification items, including identity card verification, education certificate verification, phone number verification and image verification.

3.2 Social media data

Over 40% borrowers in the company have disclosed their Sina microblog account to the platform. Sina microblog is the biggest microblog site in China, which opened in September 2009 and has had nearly 300 million users by the end of 2013. The dataset obtained includes a variable which marks whether a borrower disclosed his/her Sina microblog account. For those borrowers who have disclosed their Sina account (5239 borrowers in total), we accessed their microblog page and collected relevant data. The data we obtained from their microblog pages include social network scope metric, and engagement metric.

4. Variable Definitions

4.1 Dependent variable

Default. The dependent variable is a dummy variable. The value is 1 if the borrower defaulted on a loan and 0 if the borrower never defaulted on any loans. A borrower may default on more than one loan, but the value of the variable is still 1 in these cases. Actually, very few borrowers defaulted on two or more loans, since borrowers in default are not allowed to make a new loan request.

4.2 Independent variables

Microblog_Disclosed. It is a dummy variable, which is 1 if the borrower disclosed his Sina microblog account on the ppdai.com platform, otherwise is 0. As long as the borrower discloses his account, staff on the platform can obtain a verified hyperlink to access the borrower's microblog homepage. Through the homepage, the staff can obtain more information about the borrower and potentially contact his followers listed on his profile page.

#Followers. For a borrower who has a Sina microblog account, followers are the ones who subscribe to the borrower's microblog and follow all the updates of the borrower. The number of followers can be regarded as a proxy for the scope of borrowers' social network in the Sina microblog site. Moreover, followers can be differentiated by whether they are followed by the borrower. If two persons follow each other's microblog, they are probably friends and know each other in real life, or they are interested in each other and want to be friends. If the follower is not followed by the borrower, he/she is probably a fan rather than a friend of the borrower. Therefore, we have two sub-level proxies for the scope of borrowers' social network, that is, *#Friends* and *#Fans*.

#Microblogs. We use the number of microblogs that the borrower has posted on his microblog page as a measurement of his engagement in the social media site, since posting microblogs is the major way for a borrower to express himself and to attract followers' attention in the microblog site. Posting more microblogs costs the borrower more time and efforts.

4.3 Control Variables

Borrower's demographic characteristics. This set of control variables includes borrowers' age, gender, marital status and education. If a borrower is a male, his value of gender is 0; otherwise, the value is 1. If a borrower is single, his value of marital status is 0; otherwise, the value is 1. The value of education corresponds to the highest degree a borrower has obtained, which ranges from 1 to 6. The value 1 stands for a middle school degree or lower; the value 6 stands for a postgraduate degree; and the rest values stand for degrees between them.

Borrower's pre-verification. It is a set of dummy variables. PPdai.com recommends borrowers go through a variety of verification processes before making a loan request. The processes include verification of borrowers' identity card, education certificate, phone number and image (i.e. online visual verification). Therefore, we use a set of dummy variables to correspond to the processes respectively. The value of a dummy variable is 1, if the borrower has gone through a specific verification process, otherwise the value is 0.

5. Empirical Modeling and Results

We begin by analyzing the relationship between the default outcome and borrowers' choice to disclose their microblog account. We use a logit regression model first, and then we utilize the propensity score matching (PSM) technique and instrument variable regressions to address the endogeneity concerns.

Afterwards, based on the combined data collected from the P2P lending platform and the social media site, we analyze the effects of social media metrics with a logit regression model. The test examines if the scope of borrowers' social network and borrowers' engagement in the social media site have effects on default probability.

5.1 The Effect of Microblog Disclosure on Default Probability

In this part of analysis, the sample data is obtained from ppdai.com. There are 11047 borrower listings in our sample, and 48% borrowers disclose their Sina microblog accounts. We model a default as occurring if a payment is late by 120 days, which is suggested by ppdai.com.

5.1.1 Logistic Regression Model. Table 1 reports estimate of a logit model for the probability that a borrower defaults on a loan. With a set of control variables, we estimate

$$logit(Default) = \alpha + \beta_1 Microblog\_disclosed + \beta_2 Controls + \varepsilon$$

The results in Table 1 show that microblog disclosure is positively related to the default probability and is significant at 0.01 level. Hypothesis 1 is supported.

Table 1. Logit regression model of borrower default

Variable	Parameter	Std. error

Microblog_disclosed	-0.748***	0.055
Education	-0.150***	0.024
Marital status	-0.158***	0.059
Gender	-0.593***	0.086
Age	-0.012**	0.005
Image verified	0.090	0.056
Education verified	-0.614***	0.073
Phone# verified	-0.069	0.063
IDCertification verified	-0.159***	0.061
Constant	0.134	0.167
N	11047	
Log likelihood	-5027.467	

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

However, we find that for each covariate, the difference in averages by microblog disclosure status is significant (see Table 2), which means the data are unbalanced in covariates between the group who discloses microblog and who does not. The unbalance of the data weakens the reliability of the results of the regression model [25]. Therefore, we utilize Propensity Score Matching to adjust for the differences in covariates in the next section.

Table 2. The difference in averages of covariate by microblog disclosure status

Variable	Mean		t-test	
	Mblog Disclosed=1	Mblog Disclosed=0	t	p
Education	3.6288	3.6507	-0.99	0.323
Marital status	1.4614	1.5642	-10.86	0.000
Gender	1.1262	1.1422	-2.47	0.014
Age	29.362	30.993	-15.01	0.000
Image verified	.68253	.46551	23.61	0.000
Education verified	.28742	.26295	2.88	0.004
Phone# verified	.86058	.72898	17.29	0.000
IDCertification verified	.84845	.65350	24.21	0.000

5.1.2 PSM. The objective of PSM is to select treatment and control borrowers who resemble each other in all relevant characteristics except for microblog disclosure (the treatment), thereby creating a statistical equivalence between the two groups by balancing them on observed covariates.

The first step to perform propensity score matching analysis is to estimate the propensity scores (PS). A logit model was estimated to derive the propensity scores where the outcome variable is microblog disclosure. The model is not a behavioral one, but simply a statistical device that enables us to weight differences in observable variables between borrowers who disclose their microblog and those who do not. From the weights—the coefficients in the logit model—we can construct a propensity score for each treated and control case. The PS values summarize several characteristics of each subject into a single-index, which makes matching subjects on an n-dimensional vector of characteristics feasible. These results of the logit model along with the fit statistics are reported in Table 3.

Table 3. Logit regression model of microblog disclosure

Variable	Parameter	Std. error
Education	0.037*	0.020
Marital status	-0.235***	0.047
Gender	-0.037	0.059
Age	-0.046***	0.004
Image verified	0.595***	0.043
Education verified	-0.170***	0.052
Phone# verified	-0.563***	0.053
IDCertification verified	-0.851***	0.052

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

We then perform the process of matching the treatment and control borrowers using the estimated

propensity scores. Before applying the matching methods, we need to make sure that our treated and control units share the same support so that they are comparable. Treated cases off the common support, that is, cases whose propensity score is higher than the maximum or less than the minimum propensity score of the controls need to be excluded. In our data set, there is only one treated unit off the common support and excluded. There are a wide variety of matching methods available, such as nearest neighbor matching, radius matching and kernel matching. The primary advice to select between them is to select the method that yields the best balance [27-29]. After trying all the aforementioned methods, we find kernel matching is the optimal matching method for our study. Table 4 shows the reduction in bias on observables achieved through the kernel matching. From Table 4, it is evident that the matching achieves an appreciable reduction in bias on observables. Specifically, the absolute bias of all the covariates is less than 5%, and all the p-values are larger than 0.05.

Table 4. Summary statistics and covariate comparison before and after matching

		Mean			%reduced	t-test	
Variable	Sample	Treated	Control	%bias	|bias|	t	p>|t|
Education	Unmatched	3.6288	3.6507	-1.9		-0.99	0.323
	Matched	3.6289	3.6251	0.3	82.5	0.17	0.865
Marital status	Unmatched	1.4614	1.5642	-20.7		-10.86	0.000
	Matched	1.4613	1.4634	-0.4	97.9	-0.22	0.822
Gender	Unmatched	1.1262	1.1422	-4.7		-2.47	0.014
	Matched	1.126	1.1279	-0.6	88.2	-0.29	0.770
Age	Unmatched	29.362	30.993	-28.6		-15.01	0.000
	Matched	29.358	29.531	-3.0	89.4	-1.69	0.091
Image verified	Unmatched	.68253	.46551	45.0		23.61	0.000
	Matched	.68266	.67981	0.6	98.7	0.32	0.752
Edu verified	Unmatched	.28742	.26295	5.5		2.88	0.004
	Matched	.28747	.29716	-2.2	60.4	-1.10	0.270
Phone# verified	Unmatched	.86058	.72898	33.0		17.29	0.000
	Matched	.86074	.86102	-0.1	99.8	-0.04	0.968
IDCer. verified	Unmatched	.84845	.6535	46.3		24.21	0.000
	Matched	.84861	.84812	0.1	99.7	0.07	0.944

Table 5 shows the results of ATTs obtained before and after matching. The results support the Hypothesis 1. We find significant differences in default rate between treated and control groups.

Although the t-statistics obtained after matching is smaller than that obtained before matching, the statistics is still significant even at p=0.01 level.

Table 5. Comparisons of ATTs

Variable	Sample	Treated	Controls	ATT	S.E.	T-stat
Default	Unmatched	0.131	0.235	-0.104	0.00731	-14.27
	Matched	0.131	0.237	-0.106	0.00813	-13.04

Finally, we conduct sensitivity analysis to check for hidden bias. Since matching is based on the conditional independence or unconfoundedness assumption, if there are unobserved variables that simultaneously affect assignment into treatment (microblog disclosure) and the outcome variable (borrower's default), a hidden bias might arise [30]. Since estimating the magnitude of selection bias with non-experimental data is not possible, we address this problem with the bounding approach proposed by [30]. Instead of testing the unconfoundedness assumption itself, Rosenbaum bounds provide evidence on the sensitivity degree to which any results hinge on the untestable assumption. The results in Table 6 show that our study is not sensitive to a hidden bias until the bias doubles the odds of borrower's default.

Table 6. Sensitivity analysis: Rosenbaum critical p-values for treatment effect

Δ	p-value
1.0	<0.0001
1.1	<0.0001
1.2	<0.0001
1.3	<0.0001
1.4	<0.0001
1.5	<0.0001
1.6	<0.0001
1.7	<0.001
1.8	<0.01
1.9	<0.1
2.0	>0.1

5.1.3 Identification through Instrumental Variable. Besides the PSM technique, we also use an instrumental variable for microblog disclosure to identify causality of the model. A suitable instrument for microblog disclosure should be exogenously related to borrowers' decision on disclosing microblog but did not affect the likelihood of default. We notice that ppdai.com did not provide a function on its webpage to help borrowers disclose their microblog till the June of 2011, therefore the borrowers who registered in ppdai.com before the June of 2011 is less likely to disclose their microblog than the borrowers who registered after the date. We tested this argument and find that the relationship is strongly positive, suggesting that registration after 2011 June has a predictive power for microblog disclosure. Meanwhile, the instrument also meets the exclusion restriction. Registration after the June of 2011 could hardly affect the likelihood of default through any direct channel that is independent of microblog disclosure. The result of the IV model is reported in Table 7, where we instrument for microblog disclosure with the instruments of registration after the June of 2011. The result confirms the Hypothesis 1.

Table 7. Results of IV model

Variable	Parameter	Std. error
Microblog_disclosed	-1.896***	0.122
Controls	(Included in estimation)	

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

5.1.4 Difference-in-Difference Model Although the result of the logistic model shows that disclosing microblog account is a predictor of default probability, it does not tell the underlying cause: is it because that the borrowers being afraid of social stigma costs? We use a DID model to identify the cause. In the April of 2013, the P2P company launched a marketing campaign to encourage borrowers to disclose their microblog accounts. We estimate the effect of the campaign on the default probability of a loan whose borrower disclosed his/her social media account. The estimated model is as follows

$$\ln\left(\frac{P(Default_{it}=1)}{1-P(Default_{it}=1)}\right) = \alpha + \beta_1 Mb\_disclosed_i + \beta_2 Cmp_{it} + \beta_3 Mb\_disclosed_i \times Cmp_{it} + \beta_4 Controls_i + \varepsilon_{it}$$

The dummy variable Mb_disclosed equals 1 if the borrower of a loan has disclosed his microblog, otherwise it equals 0. The dummy variable Cmp is a time variable, which takes the value 0 and 1 for periods prior to and post the campaign. Controls represent a vector of loan characteristics, such as loan amount, interest rate, and lending period. The main parameter of interest is β_3. The result of β_3 is negative and significant (see in Table 8), suggesting that this campaign negatively influences the default probability of the loans whose borrowers have disclosed his/her social media account. One possible reason for this to happen is that these borrowers care about social stigma costs, because they may worry that after this campaign, the P2P lending company could use their microblog account as an outlet to spread the word if a default occurs, which increases their social stigma costs. With this worry in mind, they are less likely to default after the campaign.

Table 8. Results of DID Model

Variables	B	S.E.	Sig.
Mb_disclosed	*-.652*	*.051*	*.000*

Cmp	*-.199*	*.052*	*.000*
Mb_disclosed ×Cmp	*-.274*	*.087*	*.002*
Controls	(Included in estimation)		

5.2 The Effect of Microblog Metrics on Default Probability

For this analysis, we create a combined dataset in which data are obtained from ppdai.com and sina.com. We select borrowers who disclose their microblog accounts in ppdai.com, and collect the microblog metrics (e.g. #Followers, #Friends, #Fans and #Microblogs) from their profile pages in sina.com. This combined data sample includes 5239 listings.

We use a logit model to estimate the default probability of the effect of the microblog metrics on default likelihood for borrowers who have disclosed his microblog.

$$logit(Default) = \alpha + \beta_1 Microblog\_Metrics + \beta_2 Controls + \varepsilon$$

Because of the large variance and scale of the microblog metric variables, we take the natural log of them in the model. The results are presented in Table 9.

Table 9. Results of Logit Models

	M(1)	M(2)	M(3)	M(4)	M(5)	M(6)	M(7)
#Followers	-0.153***		-0.132***				-0.157***
	(0.022)		(0.035)				(0.023)
#Microblogs		-0.121***	-0.024	-0.043*	-0.047		
		(0.020)	(0.032)	(0.026)	(0.030)		
#Friends				-0.174***		-0.153***	
				(0.039)		(0.038)	
#Fans					-0.111***	-0.079***	
					(0.034)	(0.028)	
Influential							0.062
							(0.099)
Controls	(Included in estimation)						

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

Model (1) and (2) analyze the effect of #Followers and #Microblogs, respectively. The independent variable in both models is negatively related to the default probability at the 0.01 significance level. The results demonstrate that the larger the scope of the social network a borrower has in a social media site, the less likely he defaults on a loan; the more engagement a borrower has with his social media site, the less likely he defaults. Both Hypothesis 2a and 2b are supported.

However, the effect of #Microblogs is no longer significant when both #Followers and #Microblogs are included in Model (3), which indicates that #Microblogs may be closely related to #Followers. As mentioned before, followers can be categorized into friends and fans in the microblog site. Since writing microblogs are the major way for a borrower to attract fans, #Fans probably has a strong relationship with #Microblogs. While friends are usually acquaintances in real life, #Friends is likely to have a weaker relationship with #Microblogs. In other words, the information that #Microblogs conveys is a supplement to #Friends, but not to #Fans. It is confirmed by the results of Model (4) and (5).

We next examine the effect of two different types of social network, that is, friends and fans. For a borrower, both friends and fans he has in the microblog site are the sources of his social capital. Either friends or fans knowing about a borrower's default can damage his social image and cause a social stigma cost, therefore, both #Friends and #Fans should have an effect on borrower's default likelihood. However, as previous studies have demonstrated, close friends have a stronger behavioral effect on each other than strangers do [31, 32]. We therefore expected that the effect of #Friends on borrowers' default likelihood should be more intensive than that of #Fans. Model (6) shows that #Friends and #Fans are both negatively related to the default probability with $p<0.01$, but the coefficient of #Friends almost doubles relative to that of #Fans. The results indicate that although both variables are predictors of default likelihood, #Friends is a stronger signal than #Fans.

We also consider that a borrower having a large

#Followers is more likely to have a healthy financial situation as an influential person. Therefore, their low default probability may be due to their financial well-being instead of avoiding costs in social capital. To examine this probability, we ran Model (7) with an additional term for a borrower's influence, which can also be regarded as proxies for financial position. From Sina microblog site, we get not only the data of how many followers a borrower has (e.g. #Followers) but also the data of how many people the borrower is following (e.g. #Followings). It is reasonable to assume that #Followers of influential borrowers is always greater than #Following. Therefore, we created a dummy variable "Influential", whose value equals to 1 when #Followers is greater than #Following, otherwise, equals to 0. The result of Model (7) shows that #Followers remain significant while Influential is not significant. The former conjecture is denied.

5.3 Prediction Performance of the Models

We have proposed several variables as predictors of borrowers' default likelihood. In order to demonstrate their prediction power, we evaluate the proposed models with AUC, which is the area under the ROC (Receiver Operating Characteristic) curve. AUC is a standard metric for assessing models that predict classification probabilities [33]. A model that yields a higher AUC generally offers greater predictive power than a model that produces a lower AUC [34].

In Table 10, we show AUCs of the proposed models and those of the benchmark models (the models without the proposed variables). The integrated model in the last column includes the variables of both microblog_disclosed and microblog metrics (#Fans and #Friends).

Table 10. AUCs of Models

Model	Microblog_Disclosed		Microblog Metrics					Integrated Model
	Benchmark	Proposed	Benchmark	M(1)	M(2)	M(4)	M(6)	
AUC	0.6231	0.6573	0.6247	0.6522	0.6456	0.6534	0.6557	0.6636

First, all the AUCs in Table 10 are greater than 0.5, which suggests that the predictive power of the models is higher than that of random guess [34]. Second, Table 10 shows that the AUCs of all the proposed models are greater than the benchmark models, which means the proposed models have more predictive power than the benchmark models. Finally, the integrated model, which includes all the proposed variables, has the largest predictive power among all the models. These results indicate that the models with soft information on borrowers' social media outperform the models without such information.

6. Conclusion

In this paper, we study the signaling effect of social media information on borrowers' credit worthiness in P2P lending. The results suggest that social media information can be the signal of creditworthiness on two levels. On the first level, for all the borrowers in the market, their decision on whether disclosing their social media accounts or not is a predictor of their default probability. On the second level, for the borrowers who choose to disclose their social media accounts, their social media metrics, such as their social network scope and their inputs in the social media site, are predictors of default probability.

Our study contributes to the literature across IS and finance disciplines. Lenders on P2P lending marketplaces use soft information to screen borrowers [7], and our study adds to the literature on soft information [8-12] by examining a new category of soft information. Specifically, our results indicate that social media information is useful for the prediction of borrowers' default probability. To our knowledge, it is the first study that examines the usage of social media in personal finance. While most of literature regards social media as a marketing tool, we provides a new point of view by regarding social media as an information source for individual creditworthiness.

Moreover, our results provide a new insight to improve risk control in P2P lending in China. On the one hand, individuals in China do not have a well-verified credit score, such as FICO score, which enlarges the information asymmetry in Chinese P2P lending markets. On the other hand, about 80% of Internet users in China have a social media account [35]. Their social media activity provides a rich set of information that could be used by P2P lending markets for credit assessment. Our study demonstrates the validity of this approach and highlights the importance of leveraging social media information in P2P lending markets in China.

7. References

[1] Stiglitz, J. E., and Weiss, A. "Credit rationing in markets with imperfect information". The American economic

review, 1981, 393-410.
[2] Ang, J. S., Lin, J. W., and Tyler, F. "Evidence on the lack of separation between business and personal risks among small businesses". The Journal of Entrepreneurial Finance, 1995, 4(2), 197-210.
[3] Avery, R. B., Bostic, R. W., and Samolyk, K. A. "The role of personal wealth in small business finance". Journal of Banking and Finance, 1998, 22(6), 1019-1061.
[4] Manove, M., Padilla, A. J., and Pagano, M. "Collateral versus project screening: A model of lazy banks". RAND Journal of Economics, 2001, 726-744.
[5] Adams, W., Einav, L., and Levin, J. "Liquidity constraints and imperfect information in subprime lending (No. w13067)". National Bureau of Economic Research, 2007.
[6] Iyer, R., Khwaja, A. I., Luttmer, E. F., and Shue, K. "Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?". In AFA 2011 Denver Meetings Paper,2009.
[7] Khwaja, A. I., Iyer, R., Luttmer, E., andShue, K. "Screening Peers Softly: Inferring the Quality of Small Borrowers". Dissertation, Harvard College, 2013.
[8] Theseira, W. "Competition to Default: Racial Discrimination in the Market for Online Peer-to-Peer Lending". Dissertation, Wharton, 2009.
[9] Pope, D. G., and Sydnor, J. R. "What's in a picture? evidence of discrimination from Prosper. com". Journal of Human Resources, 2011,46(1), 53-92.
[10] Ravina, E. "Love and loans: the effect of beauty and personal characteristics in credit markets". Available at SSRN 1101647, 2012.
[11] Freedman, S. M., and Jin, G. Z. "Learning by doing with asymmetric information: evidence from Prosper.com (No. w16855)". National Bureau of Economic Research, 2011.
[12] Lin, M., Prabhala, N. R., and Viswanathan, S. "Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending". Management Science, 2013, 59(1), 17-35.
[13] Baumeister, R. F., and Leary, M. R. "The need to belong: desire for interpersonal attachments as a fundamental human motivation". Psychological Bulletin, 1995, 117(3), 497.
[14] de Waal, F. B. M. "Good Natured: The Origins of Right and Wrong in Humans and Other Animals" Harvard University Press. Cambridge, MA, 1996.
[15] Ahmed, E. " Shame management through reintegration". Cambridge University Press, 2001.
[16] Braithwaite, J. "Crime, shame and reintegration". Cambridge University Press, 1989.
[17] Dore, R. "Goodwill and the spirit of market capitalism". British Journal of Sociology, 1983, 459-482.
[18] Adler, P. S., and Kwon, S. W. "Social capital: Prospects for a new concept". Academy of Management Review, 2002, 27(1), 17-40.
[19] Putnam, R. D. "Bowling alone: America's declining social capital". Journal of Democracy, 1995, 6, 68.
[20] Knoke, D. "Organizational networks and corporate social capital". In R. Th. A. J. Leendersand S. M. Gabbay (Eds.), Corporate Social Capital and Liability: 17-42. Boston: Kluwer, 1999.
[21] Leana, C. R., and Van Buren, H. J. "Organizational social capital and employment practices". Academy of Management Review, 1999, 24(3), 538-555.
[22] Crocker, J., Major, B., and Steele, C. "Social stigma". In Gilbert DT, Fiske ST, Lindzey G (Eds.), The Handbook of Social Psychology, Vols. 1 and 2 (McGraw-Hill, New York), 1998, 504–553.
[23] Thorne, D., and Anderson, L. "Managing the stigma of personal bankruptcy". Sociological Focus, 2006, 39(2), 77-97.
[24] Cohen-Cole, E., and Duygan-Bump, B. "Household bankruptcy decision, the role of social stigma vs. information sharing". Working paper, Federal Reserve Bank of Boston, Boston, 2008
[25] Imbens, G. W., and Wooldridge, J. M. "Recent Developments in the Econometrics of Program Evaluation". Journal of Economic Literature, 2009, 47(1), 5-86.
[26] Rosenbaum, P. R., and Rubin, D. B. "The central role of the propensity score in observational studies for causal effects". Biometrika, 1983, 70(1), 41-55.
[27] Harder, V. S., Stuart, E. A., and Anthony, J. C. "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research". Psychological Methods, 2010, 15(3), 234.
[28] Ho, D. E., Imai, K., King, G., and Stuart, E. A. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference". Political Analysis, 2007, 15(3), 199-236.
[29] Rubin, D. B. "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials". Statistics in Medicine, 2007, 26(1), 20-36.
[30] Rosenbaum, P. R. "Observational studies". Springer New York, 2002.
[31] Bond R M, Fariss C J, Jones J J, et al. "A 61-million-person experiment in social influence and political mobilization". Nature, 2012, 489(7415), 295-298.
[32] Christakis, N. A., and Fowler, J. H. "Social contagion theory: examining dynamic social networks and human behavior". Statistics in Medicine, 2013, 32(4), 556-577.
[33] Huang, J., and Ling, C. X. "Using AUC and accuracy in evaluating learning algorithms". IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3), 299-310.
[34] Fawcett, T. "An introduction to ROC analysis". Pattern recognition letters, 2006, 27(8), 861-874.
[35] "Report on Internet Development Status in China", 2016.7. http://www.cnnic.net.cn/gywm/xwzx/rdxw/2016/201608/W020160803204144417902.pdf.