Department of Economics

Working Paper Series

Small Business Borrowing and Peer-to-Peer Lending: Evidence from Lending Club

Adam Nowak, Amanda Ross and Christopher Yencha

Working Paper No. 15-28

This paper can be found at the College of Business and Economics Working Paper Series homepage:

http://be.wvu.edu/phd_economics/working-papers.htm

Small Business Borrowing and Peer-to-Peer Lending: Evidence from Lending Club

Adam Nowak [1] Amanda Ross [2] Christopher Yencha [3]

West Virginia University West Virginia University West Virginia University

June 30, 2015

[1]College of Business & Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: adam.nowak@mail.wvu.edu

[2]College of Business & Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: amanda.ross@mail.wvu.edu

[3]College of Business & Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: cjyencha@mail.wvu.edu

Abstract

This study is interested in the ability of borrowers and lenders to signal to each other in the peer-to-peer lending market. We focus on small business loans and investigate the relationship between the loan description that a borrower provides and the impact of this description on the potential funding of the loan by investors. We find that the loan descriptions in the data can be used to predict the probability that the entire loan will be funded. In addition, we also find that an index created from a textual analysis of the loan description can be used to forecast the performance of the loan; a 1 standard deviation increase in the index will decrease the odds of default by 14%. Thus, it appears as if investors are not making investment decisions based on improper signals.

JEL Codes: D47, D53, D82, D83, G14, G21

Key words: small business borrowing, peer-to-peer lending

1 Introduction

Crowdfunding and peer-to-peer lending, where individuals raise funds by collecting money from a large group of private individuals, has existed for decades. However, since the advent of the internet, peer-to-peer lending has become a more common source of funding, as it has become easier to reach out to a wider audience. For example, GoFundMe, which allows individuals to post requests for financial support for countless purposes, has raised over $1.1 billion dollars since its inception in 2010. [1] In addition to websites that seek donations, there has been an emergence of websites that allow individuals to obtain loans from private investors. These websites allow individuals to bypass banks and work with individual investors, who may be private individuals or companies, to set interest rates and loan terms in such a way that is beneficial to both parties. One example of this type of peer-to-peer lending is Lending Club, which has effectively created its own credit marketplace to match borrowers and investors outside of the typical banking system. [2]

When interested investors go to such websites to find worthy projects to invest in, major issues exist regarding how borrowers can signal their credit worthiness, and how investors can determine which signals are the most important. While Lending Club uses its programs to assess the risk and assign a credit rating to each borrower, there is still important information regarding the viability of the project that cannot be captured in a simple credit score. Therefore, in addition to the credit ranking information, borrowers write a paragraph describing the project to try to entice investors to support their endeavor. In this paper, we focus on small business loans and, using data from the Lending Club website, we determine what key words are most likely to result in a project being funded. Then, we look at what key words are most important in terms of determining which investments are the most successful, where we define success as the loan eventually being repaid.

One of the most significant problems for individuals looking to start a small business is securing the necessary credit and capital funding to begin their endeavor (Berger and Udell (1995); Cassar (2004); Cosh et al. (2009)). When individuals have problems securing funding from banks or other traditional sources, they now have the option to go to various crowdfunding websites to obtain the necessarily funds. Since the advent of on-line peer-to-peer (P2P) lending websites, P2P lending

[1] For more information on how this website works, see http://www.gofundme.com/.

[2] We will describe the Lending Club system in more detail later in the paper. See https://www.LendingClub.com/public/how-peer-lending-works.action for additional information.

has become an increasingly popular way for entrepreneurs to secure funds to start their small business, and there is a growing literature examining how these new funding sources have affected entrepreneurial activity (Cosh et al. (2009); Mollick (2014); Ahlers et al. (2015)).

To conduct our analysis, we have obtained data from Lending Club, a P2P lending website. The data includes loans that were issued during the period of June 2007 to June 2014. Our data set includes all information that is available to the lender, including FICO scores, debt-to-income ratios, annual income, and a description of the project for which the borrower is requesting funding. We also have data on the realized performance of the loan, including whether the borrower defaulted, if the borrower is delinquent, and if the borrower is still making payments on the loan. For our analysis, we are primarily interested in how the writing of the loan description can affect the chances a given loan receives funding, and how the wording of the loan description is correlated with the ultimate success for the investor.

Funding and text data present unique challenges that require specific econometric techniques. Due to the fact that we are interested in the signaling function of loan descriptions, it is necessary to incorporate this text or character variable into the analysis. Methods for statistical analysis of text fall into the category of textual analysis. One approach for incorporating the loan description requires the researcher to pre-specify a list of words that the he or she believes are significant and create indicator variables for each word, a procedure that is referred to as the *bag of words* method in the relevant literature.[3] Related to this method are statistics for the loan description including the number of words, number of characters, and number of misspellings. A competing method to the bag of words method uses indicator variables for each word or phrase in the entire set of loan descriptions and simultaneously estimates which words or phrases (tokens) are important and how important each word is - this procedure is referred to as the *token* approach. [4] We use both methods in this paper.

After we have determined which key words are important to determine funding, we then examine which words and phrases best indicate success of the investment. The token method can then be used to estimate a binary model for funding success. Using the estimated coefficients for each

[3]Examples of the bag of words approach include Hancock et al. (2007), Pang and Lee (2008), Loughran and McDonald (2011), Da et al. (2011), and Garcia (2013)

[4]Examples of the token approach include Mitra and Gilbert (2014), Taddy (2013a), Taddy (2013b), Gentzkow and Shapiro (2010), and Gilbert and Karahalios (2010)

token, it is possible to determine a score or investment index for each loan description by summing the product of the estimated coefficients for each token present in a given loan description. In this manner, each loan description is given an index value that is meant to correlate with the likelihood of complete funding. Because investors are assumed to invest based on loan description, loans with larger values of the index contain loan descriptions that are positive signals to investors. Further, this index can be used as an explanatory variable in other analysis. In particular, we are interested in the relationship between future loan performance and signals that investors receive as captured by the index.

We find that a textual analysis of the loan descriptions can be used to predict which loans will be funded by investors. Furthermore, we find that an index created from the loan descriptions in the funding stage can be used to predict which loans will end up in default. We take these two results as evidence that 1) borrowers are able to signal to investors information about themselves and businesses that is not captured in common, quantitative measures of creditworthiness, and 2) investors are correctly reacting to the signals in that the signals can be used to predict the success of the loan.

2 Existing Research

There has been an extensive discussion on how factors, such as tax policy and various legal processes, affect entrepreneurial activity. However, the most frequently cited obstacle to individuals becoming entrepreneurs is the large amount of financing needed to initially start a business. An extensive literature has shown that the lack of an ability to access the necessary credit and capital funding is one of the main deterrents of entrepreneurs, especially for low-income individuals (Berger and Udell (1995); Cassar (2004); Cosh et al. (2009)).

Recently, there has been an increase in crowdfunding and peer-to-peer (P2P) lending. Crowdfunding allows individuals to submit an open call for loans and support for various endeavors over the internet. By drawing upon on-line social networks, entrepreneurs can obtain small loans from a larger number of individuals, versus a large loan from a sophisticated investor such as a bank or venture capitalist. The investments made by the crowd can take many forms, such as an equity purchase, loan, donation, or a pre-order on a product (Agrawal et al. (2011); Agrawal et al. (2013);

Ahlers et al. (2015); Kuppuswamy and Bayus (2014); Belleflamme et al. (2014); Mollick (2014)).

While P2P lending has become more common, little empirical research thus far has considered the impact of this new investment funding on entrepreneurial activity. Cosh et al. (2009) consider what affects rejection rates for financing from different types of investors, including private individuals. Using data from the United Kingdom, the authors find that it is small firms that are most likely to obtain financing from private individuals. Therefore, it seems plausible that the advent of P2P lending websites has made it possible for individuals to obtain the necessary finances to start up their new business.

Mollick (2014) considers the underlying dynamics of success or failure among crowdfunded ventures. Using data from Kickstarter, the author finds that the projects funded through this website are either successful by a very small margin or fail by a very large margin. In addition, the author finds that projects that are able to signal they are of a higher quality are more likely to be funded. Kuppuswamy and Bayus (2014) also use data from Kickstarter to examine the dynamics over the project funding cycle. The authors find that project support is U-shaped with most funding being received the beginning or end of the period it is posted.

However, one of the biggest issues regarding crowdfunding is determining which projects are worthwhile investments. This is true both in terms of entrepreneurs needing to signal the quality of their business, and investors determining which projects are most likely to be profitable. Ahlers et al. (2015) considered the effect of financial roadmaps, external certification, internal governance, and risk factors on fund raising success. Their results suggest that while external certification is not an important factor, the other four mechanisms are important signals for investors. Kim and Viswanathan (2013) find that initial investments by export investors in the relevant market serve as a signal to other investors, increasing overall investments in the project.

In this paper, we consider signaling markets, but focus more on the call for funding posted by the entrepreneur looking for investors. In particular, we look at funding requests posted on-line and determine what key phrases are most likely to determine which projects are funded. Mitra and Gilbert (2014) use the token method to determine which phrases are associated with successfully funded projects on Kickstarter. However, just because the project is funded does not mean it was a worthwhile endeavor. Therefore, we build on Mitra and Gilbert (2014) and investigate if key phrases that determine successfully funded projects can also be used to determine successful

investments. This is an important issue to explore given the incentive compatibility issues that arise given the de-centralized nature of P2P lending: signals from borrowers can be used by lenders in order to direct funds to the most productive projects.

Previous studies have found that information in news reports, social media and editorials can be used to predict stock performance. Engelberg et al. (2012) find the most successful short sellers are those who are able to correctly process information in publicly available news releases. Boudoukh et al. (2013) develop a methodology for classifying articles in the Dow Jones News Service that can be used in a behavioral setting. Chen et al. (2014) extract sentiment from user supplied opinions on the finance website Seeking Alpha that can be used to predict future stock returns and earnings surprises. Emphasizing the role of individuals in the public realm, Dougal et al. (2012) demonstrate that a small segment of journalists have an impact on short-term stock price movements. As a whole, these studies demonstrate that text in different forms created by different types of agents can be used to predict stock prices. Therefore, we also build upon this literature and see if text analysis can be used to determine the success of investments made through P2P lending websites.

3 Lending Club Process

P2P lending involves lending and borrowing money without interacting with traditional financial intermediaries. Such lending and borrowing of money is handled online in the United States by websites such as Lending Club and Prosper, which primarily host unsecured personal loans at interests rates dependent upon the borrower's credit as and dictated by a formulation developed by the intermediary. P2P lending services have a clear advantage over traditional financial intermediaries as the process of borrowing and lending is mostly automatic. This allows P2P lending services to function with relatively low operating costs. In addition, P2P lending services are able to connect lenders and borrowers across the country who may otherwise not have convenient access to each other.

Lending Club is one such P2P lending service that helps borrowers to obtain loans from investors on the platform. As of April 2015, Lending Club is the world's largest P2P lending service, having funded over nine billion dollars in loans. Loan amounts on Lending Club range from $1,000 all

the way up to $35,000 and these loans can be taken out for any number of reasons, such as home remodeling or debt consolidation. Typically, the loan repayment period is three years, and the loan can be paid off at any time without the borrower being penalized. Borrowers must be in fairly good credit standing, with a FICO score of at least 640, to be eligible to borrower on Lending Club's platform. Strict credit requirements are important for Lending Club to reduce default risk. Even with such strict standards, Lending Club still declines most loan applications that it receives. [5]

Lenders interested in investing in a loan can search through listings on Lending Club's website and choose the loans in which they want to invest. Information available to investors includes the interest rate on the loan, the principal, the borrower's FICO score, Lending Club's evaluation of the loan represented by a grade, the loan purpose, and a description by the borrower. When an investor decides on a loan to fund, he or she chooses the amount to fund the borrower. Investment amounts on a single note can be as little as $25 and as much as the total loan amount requested.

As long as the loan request is fully funded and passes examination by Lending Club, then the loan is issued. Borrowers are then required to pay back the loan by the end of the loan repayment period and investors collect interest in the interim. Lending Club makes its money off of these loans by charging an origination fee to the borrower and a service fee to investors. Low overhead, strict borrowing requirements, and reasonably small fees allow for Lending Club, and other P2P lending platforms, to charge interest rates that are typically lower than that of traditional financial intermediaries and generate fairly generous average returns for investors. [6]

4 Empirical Analysis

4.1 Data Description

The data used in this study comes from Lending Club, a P2P lending website. The loans are issued during the period June 2007 to June 2014. Lending Club provides for download all issued loans and all non-issued loans. Non-issued loans are those loans that do not meet Lending Club credit underwriting policy. The data set includes information on the borrower that is visible to all individuals who wish to fund the loan. The information includes FICO score, debt-to-income

[5] Information about Lending Clubs services have been gathered from https://www.Lending Club.com

[6] http://online.barrons.com/articles/SB50001424052748703496404578171401366018578?mod=googlenews_barrons

ratios, annual income, employer, employment tenure, geographic location, and other information. In addition to information on the borrower, the data set also includes information on the realized performance of the loan including whether the borrower defaulted, is delinquent, paid off the loan, or is still making payments on a loan that has not reached maturity.

Table 1 displays the data for the Lending Club loans. The largest amount possible for a Lending Club loans is $35,000. Loans carry a 36 month term or a 60 month term. The interest rate on loans range from 5.42% to 26.06%. Borrowers in the data set have an average income equal to $72,750. The debt-to-income ratio is calculated as the ratio of monthly payments on non-mortgage debt to self-reported monthly income. The data set includes the maximum and minimum FICO scores for the individual at the time the loan is requested. The FICO score in Table 1 is calculated as the simple average of the maximum and minimum.

Table 1: Summary Statistics

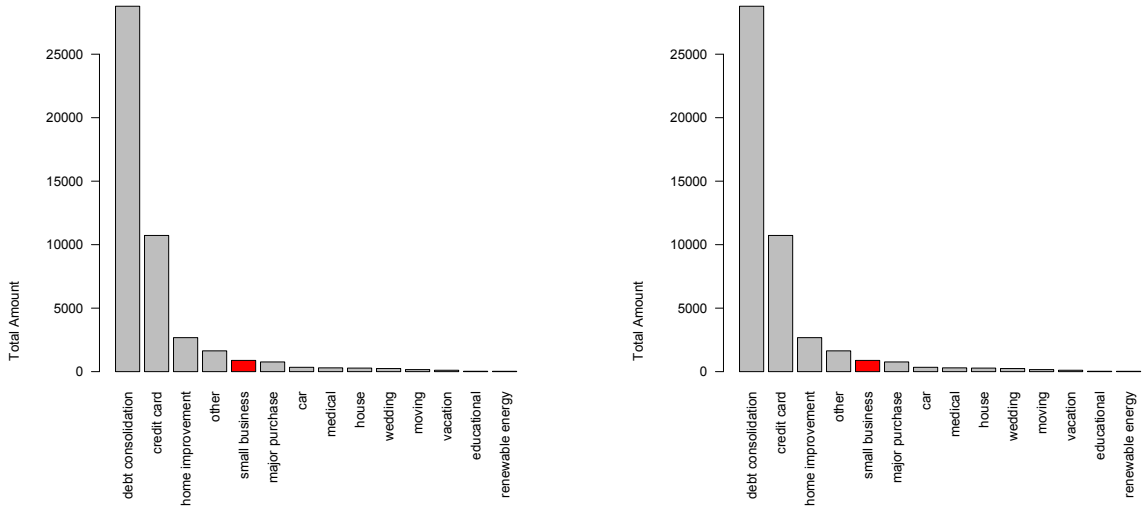
(a) All Loans

Variable	Min	Mean	Median	Max	Std Dev
LOAN AMOUNT	500.00	14091.83	12000.00	35000.00	8198.99
TOTAL FUNDING	500.00	14056.18	12000.00	35000.00	8180.81
INVESTOR FUNDING	0.00	13960.17	12000.00	35000.00	8211.67
TERM (in months)	36.00	42.20	36.00	60.00	10.50
INTEREST RATE	5.42	13.98	13.92	26.06	4.34
INSTALLMENT	15.67	429.52	378.95	1409.99	243.32
INCOME (in $1,000s)	1.89	72.75	62.00	7446.39	54.55
DTI	0.00	16.70	16.45	34.99	7.58
FICO	612.00	699.56	692.00	847.50	31.21
CHARGED OFF	0.00	0.05	0.00	1.00	0.21

(b) Small Business Loans

Variable	Min	Mean	Median	Max	Std Dev
LOAN AMOUNT	500.00	15172.81	13162.50	35000.00	9278.55
TOTAL FUNDING	500.00	14990.00	12800.00	35000.00	9163.12
INVESTOR FUNDING	0.00	14436.55	12000.00	35000.00	9276.13
TERM (in months)	36.00	42.46	36.00	60.00	10.64
INTEREST RATE	5.42	15.67	15.61	26.06	5.06
INSTALLMENT	16.25	473.45	405.69	1409.99	291.89
INCOME (in $1,000s)	6.69	84.71	70.00	1300.00	64.11
DTI	0.00	12.96	12.24	34.96	7.66
FICO	642.00	712.30	702.00	842.00	38.44
CHARGED OFF	0.00	0.16	0.00	1.00	0.36

The 5,828 small business loans have a total nominal value of $88,427,130. Figure 1 displays the total counts and nominal sums by stated purpose. Small business loans are the 5^{th} largest category by both loan counts and loan amounts. The two largest categories of Lending Club loans are debt consolidation and credit card refinancing.



(a) Loan Counts by Purpose

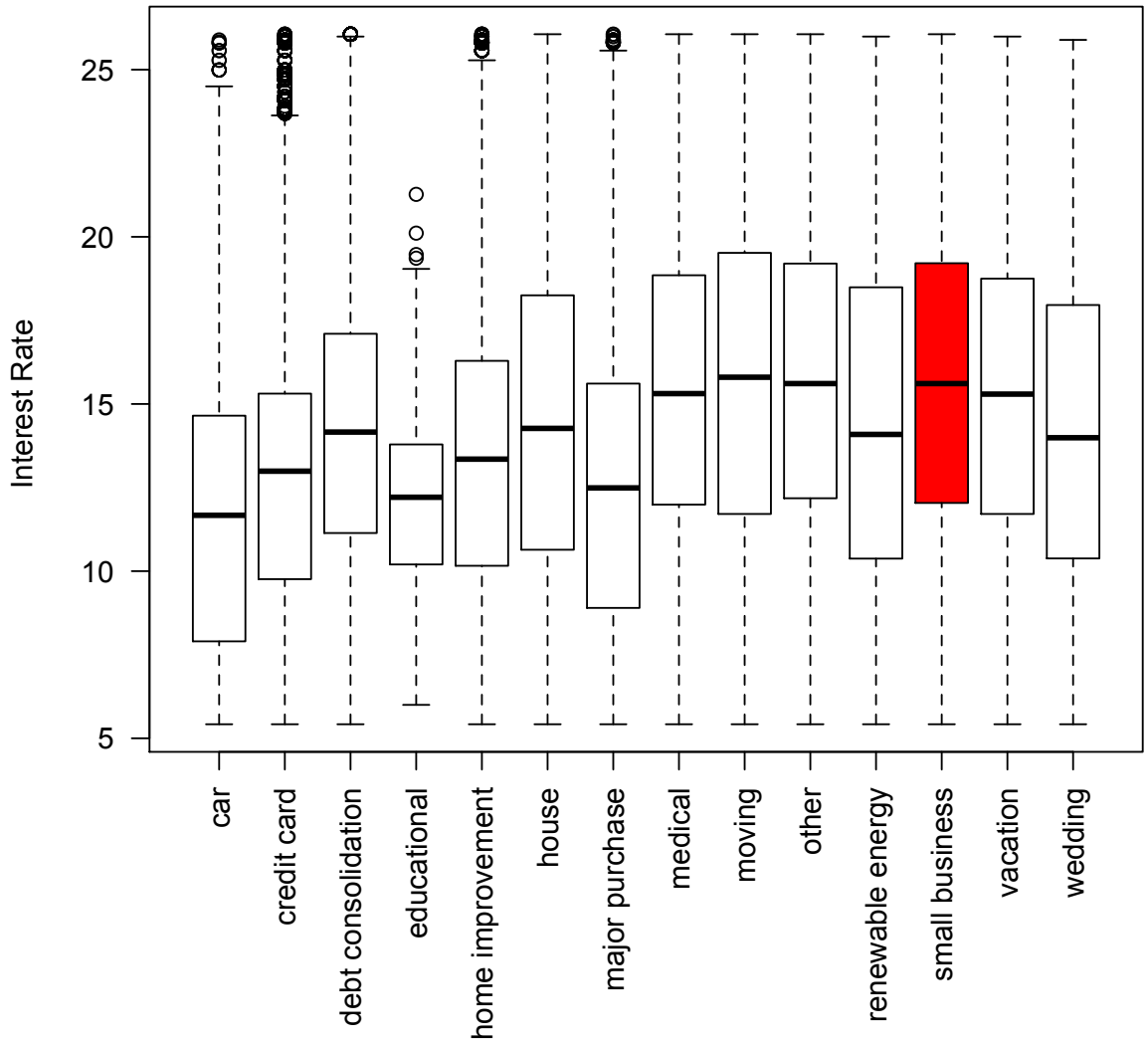
(b) Loan Amounts by Purpose

Figure 1: Total Loan Counts and Loan Amounts by Purpose

Borrows who apply to Lending Club state a requested amount and are given a credit grade $A, B, ..., F$ and quoted an interest rate. Figure 2 displays the distribution of interest rates by purpose. Small business loans have a median interest rate of 13.9%. The interquartile range for small business interest rates is 12.04% to 19.20%. The distribution of interest rates for small business loans is similar to most other categories. In the data, car loans have the lowest median interest rate at 11.67%.

In addition to providing common measures of creditworthiness, Lending Club also allows borrowers to provide a short description of their loan; however, not all loans include a description. The loan descriptions in the data set can include as many as 3,966 characters, with an average of 172 characters. This study is interested in testing whether or not the loan descriptions can be used to signal to investors about the creditworthiness of the borrower and/or the prospects of the company. Because we are interested in the signals present in the loan descriptions, we use as our sample the 2,789 small business loans that have a non-missing loan description.

Figure 2: Distribution of Interest Rates by Purpose



4.2 Methodology

We are primarily interested in examining the role of the loan description, d_i, on investor funding and loan performance. This study investigates two measures of loan funding, f_i. These measures are: an indicator variable for complete funding and the percentage of the loan funded by investors; we use these two measures of f_i due to the right-censoring nature of the data. When referring to an increase in f_i we define an increase in funding as any variable that increases $E[f_i]$.

In order to control for common determinants of funding we include several control variables. These include annual time dummies, the annual income ($INCOME$), debt-to-income ratio (DTI), and FICO score ($FICO$) of the borrower, and the interest rate ($INTERESTRATE$). In addition to the controls, we also include variables of interest that are calculated using d_i. These variables include number of words ($WORDS$), number of characters ($CHAR$), average word length (AWL), number of misspellings ($MISS$) and Simpson's Diversity Index ($SIMPSON$). We assume that funding can be written as

$$f_i = x_i'\beta + m_i'\gamma + \epsilon_i \tag{1}$$

where m_i is any of the variables of interest calculated from d_i, β and γ are parameter vectors and ϵ_i is an error term. We follow the procedure mentioned in Tweedie and Baayen (1998) in calculating a variant of the Simpson's Diversity Index to approximate the diversity of language used in a description. Originally developed for use in ecology to quantify the level of biodiversity in an environment, researchers in other fields have adapted Simpson's measure of diversity for a variety of other applications, including textual analysis Gao and Lin (2013). The formula used to calculate Simpson's Diversity Index is as follows:

$$SIMPSON_i = 1 - \frac{\sum_w n_w(n_w - 1)}{N(N-1)} \tag{2}$$

where n_w is the total number of times word w appears in d_i and N is the total number of unique words in all of the d_i. Simpson's D in this instance measures the rate at which words repeat within the loan request descriptions. A D-score of 1 indicates infinite diversity - the case in which every word within the description is unique. Likewise, a D-score of 0 suggests no diversity and a case

where every word within the description is the same. Because Simpson's D is weighted by the total number of words, comparability between measures of lexical richness beyond descriptions of only a couple of words is not an issue.

As mentioned above, the token method estimates the importance of actual words or phrases in the text. In order to incorporate the actual words and phrases used in the loan description into the analysis, we assume that each d_i can be modeled using a set of *tokens*. We use *bigrams* which are nothing more than two-word phrases created using consecutive words in d_i. An example of actual bigrams in the loan descriptions include *expand business*, *purchase equipment* and *business purchase*. For each d_i, the set of bigrams present in d_i is indicated by a $N \times 1$ vector $t_i = (b_1, b_2, ..., b_N)$ where b_n is equal to 1 if bigram n is in d_i and 0 otherwise.

Before forming the bigrams, it is common to remove all *stop words* and *stem* words in the document. Stop words are common words assumed to contain no little to no informational content. Although there is no universally accepted list, common stop words include *for, and, you, is, have, etc.*. Stemming refers to the process of mapping various variations from a common root word back to the root word itself. For example, *borrower, borrowing* and *borrowed* will all be mapped back to borrow. Of course, the list of stop words and the stemming algorithm can impact the resulting bigrams in the data. We experimented with several lists of stop words and stemming algorithms; we find that the conclusions reached in the paper are robust to the list of stop words and stemming algorithm.

Given the characteristics of the loan and the loan description, we assume that the funding for each loan can be written as

$$f_i = x_i'\beta + t_i'\theta + \epsilon_i \tag{3}$$

where x_i is a vector of borrower and loan characteristics, β is a vector of parameters, t_i is the vector of bigram indicators described above and θ is a vector of parameters. The vector x_i includes the interest rate, FICO score, debt-to-income ratio, annual income, and an indicator variable equal to 1 if the loan does not contain a description and 0 otherwise.

The coefficient θ_n can be interpreted as the change in $E[f_i]$ when bigram b_n enters the loan description. However, it is somewhat misleading to make this statement as borrowers do not write

bigrams, but rather borrowers write sentences that can be broken into a set of bigrams. Alternatively, bigrams are a byproduct of the loan description and not directly chosen or manipulated by the borrower. Textual analysis is interested in scoring the entire loan description measured by the product $Z_i = t_i'\theta$. Here, Z_i can be interpreted as the change in $E[f_i]$ based on the entire loan description. In this way the vector θ aggregates all information in the loan description into an index.

The total dollar amount funded is bounded from above by the total dollar amount requested by the borrower. As such, this number is right-censored. Therefore, when estimating Eq 3 we utilize a Cox proportional hazard model that incorporates right-censoring data. The Cox model estimates a hazard model for time until a particular event occurs. In this study, the event corresponds to investors no longer willing to fund the loan. In this estimation, the total dollar amount funded is analogous to the survival time in the traditional hazard model. The indicator variable for complete funding is equal to 1 if the entire amount of the loan was funded by investors and is 0 otherwise. In this setup, Eq 3 is estimated using both a logit model and a linear-probability model. Lastly, the percentage of the loan funded takes on a value from 0 to 1 and is right-censored. Unreported results indicate that this right-censoring does not significantly impact the conclusions reached in the paper when f_i is equal to the percentage of the loan funded.

Because there is a bigram for every single two-word phrase in the entire set of loan descriptions, the number of bigrams we use can be quite large - there are 64,096 bigrams in the data. However, we apply a conservative filter and only include those bigrams that occur in more than 0.05% of the total number of loan descriptions. This filtering results in 544 bigrams. Given the large number of bigrams and the resulting number of potential coefficients in θ, we adopt a LASSO variable-selection approach for the logit model as in Mitra and Gilbert (2014).

The estimated $\widehat{\theta}$ in Eq 3 can be used to create an investment index for the loan description. Loan descriptions with a larger investment index are more likely to be funded by investors. Examining Eq 3, the product $\widehat{Z}_i = t_i'\widehat{\theta}$ is the change in $\widehat{f}_i$ based on tokens present in d_i. If $\widehat{Z}_i > 0$ $\left(\widehat{Z}_i < 0\right)$, then the loan description is expected to increase (decrease) the amount of funding by investors. Alternatively, if d_i contains bigrams that are a positive signal to investors, then this will increase the likelihood of an investor funding the loan. If these positive signals are rational, the loans with positive signals should out-perform loans without positive signals. In our first measure of

performance, we estimate a default probability model for each loan.

$$y_i^* = x_i'\gamma + \widehat{Z}_i{}'\phi + \eta_i \tag{4}$$

where y_i^* is a latent variable for loan defaults, γ and ϕ are parameter vectors and η is a random error term. A loan defaults if $y_i^* > 0$. Eq 4 is estimated using a logit model. In the data, some loans have not reached maturity. It is possible that these loans will default, but their eventual status is not known in the data. In order to estimate Eq 4 without this sample selection issue, Eq 4 includes a quadratic trend for the age of the loan in x_i.

4.3 Results

The results for Eq's 1 and 3 are presented in Tables 2 and 4. Columns (1)-(5) in Table 2 display the coefficients from a Tobit regression using summary statistics from the d_i. Descriptions with more words, more characters and longer average word lengths tend to be less funded by investors. As the number of misspellings in the loan's description increases, investors will fund a smaller proportion of the loan. Given that spell checking software is widely and easily accessible, a borrower who is not careful with the presentation of his loan request will probably not be careful in how he uses the loaned money.

We also include a measure of lexical richness of the descriptions supplied by borrowers as measured by Simpson's D. When including a measure of lexical richness, we find that, all else equal, descriptions with greater diversity will be funded to a greater degree by investors. Descriptions with greater variety in language are not only more readable but presumably more informative. Loan requests that include information that does not repeat throughout the description could more easily attract investors. While not only weeding out loan descriptions with little to say, this measure also penalizes descriptions with frequent use of articles and prepositions. Such constructions are less crucial to the informative functions of the descriptions and may be contributing to the information overload problem reflected in columns (1)-(4). If investors are turned off by long descriptions but value lexical diversity, then it seems likely that investors are carefully parsing descriptions and looking for particular hints at the credit-worthiness of an individual.

Column (6) in 2 displays the results from the tobit LASSO regression in 3 that uses the token

Table 2: Tobit Regressions of Percentage Funding by Investors

Variable	(1)	(2)	(3)	(4)	(5)	(6)
INCOME (in $1,000s)	-0.004***	-0.004***	-0.004***	-0.004***	-0.004***	-0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
INTEREST RATE	0.012***	0.012***	0.012***	0.014***	0.013***	0.003***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
DTI	0.002*	0.002*	0.002**	0.003**	0.003***	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
FICO	0.0008***	0.0008***	0.0008***	0.0007***	0.0007***	0.0009
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
WORDS	-0.0018***					
	(0.0002)					
$WORDS^2$	2.46e-06***					
	(3.81e-07)					
CHAR		-0.0003***				
		(3.16e-05)				
$CHAR^2$		7.23e-08***				
		(1.14e-08)				
AWL			-0.0450***			
			(0.0050)			
MISS				-0.0705***		
				(0.0145)		
SIMPSON					2.234***	
					(0.525)	
Tokens	N	N	N	N	N	Y
Observations	2,789	2,789	2,789	2,789	2,789	2,789
Pseudo R-squared	0.0769	0.0749	0.0568	0.0334	0.0313	0.0931

*, **, *** significance at the 0.1, 0.05, and 0.01 level, respectively

approach to textual analysis. Standard errors are bootstrapped using 5,000 replications. Starting with a candidate set of 2,000 bigrams, the cross-validation selection procedure selects 45 bigrams. The 20 bigrams with the largest coefficients (in magnitude) are displayed in Table 3. The first two columns display the coefficients from the Tobit regression for percentage of the loan funded by investors, and the last two columns are for the logit regression for complete funding by investors. A total of 67 bigrams are selected in the Tobit specification, and 163 bigrams are selected in the logit specification.

Table 3: Bigram LASSO: Top 20

Tobit		Logit	
bigram	$\widehat{\theta}$	bigram	$\widehat{\theta}$
business.was	-0.202	thank.consideration	-0.95
what.can	-0.201	recently.purchased	-0.931
go.out	-0.183	happy.answer	-0.910
software.hardware	-0.152	income.business	-0.906
high.tech	-0.140	website.www	-0.875
lower.interest	-0.128	looking.funding	-0.860
credit.can	-0.121	debt.free	0.791
learn.more	-0.121	own.funds	-0.764
cost.effective	-0.119	advertising.business	-0.750
loan.complete	-0.115	expenses.including	0.708
funding.purchase	-0.111	over.next	-0.684
which.currently	-0.104	working.capitol	-0.674
want.buy	-0.102	money.new	-0.669
production.company	-0.100	loan.small	0.647
more.details	-0.097	looking.purchase	-0.636
bed.breakfast	-0.092	computer.equipment	-0.630
want.make	-0.089	expand.existing	-0.627
tax.preparation	-0.087	business.operations	-0.624
website.www	-0.083	business.called	-0.601
chassis.fabrication	-0.079	business.grow	-0.587

More importantly, comparing the models in Columns (1)-(6) based on overall model fit can be done using the pseudo R^2. The pseudo R^2 is calculated as the ratio of the maximized log-likelihood to the null likelihood. The token method has the largest value indicating the token procedure is most useful for predicting funding by investors. In other words, which words are in the loan description are more important than summary statistics calculated from the number of words and characters in the loan description.

Table 4: Logit Regressions of Complete Funding by Investors

Variable	(1)	(2)	(3)	(4)	(5)	(6)
INCOME	-3.86E-08	3.39E-08	3.37E-07	5.13E-07	4.12E-07	0.000
	(6.41e-07)	(6.42e-07)	(6.4e-07)	(6.27e-07)	(6.27e-07)	(0.000158)
INTEREST RATE	-0.011	-0.010	-0.008	-0.0007	-0.002	0
	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)	(0.004)
DTI	0.005	0.005	0.009	0.010*	0.010*	0
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.004)
FICO	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
WORDS	-0.011***					
	(0.001)					
$WORDS^2$	1.38e-05***					
	(2.27e-06)					
CHAR		-0.002***				
		(0.000)				
$CHAR^2$		4.31e-07***				
		(6.7e-08)				
AWL			-0.247***			
			(0.027)			
MISS				-0.310***		
				(0.078)		
SIMPSON					12.470***	
					(3.902)	
Tokens	N	N	N	N	N	Y
Observations	2,789	2,789	2,789	2,789	2,789	2,789
Log Lik	-1807	-1808	-1858	-1894	-1896	-1748

*, **, *** significance at the 0.1, 0.05, and 0.01 level, respectively

We also estimate a logit model where the binary response is either complete or incomplete funding by investors. Results are presented in Table 4. The signs of the coefficients are identical to the results in 2. More words, more characters, longer words and more misspellings are all associated with investors not completely funding the loan. The results for Column (6) are for the logistic LASSO model. The tokens that are selected as significant are in the first column of Table 3. Once again, investors are turned off by loan descriptions that have links to the borrower's website.

The previous results have shown that a textual analysis of loan descriptions can be used to predict which loans will be funded by investors. A related question is whether or not the signals that encourage investment can be used to predict performance. We provide an answer to this

question by using $\widehat{Z}_i$ as a regressor in Eq 4. As emphasized above, the presence of any given b_n is of secondary importance to the entire collection t_i and the resulting investment index, $\widehat{Z}_i$. Using the $\widehat{\theta}$ estimated in Eq 3, the investment index, $\widehat{Z}_i$, can be created. $\widehat{Z}_i$ can be used to test whether or not loan description that are associated with greater investor funding are also associated with increased performance.

The results for the binary response model for defaults in Eq 4 are displayed in Table 5. Standard errors are in parenthesis. Columns (1)-(2) use the $\widehat{\theta}$ from the Tobit and logit models, respectively. In each model, $\widehat{Z}$ is normalized to have mean 0 and unit variance. Because not all of the loans have reached full maturity, we include a second order polynomial for the age of the loan (AGE) where age is measured in months. The results in Column (1) indicate that the investment index created from the Tobit model are not statistically significant when used to predict defaults. The results in Column (2) show that a 1 standard deviation increase in the investment index will decrease the odds of default by approximately 13.5% [7]. This value is both statistically and economically significant. Thus, it appears as though the signals in loan descriptions that encourage investors to invest are also associated with a decrease in the expected odds of default.

5 Conclusions

This study investigates the ability of borrowers and lenders to signal to each other in a P2P lending market. We have shown that loan characteristics and the information that borrowers provide in the form of a loan description are useful in determining which loans are funded by investors. Simple statistics such as the number of words or characters in a loan description are shown to significantly impact the funding decision. We find that more words or characters in a loan description will decrease the probability of funding. We find that the token specification outperforms other competing models when estimating a Tobit or logit model for funding.

Furthermore, we have also shown that there is information overlap in the funding and performance of the loan. Specifically, an investment index estimated in the funding stage in the logit model is a significant predictor for future defaults. This result is encouraging for small business borrowers and lenders alike as it suggests 1) borrowers are able to signal to lenders and 2) lenders

[7] $e^{-0.144} \approx 0.865$

Table 5: Default

	(1)	(2)
INTERCEPT	7.650***	7.614***
	(1.463)	(1.461)
AGE	-0.699***	-0.676***
	(0.248)	(0.248)
AGE^2	-0.173**	-0.183**
	(0.077)	(0.077)
INCOME (in $1,000s)	-0.003***	-0.003***
	(0.001)	(0.001)
DTI	0.004	0.004
	(0.007)	(0.007)
INTEREST RATE	-0.049***	-0.049***
	(0.014)	(0.014)
FICO	-0.009***	-0.009***
	(0.002)	(0.002)
$\widehat{Z}$	-0.003	-0.144***
	(0.056)	(0.050)
N	2,789	2,789
Deviance	2283.310	2291.504

*, **, *** significance at the 0.1, 0.05, and 0.01 level, respectively

are correctly responding to these signals in that the loan descriptions with more positive signals are less likely to default.

References

Agrawal, A. K., Catalini, C., and Goldfarb, A. (2011). The geography of crowdfunding. Technical report, National Bureau of Economic Research.

Agrawal, A. K., Catalini, C., and Goldfarb, A. (2013). Some simple economics of crowdfunding. Technical report, National Bureau of Economic Research.

Ahlers, G. K., Cumming, D., Günther, C., and Schweizer, D. (2015). Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice*.

Belleflamme, P., Lambert, T., and Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, 29(5):585–609.

Berger, A. N. and Udell, G. F. (1995). Relationship lending and lines of credit in small firm finance. *Journal of Business*, pages 351–381.

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2013). Which news moves stock prices? a textual analysis. Technical report, National Bureau of Economic Research.

Cassar, G. (2004). The financing of business start-ups. *Journal of Business Venturing*, 19(2):261–283.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5):1367–1403.

Cosh, A., Cumming, D., and Hughes, A. (2009). Outside enterpreneurial capital*. *The Economic Journal*, 119(540):1494–1533.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.

Dougal, C., Engelberg, J., Garcia, D., and Parsons, C. A. (2012). Journalists and the stock market. *Review of Financial Studies*, page hhr133.

Engelberg, J. E., Reed, A. V., and Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics*, 105(2):260–278.

Gao, Q. and Lin, M. (2013). Linguistic features and peer-to-peer loan quality: A machine learning approach. *Available at SSRN 2446114*.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300.

Gentzkow, M. and Shapiro, J. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Gilbert, E. and Karahalios, K. (2010). Widespread worry and the stock market. In *ICWSM*, pages 59–65.

Hancock, J. T., Landrigan, C., and Silver, C. (2007). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 929–932. ACM.

Kim, K. and Viswanathan, S. (2013). The experts in the crowd: the role of reputable investors in a crowdfunding market. *SSRN Electronic Journal. Available at: http://ssrn. com/abstract*, 2258243.

Kuppuswamy, V. and Bayus, B. L. (2014). Crowdfunding creative ideas: The dynamics of project backers in kickstarter. *UNC Kenan-Flagler Research Paper*, (2013-15).

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Mitra, T. and Gilbert, E. (2014). The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 49–61. ACM.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, 29(1):1–16.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Taddy, M. (2013a). Measuring political sentiment on twitter: Factor optimal design for multinomial inverse regression. *Technometrics*, 55(4):415–425.

Taddy, M. (2013b). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Tweedie, F. J. and Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

A Appendix: Variable Description

Variable descriptions are from the Lending Club data dictionary where applicable.

Variable	Description
LOAN AMOUNT	The listed amount of the loan applied for by the borrower.
TOTAL FUNDING	The total amount committed to that loan at that point in time.
INVESTOR FUNDING	The total amount committed by investors for that loan at that point in time.
TERM	The number of payments on the loan. Values are in months and can be either 36 or 60.
INTEREST RATE	Interest rate on the loan.
INSTALLMENT	The monthly payment owed by the borrower if the loan originates.
INCOME	The annual income provided by the borrower during registration, measured in $1,000s.
DTI	A ratio calculated using the borrowers total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
FICO	The average of the last upper and lower boundaries of the range of the borrowers FICO.
CHARGED OFF	An indicator variable equal to 1 if the borrower defaults on the loan and 0 otherwise.