Loan Fulfilment In Peer-to-Peer (P2P) Lending As A Measure Of State-Level Economic Health

ECON-GA 3200 Special Project in Economic Research

Professor Michel Leonard

Spring 2017

Kent Bhupathi | Asad Sami | Joon Kim[1]

Abstract

In this paper, we address the following question: can loan fulfilment in the peer-to-peer lending market be an effective measure of state-level economic health in the US? It is our belief, that when aggregated, these loans have the potential to explain macroeconomic forces because the very nature of these loans represent a truly free movement of capital. It is our understanding for such economies that there will be positive feedback across positive variables — e.g. higher output will lead to better credit conditions, as would higher rates of employment, etc. However, our findings clash with our expectations — the direction of all 'economic health' independent variables turn out to be the opposite. We explain this clash by discussing the user base of P2P lending and elaborating on how and where these loans are used.

[1] This paper has been made possible by the support of New York University. We would like to express our gratitude to Professor Leonard for providing us with his valuable feedback and advice.

I. The Question

Can loan fulfilment in the peer-to-peer lending market be an effective measure of state-level economic health in the US?

Peer-to-peer lending, abbreviated P2P lending, is the practice of monetary lending to individuals or businesses through online services that match lenders directly with borrowers[2]. Due to greater risk aversion in financial services in recent years, that has caused greater selectivity in conventional bank lending[3], the market for 'alternative finance'[4] began to gain popularity at its expense, which paved the way for P2P lending companies to start-up quite rapidly. Moreover, on this note, a finding by Transparency Market Research suggests that the global P2P lending market will be worth $898 billion by the year 2024, which would be an increase of roughly 3350%, since 2014. In this, the P2P market embodies a great deal of potential going forward, from which we have drawn inspiration for this project.

P2P companies started forming rapidly, following the 2008/09 financial crisis, because banks were no longer viewed as a reliable source for loans. P2P platforms offer major advantages over established banks, and their innovative use of technology (software-as-a-service, SaaS) provides greater transparency, flexibility and convenience. Additionally, executives from traditional financial institutions are joining P2P companies as board members, lenders and investors, indicating that the new financial model is establishing itself in the mainstream. Naturally, the implications of such a phenomenon on the national economy are interesting to study. This paper sets its focus on a simple loaning characteristic of the P2P model — namely, trying to answer whether the likelihood of loan fulfilment[5] can have economic consequences beyond its niche market.

The largest P2P company in the US is Lending Club, which funded $24.6 billion worth of loans by the end of 2016, and holds approximately 55% of P2P market share within the US, with several more such firms up-and-coming to expand this market[6]. We see the P2P market as a 'loaning system', like any other, and consider analyzing the potential link of such a system with economic health. The relevance of the P2P lending market is evident in its exhaustive array of purposes, where the

[2] This type of lending can be construed as a customer-to-customer (C2C) type lending, often present in micro-loan scenarios.
[3] *HSBC Chief Warns of Growing Risk Aversion Among Bankers*, Margot Patrick, 2014.
[4] I.e. financing from external sources other than banks or stock and bond markets (*Small Business Association*).
[5] Loan fulfilment means that the loan has been paid off in full. In opposition to testing for the rate of default (*a more conventional research interest*), we were interested in analyzing a positive spin, *so-to-speak*, of the final status of a loan, namely whether a loan was successfully paid off in full, because we wanted to understand how this unique dataset could exhibit positive feedback.
[6] The second largest P2P firm within the US market is Prosper.com, which holds roughly 20% of market share.

loans are used to pay off anything from student loans to credit cards, refinancing options to home improvement, and even car financing to home buying.

Fortunately, Lending Club makes all its data available to the public, which can be used to answer questions that will enhance our understanding of consumer behavior and, consequently, economic health. By *economic health*, we mean conventionally aggregated and averaged, regional economic conditions, such as unemployment, consumption, output, among others.

Is there any potential in the P2P lending market to explain macroeconomic forces? Given past sensitivities in the US's credit market (credit crunches) and their impact on the national economy[7], what role could the P2P lending market play in the economy? Can this market be used to analyze yet another credit phenomenon, if any? To what extent is this market representative of the entire consumer credit market? These inquiries of interest, in addition to the central question at the beginning of the paper, helped to give this project a methodological focus, which manifested into two phases:

1) Phase one looks at the probability of a P2P loan being paid off in full — in that it analyzes the determinants of loan fulfilment. Firstly, a pool of financially pertinent variables to loan performance is selected from Lending Club's expansive dataset (see data section), and then probabilistic econometric techniques (see methodology section) are used to estimate the likelihood that each loan within the dataset would be paid off in-full.

2) In phase two, we aggregated and averaged these likelihoods, by US state and into a panel-data form, so that we could econometrically test them against conventionally used state-level economic conditions, to check for consistency in economic intuition, in a macroeconomic context.

II. Expectations

P2P lending eliminates financial intermediaries (like banks), similarly to how eBay removes the middleman between buyers and sellers. To simplify our understanding of P2P, we like to describe it as The Economist describes it: 'From the people, for the people'[8]. An immediate consequence of credit as easy as this is that it increases spending, thus increasing income levels across the economy. This, in turn, leads to greater productivity and faster growth. It also, however, leads to the creation of debt cycles.

Intuitively, good credit leads to better economic conditions. This is an immediate expectation. Better loan fulfilment would mean that credit is performing well (nominally), in the sense of demand for loans. P2P lending can account for a

[7] *Impact from credit crunch will be huge, study says*, Greg Robb, 2008, MarketWatch.
[8] *From the people, for the people*, 2015, The Economist.

smoother flow of money through the economy to ensure that periodic starts and stops are not affected by variations in the cash flow. Furthermore, it ensures smooth operation for an individual. Hence, we first postulate that higher likelihoods of loan fulfilment will be linked to better economic indicators (at the state level).

It is also important to study the 'noise', both qualitatively and quantitatively. How are people of a state selecting between different loan instruments? Who are these people? Which states are prevailing in P2P lending relative to other states? Is it truly too simple-minded to expect that positive performance of a loan is positively related to output?

It is our belief, that when aggregated, these loans have the potential to explain macroeconomic forces because the very nature of these loans represent a truly free movement of capital. It is our understanding for such economies that there will be positive feedback across positive variable — e.g. higher output will lead to better credit conditions, as would higher rates of employment, etc. Table 1 and 2, within the appendix, outline the expected direction of each variable coefficient (see data description section for the variables).

III. Findings

i. Phase One

The objective of phase one is to obtain the most accurate predictive probabilities for loans to be used in the second phase analysis — the economic health viability (state-level) phase. We implemented the IV (instrumental variable) probit model to obtain the most accurate predictive probabilities. Please refer to the methodology section for more detail.

Table 3 in the appendix states our findings for phase one. The result of the IV probit model coefficients cannot be interpreted directly from the output. Instead, we examine the marginal effects of the probit, to understand the impact of each variable on the dependent variable, holding other variables constant.

We obtained the results as expected, in that the signs of all variables match conventional credit intuition. For example, a year increase in the length of employment increases the probability of a loan being paid in full by 0.9%.

We first statistically checked, through hypothesis testing, all thirty-nine of Lending Club's variables (see data), to examine their ability to predict the probability that a loan is paid in full. However, most of the variables turned out to be statistically insignificant. There are some possible reasons for this. Firstly, Lending Club relies on people's honesty in providing some of their personal information, such as monthly income. In some cases of personal reporting, people would note their income to be at Lending Club's maximum income category, over 1 million dollars,

which cannot possibly be true, given the context. In this, monthly income was never a viable variable to use; and, in the case of when we did use it, our econometric models were unable to converge.

As the most interesting finding from this result, owning a home does not affect the probability of loan fulfilment. On the other hand, owning a home as a mortgage increases the probability. One reason for the positive relationship is that, since the financial crisis, the standards of issuing mortgages became stricter, making them available to those who have relatively strong financial records. Therefore, owning a home as a mortgage positively affects the probability of a loan meeting fulfilment.

After conducting the IV probit model, we calculated the area under the ROC (receiver operating characteristic) curve[9]. For our model, we have an area under the ROC of 0.6693, which classifies well given our limited data. Since we have the satisfying results, we calculated the predicted probabilities for each loan (see methodology section) to be used for the second phase panel analysis.

ii. Phase Two

Table 4 in the appendix reports our findings. Note that this section simply states our findings whereas the next section elaborates on the interpretations and their interplay with our expectations.

Our findings of the second phase model did not coincide with our expectations – the signs of the all variables came out to be the opposite to what we expected. As mentioned in the expectations section, we would expect that the probability of loan fulfilment would be positively related with healthier economic conditions. However, in our case, this came out to be not true. For instance, a 1% increase in unemployment rate increases the probability of a loan being fully paid by 0.42%. Similarly, a $1 million increase in GDP decreases the probability of a loan being fully paid by 1.88%. Economic confidence index turned out to be statistically insignificant at the 10% level. Moreover, a unit increase in price parity index increases the probability of a loan being fully paid by 0.14%.

[9] The ROC curve illustrates the performance of a binary classifier by plotting the true-positive rate and false-positive rate at each threshold. The area under the ROC curve gives an idea of whether the model is correctly classifying 1 and 0, where the value of the area under the ROC curve ranges from 0.5 to 1, with 0.5 being random choice and 1 being perfect classification.

IV. Interaction Between Findings and Expectations

Why do our findings clash with our expectations?

One approach to explaining this difference is through a more thorough understanding of the P2P user base, which we call the 'Positive Crowd Out':

We must acknowledge that a certain population demographic uses P2P lending, namely those who either cannot borrow from the traditional financial institutions or those who want to take out a micro-loan. These users are vulnerable to the macroeconomic conditions, as well as their personal, financial strength. During economically thriving times, the users with relatively strong financial status will tend not to borrow from the P2P lending market, due to its relatively high interest rates. So, when GDP is rising, those with relatively strong financial status will move to the traditional banking sector for a loan, whereas the remaining users of the P2P lending market will be those who have relatively worse financial status. Once those who are financially worse off are the only active users of the market, the probability of loan fulfilment status will then decrease regardless of the current economic conditions. Regarding this explanation, healthier economic conditions thusly do not directly cause the probability of the loan fulfilment positively, but rather indirectly removes the relatively financially strong users out of the market. Therefore, the probability of loan fulfilment status and macroeconomic indicators, *at least at the State-level*, move in the opposite directions.

Another approach to explaining these differences can be attributed to the *use* of these loans, which we call 'Re-debting':

Per Lending Club's dataset, approximately 70% of the loans are used for refinancing — that is, users are simply trying to replace an existing debt obligation with another debt obligation, under different, favorable terms. Given this refinancing phenomenon, loan fulfilment and macroeconomic forces could potentially be ambiguous, since there is merely debt substitution (*wealth movement*), rather than wealth creation.

Moreover, financial intuition would lead us to assume that users of the P2P lending market would pay off their obligations more readily as their income(s) increased. However, in our experiment, this does not seem to be the case. Most of the users tend to have a relatively weak financial status. Given this, and considering our regression results, it is our understanding that with the increase in income, P2P lending users will more likely spend their extra income on home/life essentials, rather than spend it to pay off their loans, in the short-run. Combined with the 'Positive-Crowd Out' effect, 'Re-debting' behavior has the potential to amplify the chances of loan fulfilment moving in the opposite direction to economically prosper environments. Therefore, it appears to be the case that our unexpected results are not without reason, and instead present to us a consumer behavior phenomenon that is rarely observed, in part due to the selection biases in lending by traditional banking.

V. Data Description

i. Phase One

Our research methods can be broken down into two different phases — or, models: the first model (an instrumental variable probit) only makes use of data from Lending Club, whereas the second model (a series of fixed and random effects panel-based regressions) makes use of the estimates from the first model as well as macroeconomic data from various, online sources.

For the first model, we use Lending Club's data, only. The original dataset from Lending Club contains approximately twenty million records of loans (a panel dataset for each loan — over time), ranging from 2007 to 2017, and thirty-nine characteristic variables that encompass initial borrower inputs and investor decisions. Since the data was in a panel-data form, we reduced the size of the dataset by taking the first month of each loan. We chose to only consider the first month of data for each loan because:

- There are twenty million rows in the original data set, and that is computationally expensive to analyze repeatedly.
- Almost all borrower inputs were constant throughout the life span of the loan. For example, a borrower's FICO score was fixed at the time the loan was initially issued. In this, we would not gain much information from using Lending Club's panel data, but rather its cross-sectional data.

Furthermore, we discarded all loans that are currently performing because we do not know the future outcome of these loans. After the abovementioned data clean-up, the total number of observations was reduced to 550,551 observations. Of all the thirty-nine descriptor variables, we chose to use only six variables: interest rate, principal amount, employment length, and two dummy variables for types of home ownership. These variables are explained in the following points:

- Interest Rate: It is the fixed interest rate for each loan. The interest rate is set by Lending Club, based on its selection criterion.
- Employment Length: it measures the total number of years that a borrower has been with its current employer — it ranges from 0 to 10.
- Home Ownership, 'Own': a dummy variable indicating whether the loan issuer owns a home or not. It takes the value of 1 if the issuer owns a home and 0 otherwise.
- Home Ownership, 'Mortgage': a dummy variable indicating whether the loan issuer owns a home through a mortgage. It takes the value of 1 if the issuer owns a home as a mortgage, and 0 otherwise.
- FICO: the FICO score is a personal credit score, which ranges, in this dataset, from 610 to 850.
- Public record: total number of major public records held by the borrower. Includes foreclosures, bankruptcy and civil judgements. It ranges from 0 to 9.

ii. Phase Two

Instead of using all fifty states, we decided to use only seventeen states that make up 75% of the total loans. The reason for only using these states is to make the dataset more manageable to work with in the second phase. In addition, some loans have not been issued from some states until 2014, which would heavily disrupt the panel analysis for the second phase. Therefore, we decided to only use seventeen states. They are: Arizona, California, Colorado, Florida, Georgia, Illinois, Maryland, Massachusetts, Michigan, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Texas, Virginia, and Washington.

The purpose of the second model is to test for the correlations between the probabilities of loan fulfilment (the dependent variable) and an array of state-level economic variables (the independent variables). After obtaining the estimated probabilities for each loan, *in phase one*, we aggregated the loans by each of the seventeen abovementioned states and then took the average for each month. Once the data was structured for phase two, we had an unbalanced panel dataset, ranging from 2007 to 2017, for each state. Please note that we did not use all the listed variables for the reasons explained in the Findings section.

The state-level 'economic health' variables we used are:

- Unemployment Rate: it measures the monthly unemployment rate. We got the data from the Bureau of Labor Statistics.
- GDP: Gross Domestic Product for each state. Since GDP is published quarterly, we decided to keep it constant throughout each quarter. The first quarter of 2017 has not been published. We used the data from the Bureau of Economic Analysis.
- Economic Confidence Index: it is a survey that measures people's view of current and future economic conditions. The minimum and maximum theoretical values of the index are -100 and 100, respectively. A value above zero indicates that more people have a positive than a negative view of the economy; values below zero indicate the negative economic view. The survey started in 2008, and therefore we miss values of 2007. We obtained the data from the Gallup.
- Regional Price Parities Index: measures price level differences across regions for one period. It compares the average price level of a state with the national average price for all states. Having 100 as the national average price level, the index compares the price level across the states. If a value is above 100, it indicates the price level of a State at that period is higher than the national average price level. We used 'All items' to calculate the price index, which includes goods, rents, and others. Since the data was available on annual basis, we kept it constant throughout each year. Also, the data was only available between 2008 and 2014. The data was obtained from the Bureau of Economic Analysis.
- Median Income: measures the quarterly median income. Because the income tends to show linear movement throughout time, we interpolated between

quarters to obtain monthly data. The data was available up to fourth quarter of 2015. It was obtained from the Henry J. Kaiser Family Foundation.

- Building Permits: Privately-owned housing units authorized by building permits. Think of this as housing starts. The original data was semi-annual basis. So, we linearly interpolated the missing values. The data was obtained from US Census Bureau.

VI. Methodology and Regressions

i. Phase One

The reason for using the IV (instrumental variable) probit model is to account for the possibility of endogeneity effects for one of the exogenous variables, namely the interest rate. Since the presence of endogenous variables, as the independent variables, gives biased result, we decided to use the IV method within our probabilistic regression.

The interest rate is a function of several factors, such as financial history records, current monetary policy, etc. Since the inclusion of the FICO score and the interest rate in the same model causes multicollinearity concerns, we decided to use the FICO score and the number of public record as the instruments for the interest rate. From this, we could more accurately analyze the effects of the independent variables and obtain a more accurate predictive probability model. Below is the mathematical model for the IV probit model:

$$Y_{2i} = 1(X_i\beta + y_{1i}\alpha + \epsilon_i > 0) \quad \text{(I)}$$

$$Y_{1i} = X_i\gamma + Z_i\theta + v_i \quad \text{(II)}$$

Where,
X_i is the control variables
Z_i is the instrument variables
Y_i is the variable of the interest.
To calculate the IV probit model, firstly calculate equation (II), and estimate Y_{1i}. Then, use the estimated Y_{1i} to calculate the equation (I). In this way, the endogeneity problem is minimized.

ii. Phase Two

When there is reason to believe that differences across entities have influence on the dependent variable (*as is expected to be the case with States that share a common economic union*), then the random effects panel-data model becomes appropriate. The general formula for the random effects model is displayed on the next page:

$$Y_{it} = \beta_n X_{nit} + \alpha + u_{it} + \epsilon_{it} \quad \text{(III)}$$

Where,
Y_{it} is the dependent variable, where i = states, which are: Arizona, California, and t = time;
X_{nit} represents one independent variable, where n = independent variable number;
β_n is the coefficient for a respective independent variable n;
u_{it} represents the between-effects error and ε_{it} represents the within-effects error, both of which are embedded within the independent variable coefficients.

Since all individual characteristics that may or may not influence the predictor variables need to be specified before running the model, the problem then becomes that some variables may not be available (*substituted with proxies*), which therefore lead to omitted variable bias in the model, should such variables be removed. The example by which this was true to our research was the price_index variable, which substituted for purchasing parity across States, and although a statistically insignificant variable, it provided structure for the GDP to remain significant.

VII. Appendix

Table 1: Expected Dirction of Variables (Phase One)

Variables	Expectations	Result
Interest Rate	-	-
Principal amount	-	-
Employment length	+	+
Home_Own	-	+
Home_Mortgage	+	+

Table 2: Expected Directions of Variable Coefficients (Phase Two)

Variables	Expectations	Result
Unemployment Rate	-	+
GDP	+	-
Economic Confidence Index	+	-
Price Parity Index	-	+
Median Income	+	-

Table 3: Marginal Effect IV Probit Output

Variable	Coefficient	Robust Std. Error	Z-score	P-value
Interest Rate***	-12.243	0.873	-140	0.000
Principal amount***	-2.08E-06	2.43E-07	-8.54	0.000
Employment length***	0.009	0.0005	18.37	0.000
Home_Own	0.008	0.006	1.26	0.209
Home_Mortgage***	0.117	0.005	28.01	0.000
Constant***	2.278	0.012	194	0.000

Note: * indicates significance at the 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level

Table 4: Panel Output

Variable	Coefficient	Robust Std. Error	Z-score	P-value
Unemployment***	0.0042	0.0011	4.01	0.000
GDP**	-1.88E-08	-9.36E-09	-2.01	0.044
Economic Confidence Index***	-0.0012	0.0001	-7.19	0.000
Price Parity Index	0.0014	0.0009	1.59	0.111
Median Income**	-1.67E-06	7.40E-07	-2.26	0.024
Constant***	0.6476	0.0591	10.96	0.000

Note: * indicates significance at the 10% level, ** indicates significance at 5% level, and *** indicates significance at 1% level