Network Analysis of Peer-to-Peer Lending Networks
CS224W Project Final Paper

Ankit Kumar Matan Zinger

ankitkr@stanford.edu mzinger@stanford.edu

1. Introduction

Peer-to-peer lending marketplaces on the web have been growing over the last decade, providing a platform for individuals to borrow and lend money directly from and to one another. Research has been conducted over many aspects of these websites, some of which view these online communities as social networks, and analyzing different characteristics and dynamics in these networks. The problems of interest in this field includes predicting success of a listing (funding), probability of loan repayment, identification of good borrower characteristics, risk minimization, investment diversification etc. We have chosen to focus on this field as well.

In our project we analyze the peer to peer lending network of prosper.com, a leading lending website, utilizing different network analysis techniques over different signals. We later discuss the insights and contributions made by the approaches as well as additional directions that could be explored. There are three main problems of interest we explore; first is identifying listings that are going to be fully funded. Second, one may be interested to identify the level of risk involved in a listing, in term of probability of repayment. That helps a lender to determine the appropriate interest rate, as well as diversification of their portfolio. Third, lenders could benefit from a recommender system that could intelligently suggests listings to them catered to their style of investment.

We are interested in extracting features from the lenders network, in order to work on these three problems. We intend to analyze the lender networks to predict listing funding, loan repayment, and develop a personalized listing recommendation mechanism.

2. Prior Work

Krumme and Herrero[1] describe a study of 350,000 loan listings from prosper.com, along with the information about their bids, and the social connections (“friends”) between borrowers. Alongside a study on the dynamics of biddings, that shows a herding behavior, the authors also study the impact of social connections. They show that being a member of a group increases both the probability of a borrower to successfully receive funding for her listing as well as the probability of the loan to be repaid. Having a large number of friends (degree greater than 20) in the social graph is shown to be correlated with high credit score (therefore may indicate that a borrower is trustworthy). The authors note that additional work is required to further understand the interaction of the social and economic networks.

Lin et. al.[2] studied the impact of different social network characteristics in Prosper.com on the probability of a listing to be funded, on the interest rate, and on the probability of the loan to default. The analyzed networks are the graph where edges express ‘friendship’ relation between borrowers, and the graph that represent group memberships. While structural aspects (such as node degree, connectivity, and clustering coefficient) of these social graphs were found to be of less significance to the listing outcome, relational aspects of the social connections found to be significant.

For example, for borrowers of lower credit brackets, their number of friends who are lenders, and of which the number that chose to bid or not to bid on the borrower’s listing, were all found to have a significant correlation with all three studied parameters of the listing outcome.

The paper concludes that a borrower’s social capital is not measured by the size of its network, but rather by its quality. This paper also observes the information gain from the group memberships of the borrowers and shows correlations with group size, group type (alum, religious) with funding probabilities.

Both these articles show that the borrower's social relationship can serve as an effective indicator of her listing's outcomes, a fact that motivates the work we present in this project.
Ceyhan, Shi and Leskovec[3] have studied the temporal dynamics of the bidding in a peer to peer lending marketplace (the shape of the accumulated bidding curve over time), and show that it carries information of whether a listing will be funded successfully, and whether it would be paid back or default. First, the authors observe the skewed trajectory of a listing's accumulated bids over time, and show that a listing experiences herding when it's accumulated bids exceed the required amount, and close to its deadline. The paper establishes that the time series bid pattern of listings can be fit closely by a sigmoid which parameters (rate at which the amount accumulated in the listing and the time at which the jump happens) have are correlated with the listing's outcome. They extract these features and use them in a classifier, thus improving the accuracy in predicting both lising funding probability and the probability of loan to be repaid, over the baseline that only used the borrower's credit features.

3. Motivation

Previous research conducted on peer-to-peer lending marketplaces have focused mainly on determining the success probability of a loan by analyzing signals regarding the quality of a borrower, such as her credit rating, debt to income ratio. Some work studies the social network roles of the borrowers' friends in order to determine their credibility. The paper by Leskovec et. al. have focused on drawing signals from lenders' perception of a loan by looking at their reactions (bidding dynamics) to the listing over the duration of time it was on auction. By adding the social signals described in article[2] to the model described by Leskovec et al., we might be able to gain additional information and improve the prediction accuracy.
In the aforementioned literature, we did not recognize a discussion on analyzing the joint network of both lenders and borrowers. One can see multiple networks in a p2p lending market - first being the social network of all borrowers and lenders, and second being the economic network formed by bidders bidding on listings posted by networks. Furthermore, we can formulate implicit networks by looking at co-bids, i.e. networks where lenders are related where they have bid on same listings multiple times in past. There is information to be gained from these networks that can be used in interesting ways. In this project we plan to formulate such networks and look at what inferences we can draw from them.

4. Dataset

For this project, we examine the lending data available from the peer to peer marketplace Prosper.com. We have the bidding data from prosper from Nov 2005 to Aug 2009 comprising of about 6 million bids, 900k members, 230k listings. Here is an example listing in the Prosper network:

```
Amount funded: 1500              Amount requested: 1500           Bid count: 7
Borrower rate: 0.059             Lender rate: 0.0595
Borrower maximum rate: 0.06      Bid maximum rate: 0.059
Credit grade: AA                 Debt to income ratio: 0.01747    Has verified bank account: 1
                Description: “I’m asking for a loan to pay off  credit card debt.”
Borrow city: SAN FRANCISCO       Borrower state: CA
Category: 0                      Duration: 14                     Status: Completed
```

Out of these listings about 20k have been funded. Out of the funded listings about 5k listings were paid back in full and about the same number defaulted.
There are two networks of interest in the lending marketplace. First, there is the explicit social network where borrowers and lenders have friends and can endorse each other. Second is the implicit bidding network, where each directed edge from a lender to a borrower represents a lender bidding on at least one of the borrower's listings. We present the degree distribution plots in both these networks. The social

network is very sparse (mostly tree like); few nodes have high degrees while most nodes have degree of 0 or 1. Figure 1 shows the degree distribution of the social network and shows it to be in agreement with the power law. In figure 2, we show the in and out degree distributions of the implicit 'bidding' network. The out degree distribution i.e. the distribution of number of people lender has bid on clearly follows power law (after sufficient x_{min}), while the indegree distribution of borrowers, i.e. how many people have bid on this borrower probably has a power law distribution with exponential tail cut off as evident from their CCDFs (Figure 3).

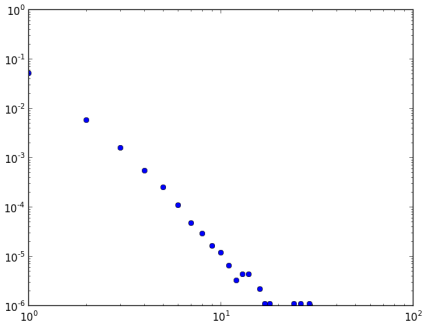

Fig 1. Degree distribution of social network

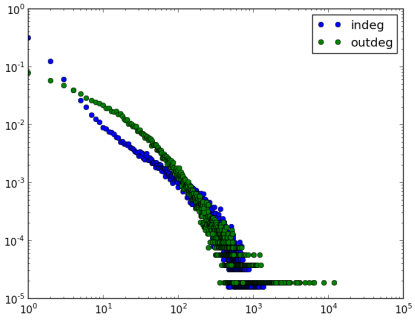

Fig 2. Degree distributions of the bidding network

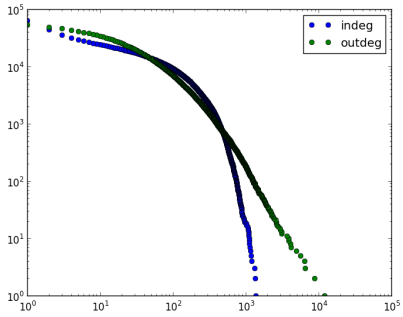

Fig 3. CCDFs of the bidding network

5. Predicting Listing Funding / Loan repayment

5.1. Our Approaches

The general goal we set in our project is learning different social and network characteristics of the peer-to-peer lending marketplace. Those would be used to extract features that would allow us to predict both the probability of a listing to get funded, and furthermore - the probability of a loan to get repaid in full. This section describes multiple approaches we have tested for that goal (some based on previous work, e.g. bid dynamics features). We describe the set of features extracted in each of these approaches.

5.2. Baseline Features

As a baseline for our model, we used the following features: maximum interest rate the borrower is willing to accept, the borrower debt-to-income ratio, total amount requested, whether the borrower is a homeowner, the length of the listing description. These are the same 5 base features used by Ceyhan, Shi & Leskovec[3].

5.3. Features from Crowd Bidding dynamics

This corresponds to the features described by paper[3]. We tried to reproduce the analysis of the herding patterns associated with listings, and extracting relevant features. For a listing L, let a_i be the amount bid by the i^{th} bid on the listing. Assuming the time scale of 0 to 1 where time of 0 means the time when the listing received the first bid and time of 1 as the time when listing received the last bid, we can look at the graph of cumulative amount bid on the listing as a function of time. i.e. plot sigma a_i (i from 0 to t) A where A is the total amount accumulated on the listing L at t = 1.

We plotted all such graph of time series accumulation of bid amount for every listing. We then try to fit all the curves by the sigmoid function parameterized by q and Φ. i.e. $y = \frac{1}{1 + e^{-q(t-\varphi)}}$

The hypothesis is that the parameters of this curve fitting, q i.e. the measure of herding / how quickly the bids are accumulating money on the listing and Φ i.e. when the herding starts have information on what the market feels about the listing and can help predict the funding probability / loan repayment probability.

We wrote a logistic regression classifier that takes into account these features and we got improvements in classification accuracy as described in table. Figure 4 shows the variation of q vs Φ for with regards to

funding, and figure 5 shows the same for defaulted vs paid back loans. Points in blue have positive outcomes, red are negative.

From figures 4 and 5, it appears that the parameter q is not very significant in determining probability of funding / default. However, the parameter phi seems to be of more impact. The first plot shows that the listing for which the herding starts late in their life cycle have higher probability of getting funded. The second plot shows that the loans for which the herding started late in their life cycles have a higher probability of default. The first observation seems to be somewhat counter-intuitive but that is what we have empirically observed from the data.

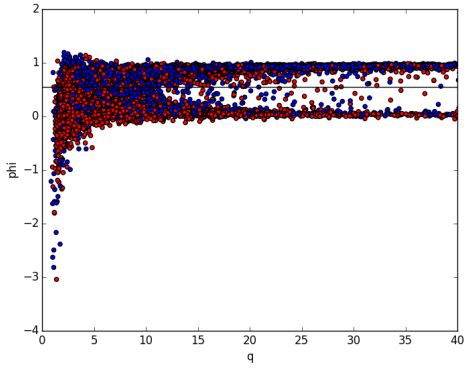

Fig 4. Plot of sigmoid parameters for funded vs non-funded

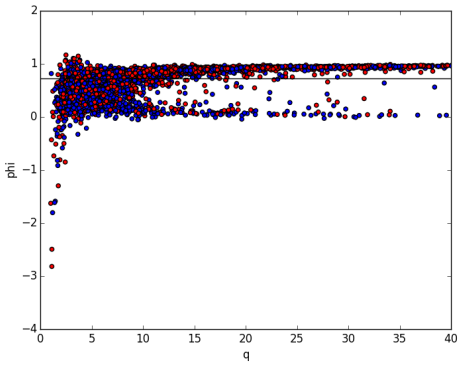

Fig 5. Sigmoid parameters for paid vs defaulted loans

5.4. Detecting Expert Lenders

This is a different approach than paper *[3]* describe. Instead of extracting the crowd's opinion on a listing, we are trying to extract the opinions of "experts". Our hypothesis is that there are multiple types of lenders - some who are layman in the field of risk assessment, and their bids are being influenced mostly by the credit parameters provided per listing, their personal preferences with regards to risk, and the actions of other lenders in the network (as previous work suggests, there is a strong herding behavior). However, we suspect there might be a second type of lenders, that carefully examine listings beyond basic credit properties, and are able to distinguish loans that are likely to be paid or default better than the crowd. The presence (or absence) of these expert lenders in a listing may hint its outcome.

One may expect a correlation between higher interest rate and higher amounts of money that a lender lost due to defaulted loans; however, when examining Figure 6., we observe large portion of lenders with high rates and low fraction of lost money. When examining the distribution of lenders by avg. interest rates (Figure 7), we notice a long tail of lenders with the high rates: the expectancy is 1.69%, the 75th percentile is relatively close (1.93%), but the tail continue all the way to 3.575%. These observations inspire us to detect those lenders with significantly better performance than the network average.

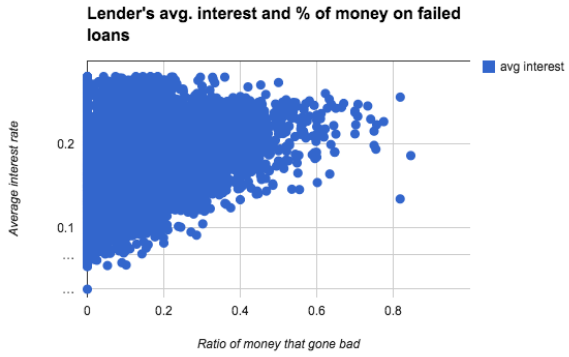

Fig 6. Average lender interest rate, by the fraction of lost money

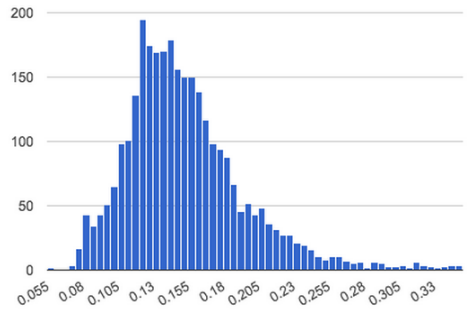

Fig 7. Distribution of lenders by average interest rates

Initially, we have defined “experts” as lenders who have never participated in a loan that was defaulted or late, and that have participated in at least 8 loans that were paid in full. We have learned the group of lenders that correspond to this definition from the training listings set, and added 5 binary features for each listing, to indicate whether at least *i* expert lenders are participating in the listing (for $i \in \{1,...5\}$). That definition outperformed the baseline for high credit grades in predicting listing’s funding, but did relatively poorly in predicting loan repayment (56% accuracy).

The second approach we tried is to define the following vector per each listing $< Bid_m \mid for\ m\ in\ Lenders >$: A binary feature per each of 52,970 lenders, that indicates whether the lender bid on the listing (not necessarily “winning” the bid and ending up participating in the loand). Similar to how term-vectors are used to classify e-mails as spam, we used “lender-vectors” to provide a score for a listing.

Naive-bayes assumes independence between the features, and indeed our hypothesis assumes expert lenders make decisions without being affected by the general lenders crowd.

By training a model over 3728 listings (half were repaid and half were late or default), and testing over 1000 loans that happened later in time, the classifier produced 64% accuracy, without even incorporating the baseline features. By inspecting the top 15 lenders that were scored as most likely to be associated with a successful loan, we noticed that all of them have bid on at least 60 loans, had less than 10% of their previous investments on loans that gone bad, and get at least 0.1 interest rates by average.

We used that as the new criteria for determining expert lenders, and extracted features for having at least {1,2,4,8,16} smart lenders bidding on a listing.

5.5. Borrowers Social Features

The work by Mingfeng et al.[2] has shown a significant impact for what they refer to as “social capital” in predicting listing’s funding and repayment. That term was defined as the social relationships a borrower maintains in the embedded social network of the lending platform, as well as the nature of these friends.

Motivated by the conclusion of the said study, we analyzed that social network of Prosper.com, which as previously discussed is a very sparse tree-like graph, with most members not having any friends, and extracted the following features per each borrower:

The features we extracted are the following: number of friends the borrowers have, number of friends with roles that a borrower has, number of lender friends a borrower has, the number of lender friends who have previously invested in some listing by the borrower, the number of lender friends who have not invested in the lender in the past. We observe that most of the values for these features are 0 as expected.

In order to utilize these features, for each listing we added the social features of the borrower who initiated the listing.

6. Evaluation

6.1 Predicting Listing Funding

For each rating, we have randomly downsampled the set of listings such that a listing expires or being funded with equal probabilities, and randomly split the set of listings into a training set (80% of the data) and a test set (the remaining 20%), both with equal funding probabilities. We trained a logistic regression model for each rating, using different combinations of features. The prediction accuracy results are below:

Rating	# Listings	Baseline	Experts	Social	Crowd	Experts + Base	Social + Base	Crowd + Base	All Features
AA	4450	0.64	0.80	0.61	0.56	0.89	0.64	0.65	0.86
A	4240	0.62	0.85	0.6	0.5	0.88	0.64	0.64	0.84
B	5210	0.63	0.84	0.58	0.54	0.86	0.67	0.68	0.82
C	6710	0.67	0.69	0.56	0.58	0.73	0.66	0.64	0.74
D	6150	0.66	0.59	0.6	0.64	0.74	0.65	0.69	0.73
E	4220	0.67	0.5	0.6	0.72	0.68	0.68	0.75	0.75
HR	4310	0.76	0.49	0.63	0.74	0.77	0.74	0.77	0.77

The expert-lenders feature seem to outperform the baseline for high credit grades (AA, A and B); when combined with the baseline features, an improvement is gained also for the C & D categories. Interestingly enough, the crowd-based features gain improvement over the baseline only for the E credit category, and in that sense they "complete" the expert-lenders signal, which did not prove very effective for this category. When combined, they show an improvement in accuracy over the baseline for every credit grade except HR. The social-network features do not provide any significant improvement over the baseline.

6.2. Predicting Loan Payment

We have repeated the same process for predicting loan payments, having the classification classes be "Payment" or "Default" (ignoring loans in-progress).

# Loans	Baseline	Experts	Social	Crowd	Experts+Base	Social+Base	Crowd+Base	All Features
2910	0.7	0.64	0.55	0.54	0.74	0.71	0.73	0.72

All three signals provide a slight improvement in the testing accuracy, but it seems insignificant in all cases. The expert-lenders signal provided the highest gain in accuracy.

7. Listing Recommendation System

We also tackle another problem in this project - the problem if recommending listings to lenders. The objective here is to come up with a listing recommender system that can provide recommendations to lenders looking for listings to bid on. The idea is to cater the recommendations to the investment style / taste of a lender, and surface the top listings that are currently open for bidding and are most likely to be of interest to them. In this system, we are mainly trying personalizing the listing recommendation system with little emphasis on listings' predicted probability of funding / payment or borrowers features. The idea is to analyze the implicit lender network in the peer to peer marketplace to personalize potential listing recommendations.

Here we analyze the implicit network formed when a member of marketplace bids on a listing posted by another member. In such a directed graph with nodes as members of the marketplace, each directed edge points from a lender to a borrower for *some* listing i.e. there is an edge between members A and B if A has bid on *some* listing posted by B. Please note this graph is a directed graph and not a multigraph.

In such an implicit network, we try the Simrank[4] approach where we try to find lenders similar to a lender to capture the investment taste of a lender which we later use for developing the personalized recommender system. Similar to Simrank, we compute the personalized pagerank of other nodes w.r.t. to a node in the graph and take it as the similarity of other lenders to the chosen lender. More specifically, for any lender l in the network we compute personalized pagerank of all nodes with teleport set {l}, giving us a measure of similarity of other lenders in the network to l. Please note that during pagerank propagation, we take the graph as undirected i.e. ignore the edge direction.
Now, the recommender system works as follows. Given that we maintain a list of top K (20) lenders similar to any lender in the network (list maybe computed after every fixed time interval) at any time, we rank all the open listings at that time by counting the number of bids a listing has from the similar lenders. Ties are broken by favoring listings with greater overall bids. We choose top N (10) listings and show them to the user as personalized listing recommendations.

7.1. Evaluation

Unfortunately there is no *correct* evaluation of our personalized recommendation system as we do not actually have control on the website and cannot actually show users those recommendations and analyze their responses. Here we attempt to evaluate our recommendation system just by the historic data we have (i.e. the data of what was the actual bid a lender made at any time and what our recommendation system would have recommended to him at that time) and splitting it into train and test set. Please note however that this evaluation is not entirely accurate and is biased against by any recommendation system the website might actually originally have.
To evaluate our system, we first randomly sample 4k members of the marketplace (say set T). Now, for any lender l in T, we look at all the times in sequence t_i where he made a bid. We take 80% of those bid times for training and 20% for testing i.e. we take only 80% of the bids to compute the implicit bidding network to compute pageranks (to prevent any bias coming into evaluation by computing pageranks knowing the test bid times outcomes). In such a network we compute the personalized pageranks w.r.t. each of these *test* members i.e. compute top 20 similar lenders to each member in T. Now that we have a recommendation system trained, we evaluate the system on the 20% bid times of the members in T i.e. for any member u in T and any test bid time t, we compute the recommendation list of top 10 listings and evaluate whether the u actually bid on any of the recommended listings at test bid time t.
Our evaluation metrics are accuracy i.e. fraction of times user bid on one of the listings suggested by the recommender system and average click position i.e. the position from top (indexed at 0) where the average click position on our recommendation list is.
We also evaluate two baselines for comparison. The first baseline is the *random* recommender which recommends a list of randomly chosen open listings to the user. The second baseline is the *most-popular* recommender which sorts all open listings at a time by the overall number of bids it has accumulated till now and recommends top listings to any user.
Please note however that for some members, we do not have sufficient data i.e. the top similar lenders have not made sufficient bids to show up 10 recommendations. In such a case, if we compare the accuracy of our recommender system with any baseline - it would just lose because of showing less recommendations. Hence, in addition to the above approach, we also show metrics for an approach where after computing the personalized recommendations as previously explained, we fill up the remaining space with the *most-popular* recommendation scheme as mentioned above.

The following table shows up the metrics we get for the baseline systems and our recommender systems. Our approach is much better than the random recommender and achieves about 10.7% accuracy. Combining our approach with the most popular at tail gives about 24.3% accuracy which is slightly lower

than the most popular system (25.4%). As previously explained this may be because of any recommendation strategy the website may have already had.

Recommendation System	Accuracy	Average Click Rank
Random	1.39%	4.38
Most Popular	25.43%	3.77
Simrank based	10.66%	3.73
Simrank + Most popular	24.34%	3.82

8. Conclusion

We have inspected multiple signals that are available due to the network setting of the peer-to-peer lending platform, signals that are not available in transitional credit systems. We have used these signals to predict the outcome of loans, as well as listings that a lender would most likely choose to participate in.
The dynamics in bidding of the general lenders crowd was proven as significant to predict whether a listing would be funded, mostly for listings published by borrowers of low credit rate. The bidding patterns of lenders who we categorized as experts were effective in predicting funding of listing by borrowers of middle and high credit rates, The social relationships of borrowers did not prove to carry a significant signal for this purpose at any credit rate.
In predicting loan repayment, all three provided a slight improvement over the baseline, but were hardly as effective as they were for the listing funding prediction task. The expert lenders signals provided the highest improvement.
For the purpose of recommending listing, we have built a mechanism based on similarities in the historical bidding choices of lenders. An actual evaluation would have included providing these recommendation to users in real time, and measuring the changes in the rate they are bidding on recommended bids, versus the same metric when surfacing recommendations from the baseline algorithms. However, since our analysis was only based on passive analysis of the public data provided by prosper.com, we could not have experimented with presenting recommendation. Therefore our evaluation analyzed the datasets as if lenders were exposed to all open listing at the network at any point in time, and found a significant improvement over the baseline random recommendation. We hypothesize that the it did not outperform the “most popular” algorithm might be due to prosper.com recommending listings to lenders according to the “most popular” scheme.

9. Future Work

Both in the recommendation mechanism, as well as the expert lenders identification, we have considered a lender to be connected to a borrower or a listing if it had bid on it. In that fashion, a lender who bid $50 on a listing they considered risky but with high gain potential, and a lender who bid $500 on that listing since they found it attractive, will be considered as having similar preferences. Possible extensions of our work may involve using weighted connections, where the weight could be determined by the bidding amount, and also by the suggested interest rate, to better extract information from these relationships.

Furthermore, while identifying similar lenders for listing recommendations, we have considered the bidding network as an undirected graph and have coalesced multiple bids between a lender - borrower pair into a single edge. It can be interesting to see the performance of considering the network as a directed multigraph as number of bids between a lender - borrower pair has extra information.

Third, we could not observe a significant impact for features directly from the social network, but we feel they still have a potential to be proved as significant. One potential issue being the fact that most members of the network hardly maintained any social relationships in the network, and so one extension of our work could be testing the same results when limiting only to listings by borrowers with at least one social connection. Second potential issue might be the fact that the features we extracted do not have linear connections, and so feature engineering might be required to represent them in additional fashions.

Individual Contributions

Data pre-processing, extraction of features from the crowd-bidding dynamics, the recommendation system and its evaluation were done mostly by Ankit. The analysis of expert lenders, prediction of funding by credit grade, and prediction of loan repayment were done mostly by Matan. The extraction and evaluation of social network features, as well as the writing of different documents submitted during the project were done by both.

References

[1] Lending behavior and community structure in an online peer-to-peer economic network (Krumme & Herrero, 2009)
[2] Judging Borrowers by the Company They Keep: Social Networks and Adverse Selection in Online Peer-to-Peer Lending (Lin, Viswanathan, Prabhala, 2009)
[3] Dynamics of Bidding in a P2P Lending Service: Effects of Herding and Predicting Loan Success (Ceyhan, Shi, Leskovec, 2011)
[4] SimRank: A Measure of Structural-Context Similarity (Glen Jeh, Jennifer Widom, 2002)