Analysis of Peer-to-Peer Money Lending Network

Yen-Ting Liu
Stanford University
eggegg@stanford.edu

Lun-Kai Hsu
Stanford University
luffykai@stanford.edu

Jocelin Ho
Stanford University
jocelin@stanford.edu

ABSTRACT

Online money lending has become more popular these days. Instead of borrowing money from bank, now users can borrow money from other users with interest rate proposed by himself. Users can also invest in other users, bidding with interest rate he proposed as well. In this paper, we analyze the properties of this money lending network. First, we analyze the degree distribution, robustness/connectivity, PageRank, and HUB/AUTH score. Then, we model the borrower and lender's behavior by extracting and visualizing features. One of the interesting insight we found is that users bidding with very high or very low interest rate usually don't bid frequently. Finally, we use machine learning tools to predict several properties, including actual amount borrowed, credit level, and interest rate. According to the result, we found that the actual amount borrowed usually has not much to do with the money proposed by the borrower. It depends more on the previous record and credit level of the borrower. Finally, we conclude the paper with conclusion and future work.

INTRODUCTION

Prosper is America's first peer-to-peer online money lending network, with more than 2 million members and over $2,000,000,000 in funded loans. Borrowers choose a loan amount, purpose and post a loan listing. Investors review loan listings and invest the ones that meet their criteria. Once the process is complete, borrowers make fixed monthly payments and investors receive a portion of those payments plus interests directly to their Prosper account.

In this project, we would like to analyze the structure of this peer-to-peer money lending network. There are several problems that we are interested in: 1) Discover the internal structure of the network. For example, the degree distribution, PageRank of each user, as well as the robustness of the network. 2) Predict the characteristics of the users (node). Can we predict the credit level of a user by network structure? 3) Predict the feature of the lending relationship (edge). For instance, how possible it is for user A to lend money to user B, of what amount would the borrower receive? and what might be the interest rate of the transaction? By analyzing the network structure, we could understand the underlying relation among the users, and also make predictions to new transactions.

The goals of the project are listed below:

1. Analyze the network structure
   (a) Standard network features: degree distribution, cluster coefficient, and network diameter
   (b) Detect the underlying clusters among users
2. Predict the edge synthesis process
   (a) Predict the amount of money a user would lend to/borrow from others
   (b) Predict the amount of money a user would bid
   (c) Predict the interest of a certain transaction
   (d) Analyze the relation between purpose and the amount of money/interests/success rate
3. Analyze the characteristic of users
   (a) Predict users credit score
   (b) Analyze the relation between success rate and the users credit score Rank the user by PageRank/HITS

RELATED WORK

To analyze our money lending network, we selected papers that present algorithms for discovering internal clusters and predicting edges in an evolving network. Ceyhan et al.[1] analyzed the same money lending dataset as our work. The work took temporal effect into account, predicting how interests, and the probability of winning a bid evolved over time. In addition, it analyzed what the lenders pattern of bidding is, whether they bid at a constant frequency over the entire time period when the listing is open, or whether there are specific time periods when lenders are more likely to bid. Flake et al.[2] described an algorithm for undirected graph clustering. They presented a new clustering algorithm based on minimum cut trees, creating clusters that have small inter-cluster cuts and large intra-cluster cuts, which is a strong criteria of clustering. Liben-Nowell et al.[5] provided several methods to predict possible links. The paper described approaches to predict links by analyzing the proximity of nodes in the network. It is more likely to link between two nodes when the similarity is high. Leskovec et al.[4] proposed a maximum-likelihood based model and discovered that edge locality plays an important role in evolution of networks. Especially, the edge initiation process are accelerating with node degree, and this leads to power law out degree distribution.

DATASET

Prosper peer to peer money lending dataset [1] is in XML format which accompany with an XSD file for styling. The data size is 2.5G for bidding and 1.7G for everything else. Propser also provides daily differential data for incremental network analysis. The dataset contains various information for the transactions, including the amount, time, location, short description, bid, rate, credit, etc. There are also some features for the users such as: is the borrower a homeowner, does the borrower has a verified bank account, whats his monthly loan payment, etc. Normally we need a XML parser for processing such data. However, parsing XML tree requires memory for storing the whole XML file and the additional parsing structure which is not very efficient. We develop a line-by-line

[1] https://www.prosper.com/tools/DataExport.aspx

parsing method for grepping data from various XML entities. The generated data are ID list of different entries, hence we need another ID matching process to generate the final network data.

NETWORK ANALYSIS

Lending Network Property

The peer-to-peer lending network has 132,264 nodes and 2,751,254 edges. 52,558 (39.7%) users are lender and 85,422 (64.6%) users are borrowers. Some of the users are both lenders and borrowers (4.3%), so it doesnt sum up to 100%. We define an edge from node A to node B represent that user A had lent money to user B. Therefore the whole network is a directed graph. The two figures above are in degree and out degree distribution. As we can see, the distribution follows power law, which is common in real-life networks. The (undirected) diameter of the network is 7. The size (in percentage of nodes) of the largest strongly connected component is 0.02. This is expected since there are less nodes (4.3%) that are both lender and borrower. The size of the largest weakly connected component is 0.99, hence there is almost no isolated node in this graph.

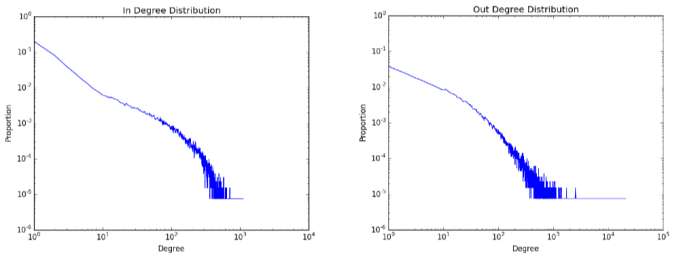

Figure 1. Degree distribution, the left is in-degree distribution and the right is out-degree distribution

Robustness/Connectivity

In order to test robustness and connectivity, we perform attacking procedure on our network. Similar to the paper by Albert, Jeong and Barabasi, Error and Attack Tolerance of Complex Networks (Nature, 2000), we deleted nodes in batch, and measured diameter and largest connected component of the remaining network. To simulate attack, we deleted nodes in decreasing order of their degree. That is, a node with high degree will be deleted earlier than a node with lower degree. We deleted 1000 nodes at a time, and stopped when there is only 50% of the nodes left. The result can be found in the following figures. Fig. 2 used diameter as measurement, while Fig. 3 used largest connected component.

According to the graph measured with diameter, our network has very similar performance to random network. It experiences a rise when deleting around 25% of the nodes, and then a significant drop when more nodes are removed. This might due to the reason that after removing higher degree nodes, the average distance between nodes will increase. However, after a certain point, the graph will break down into pieces, which gives us the drop in the graph. Since the rise happened at around 25%, we can infer that our network is not autonomous

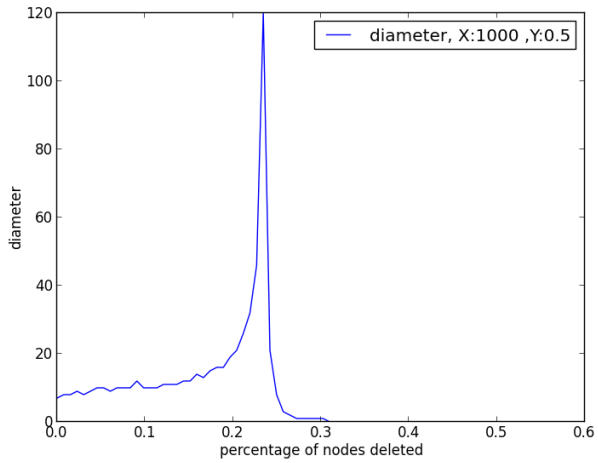

Figure 2. Robustness of the network measured by diameter

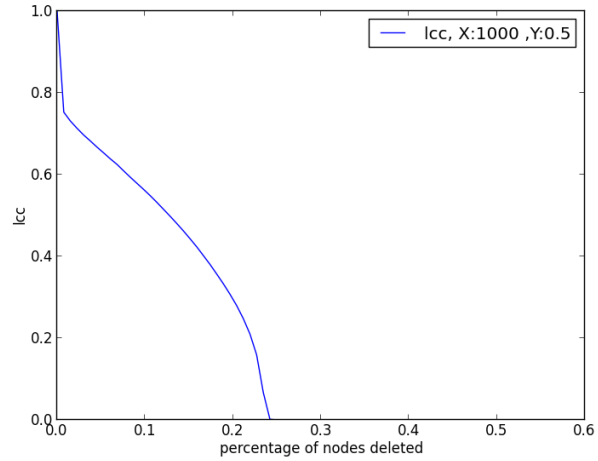

Figure 3. Robustness of the network measured by lcc

system nor with preferential attachment, which normally has a rise at around 10% or less.

In the graph measured with largest connected component, the curve experience a significant drop at the beginning, then the drop become smoother very soon. We assume that this network contained one large connected component, which included a lot of high degree nodes, while the other part of the network is similar to a random graph. Therefore, after deleting high degree nodes, the largest connected component would shrink rapidly, but soon it will slow down. However, what we found interesting is that removing these high degree nodes wont affect diameter.

Link Analysis

In this section, we conduct link analysis on the the peer-to-peer lending network. For a simple model, we can consider the formation of lending edges as a sign of trust. If an user succeeded in borrowing money from various other users, the user is more trustworthy. Link analysis algorithms are handy

under the trust model. For example, we can use the PageRank[6] score as the trust value. The higher the PageRanks score is, the better credit score the user will get. HITS[3] provides another view on this problem. The users who borrows a lot may have a higher hubs score while the users may have higher authority score for lending events.

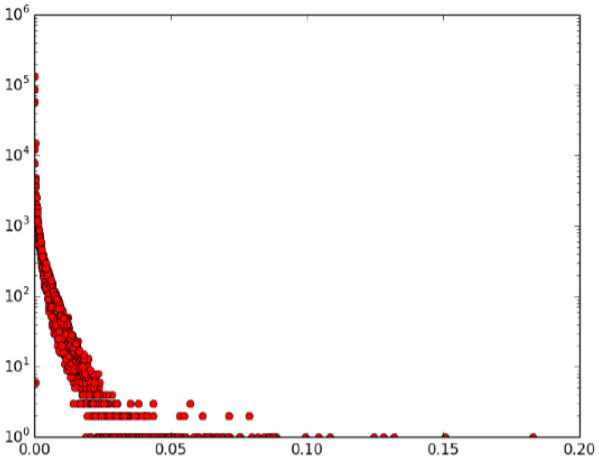

Figure 4. Histogram of PageRank scores

Fig.4 shows the histogram of PageRank score. The maximum PageRank score is less than 0.20. The distribution roughly follows the power law, which is common for real-world data. There are very few users who have PageRank score higher than 0.05. Most of the users have score ranging from 0.0 to 0.3. Since we would include the PageRank score as a feature in prediction tasks, we take the logarithm of the score to reflect its power law nature.

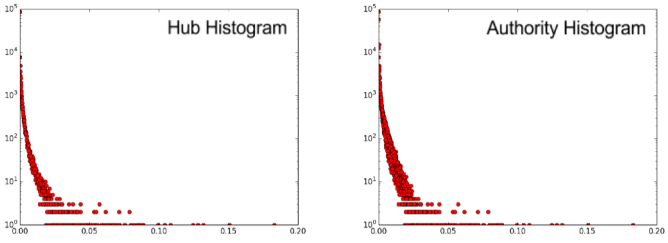

Figure 5. Histogram of Hub and Authority scores

Fig.5 shows the histogram of hub scores and authority scores. As we can see, both hub and authority scores have similar distribution as the PageRank. Top lenders and top borrowers contributes most of the transactions. Therefore, it is essential to identify these top borrowers and lenders.

Modeling Borrowers

To have a better understanding of what types of borrowers can receive more money, we analyze the successful transactions with different interest rate and different borrower credit score. Fig. refxd1 shows the average amount of money a user can borrow with different interest rate.

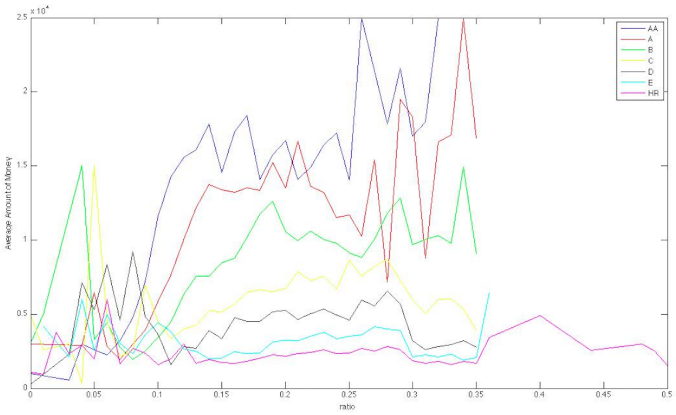

Figure 6. The average amount of money a person can borrow

First, since blue line (AA) is above red line (A), red line is above green line (B), and so on, we can observe that borrowers with better credit score can normally get more money than borrowers with worse credit score. This is consistent with intuition since lenders would like to lend more money to those with better credit score.

In addition, the amount of money increases as the interest rate increases regardless of the credit score. This is because the more money the user borrow, the higher risk the lender has. Therefore the interest rate should be higher to compensate for the lenders.

Another interesting fact is that only the users with credit score "HR", which is the worst credit level, borrowed money with interest rate larger than 0.36.

Modeling Lenders

It would be interesting and helpful for the borrowers if we can model the behaviors of different lenders. To achieve the goal, we extracted and visualized 3 features of the lenders: the total times the user lends money, the average interest rate of each lending, and the average amount of money lent.

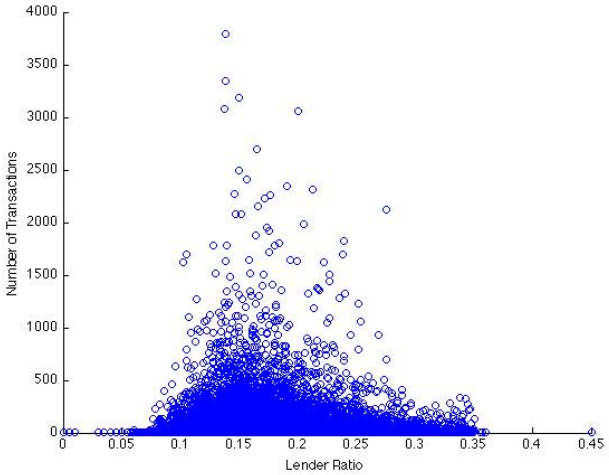

Figure 7. Number of transactions v.s. Average ratio

As we can see in Fig. 7, each dot represents a unique lender. The y-axis is the number of transactions the user involved, and the x-axis is the average interest rate when the user lends money. There is a peak appearing around interest rate equals 0.15. Note that this is not a distribution or curve but a scatter plot, so this is a very interesting observation: if the lender involves in a large amount of transactions, it's more likely that the average interest rate of the lender lies within 0.15 to 0.2. In other word, there doesn't exist lenders keep lending money with very high (> 0.3) or very low (< 0.08) interest rate. One interpretation is that those who lend money frequently are very reasonable and doing so in a safe range interest rate. On the other hand, those who lend money with high or low interest rate might not get good return and thus stop this strategy.

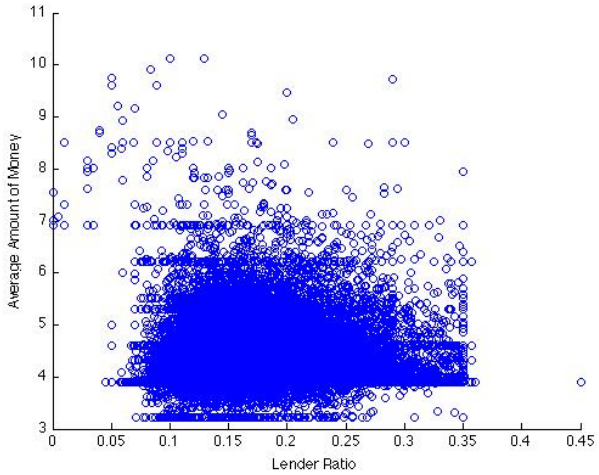

Figure 8. Average amount of money v.s. Average ratio

In Fig. 8, each dot represents a unique lender. The y-axis is the average amount of money (after log) the user lent each time, and the x-axis is the average interest rate when the user lent money. We can observe that most lenders have rate around $0.1 \sim 0.3$, and lend money less than 1000 dollars.

We can also observe that those who lent with high amount of money (> 7 on log scale) tend to have lower interest rate (< 0.2). The reason is that higher rate implies higher risk, so we will be more careful when lending a great amount of money with high interest rate.

Another observation is that there is no lender who lent a small amount of money (< 7 on log scale) with low interest rate (< 0.05). This is very intuitive since the lender wouldn't earn too much.

In Fig. 9, each dot also represents a unique lender. The y-axis is the number of transactions the user involved, and the x-axis is the average amount of money (after log) the user lent each time. We can observe that lenders who lent more money each time involved in fewer transactions. One explanation is that there is no such lender with a great amount of money thatt can not only lend to many users but also with high amount. Therefore, the lenders can be categorized into 2 types: (1) high frequency, low amount, which locate at upper left part of the figure, and (2) low frequency, high amount, which locate at bottom right part of the figure.

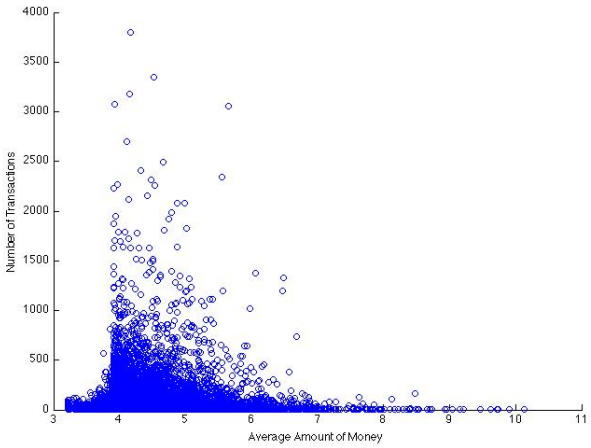

Figure 9. Number of transactions v.s. Average amount of money

Another thing worth noticing is that there is a blank area on the left of the figure, meaning that there are no lenders keep lending money with small amount (< 4 on log scale). This phenomena is hard to explain and may need further analysis to figure out the underlying reason.

NETWORK INFERENCE

Predict actual money borrowed of a bid

Whenever a listing is posted, there are many factors that can affect the actual amount the bidder would get in the end, including the interest rate, the current credit level, as well as previous bidding record of the bidder. Therefore, we are interested in building a model simulating the process. Given a new bid, we try to use machine learning model to predict what the final amount the bidder could get.

We use the following features as our input, and then test with several methods.

Given the features, we are interested to know whether the actual amount borrowed is related to the proposed money. That is, if we can predict a certain ratio of the proposed money a bidder could get. In addition, we also would like to know if we can directly predict the actual amount. The result can be found in the following table, where we used relative absolute error as our measurement, and tested with 10 fold cross validation:

From the result above, we can tell that it is not possible to predict the ratio of the proposed money the bidder could get. The ratio is not related to the features we proposed above. On the other hand, we can directly predict the actual money borrowed with very low error. That means the actual money a borrower can get has nothing to do with his/her proposed money. There's a high chance that no matter how much the user proposed, he/she could only get similar amount of money, depending on his/her previous bidding record.

	TP Rate	FP Rate	Precision	Recall	F-Measure
Random Forest (10 trees)	0.73	0.14	0.72	0.73	0.72
Random Forest (30 trees)	0.73	0.14	0.72	0.73	0.72
Naive Bayes	0.65	0.02	0.75	0.65	0.69
Logistics	0.72	0.26	0.66	0.72	0.67
Adaboost(DecisionStump)	0.67	0.67	0.45	0.67	0.54

Table 2. Result of predicting credit level

Feature	Description
Proposed amount	The amount the bidder proposed
Borrower rate	The borrower rate for this loan
Credit level	The borrower's current credit
Avg proposed amount	The average amount the bidder has proposed so far
Avg lender number	The average number of lenders that lend money to the bidder
Avg lending money per lender	The average money a lender lends to the bidder
Avg actual amount	The average amount the bidder actually received
Mode of credit level	The credit level the bidder gets the most
PageRank	The PageRank of the borrower
HUB score	The HUB score of the borrower
AUTH score	The AUTH score of the borrower

Table 1. List of features

	Predict Ratio	Predict Amount
Linear Regression	3.47%	98.58%
Random Forest (10 trees)	5.10%	78.53%
Random Forest (50 trees)	1.87%	77.01%

Table 3. Relative absolute error of predicting final amount borrowed

Predict the mode of user's credit level

The user's credit level would change over time. Therefore, the user would have different credit level associated with different bids. We are interested in predicting the mode of the user's credit level by the features in Table. 1, to see if the mode of the credit level could correctly reflect the user's previous record.

We used the features in Table 1, and tested with various methods. The result is in Table 2.

Predict the interest rate of a loan

We imposed similar method for predicting the interest rate of a certain loan. All features from Table 1 are included for the prediction. The result can be found in the Table 4, where we also used relative absolute error as measurement, and tested with 10 fold cross validation. From the training process, we believe that the lender ratio is highly related to the borrower ratio. Other effects such as credit level and the total amount of borrowing also played a role in the final interest rate.

CONCLUSION

In this project, we achieved all the goals listed in the project proposal. We understand the network structure by study the network properties. Models for borrower and lenders are built by combining network feature (node degree) with the loan property (interest rate, borrowing amount ...). Then we utilized the network features we learned to build models for predicting the property of a transaction event. Network features are proved useful through the whole project.

	Relative absolute error
Linear Regression	5.13%
Random Forest (10 trees)	5.54%
Random Forest (100 trees)	4.90%

Table 4. Relative absolute error of predicting interest rate

INDIVIDUAL CONTRIBUTION

Yen-Ting Processing the data, preliminary data analysis, predicting interest rate

Lun-Kai Link, analysis, borrower and lender modeling

Jocelin Analyzing robustness/connectivity, predicting actual money borrowed and credit score

REFERENCES

1. Ceyhan, S., Shi, X., and Leskovec, J. Dynamics of bidding in a p2p lending service: Effects of herding and predicting loan success. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, ACM (2011), 547–556.
2. Flake, G. W., Tarjan, R. E., and Tsioutsiouliklis, K. Graph clustering and minimum cut trees. *Internet Math. 1*, 4 (2003), 385–408.
3. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM 46*, 5 (Sept. 1999), 604–632.
4. Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, ACM (New York, NY, USA, 2008), 462–470.
5. Liben-Nowell, D., and Kleinberg, J. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, ACM (New York, NY, USA, 2003), 556–559.
6. Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.