Network Effects on Interest Rates in Online Social Lending*

Ulrik Brandes Jürgen Lerner Bobo Nick Steffen Rendle

Abstract: Peer-to-peer lending platforms such as Prosper Marketplace facilitate lending by matching potential borrowers with potential lenders using an auction mechanism. Borrowers can post loan requests and lenders can bid on a request specifying a minimum interest rate and an amount they are willing to lend. In this paper we analyze the impact of social information on the interest rate of funded loans and on the credit default risk. While externally determined credit grades strongly influence these outcomes, there is large variation within classes. We hypothesize that potential lenders use *social information*—including group membership and endorsements from other users—to trust some borrowers more than it is suggested by their credit grade. Furthermore we analyze whether this social behavior is rational in the sense that it leads to lower credit default risk for those that get lower interest rates.

Keywords: social network analysis; `prosper.com`; financial networks

1 Introduction

Can peers estimate a borrower's credit worthiness better than banks or other professional lenders? Online peer-to-peer lending sites such as *Prosper Marketplace*[1] and *Lending Club*[2] seem to hinge on this belief. In such sites, potential borrowers are not only characterized by "financial" variables such as credit grade, debt-to-income ratio, or homeownership but also by "social" information that might give additional hints to potential lenders. For instance, users can present pictures of themselves, their homes, family, or pets; they can write about their work and leisure activities and about why they can be trusted to pay back their loans; users can join groups; and they can receive endorsements from other users writing about their relationship to the prospective borrower. The information about borrowers is, thus, much richer—but also harder to process automatically—than the information from which professional lenders, such as banks, traditionally estimate credit worthiness.

In this paper we analyze the impact of social information on the interest rate of loans and on the credit default risk, i. e., the probability that a loan does not get payed back. After presenting how the Prosper Marketplace works (Sect. 2), we analyze in Sect. 3 whether social information has any impact at all. In all models we deal with the question whether those borrowers that get lower interest rates (i. e., the users that are considered as more

*A preliminary version of this work has been presented at the 31st INSNA Sunbelt Social Networks Conference, February 8–13, 2011, in St. Pete, FL, USA. There are no published proceedings.

[1] `http://www.prosper.com/`

[2] `http://www.lendingclub.com/`

trustworthy) also have lower default risks; if so, the behavior of lenders is considered as economically rational. Building on what we learned from the simple models in Sect. 3 we develop more sophisticated models for predicting the interest rate in Sect. 4.

There are at least two different sources of motivation for conducting this study. First, the well-know economic argument for why the market should be able to estimate credit default risks is the avoidance of *market failure*. If differences in the credit worthiness of prospective borrowers could not be assessed at all, the only possibility would be to assign every borrower the same (average) interest rate. While this would be very favorable for the high-risk borrowers, it would be too expensive for the low-risk borrowers who, in turn, leave the market in search for more adequate loans. The loss of "good" customers (i. e. low-risk borrowers) would finally lead to increasing default risk and, necessarily, increasing interest rates; eventually, the market could deteriorate until trading on Prosper is not profitable for anyone. Second the dataset from `prosper.com` is a good case study to test methods for the estimation and prediction of social network effects. Since here the outcome variables (interest rates and credit defaults) are quantitative and directly available in the data, the study does not suffer from difficulties in measuring variables that are typical for social network analysis, such as customer satisfaction, happiness, job performance, health, etc.

We emphasize that in the meanwhile Prosper changed its business model—after we collected the data, but before publication of this article. In the new model interest rates are no longer determined by an auction mechanism but are fixed by Prosper depending on credit grades and other borrower characteristics. Note that—while our results, thus, no longer apply to the current Prosper Marketplace—they nevertheless provide interesting insights into the functioning (or dis-functioning) of an economic system that partially hinges on social mechanisms.

In previous work using data from `prosper.com`, Freedman and Jin [FJ08] also point out relations between (among others) interest rates and credit grades and, moreover, compare them with loans contracted from professional, institutionalized lenders such as banks. Chen, Gosh and Lambert [CGL09] propose and analyze a game-theoretic model for the Prosper auction mechanism. Greiner and Wang [GW09] analyze the influence of *social capital* on the probability of funding, interest rates, and default rates. Mixture models to cluster Prosper data have been applied by Herrero-Lopez [HL09]. Several such studies have found a positive effect of social information on low interest rates. In contrast to [GW09], however, we also find a less desirable effect on credit default risks.

In the area of predicting modeling there are also several approaches that take social information into account. For instance, [JE10] and [MKL09] show that for predicting future scores on movies, it is beneficial to take social information about friendships into account. The method proposed by Ma, King and Lyu [MKL09] adds indicators for friends to a linear prediction model with factorized interactions. Another approach for this task is proposed by [JE10], where social information is integrated by assuming similar prior-distributions over factors for persons that are friends. In Section 4.2.2 of this paper, we follow the first approach and integrate endorsement information using additional predictors where interactions are factorized.

2 The Auction Mechanism: Bidding on Loan Requests

We continue with introducing the historical auction mechanism at the Prosper Marketplace, i. e., the algorithm that was used to match borrowers and lenders and to determine interest rates.[3] For illustration, we use the hypothetical example show in Fig. 1.

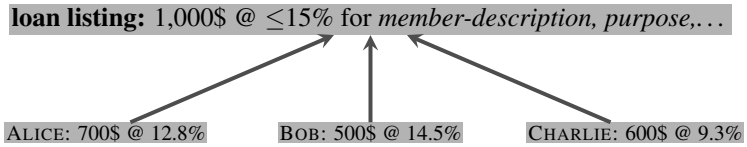

Figure 1: Illustration of the auction mechanism at `prosper.com`. A member posts a loan listing for 1000$ at a maximum rate of 15%. Alice, Bob, and Charlie bid on this listing specifying different participation amounts and minimum rates. When the bidding phase ends at this point, the loan is granted for 12.8%; Charlie participates with 600$, Alice with 400$, and Bob does not participate at all. Note that in this example Charlie also gets 12.8% interest rate.

Users with a Prosper account can post *loan listings* at the Prosper Website in which they specify the requested amount and a maximum interest rate; further these potential borrowers can provide textual descriptions for the purpose of the loan, information about themselves, arguments why they are able to repay the credit and so on. These listings are visible on the Prosper Website, together with the credit grade of the member (and some other variables) and information about all bids that are already placed on them. For a certain duration, other members can *bid* on loan listings specifying a participation amount and a minimum interest rate for which they are willing to lend the money. Before a bid can be placed, members have to deposit the amount on their Prosper account; this ensures that there are no fake biddings that intent only to lower the interest rate. After the bidding phase ends, it is determined whether the listing becomes a loan and (if so) the interest rate and the participation amounts of the bidders:

- A listing becomes a *loan* if all bidders together are willing to lend the requested amount. Thus, borrowers get either the full loan or nothing.
- The "cheap" lenders, i. e., those that demand lower rates, participate first until the requested amount is filled up. In the above example, Charlie (who demands 9.3%) participates with the whole amount he offered, Alice (who demands 12.8%) participates with 400$ (thus less than she was willing to lend); then the loan is fully funded and Bob (who demands 14.5%) does not participate at all.
- The interest rate of the loan is the minimum rate of those bids that do not fully or not at all participate in the loan. All lenders get the same rate—even if they were willing to lend the money for less. This rule avoids strategic bidding where lenders demand more than necessary just because they think that the loan does not become fully funded for the rate that they would be willing to lend the money. Consider

[3] To enhance readability we no longer write of the *historical* mechanism in the remainder of this paper.

the example in Fig. 1: if Charlie is willing to lend the money for 9.3% he does not have to worry about whether demanding a higher rate would increase his gain; if the loan is only granted for a higher rate, he will get this higher rate. In this example, the interest rate is 12.8%, i. e., the interest rate of Alice' bid which does not fully participate. If the requested amount was 1,300$, then, assuming the same set of bids, the interest rate would be 14.5%, i. e., the interest rate of Bob's bid which would then be the cheapest one that does not participate in the loan.

In case that a listing becomes a loan, Prosper transfers the amount, minus a 1% fee, to the borrower's account. The duration of all loans is three years and borrowers have to make monthly repayments to Prosper which forwards the money to the lenders, proportional to their participation ratio. When borrowers fail to make these repayments, Prosper starts a debt-collection process similar to what a bank would do: that is, after a certain delay the right to collect the money is transferred to a debt collector. Prosper covers the risk of credit default only in rare cases, for instance, if a user could log-in under a false identity. In most cases the risk is entirely with the lenders; thus, these have economical interest in lending money only to creditworthy borrowers.

The whole bidding process is very transparent to visitors of the Prosper Website (whether they are logged-in or not) and also to analysts. Specifically, Prosper offers to download historical data[4] about members, groups, loan listings, bids, and loans in a very fine-grained level. For this paper we used data that we downloaded in October 2010.

3 Linear Models

In this section we model the interest rate of fully funded loans by linear regression and the credit default risk (CDR) by logistic regression. We start with models build from the "traditional" predictors *credit grade*, *debt-to-income ratio*, and a binary indicator for *homeownership*. In Section 3.2 we extend these models by the "social" variables *group membership* and (functions of) *endorsements* from other members.

3.1 Traditional Predictors

Whenever a loan listing is posted on `prosper.com`, potential lenders can see the *credit grade* of the borrower. This categorical variable is determined by an external credit agency and is, thus, similar to the main criteria for credit-worthiness that is used by banks or telecommunication agencies. The credit grades range from `AA` (best rating) down to `HR` (for *high risk*); `NC` means that no credit history is available for the borrower so that the agency cannot determine the grade; *N/A* means that the variable for the credit grade was missing in the data that we downloaded from `prosper.com`. Figure 2 shows the distribution of interest rates (mean and standard deviation) over the different grades and gives

[4] See `http://www.prosper.com/tools/DataExport.aspx`

the number of loans in each category. The average interest rate goes up in a roughly linear fashion when we move from AA (mean of 9.77%) down to E (mean of 25.11%) but then does not increase for the high-risk borrowers. For the following analysis we remove all loans with missing credit grade or credit grade equal to NC which leaves us with 28,874 observations.

CG	AA	A	B	C	D	E	HR	NC	*N/A*	*uncond.*
mean	9.77	12.45	15.24	17.81	20.99	25.11	25.02	21.47	19.85	18.4
sdev	3.31	4.21	4.33	5.67	5.96	5.73	6.72	6.06	9.55	7.83
N	3,530	3,323	4,397	5,646	5,156	3,298	3,524	143	6,143	35,160

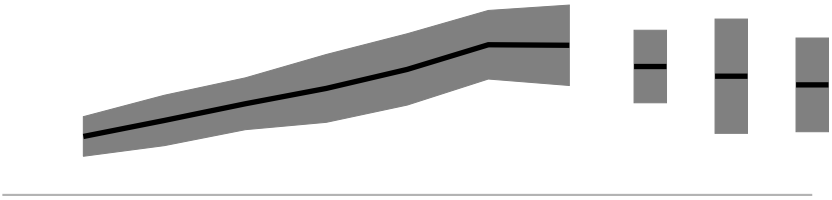

Figure 2: Distribution of interest rates (in %) by credit grade of the borrower. The values in the column marked by *uncond* are the summary statistics for all 35,160 loans irrespective of their grade. The thick black line shows the (piecewise linearly interpolated) mean; the gray-shaded area indicates the mean±standard deviation. The horizontal line at the bottom of the diagram marks the origin of the y-axis, i. e., the 0% level.

Analyzing how much is gained by the different variables, it turns out that (as it could be expected) the credit grade is the strongest predictor for the interest rate of loans. The table below shows the residual standard error for some models based on what we call financial information. The model M_1 that estimates the interest rate independently for each credit grade has a residual standard error of 5.3 and, thus, considerable lower than the baseline model M_0 that estimates the unconditional average rate for all loans. In model M_2 we replace the categorical variable for credit grade by a numeric variable which assumes the value of 6 for grade AA, 5 for A, down to 0 for HR and require a linear relation with the interest rate; it can be see that the RSE increases only slightly to 5.35. Adding the two variables *debt-to-income ratio* and *homeownership* yields a RSE of 5.18. All following models in this section include the model M_3 build from this financial information.

model	RSE	$1 - R^2$
M_0: `rate` $\sim$ `const.`	7.38	1.0
financial info.		
M_1: M_0 + `CreditGrade`	5.30	0.52
M_2: M_0 + `CreditGradeLinear`	5.35	0.53
M_3: M_2 + `D/I ratio` + `home`	5.18	0.49

For estimating the credit default risk (CDR) we use the `Status` variable of loans, whose summary distribution is shown in the table below. Specifically, a loan with status equal to `Paid` is considered as a non-default instance and loans with status equal to `Defaulted (Bankruptcy)` or equal to `Defaulted (Delinquency)` are considered as yes-

instances. When estimating the CDR we drop loans whose status is equal to `Current`, `Late` or any other value, since for these it is not evident whether they will be repaid or not.

status	# loans	interpretation
`Paid`	13,510	`default=no`
`Defaulted (Bankruptcy)`	1,007	`default=yes`
`Defaulted (Delinquency)`	1,428	`default=yes`
`Current`	10,824	*dropped*
`Late`	666	*dropped*
other	7,725	*dropped*

To assess whether lender behavior is economically rational we estimate the parameters of the model M_3 for the interest rate (by linear regression) and CDR (by logistic regression). The parameters and their standard errors (in brackets) are shown in the table below. It turns out that—as far as we consider only these variables—lenders do adapt the interest rate in a rational manner. Those with better credit grades get lower interest rates and have a lower default risk; those with higher debt-to-income ratio, as well as homeowners, have to pay higher rates but also default more often.

	effect on rate	effect on CDR
CGradeLin	-2.88 (0.179)	-0.487 (0.015)
D/I Ratio	0.29 (0.033)	0.084 (0.021)
Homeowner	0.94 (0.067)	0.513 (0.052)
constant	17.38 (0.044)	-2.142 (0.038)

We noted above that this model for the interest rate has a RSE of slightly more than 5%. In this context, paying 5% more or less seems to make a big difference; thus, the model so far is quite coarse. Before refining it we turn to the question whether the difference of the observed rate to the predicted rate is random or rather can be justified by the CDR. To assess this we define a new variable `Spread` that encodes for each loan the difference `observed rate` − `predicted rate` (where predicted rate refers to the rate predicted by the model M_3 above). If a loan has a positive value in this variable, then the borrower has to pay more for the credit than our model would suggest; if, in turn, these loans are also associated with a higher risk, then the spread is not random but appears to be economically rational and we can conclude that lenders assign the rates "better" than our model M_3. The table below—more specifically the positive parameter associated with the spread variable—shows that this is indeed the case.

	interest rate	CDR	CDR
CGradeLin	-2.88 (0.179)	-0.487 (0.015)	-0.51 (0.015)
D/I Ratio	0.29 (0.033)	0.084 (0.021)	0.10 (0.021)
Homeowner	0.94 (0.067)	0.513 (0.052)	0.58 (0.053)
Spread	·	·	**7.69 (0.460)**
constant	17.38 (0.044)	-2.142 (0.038)	-2.16 (0.039)

3.2 Bringing-in Social Information

The first "social information" that we add to our models is a binary variable encoding whether the borrower belongs to any group or not. Note that this information—as well as any other social information—is visible to potential lenders. It is assumed that membership to groups is an indicator of trust worthiness. Thus, group members should get access to loans for lower rates and also have lower default risk. The empirical analysis (see the table below) shows that these expectations are only partially supported. Group members do get lower rates (by almost 2%) but, empirically, their default risk is higher. Thus, the group variable points for the first time to an irrational behavior of lenders: they are willing to lend money for lower rates to group members how actually should pay more to cover the increased risk.

	rate	CDR	CDR
CGradeLin	-3.00 (0.050)	-0.472 (0.015)	-0.493 (0.016)
D/I Ratio	0.35 (0.033)	0.079 (0.021)	0.093 (0.021)
Homeowner	0.93 (0.066)	0.514 (0.052)	0.587 (0.053)
Group	**-1.94 (0.064)**	**0.229 (0.049)**	**0.258 (0.050)**
Spread	·	·	8.609 (0.481)
(const)	18.15 (0.050)	-2.258 (0.046)	-2.308 (0.047)

We can shed more light on this by differentiating between the probability of a *delinquent* default and the probability of a default due to *bankruptcy*. The CDR restricted to delinquent defaults actually yields the same pattern of higher risk for group members.

	rate	CDR (del.)	CDR (del.)
financial info.			
Group	**-1.94 (0.064)**	**0.67 (0.070)**	**0.73 (0.071)**
Spread	·	·	8.34 (0.606)

On the other hand, when only bankrupt defaults are considered, group members do have lower CDR (as originally expected).

	rate	CDR (bankr.)	CDR (bankr.)
financial info.			
Group	**-1.94 (0.064)**	**-0.28 (0.070)**	**-0.26 (0.070)**
Spread	·	·	9.03 (0.676)

These findings could point to an intentional abuse of group membership in the sense that borrowers who never had the intention to pay back the credit (i. e., those that will eventually result in a delinquent default) sneak into groups to gain reputation. Such findings could be an explanation for the decision of `prosper.com` to abandon the auction mechanism.

A second variable based on social information is the number of endorsements received from other members. As the empirical analysis shows this indicator has the expected

influence: the higher the number of endorsements the lower the interest rate and the lower the credit default risk.

	rate	CDR	CDR
financial info.			
Group	-1.90 (0.065)	0.27 (0.050)	0.30 (0.051)
# endorsements	**-0.17 (0.045)**	**-0.34 (0.043)**	**-0.32 (0.043)**
Spread	·	·	8.46 (0.482)

The number of endorsements—or, formulated in the language of social network analysis, the *indegree* in the endorsement network—seems to be a quite coarse measure since it treats all endorsements equally, independently of the endorser. Following established ideas in social network analysis, we might treat an endorsement as more valuable if it comes from a member who receives many endorsements herself. Iterating this idea yields the well-known *eigenvector centrality* which is however (for technical reasons) inappropriate for networks that are not strongly connected. A measure also build on the idea of a *feedback centrality* that is more robust for unconnected networks is the well-known *page rank*. Using page rank as an explanatory variable (see the table below) instead of indegree (a. k. a. number of endorsements received) shows a quite similar pattern: members with higher page rank get lower interest rates[5] and have lower default risk. From the table below we cannot conclude whether page rank explains interest rates and default risks better than indegree or vice versa.

	rate	CDR	CDR
financial info.			
Group	-1.94 (0.064)	0.23 (0.050)	0.26 (0.050)
pagerank	**-8.14 (4.856)**	**-17.37 (4.925)**	**-16.76 (4.970)**
Spread	·	·	8.58 (0.481)

To shed light on the relative explanatory power of page rank vs. indegree we include both variables in the same model (see table below). What we find out is that, controlling for indegree, page rank has no significant influence neither on the interest rate nor on the credit default risk, while the influence of indegree does not change. (Note that the standard errors in the row associated with page rank are much larger than the absolute values of the parameters.) The effect of page rank, reported in the table above, is probably only due to its correlation with indegree; controlling for this variable, page rank has no influence at all. Thus, in contrast to our intuition, the more sophisticated measure does not turn out to be a better predictor.

[5]Note however that a parameter of -8.14 with a standard error of 4.856 implies only a significance level of $< 10\%$ while all other parameters discussed so far are significant on the $< 5\%$ level (and often much more).

	rate	CDR	CDR
financial info.			
Group	-1.90 (0.065)	0.27 (0.050)	0.30 (0.051)
indegree	**-0.17 (0.050)**	**-0.33 (0.047)**	**-0.31 (0.047)**
pagerank	**-0.33 (5.359)**	**-1.76 (5.265)**	**-1.83 (5.331)**
Spread	·	·	8.46 (0.482)

The analysis so far showed that social information does have an influence on the outcome variables; on the other hand it still suffers some serious drawbacks. The models above assumed homogeneity over all members and groups; they did not leave any room for patterns where some members trust some groups (or some other members) more than others. In the next section we introduce and apply models that can deal with such unobserved (latent) clusters of users and groups.

4 Predictive Modeling

The analysis so far investigates the influence of several predictors on the credit rate using a linear regression model. We have discussed the model parameters that indicate whether a predictor has a positive or negative (or possible none) impact on the target (e.g. credit rate or risk). In the following, we want to investigate more complicated modes for predicting interest rates on unobserved future data. The focus here is not on analyzing the model parameters but on predictive accuracy. Thus, our proposed models can deal with a large number of model parameters.

4.1 Data and Methodology

For modeling interest rates, we use the following data and notation. Let the domains be:

- $L = \{l_1, l_2, \ldots\}$ is the set of all loans. In our case the data of $27,165$ loans from Prosper.
- $CG = \{AA, A, \ldots, HR\}$ the set of credit grades.
- $USER = \{u_1, u_2, \ldots\}$ is the set of all users/ members of the prosper platform.
- $GROUP = \{g_1, g_2, \ldots\}$ the set of groups.

The relations involved are:

- $y : L \to \mathbb{R}$ the interest rate that we want to model. We assume that it is only partially observed and the missing values should be predicted.
- $cg : L \to CG$ the credit grades of a loan.

- *home* : $L \to \{0, 1\}$ whether or not the applicant for a loan is a home owner.
- *group* : $L \to \mathcal{P}(GROUP)$ the groups a loan is assigned to.
- *end* : $L \to \mathcal{P}(USER)$ the members that endorse the applicant of a loan. Note that endorsement is a relation between two users, which we join here with the applicant of a loan.

We set up a forecasting experiment for y, where based on observations of the past future interest rates should be predicted. We split L into two disjoint datasets $L = L_T \cup L_V$ where L_T is the set containing all observations before a specific date and L_V all observations after this date. In our experiments, we choose June 30, 2008 as reference date for splitting which results in $23,654$ training cases and 3511 test cases. The model is then trained only using the data in L_T and a small error on the withheld future data L_V is desired. The error is measured by root mean square error (RMSE) between predicted $\hat{y}$ interest rate and true interest rate over the withheld data L_V:

$$RMSE = \sqrt{\frac{1}{|L_V|} \sum_{l \in L_V} (y(l) - \hat{y}(l))^2} \tag{1}$$

4.2 Models

In the following, we will discuss models for the interest rate y. We start with the basic model discussed and analyzed before. Then we extend this model using additional predictors and finally we add factorized interactions between the predictors.

4.2.1 Basic Model

The model discussed in Section 3.2 is our baseline. It can be formalized as:

$$\hat{y}^{BASIC}(l) := \beta_0 + \beta_{cg(l)} + \beta_{home} home(l) + \beta_{group} |\{group(l)\}| + \beta_{end} |\{end(l)\}| \tag{2}$$

where $\beta \in \mathbb{R}^{11}$ are the model parameters to estimate. These are: the offset β_0, 7 parameters for each of the credit grades and one each for the home owner, for the group membership and for the endorsements. As discussed before, this model looses information by aggregating over variables. E.g. the group and endorsement relations are modeled each by one indicator in total. In the following, we want to discuss more expressive models.

4.2.2 Extended Model

When we are interested in predictive accuracy, we can set up more complicated models with much more model parameters. In the model so far, we aggregate the information about groups in one binary indicator which models the influence of *having* a group on the interest rate. However, it makes sense to assume that different groups have different

influence. E.g. lending money for a business investment might lead to a lower interest rate then lending money for personal expenses like a holiday. So the first generalization is to model each group with an own indicator. For a specific loan, the model contains an individual term for the group the loan is assigned to:

$$\hat{y}^{GROUP}(l) := \frac{1}{\max(1, |group(l)|)} \sum_{g \in group(l)} w_g \tag{3}$$

There are 696 different groups in our loan data, so this results in 696 additional parameters to estimate.

Secondly, instead of counting the endorsements that a user got from other members, it makes sense that some members have higher reputation than others which should influence the interest rate. Again, we model this explicitly:

$$\hat{y}^{END}(l) := \frac{1}{\max(1, |end(l)|)} \sum_{u \in end(l)} w_u \tag{4}$$

We have endorsements from 5245 members, so again there are 5245 additional model parameters.

In total our extended model is:

$$\hat{y}^{EXT}(l) := \hat{y}^{BASIC}(l) + \hat{y}^{GROUP}(l) + \hat{y}^{END}(l) \tag{5}$$

4.2.3 Factorized Interactions

The effects that we have described so far can also interact. For example an endorsement of one member might only be of high reputation in a certain group – e.g. an endorsement of an investor in startup companies should have a higher influence on loans that are assigned to a business group than for a group about cars. The typical approach is to model such interactions with individual model parameters, e.g.:

$$\hat{y}(l) := \frac{1}{\max(1, |group(l)|)} \frac{1}{\max(1, |end(l)|)} \sum_{g \in group(l)} \sum_{u \in end(l)} w_{u,g} \tag{6}$$

However, this is likely to fail for social network data where the number of interactions is much larger than the number of observations. E.g. in our case this would result in $5245 \cdot 696 = 3,650,520$ independent model parameters where the number of observations is only $27,165$. To solve this issue, the interaction parameters can be factorized which results in a much lower number of parameters and the independence of model parameters is broken [RFST10]. So instead of modeling $w_{u,g}$ with one parameter, one can factorize the interactions

$$\hat{w}_{u,g} := \sum_{f=1}^{k} v_{u,f}\, v_{g,f}, \tag{7}$$

where $\mathbf{V} \in \mathbb{R}^{(5245+696)\times k}$ is a matrix containing factors. This way each group or member is described with a set of k latent factors. The dot product of the factors of group with a member is learned such that it leads to the desired interaction. The main advantage of this approach is that if an interaction is unobserved in S_T but appears in S_V, e.g. between a group g and a member u, one can still predict the interaction if g and u have been observed individually. For example if the interaction of u with g' is observed and one knows that g and g' are similar (have similar interactions with other members) that means they have similar factors, one can infer that the interaction between m and g is similar to the observed one of u with g'.

Instead of deriving a solver for this specific task, we apply a Factorization Machine (FM) [Ren10]. FMs are an extension of linear models that include interactions like polynomial models but where all interactions are factorized as described before. It has been shown that FMs include many of the most successful factorization models, e.g. for collaborative filtering aka recommender systems [Kor08, RS05]. We use the same predictors as in the previous section but add all pairwise interactions. That means among others the described interactions between endorsements and groups are included in this model.

4.3 Priors on Model Parameters

The previous two sections added a large number of model parameters Θ to the basic model. In addition to the basic 10 predictors, there are $5245+696 = 5941$ indicators for individual groups and endorsements. Moreover all pairwise interactions are modeled.

Extending the expressiveness of a model runs the risk of overfitting. That means adding indicators will decrease the error on the data where the model is fitted (here S_T) but at the same time it might also increase the error on the withheld data S_V. With a high number of predictors, overfitting is very likely.

A popular strategy to prevent overfitting, is to model prior knowledge about the model parameters. The most simple and very successful one is to place Gaussian priors on each model parameter $\theta \sim \mathcal{N}(0, \frac{1}{\lambda})$ – also known as ridge regression or maximum-margin. The maximum likelihood estimator (also called maximum a posteriori estimator in this case) corresponds to the following regularized optimization task:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{l \in L_T} (y(l) - \hat{y}(l))^2 + \lambda \sum_{\theta \in \Theta} \theta^2 \tag{8}$$

That means besides the standard least square objective, the second objective is to have small model parameters. The value of λ is a hyperparameter that can be found e.g. by cross-validation.

4.4 Results and Discussion

We run the three models discussed so far on the forecasting problem where the interest rates for loans should be predicted based on historical data. The models are learned each on the loans L_T which were granted up to June 2008. The predictive accuracy is measured in terms of RMSE for future loans L_V from July to October 2008:

Method	#Predictors	Prediction error (on L_V)
Basic model	10	0.0686
Extended model	5951	0.0681
Factorized Interactions	5951	0.0674

The table shows that the models including individual group and endorsement predictors have a slightly higher accuracy on withheld data. The basic model with 10 predictors has an error of 0.0686. With additional indicators that model individual groups as well as endorsements of members, the RMSE lowers to 0.0681. And the factorized interactions lead to 0.0674.

In total, the improvement of the more complicated models are rather small. We assume that the reason is the high sparseness of the network data: (1) In the whole dataset of $27,165$ loans there are only $24,829$ distinct applicants. (2) There are only few loans where the applicant has endorsements. These two facts make it difficult to see large effects from the endorsement network on the overall prediction score. Furthermore, for each loan there is at most one group. This makes it hard to infer over several groups, e.g. to detect group-similarity, because groups are typically not connected neither over loans nor over users.

5 Conclusion and Future Work

This study demonstrates that social information does have an influence on outcome variables in peer-to-peer lending—both in explanatory and in predictive analyses. The way lenders process this information sometimes leads to irrational behavior; for instance, group members get lower rates but, empirically, have higher default risk. Since this remains true even when excluding credit defaults due to bankruptcy, it also suggests an intentional abuse of group membership by some borrowers and may thus explain the abandoning of auctioning at Prosper.

An issue for future work is to extend the predictive models from Sect. 4. Besides the rather small data of granted loans there is other network information available such as bidding events that are several orders of magnitude larger than the loan data. Integrating this information might lead to a more connected network which is supposed to lead to better propagation of information and estimation of model parameters.

References

[CGL09] Ning Chen, Arpita Ghosh, and Nicolas Lambert. Social lending. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, pages 335–344, New York, NY, USA, 2009. ACM.

[FJ08] S. Freedman and G. Z. Jin. Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper.com. Technical Report 08–43, NET Institute, 2008.

[GW09] Martina E. Greiner and Hui Wang. The Role of Social Capital in People-to-People Lending Marketplaces. In *ICIS 2009 Proceedings, Paper 29*, 2009.

[HL09] Sergio Herrero-Lopez. Social Interactions in P2P Lending. In *3rd SNA-KDD Workshop*, 2009.

[JE10] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 135–142, New York, NY, USA, 2010. ACM.

[Kor08] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, New York, NY, USA, 2008. ACM.

[MKL09] Hao Ma, Irwin King, and Michael R. Lyu. Learning to recommend with social trust ensemble. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 203–210, New York, NY, USA, 2009. ACM.

[Ren10] Steffen Rendle. Factorization Machines. In *Proceedings of the 10th IEEE International Conference on Data Mining*. IEEE Computer Society, 2010.

[RFST10] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing Personalized Markov Chains for Next-Basket Recommendation. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 811–820, New York, NY, USA, 2010. ACM.

[RS05] Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

ISBN 978-3-88579-286-4