Fairness of Classification Using Users' Social Relationships in Online Peer-To-Peer Lending

Yanying Li*, Yue Ning*, Rong Liu+, Ying Wu+, Wendy Hui Wang*
*Department of Computer Science, +School of Business
Stevens Institute of Technology
Hoboken, New Jersey
yli158,yue.ning,rliu20,ywu4,hwang4@stevens.edu

ABSTRACT

Peer-to-peer (P2P) lending marketplaces on the Web have been growing over the last decade. By providing online platforms, P2P lending enables individuals to borrow and lend money directly from and to one another. Since the applicants on P2P lending platforms may lack sufficient financial history for assessment, quite a few P2P lending service providers have been utilizing the applicants' social relationships to improve the risk prediction accuracy of loan applications. However, utilizing the information of applicants' social relationships may introduce discrimination in prediction. In this paper, we analyze and evaluate the impact of the applicants' social relationships on the fairness of risk prediction for P2P lending. We investigate over a million loan records collected from Prosper.com, one of the leading P2P lending companies in the world. We construct the Prosper social network of loan borrowers and lenders, and generate the social features of applicants by adapting a state-of-the-art social credit scoring scheme to the Prosper social network. We consider two types of fairness notions in the literature, namely *individual fairness* and *counterfactual fairness*. Our results demonstrate that the social score harms both individual and counterfactual fairness of classification. To address this issue, we design two new algorithms that mitigate bias by generalizing social features. Our experimental results show that our mitigation algorithms can reduce bias while utilizing social scores effectively.

KEYWORDS

Algorithmic fairness, machine learning, social network

ACM Reference Format:
Yanying Li*, Yue Ning*, Rong Liu+, Ying Wu+, Wendy Hui Wang*. 2020. Fairness of Classification Using Users' Social Relationships in Online Peer-To-Peer Lending. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3366424.3383557

1 INTRODUCTION

Credit evaluation and approval is the process a business or an individual must go through to become eligible for a loan or to pay for goods and services over an extended period. *Creditworthiness*, an assessment of the likelihood that a borrower will default on a loan, is one of many factors defining a lender's credit policies. Traditionally, a creditworthiness evaluation is based on an individual's financial history, primarily their payment records, current debt profiles, and credit history. Machine learning (ML) algorithms, such as classification models, rely on a measured creditworthiness to predict if a loan or a credit application will get approved or not.

A widespread problem of traditional creditworthiness evaluation is that first-time applicants and thinner-file borrowers such as students, foreign nationals, and populations of under-banked individuals are highly likely to face rejections due to a lack of financial history for assessment of their creditworthiness. In the past few years, the credit scoring industry has witnessed a dramatic change in utilizing users' social data to assess consumer creditworthiness [7, 18, 24]. For example, Lenddo has reportedly assigned credit scores based on user information such as education, employment history, and their social network friends [29]. Similar to Lenddo, a growing number of innovative lenders (e.g., FriendlyScore [1] and LendingClub [2]) are exploring the use of borrowers' social networking information in their credit underwriting process. These firms claim that their social-network-based credit scoring and financing practices can broaden opportunities for a larger portion of the population and may benefit low-income consumers who would otherwise find it hard to obtain credit.

Recently, the practice of online peer-to-peer (P2P) lending has become popular. It also relies heavily on borrowers' social information for creditworthiness assessment. For example, Prosper.com, the first P2P lending website in the US, encourages borrowers and lenders to form online groups and establish friendships with other Prosper members. It also allows group leaders and Prosper friends to offer endorsements and highlight bids from group members. A recent study [11] has shown that loans with friend endorsements and friend bids tend to have less missed payments and yield significantly higher rates of return than other loans. Another study [25] also shows that utilizing alternative data such as borrowers' social relationships can significantly improve the prediction accuracy of borrowers' default behavior and increase platform profits.

Although using borrowers' social relationships (either on Internet or embedded in the lending platform) can improve the prediction accuracy of loan approvals, it also raises a potential risk of discrimination and exclusion triggered by social financing. An algorithm that assumes financially responsible people socialize with other financially responsible people may incorporate systemic biases, and thus denies loans to individuals who are themselves creditworthy but lack creditworthy connections. While the Equal

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383557

Credit Opportunity Act (ECOA) by the Federal Trade Commission (FTC) of USA has prohibited credit discrimination on the basis of applicants' race, color, religion, national origin, sex, marital status and age, there have been little efforts of assessment and evaluation of the potential discrimination by considering social relationships as a feature for classification. Whether social relationships can introduce discrimination to ML-based decision making is largely neglected and remains questionable [8, 27].

In this paper, we analyze and evaluate the impact of utilizing borrowers' social relationships on classification fairness in P2P lending. We performed a study over a million loan listings on Prosper.com from November 2005 to September 2011. Since Prosper dataset does not contain any demographic information of borrowers, we cannot evaluate *group fairness* (i.e., whether the protected subgroups such as racial or gender groups are fairly treated). Therefore, in this paper, we mainly consider two types of fairness for individuals: (1) *Individual fairness* [9] that requires similar objects receive similar prediction results (according to a given similarity metric); and (2) *Counterfactual fairness* [12, 22] that defines fairness based on counterfactual examples (i.e., how would the prediction change if the attribute referenced in the example were different?).

To the best of our knowledge, we are the first to investigate if incorporating users' social relationships into machine learning algorithms introduces discrimination and bias to prediction. We summarize our main contributions and findings as below. First, we build a social network of Prosper dataset, and follow the state-of-the-art social financing models [17, 32] to derive a *social score* of borrowers based on Prosper social network. This social score is used as the social feature in our classification models. Second, we perform an extensive set of experiments on Prosper dataset to evaluate if the social score brings discrimination to individual fairness. Our results show that some similar loans (same non-social features and very close social scores) receive different classification results. This shows that individual fairness can be easily harmed by incorporating the social score in classification. Third, we also perform experiments to evaluate counterfactual fairness based on counterfactual examples constructed from different types of social relationships, e.g., a borrower who has many friends of low default risk versus someone who has many friends of high default risk. We evaluation if the classifier treats these different counterfactual examples fairly. Our study shows that the classifier does not treat these different counterfactual examples fairly. Fourth, to mitigate the bias introduced by the social score, we design two new methods that *generalize* continuous social scores into discrete values. Our results show that different generalization granularities lead to different degrees of bias mitigation for both individual and counterfactual fairness.

The rest of the paper is organized as follows. Section 2 presents the preliminaries. Section 3 explains how to compute social scores. Section 4 defines our fairness measurement metrics. Section 5 discusses the experimental setup. Section 6 presents the main experimental results of fairness evaluation. Section 7 presents our bias mitigation methods and their performance. Section 8 introduces the related work. Finally, Section 9 concludes the paper.

2 PRELIMINARIES

2.1 Classification Methods for Loan Approval

The assessment of whether a loan application can be approved or denied is accomplished by estimating the loan's default probability through analyzing a historical dataset and then classifying the loan into one of two categories: (a) *high risk* - likely to default on the loan (i.e., be charged off/failure to pay in full) and (b) *low risk* - likely to be paid off in full.

Typically, the classification algorithm takes customers' personal information (age, gender, martial status, job, income, etc.), credit information (monthly payment amount, interest rate, etc.), credit history (payment history and delinquencies, amount of current debt, types of credit in use, etc.), and bank account behavior (average monthly saving amount, maximum and minimum levels of balance, number of missed payments, etc.). We call these features *non-social*. Besides these non-social features, the borrowers' social information can be used as *social* features. In this paper, we consider a social feature that is modeled from the borrowers' social relationships embedded in the P2P lending platforms. How the values of these social features are calculated will be discussed in Section 5. The social features will be employed together with non-social features by a classification algorithm for loan approval/denial decision-making. In this paper, we explore a few classification algorithms, including random forest, k-nearest neighbors, logistic regression, naive Bayes, SVM, AdaBoost, gradient boosting, and neural networks. More details of the classification algorithms we used can be found in Section 5.

Formally, given a labeled dataset D where each record represents an individual loan application, each record consists of k features $X = \{x_1, \ldots, x_k\}$. The class label $y \in \{0, 1\}$ is the variable that the model tries to predict for each loan application. A positive class label $y = 1$ expresses that the loan application is of low risk, while a negative class indicates a high risk loan application. We consider a classifier $\mathcal{H}$ that produces a prediction $\hat{y}$, with the aim to minimize some notion of error between y and $\hat{y}$. For notation simplicity, we restrict the definitions to a single binary class, but they can be easily generalized to multi-class classification problems. We use S and T to denote the social and non-social features respectively.

2.2 Individual Fairness

Over the past few years, the machine learning community has proposed a multitude of formal, mathematical definitions of fairness. These fairness definitions can be categorized into two broad classes, namely *group* fairness and *individual* fairness. Group fairness is concerned with a small number of protected subgroups (such as racial or gender groups) and requires that some statistic of interest should be approximately equalized across groups. Standard choices for these statistics include positive classification rates [5], false positive or false negative rates [16, 21] and positive predictive values [6]. On the other hand, individual fairness [9] prevents discrimination against individuals and requires similar individuals are treated similarly.

Given the fact that social information is typically associated with individuals, the fairness of social-score-based classification can be defined without reference to groups. Therefore, we adapt the definition of individual fairness [9] to our setting. At a high level, individual fairness requires that similar individuals should receive the same classification results. Next, we formally define individual fairness. Given two records r, r', they are similar (denoted as $r \approx r'$) if $d(r, r') \leq \epsilon$, where d is a distance metric, and ϵ is a user-specified threshold.

Definition 2.1 (Individual fairness [9]). A predictor achieves *individual fairness* if and only if for any two similar records r, r', they must satisfy that $\mathcal{H}(r) \approx \mathcal{H}(r')$.

In this paper, since $\mathcal{H}$ is a binary classifier, we require $\mathcal{H}(r) = \mathcal{H}(r')$ instead of requiring $\mathcal{H}(r) \approx \mathcal{H}(r')$. Dwork *et. al* [9] have shown that the notion of individual fairness can be captured by (D, d)-Lipschitz property, which states that $D(\mathcal{H}(r), \mathcal{H}(r')) \leq d(r, r')$, where D is a distance measure for distributions. In general, individual fairness is agnostic with respect to its notion of similarity metric, since there is no unified way of defining similarity.

2.3 Counterfactual Fairness

Counterfactual fairness investigates how the prediction would change if the concerned features were changed to different values. These different values are called the *counterfactual examples*. In particular, let $\Phi(r)$ denote the set of counterfactual examples associated with an example r. Counterfactual fairness requires the predictions of a model for all counterfactual examples are within a specified error. Formally,

Definition 2.2 (Counterfactual fairness based on counterfactual examples [12]). A classifier $\mathcal{H}$ is counterfactually fair with respect to a counterfactual generation function Φ and some error rate θ if

$$|\mathcal{H}(r) - \mathcal{H}(r')| \leq \theta, \ \forall r \in R, r' \in \Phi(r),$$

where θ is a user-defined threshold. Since $\mathcal{H}$ is a binary classifier, we require $\theta = 0$.

3 COMPUTATION OF SOCIAL SCORE

Most of the existing social financing models [3, 15, 32] follow the same strategy of computing a *social score* to measure a borrower's "position in a social structure based on esteem that is bestowed by others" [17] using his/her social network information. In this paper, we consider the latest social scoring scheme [32] of utilizing the social relationship of borrowers to compute the social score. The scoring scheme categorizes the borrowers into two types: *positive* and *negative*. The positive borrowers have a low risk of loan default, while negative ones have a higher default risk. The intuition behind this social scoring scheme is that a borrower who is connected with more positive-type friends should be more likely to be a positive type, and thus receive a high social score. Based on this intuition, the scoring scheme measures the social score as the probability that a borrower is of positive type given his/her social network. We must note that although the borrowers have been categorized into positive or negative types based on their loans (Section 5.2), this categorization only delivers a binary decision, and it does not consider the social networks of borrowers. A numerical social scoring system better quantifies the belief that a borrower belongs to the positive or negative type by taking the social networking information into consideration. Next, we explain how to compute the social score in detail.

Given a social network G, the social score s_i of a borrower $u_i \in G$ is calculated as the probability of u_i being positive type given its connections in G:

$$s_i = P(u_i = pos|\mathbf{Y_i}) = \frac{1}{1 + \left(\frac{\lambda}{1-\lambda}\right)^{g_i} \left(\frac{\lambda p+(1-\lambda)}{\lambda+(1-\lambda)p}\right)^{L_i} \left(\frac{\lambda+(1-\lambda)p}{\lambda p+(1-\lambda)}\right)^{H_i}}, \tag{1}$$

where $\mathbf{Y_i}$ is the 1-hop neighbors of u_i in G, $g_i \in \{-1, 1\}$ is the observed signal of u_i, which is calculated by the loan history (more details in Section 5.2), L_i and H_i are the number of negative-type and positive-type borrowers in $\mathbf{Y_i}$, λ is the probability of wrong observations, and p is the probability that two users of different observed types are connected in G. We note that λ must be set as $\lambda < 0.5$, to ensure $\frac{\lambda p+(1-\lambda)}{\lambda+(1-\lambda)p} > 1$, and $\frac{\lambda+(1-\lambda)p}{\lambda p+(1-\lambda)} < 1$. By such setup, the number of negative- and positive-type friends for a borrower's social connections affects the assessment of that borrower's creditworthiness in different directions. In particular, the social score decreases when L_i increases (i.e., u_i has more negative-type friends), and increases when H_i increases (i.e., u_i has more positive-type friends). When $L_i \to 0$, and $H_i \to \infty$, the social score $s \to 1$.

4 FAIRNESS MEASUREMENT

In this paper, we mainly focus on individual and counterfactual fairness. In this section, we explain the evaluation metrics of individual and counterfactual fairness that we use.

Individual fairness. One challenge of evaluating individual fairness is the definition of the similarity metric, as there is no unified way of defining similarity of individuals. Therefore, in this paper, we consider the most conservative similarity function that accepts individuals whose social features only differ slightly as similar. Formally, let S and T be the social and non-social features. We use lowercase s and t to denote the value of the variables S and T.

Definition 4.1 (Similarity Function). Given two individual records r and r', we say r and r' are similar, denoted as $r \approx r'$, if: (1) $t = t'$; and (2) $|s - s'| \leq \epsilon$, where ϵ is a user-specified threshold.

This similarity function guarantees that any pair of similar records must have the same non-social feature values, and thus must always have the same prediction results if only non-social features are used for classification. Therefore, any two similar records that receive different prediction results after taking the social score into consideration can be considered as discrimination incurred by the social score. Based on this reasoning, we measure the *bias* as the percentage of records receiving unfair treatment (i.e., their similar peers receive different classification results). Formally,

Definition 4.2 (Bias). Given a set of records R and a classification algorithm $\mathcal{H}$ on R, let $B = \{r \in R | \exists r' \in R \text{ such that } r \approx r', \hat{y} \neq \hat{y}'\}$ (i.e., the set of similar records that receive different classification results). We measure the bias b of $\mathcal{H}$ as $b = \frac{|B|}{|R|}$.

Apparently, our bias measurement eliminates the impact of non-social features on individual fairness, as those similar individuals must have the same non-social feature values, and thus must receive the same classification results (and must be fair).

Counterfactual fairness. We use the *counterfactual token fairness gap* (CFGAP) metric [12] to evaluate counterfactual fairness with respect to a given counterfactual generation function. Formally, for a single example x, the counterfactual token fairness gap is measured as the average gap in prediction over all of the counterfactual pairs for that example r:

$$\text{CFGAP}(r) = \mathbb{E}_{r' \in \Phi(r)} |\mathcal{H}(r) - \mathcal{H}(r')|, \tag{2}$$

where $\Phi(r)$ denotes the set of counterfactual examples associated with an example r. Over an entire dataset, the gap is the average of all examples that have valid counterfactuals. Formally, given a testing dataset R, the counterfactual token fairness gap (CFGAP) of R is measured as:

$$\text{CFGAP}(R) = \frac{\sum_{r \in R} \text{CFGAP}(r)}{|R|}. \tag{3}$$

We will explain how to generate counterfactual examples of the social score in Section 6.5.

5 EXPERIMENTAL SETUP

5.1 Dataset

We use the Prosper Loans Network Dataset[1], which contains the loan data collected from Prosper Inc[2], America's first peer-to-peer online money lending network which has more than two million members and over twenty billion US dollars in funded loans. The Prosper dataset contains 1,048,575 loan records occurred from November 2005 to September 2011. Each record contains nine features:

- *Lender ID*: the ID of the member who contributed to this loan;
- *Borrower ID*: the ID of the member who received funds from this loan;
- *Timestamp*: the timestamp of the loan;
- *Amount*: the amount of the loan;
- *Status*: the status of the loan. It has 11 discrete valuess: *paid*, *payoff*, *repurchased*, *late*, *defaulted*, *current*, *1 month late*, *2 months late*, *3 months late*, *charge-off*, and *cancelled*.
- *Lender rate (rate1)*: the interest rate that the lender will receive;
- *Borrower rate (rate2)*: the interest rate that the borrower will pay, usually the same as lender rate;
- *Rating*: the rating of the loan is assigned with one of the following values: AA, A, B, C, D, E, HR (in descending order). There are 1762 loan records that have missing rating values. These missing values were denoted as NC.

The dataset contains 46538 (67.59%) lenders and 26268 (38.15%) borrowers, in which 3957 (5.75%) users as both lenders and borrowers.

5.2 Classification Setup

Ground truth of loan classification. The ground truth of the label Y (i.e., high/low-risk of loans) is generated from the *Status* feature. By consulting with an expert in bank finance and the Prosper Q&A webpage[3], we divide the given eleven distinct values of *Status* feature into the following two classes:

- *high-risk* ($y = 0$): Status = "late", "defaulted", "current", "1 month late", "2 months late", "3 months late", "charge-off", and "cancelled"[4];
- *low-risk* ($y = 1$): Status = "paid", "payoff", and "repurchased".

There are 405,486 (38.67%) loans labeled as high-risk, and 643,089 (61.33%) loans labeled as low-risk.

Categorization of borrowers based on loans. We categorize the borrowers into positive/negative type as following. For each borrower u_i, we count the number of high-risk loans h_i as well as the number of low-risk loans l_i that u_i has. If $h_i < l_i$, we label u_i as positive type, otherwise, we label u_i as negative type. We use the positive/negative type as the value of g_i, the *observed signal* used in Formula 1 for the calculation of social scores. In particular, $g_i = -1$ when the user is a negative type and $g_i = 1$ when the user is a positive type. We must note that the types of borrowers are different from the types of loans. Our goal is to predict the loan types in the testing data, with the knowledge of the type of borrowers collected from the training data.

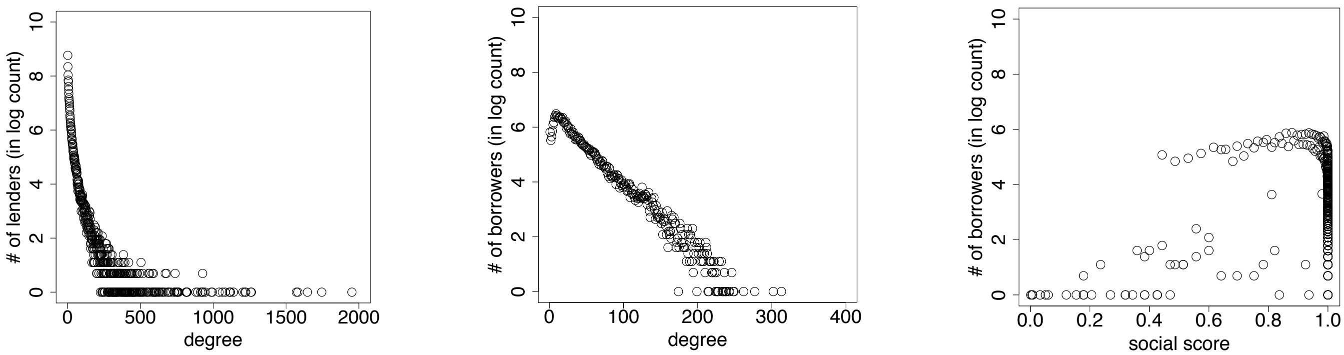
Training and testing data. We randomly pick 80% of the dataset (838,860 records) for training, and the remaining 20% of the dataset (209,715 records) for testing. In these 209,715 records, 81,493 (38.85%) records are high-risk, and 128,222 (61.15%) records are low-risk.

Classification models. We do not consider the *timestamp* feature in classification. We use the features *amount*, *rate1*, *rate2*, *rating* (as non-social features), and *social score* (as social feature) for classification. How the social score is computed was explained in Section 3. We investigated a few classification algorithms, including random forest, k-nearest neighbors (KNN), XGBoost, logistic regression, naive Bayes, SVC, and neural networks. We only report random forest, KNN, and XGBoost given their better performances. We use the implementation of these algorithms from sklearn[5].

5.3 Evaluation Metrics

We measure the classification accuracy of the whole testing dataset as well as each class (i.e., high-risk and low-risk). We use |TP|, |TN|, |FP|, and |FN| to denote the number of true positive, true negative, false positive, and false negative loans respectively. True positive loans are the low-risk loans that are predicted as low-risk, true negative loans are the high-risk loans that are predicted as high-risk, false positive loans are the high-risk loans that are predicted as low-risk, and false negative loans are the low-risk loans that are predicted as high-risk. The classification accuracy Acc of the whole testing dataset is measured as $Acc = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$. For the high-risk records in the testing dataset, we measure the classification accuracy Acc_H as: $Acc_H = \frac{|TN|}{|TN|+|FP|}$. Similarity, for the low-risk records in the testing dataset, we measure their classification accuracy Acc_L as $Acc_L = \frac{|TP|}{|TP|+|FN|}$. We also measure the precision and recall. Precision is measured as $Pre = \frac{|TP|}{|TP|+|FP|}$, and recall is measured as $Rec = \frac{|TP|}{|TP|+|FN|}$, which is the same as Acc_L.

[1]http://mlg.ucd.ie/datasets/prosper.html
[2]https://www.prosper.com/
[3]https://prosper.zendesk.com/hc/en-us/articles/210013083-Where-can-I-download-Prosper-loan-data-
[4]More details about delinquency status of Prosper loans can be found at: https://prosper.zendesk.com/hc/en-us/articles/208500186-How-can-I-review-the-status-of-a-late-loan-.
[5]https://scikit-learn.org/stable/supervised_learning.html#supervised-learning



(a) Lender node degree distribution **(b) Borrower node degree distribution** **(c) Freq. distribution of social score**

Figure 1: Distribution analysis of Prosper social network graph and its social scores

6 FAIRNESS EVALUATION AND EXPLANATION

Since the Prosper dataset does not contain any demographic information of borrowers and lenders, it does not support the evaluation of group fairness that typically relies on the demographic features (e.g., gender and race). Thus we only focus on the evaluation of individual and counterfactual fairness. The goal of our case study of Prosper dataset is to understand the followings:

- Whether using the social score in classification improves the prediction accuracy. If it does, how important the social score is to the prediction accuracy;
- Whether utilizing the social score leads to violation of individual and counterfactual fairness. If it does, what is the reason of such violation?

6.1 Social Scores of Prosper Dataset

6.1.1 Prosper Social Network Graph. Since Prosper dataset does not contain any personal information of the borrowers and lenders, we cannot link it with any external social media data (e.g., Facebook and Twitter). Therefore, we follow the state-of-the-art work [32] to construct a social network of borrowers and lenders embedded in the Prosper platform. It has been shown that social networks play a significant role in predicting the repayment probability of borrowers [23]. Formally, each borrower or lender user corresponds to a vertex in the graph. There is an un-weighted edge directed from the vertex v_A to vertex v_B if user A had lent money to user B. The graph has 68,849 vertices and 1,037,284 edges. The number of edges is inconsistent with the number of transactions (1,048,575) because there are some lenders who contribute to the same borrowers multiple times. We analyze the degree distributions of lenders and borrowers in the Prosper graph, i.e., the number of loans that a lender has contributed or a borrower has received, as shown in Figures 1a and 1b respectively. The highest out degree of lender nodes is 1952 (Figure 1a), i.e., a lender has contributed to 1952 loans at maximum. The highest degree of borrowers is 313 (Figure 1b), i.e., a borrower has received funds from 313 loans at maximum. The degree distributions of lenders follow the power law distribution.

6.1.2 Social Scores of Prosper Dataset. To compute the social scores, we set $\lambda = 0.4$ as suggested by [32]. We calculate p as the fraction of edges in the Prosper social network graph that connect borrowers of different types. It turned out that p = 0.39 for the Prosper dataset. The frequency distribution of the social scores is shown in Figure 1c. It can be observed that the distribution of social scores is much skewed. The maximum, minimum, and average of all social scores of Prosper dataset are 1, 0.0016, and 0.9362 respectively. The standard deviation is 0.1129. Most of the social scores are scattered in the range [0.5, 1]. There are 80 borrowers out of 26,268 borrowers who are associated with the social score 1. Furthermore, 14,022 borrowers have social scores that are greater than 0.99.

We also observe the distribution of social score s is correlated with the number of negative-type and positive-type neighbors L_i and H_i. The association rules between s and L_i/H_i are listed below:

- When $L_i \geq H_i$ the social score $s \in (0, 0.5)$.
- When $0 < H_i - L_i \leq 10$, $s \in [0.5, 0.9)$;
- When $10 < H_i - L_i \leq 25$, $s \in [0.9, 0.99)$;
- When $H_i - L_i > 25$, $s \in [0.99, 1]$;

The social scores also depend on the observed signal g_i. Among the 26,268 borrowers, 10,302 are observed as negative-type, while the remaining are observed as positive-type. We checked the Prosper social relations of those 80 borrowers whose social score is 1. For all of them, their number of positive-type friends largely dominates the number of negative-type friends. For example, some of them have 239 positive-type friends and 22 negative-type friends, and some have 141 positive-type friends and no negative-type friends. We have to note that not all these 80 borrowers are observed as positive type, although their social score is 1. 58 of them are observed as positive-type, while the other 22 are observed as negative-type.

6.2 Importance of Social Feature

First, we evaluate if using the social score as a feature indeed can improve the classification accuracy. We run a number of classification algorithms (listed in Section 5) with and without the social score. All of these algorithms witnessed at least 10% accuracy improvement by using the social score. Among all these classification algorithms, random forest, k-nearest neighbors, and XGBoost witnessed the best accuracy improvement by the social feature. Therefore, in the rest of the paper, we mainly focus on these three classification algorithms. We measure the classification accuracy of the whole testing dataset as well as the accuracy for low-risk and high-risk loans in the testing dataset separately. Recall that in the testing data, 38.85% records are high risk and 61.15% records are low risk. The accuracy,

Table 1: Classification performance (recall, precision, and accuracy) with vs. without social score

Classification model	without social score					with social score				
	Precision	Recall	Overall *Acc*	High risk *Acc*	Low risk *Acc*	Precision	Recall	Overall *Acc*	High risk *Acc*	Low risk *Acc*
Random forest	0.77	0.84	0.75	0.6	0.84	0.92	0.95	0.92	0.87	0.95
K-nearest neighbors	0.74	0.8	0.7	0.55	0.8	0.88	0.91	0.87	0.81	0.91
XGBoost	0.74	0.84	0.72	0.53	0.84	0.93	0.95	0.92	0.88	0.95

Table 2: Feature importance before & after using social score

Feature	Feature importance	
	without social score	with social score
amount	0.2	0.04
rate1	0.33	0.19
rate2	0.4	0.23
rating	0.07	0.05
social score	N/A	0.5

Table 3: Dependency between non-social features and social score

	Non-social feature			
	amount	rate1	rate2	rating
Pearson correlation	-0.09	-0.22	-0.22	-0.31
Mutual information	0.26	1.5	1.5	0.23
Causal relation (non-social→ social)	0.0015	-0.0023	-0.0019	-0.0017
Causal relation (social→ non-social)	-0.0027	0.0019	0.0019	0.0019

precision, and recall of the three classification algorithms are listed in Table 1. We observe that accuracy, precision, and recall on the high-risk loans are improved by the social score much more than the low-risk loans, although the accuracy, precision, and recall on the low-risk loans still remains higher than that of the high-risk loans.

To have a better understanding of the importance of the social score to prediction accuracy, we measure feature importance output by random forest before and after using the social score, and show the results in Table 2. The observation is that the importance of the social score dominates all the non-social features. This convinces the use of the social score for classification. Due to the importance of the social score, the classification results are highly sensitive to the social score. Thus involving the social score incurs high risk of fairness violation. This leads to the trade-off between prediction accuracy and fairness, which we will investigate later.

6.3 Dependence between Social and Non-social Features

In this section, we evaluate three types of dependence between non-social features and the social score: (1) linear dependence evaluated by Pearson correlation; (2) non-linear dependence evaluated by mutual information; and (3) causal relation.

Pearson correlation. Table 3 shows the Pearson correlation between the social score and each non-social feature. The main observation is that the absolute value of Pearson correlation between social and each non-social feature does not exceed 0.3. In other words, the linear correlation between social and non-social features is weak.

Pairwise mutual information. Table 3 shows the pairwise mutual information between the non-social features and the social score. Apparently, the mutual information between any non-social feature and the social score is always less than or around 1. Thus, little information (about 1 bit) can be obtained about the social score through observing the non-social feature.

Causal relation. One way to understand how the social relationships impact the fairness of classification is through causal inference [22]. Formally, given two random variables X and X', X *causes* X' if there exists a mechanism F that transforms the values taken by the cause X into the values taken by the effect X'. Mathematically it is denoted as $X' \leftarrow X$. Intuitively, if the value of the cause X is changed, then a change in the value of the effect X' would follow. The change is not symmetric (i.e., the change of the value of the effect X' is not followed by a change in the cause X). We use the causal discovery tool to measure the pairwise causal relation between each non-social feature and the social feature in both directions. [6] For any two given attributes x and x', the directed causal relation between x and x' is measured in the domain [-1,1]. The causal relation valued 1 means that x causes x', -1 means x' causes x, and 0 means there is no causal relation between x and x'. We report the results of the causal relation in Table 3. The pairwise causal relation between any non-social and the social feature is always close to 0. In other words, the causal relation between the social feature and the non-social features is weak.

To summarize, the dependence studies show that the social score is not correlated to the non-social features. This suggests that removing the social feature can eliminate the discrimination that it brings towards the classification results. However, the social score also is the most important feature for classification. Thus it cannot be simply removed for the concern of classification accuracy. In Section 7, we will discuss how to mitigate bias without removing the social feature from the model.

6.4 Evaluation of Individual Fairness

We perform two sets of experiments to evaluate the impact of the social feature on individual fairness.

- **One classification model** We use random forest as the classification model, and consider the similarity function (Def. 4.1) with various similarity threshold values. We aim to study if the social feature impacts the individual fairness for this particular setting of similarity function.

[6] https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox

- **Multiple classification models.** We consider three classification models: random forest, XGBoost, and k-nearest neighbors. Our goal is to study if adding social feature will bring bias for all the three classification models.

We must emphasize that our similarity function (Def. 4.1) guarantees that similar records always receive the same classification results before considering the social score. Thus any pair of similar records that receive different classification results when the social score is taken into consideration will act as the evidence of bias incurred by the social feature.

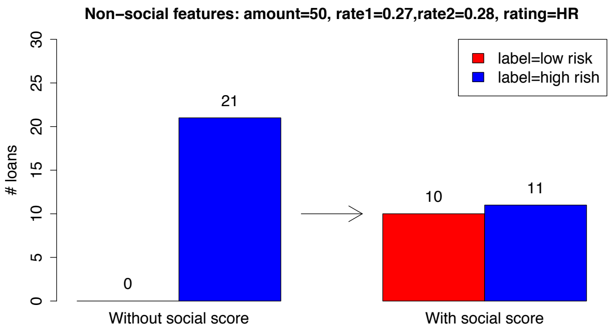


Figure 2: An example of loan records that violate individual fairness

6.4.1 One Classification Model. We use Figure 2 to illustrate 21 loans whose prediction results changed after using the social score. We use the similarity threshold $\epsilon = 0.3011$ for the similarity function (we will explain why we choose 0.3011 for ϵ later in this part). These 21 loans are associated with the same non-social values *amount* = 50, *rate1* = 0.27 *rate2* = 0.28, *rating* = "HR". All of them were classified as high risk before using the social score as shown in the left part of Figure 2. However, after taking the social score into consideration, 10 of them are classified as low risk, while the rest 11 loans remain as high risk as shown in the right part of Figure 2. Since all of these loans are of the same values on non-social features, the difference of their social scores determines if they are similar or not. The 10 low-risk loans are associated with four different social scores, namely 0.8966, 0.9542, 0.9676 and 0.9959; while the social scores of the rest 11 high-risk loans are associated with five different social scores, namely 0.6948, 0.8212, 0.8671, 0.9741 and 0.9945. We have to note that it is not necessary that high social scores always lead to low-risk decision in the prediction results. There are $10 \times 11 = 110$ loan pairs, each containing one low-risk and one high-risk loan chosen from these 21 loans. We measured the difference of social scores between any pair of these 110 pairs. The maximum score difference is 0.3011, the same as the threshold ϵ. Recall that these loans have the same values on the non-social features. Thus any pair of them must be similar. Indeed, for any ϵ value such that $\epsilon > 0.3011$ (i.e., the maximum score difference), all the 110 paired loans must be considered as similar. Since these similar loans receive different classification results, they will be the evidence that the social score introduces discrimination to individual fairness.

Next, we change the similarity thresholds. We found 20 records associated with the same non-social values *amount* = 100, *rate1* = 0.215, *rate2* = 0.22, *rating* = "D". All of them were classified as low risk before using the social feature. However, after adding the social feature, 10 of them are classified as high risk, while the other 10 remain as high risk. The 10 high-risk loans are associated with five unique social scores, namely 0.9169, 0.9294,0.9816,0.9845, and 0.9998, and the 10 low-risk loans are associated with two social scores: 0.9987 and 0.9999. The minimum distance of the social score of any pair of these 10 high-risk and 10 low-risk loans is 0.0001, and the maximum distance of the social score is 0.0705. There are $10 \times 10 = 100$ pairs of loans, one of high-risk and the other of low-risk. Since these loans have the same values on non-social features, any value of the threshold such that $\epsilon > 0.0705$ make the loans in each of these 100 pairs similar. However, they receive different classification results. Therefore, the classification of these loans violates individual fairness.

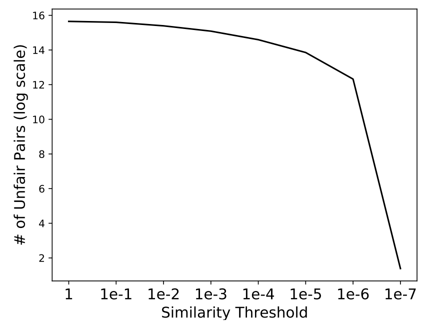


Figure 3: # of unfair pairs w.r.t. various similarity thresholds

6.4.2 Multiple Classification Models. We consider three classification models, namely random forest, XGBoost, and k-nearest neighbors. For each model, we use three different similarity thresholds. We find the set of loans whose classification results become unfair for each similarity threshold. Then we intersect the three sets of unfair loans. Apparently, the intersection results include those loans whose classification results are changed in *all* the three classification models. It turned out there are 19519 records changed after using social scores for all models. Among them, 56 loans associated with the same non-social values *amount* = 200, *rate1* = 0.2, *rate2* = 0.2, *rating* = "C". All of them are all classified as low risk before using the social feature; after adding the social feature, 33 of them are classified as high risk by all the three classification models, while 23 of them remain the same as low risk, and 1 of them get different results from different models. The 33 high-risk loans are associated with 11 social scores 0.7638,0.8212,0.9571,0.9637,0.9741,0.9923,0.9981,0.9986,0.9988,0.9999 and 1, and the 23 low-risk loans are associated with 12 unique social scores 0.8966, 0.9767, 0.9902, 0.9917, 0.9931, 0.9951, 0.9959, 0.9965,0.9975,0.9987,0.9996,and 0.9998. There are $33 \times 23 = 759$ pairs of loans, one of high-risk and the other of low-risk. The maximum distance of the social score of all 759 pairs is 0.236. Apparently, any pair of these loans in 759 pairs is considered similar if the threshold $\epsilon \geq 0.236$, the maximum social score difference.

6.4.3 Analysis of Similarity Threshold. In particular, for one classification model setting, we vary the similarity threshold of social scores, and count the number of *unfair* pairs that receive the same prediction results before using social scores but classified differently after using social scores. The results are shown in Figure 3.

We show the log scale (base=e) of the count number as it is large (6, 263, 990 for threshold=1), The results show that the prediction are very sensitive to social scores. Even when the threshold is as small as $1e-6$, there are still 224,497 unfair pairs.

6.5 Evaluation of Counterfactual Fairness

Generation of counterfactual examples. Since the social score is a numerical value, we cannot use all possible values of the social score as the counterfactual examples. Therefore, we generate the counterfactual examples of the social network structure instead. In particular, for a given user u_i, let H_i and L_i be the number of positive-type and negative-type friends of u_i in his/her original Prosper social network. We consider the set of counterfactual examples $\Phi(u_i)$ that consists of three types of counterfactual social networks of u_i:

- *Opposite social type*: we switch the values of H_i and L_i. Intuitively, if u_i have more positive-type (negative-type, resp.) friends in the original social network, his/her counterfactual social network will have more negative-type (positive-type, resp.) friends.
- *Less-active social type*: we set $H_i = H_i/2$ and $L_i = L_i/2$.
- *More-active social type*: we set $H_i = H_i * 2$ and $L_i = L_i * 2$.

Evaluation of counterfactual fairness. We measured CFGAP of the three counterfactual examples. The result of CFGAP is 0.07. Apparently it violates counterfactual fairness (Definition 2.2) as $CFGAP > 0$. This shows that the social score brings non-negligible amounts of discrimination to the classification results. We also measured CFGAP for each individual counterfactual example. The CFGAP of opposite, less-active, and more-active counterfactual examples are 0.03, 0.07, and 0.11 respectively. This shows that changing the social type from positive/negative-type to negative/positive-type has the largest impact on counterfactual fairness. Furthermore, shrinking the social network size also has moderate impact on counterfactual fairness, as each friend plays a more important role when there are fewer friends. Moreover, enlarging the social network size has the least impact, as increasing H_i and L_i results in relatively smaller change of the social score.

7 BIAS MITIGATION METHODS

As shown by the empirical study in Section 6, using the social score in classification can bring discrimination against both individual and counterfactual fairness. Since the social score has weak correlations with the non-social features (as shown in Section 6.3), an easy solution of bias mitigation is to remove the social scores. However, given the importance of the social scores to the prediction accuracy, it is not an ideal solution to remove social scores completely from the learning process. An alternative solution is to add fairness constraints to the objective function [20, 35], but this solution is expected to hurt the prediction accuracy significantly if the constraint is too rigid.

In this paper, we design a new bias mitigation method that uses a *generalized* social score instead of the original one in classification. Intuitively, all the original social scores are split into ℓ continuous ranges, where each range corresponds to a discrete value (e.g., *low*, *medium*, and *high*). Next, we present the details of our generalization schemes (Section 7.1) followed by an empirical study of our schemes (Section 7.2).

7.1 Generalization Schemes

We design two generalization schemes to generate the generalized social scores: (1) the *equal-width* generalization scheme; and (2) the *equal-size* generalization scheme. Both generalization schemes map the given k unique social scores to $\ell < k$ ranges, where these ranges either are of the same width (equal-width) or contain the same number of social scores (equal-size). We explain the key ideas of both generalization schemes as below.

Equal-width generalization scheme. Using this scheme, all unique social scores are assigned to ℓ continuous ranges that are of the same width. More precisely, each range is of width $r = (max - min)/\ell$, where min and max are the minimum and maximum of the k given social scores. Each range corresponds to a discrete generalized value. As an example, consider the social scores whose distribution is shown in Figure 4a, Figure 4b shows one of its equal-width generalization scheme of $\ell = 10$ ranges. All 10 ranges are of the same width.

Equal-size generalization scheme. The equal-width generalization scheme cannot deal well with input data of skewed distribution. To deal with the social scores of skewed distribution, we design the *equal-size* generalization scheme by which the social scores are split into ℓ ranges, where each range contain similar number of social scores (including the repeated ones). The ranges can be generated by constructing an equal-height histogram of the social scores. Figure 4c shows one example of the equal-size generalization scheme for the social scores in Figure 4a.

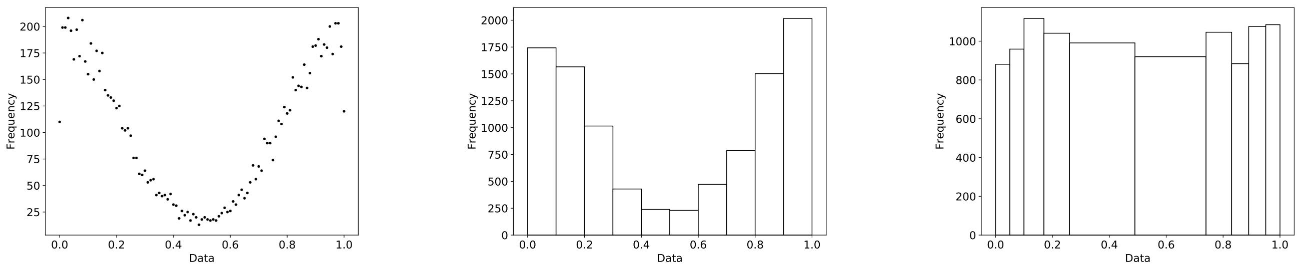
For both schemes, intuitively, fewer ranges lead to more generalized social scores. More generalized social scores lead to less accurate but more fair prediction. Therefore, we can address the trade-off between accuracy and fairness by controlling k, the number of generalization ranges.

7.2 Evaluation of Bias Mitigation

In this section, we present the evaluation results of the two bias mitigation methods for both individual and counterfactual fairness.

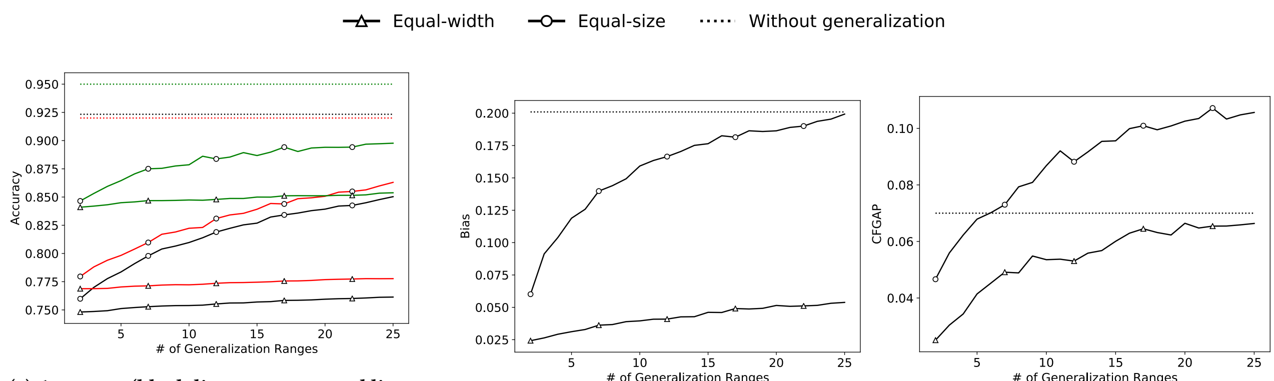
Accuracy. We measure the classification accuracy on the generalized data, and show the results in Figure 5a. Unsurprisingly the accuracy degrades after generalization for both schemes. However, the equal-size generalization scheme witnesses much less accuracy loss than the equal-width scheme, as it handles better with skewed data distribution than the equal-width scheme.

Individual fairness. We vary the number of generalized ranges for both equal-width and equal-size generalization schemes, and measure bias (Def. 4.2) by these two generalization schemes. We choose up to 25 generalization ranges. The results are shown in Figure 5b. The first observation is that more generalization ranges always lead to less generalized values, and thus higher accuracy as well as higher bias. This is straightforward due to the trade-off between accuracy and fairness. Second, both equal-size and equal-width generalizations schemes witness decrease in bias. This demonstrates the effectiveness of these schemes for bias mitigation. Furthermore, the equal-width generalization scheme always has smaller bias than the equal-size scheme. To explain this, we compared the distribution of social scores before and after generalization for both schemes. It turned out that some ranges by the equal-size generalization scheme are very small (e.g., of width 0.01). By those small ranges, some social scores that are similar



(a) Distribution of input data (b) Equal-width histogram (c) Equal-size histogram

Figure 4: Illustration of two generalization schemes (# of generalization ranges = 10)



(a) Accuracy (black line: accuracy; red line: precision; green line: recall) (b) Bias for individual fairness (c) CFGAP for counterfactual fairness

Figure 5: Equal-width vs. equal-size generalization scheme

(e.g., 0.5 and 0.53) are put into different ranges, and consequently are generalized as different discrete values. Such change leads to different classification results and counts as bias. However, for the equal-width scheme, due to the equal width of all ranges, the distribution of social scores is similar before and after generalization. Thus the bias is smaller than the equal-size scheme. We must note that due to the trade-off between accuracy and fairness, although the equal-size scheme loses to the equal-width scheme in fairness, it wins in accuracy as shown in Figure 5a.

Counterfactual fairness. We still use the three counterfactual examples defined in Section 6.5. In particular, we change the social network structure for the three types of counterfactual examples, re-compute the social scores of these counterfactual social graphs, and generalize the social scores after re-computation. Then we measure CFGAP (Def. 2) based on the generalized social scores for both generalization schemes. The results are shown in Figure 5c. Our main observation is that the counterfactual fairness result is similar to that of individual fairness - the equal-size scheme delivers worse CFGAP than the equal-width scheme. Indeed, when there are more generalization, the CFGAP of the equal-size scheme can be larger than it is before generalization. However, the CFGAP of the equal-width scheme is always smaller than it is before generalization. The reason behind this observation is similar to our analysis for individual fairness - the equal-size scheme changes the distribution of social scores much more significantly than the equal-width scheme.

To summarize, there always exists the trade-off between accuracy and fairness. The equal-size generalization scheme is preferred if accuracy is considered with higher importance than fairness. Otherwise, the equal-width generalization scheme is a better candidate than the equal-size scheme given its effectiveness in bias mitigation.

8 RELATED WORK

Financial machine learning using social networking data. Social networking data has been used in various financial machine learning applications, including risk assessment for identify theft and fraud [26], financial performance analysis and prediction [30], credit scoring [23, 34], to name a few. In this paper, we mainly focus on the application of credit scoring by using the borrowers' social network information in P2P lending platforms, and study the fairness problem under this context.

Algorithmic fairness. Several competing notions of fairness have been recently proposed in the machine learning literature. The definition of fairness can be categorized into three types [28]: 1) it is not based on protected attributes such as gender or race (*fair treatment*), 2) it does not disproportionately benefit or hurt individuals (*fair impact*), and 3) given the target outcomes, it enforces equal discrepancies between decisions and target outcomes across groups of individuals based on their protected characteristic (*fair supervised performance*). Fair treatment can be implemented via fairness through unawareness [14] which ignores the protected attributes. Examples of fair impact constraints include 80% rule

[10] and demographic parity [4, 20]. Examples of fair supervised performance constraints include equal opportunity and equal odds [16] and de-correlation [35]. Most of these definitions focus on fairness of groups (i.e., individuals who share the same value on the protected attributes). Individual fairness [9, 22, 31] is defined as a non-preferential treatment towards an individual. Counterfactual fairness [12, 22] evaluates fairness in terms of causal inference and counterfactual examples. In this paper, we mainly focus on both individual and counterfactual fairness.

Bias mitigation algorithms. Broadly, the bias mitigation algorithms fall into three categories: (1) *pre-processing*: the bias in the training data is mitigated [4, 10, 19]; (2) *in-processing*: the machine learning model is modified by adding fairness as additional constraint [5, 13, 35]; and (3) *post-processing*: the results of a previously trained classifier are modified to achieve the desired results on different groups [16, 33]. Most of these methods mainly consider group fairness. Our bias mitigation methods are the first to address individual fairness and counterfactual fairness.

9 CONCLUSION AND FUTURE WORK

In this paper, we study if involving social relationships in classification tasks introduces any discrimination in the classification results. We construct a social network graph on the Prosper dataset, and implement a well-used social scoring scheme [23] to derive the social feature from the Prosper social network. We evaluate both individual and counterfactual fairness of the loan classification results with the social feature taken into consideration. Our experimental results show that although the social score can improve the prediction accuracy significantly, it introduces discrimination to both individual and counterfactual fairness, due to the high sensitivity of the classification results to the social score. This leads to the trade-off between prediction accuracy and fairness. Thus we design new bias mitigation methods to reduce the bias of prediction incurred by using social features. Our experimental results demonstrate the effectiveness of our bias mitigation approaches.

Our future work will focus on both extending the ideas in this paper to other types of fairness notions, models and domains, and providing theoretical performance guarantees. We will consider group fairness. In this paper, we did not evaluate group fairness due to the lack of demographic data of users in the Prosper dataset. However, in general, since the social relationships are highly correlated with the protected attributes (e.g., race and gender), the group fairness is expected to be affected significantly by the social features that model the social relationships. Additional case studies can be performed when more suitable data becomes available.

REFERENCES

[1] Friendlyscore inc. https://friendlyscore.com/.
[2] Lendingclub corporation: Peer-to-peer lending platform. https://www.lendingclub.com/.
[3] Dries F Benoit and Dirk Van den Poel. Improving customer retention in financial services using kinship network information. *Expert Systems with Applications*, 39(13):11435–11442, 2012.
[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.
[5] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
[7] Michael Chui et al. Social media to boost financial services. *InFinance: The Magazine for Finsia Members*, 127(2):34, 2013.
[8] Pam Dixon and Robert Gellman. The scoring of america: How. In *World Privacy Forum*, 2014.
[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
[10] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.
[11] Seth Freedman and Ginger Zhe Jin. Do social networks solve information problems for peer-to-peer lending? evidence from prosper. com. 2008.
[12] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.
[13] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.
[14] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.
[15] Michael Haenlein. A social network analysis of customer-level revenue distribution. *Marketing Letters*, 22(1):15–29, 2011.
[16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
[17] Yansong Hu and Christophe Van den Bulte. Nonmonotonic status effects in new product adoption. *Marketing Science*, 33(4):509–533, 2014.
[18] Patrick Jenkins. Big data lends new zest to banks' credit judgments. *Financial Times*, 23, 2014.
[19] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.
[20] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.
[21] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, pages 43:1–43:23, 2017.
[22] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.
[23] Mingfeng Lin, Nagpurnanand R Prabhala, and Siva Viswanathan. Judging borrowers by the company they keep: Social networks and adverse selection in online peer-to-peer lending. *Ssrn Elibrary*, 2009.
[24] Steve Lohr. Banking start-ups adopt new tools for lending. *New York Times*, 18, 2015.
[25] Tian Lu, Yingjie Zhang, and Beibei Li. The value of alternative data in credit risk prediction: Evidence from a large field experiment. In *Proceedings of the 40th International Conference on Information Systems, ICIS 2019, Munich, Germany, December 15-18, 2019*, 2019.
[26] Sunil Madhu, Giacomo Pallotti, Edward J Romano, and Alexander K Chavez. Risk assessment using social networking data, March 29 2016. US Patent 9,300,676.
[27] Frank Pasquale. *The black box society*. Harvard University Press, 2015.
[28] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*, pages 677–688, 2017.
[29] Evelyn M Rusli. Bad credit? start tweeting. *WALL Street Journal, April*, 1, 2013.
[30] Dara Schniederjans, Edita S Cao, and Marc Schniederjans. Enhancing financial performance with social media: An impression management perspective. *Decision Support Systems*, 55(4):911–918, 2013.
[31] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.
[32] Yanhao Wei, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas. Credit scoring with social network data. *Marketing Science*, 35(2):234–258, 2015.
[33] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.
[34] Xiujuan Xu, Chunguang Zhou, and Zhe Wang. Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications*, 36(2):2625–2632, 2009.
[35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017.