A Social Relationships Enhanced Credit Risk Assessment Approach

Caihong Sun, Chengcheng Deng, Wei Xu, and Jun Su(✉)

School of Information, Renmin University of China, Beijing 100872, China
{chsun, weixu, sujun}@ruc.edu.cn,
dengchengchengruc@163.com

Abstract. With the rapid growth of personal loan applications, credit risk assessment has become very crucial both in academic and industrial domain. Research literatures show that besides "hard" information, such as individual socio-demographic information and loan application information, "soft" information such as social relationships of the borrowers is a key factor to the credit risk assessment as social capital. In social networks, a user's position and its influence are affected not only by the direct relationships (its friends) but also the indirect relationships (friends' friends). A user's importance and influence in his communities are attractive and valuable for credit assessment. But due to data deficiency in real life, social relationships are rarely considered in lending markets. By leveraging data from various sources, we proposed a social relationship enhanced credit risk assessment system, by building a social network from users' geolocation data, extracting social relationship features at three different levels: ego, community and global level to capture a user's position and influence from direct relationships, community and whole network perspectives. A real-life loan granting dataset is utilized for verifying the performance of the system. The experiment results show that, by combining the conventional financial indicators along with the proposed social network features, our system outperforms benchmark methods. Novel social network features we proposed make a good contribution to the loan default prediction. The research highlights the power of social relationships in detecting the default loans.

Keywords: Credit risk assessment · Feature engineering · Social network · Community detection

1 Introduction

Credit risk assessment is a crucial process for financial institutions for their operations and sustainable growth. Both the increasing volume of individual unsecured loans and fast-growing default risk highlight the importance of credit risk assessment. For instance, in China, the number of bad loans is 1.83% by the end of 2018 based on the China Banking Insurance Regulatory Commission's (CBIRC) report, higher than 1.74% in 2017. How to assess the credit risk more accurately has been a crucial and hot topic both in academic and industrial domain.

At the beginning, the assessment of credit risk mainly depended on subjective judgments such as the 5Cs rules which referring as character, capacity, capital,

C. Douligeris et al. (Eds.): KSEM 2019, LNAI 11775, pp. 222–233, 2019.
https://doi.org/10.1007/978-3-030-29551-6_20

condition and collateral, and LAPP which referring as liquidity, activity, profitability and potentialities. However, the booming of the credit industry made it impossible to assess thousands of applicants completely manually but to automate the process. Hence various machine learning credit scoring models have emerged to help financial institutions enforcing efficient credit approval [1, 2]. The main concern of credit risk assessment is how to classify the applicants into two types of groups: default and non-default. Then, the evaluator may decide to reject the loan application or approve it. For these binary classification problems, besides classification models, the feature selection is a key factor for the accuracy of models. In a traditional credit risk assessment system, individual socio-demographic information and loan application information are designed as input for feature engineering; however, in online lending market, such "hard" information is often deficient and not easy to be acquired and "soft" features are needed to enhance the performance. As Seth and Ginger [3] argued that "soft" information such as social relationships among borrowers, could make up for the devoid of some "hard" information. The importance of social relationships, i.e., social network, in credit assessment, has been proved by many researchers in online P2P lending market [2, 3, 11–13]. Due to deficiency of social network data in real life, social network features are not commonly used in credit risk assessment model yet. In this paper, we proposed a method to build social network from users' geolocation log, and then proposed novel social network features from the community, i.e. social group perspectives to enhance prediction ability.

We propose an assessment method that incorporates both conventional data, such as individual socio-demographic information and loan application information, and data of applicants' social relationships. Our method acquires social relationships by establishing social network from applicants' geolocations information. By proposing a social relationship enhanced credit risk assessment system, the main contribution of this research is in two areas: first, this work demonstrates how social relationships could be obtained from geolocation information, and different types of data (structured data and network data) could be fused to train machine learning algorithms. Second, we propose several novel network features at community level, to capture users' social group features. Five frequently used classification models are employed in our experiments and the experiment results show that social network features matter in credit risk assessment systems.

This paper proceeds as follows. Section 2 outlines previous studies related to this work. Section 3 introduces our system framework and the extraction method for social relationships. Section 4 analyzes the empirical results with using this method and Sect. 5 provides conclusions, discusses the limitations and identifies future work.

2 Related Work

Our work focuses on introducing social relationship to enhance the effectiveness of credit assessment risk. We mainly talk about literatures in two areas: features of credit risk assessment and measurements of social networks.

2.1 Features of Credit Risk Assessment

The accuracy of a predictive model depends heavily on the feature extraction and engineering. When it comes to credit risk assessment, three types of features have been explored in academic domain: personal, behavioral and social network features.

Personal features are about the personal characteristics of applicants (borrowers). In traditional credit market, banks utilized a set of credit scoring mechanism to evaluate the credit level of applicants, such as 5Cs rule and LAPP. Many researchers have used personal features in their predictive models, including age, income, telephone, marital condition, career, gender, family size, credit history, assets, saving account, residence, the number of credit cards etc. [4–9].

Behavioral features are about what the applicants have done, which include data about financial and transaction history. Avery et al. [4] used both personal information and transaction history (customers' usage history for a six months period and its aggregation), the local economic situation and personal circumstances (like medical emergency) into a linear probability regression to assess credit risk. Sustersic et al. [8] considered transaction history such as cash inflows, outflows and their aggregation into a NN algorithm to predict the credit risk. Financial histories and transaction data are considered as important features for their credit risk assessment [5–7, 9].

Besides above "hard" information, social network features, capturing the social relationships of the applicants, are mainly introduced in online P2P lending market. Social network is a kind of soft information which can be produced and used without financial intermediaries [13]. Granovette [10] argued that when solving financial problems, it should consider not only individual himself, but also the social embeddedness of the individuals, i.e., the individuals' social capital. Lin [11] found that in internet lending markets, social relationships did affect the borrower's behavior after lending. Lin et al. used the data from Proper.com, which is an American internet financial credit market platform, and found that social network information is highly related to the borrowers' credit [12]. Further studies showed that the more trustworthy their friends are, the more trustable the applicants are. In online P2P credit market, information of social networks (or social capital) could help predict the default risk [11, 13].

To sum up, besides "hard" information, social network features are useful for credit risk assessment [11–13]. But since social relationships are hard to acquire in offline financial market, few previous studies incorporate social network features into credit risk predictive models. We use information technology to harden the soft information. By using users' geolocation information reported by their mobile phones, we build applicants' social relationships and incorporate novel social network features into the credit risk predictive models.

2.2 Social Network Measurements

Social network measurements could be classified at different levels: ego network, community network, and whole network. Measurements of whole network tell the position and roles of a node in global level. While "Ego" is an individual "focal" node, if we want to understand variation in the behavior of individuals, we need to take a closer look at their local circumstances, neighborhood [14]. "Neighborhood" is the

collection of ego and all nodes to whom ego has a connection at some path length (in our case we set path length to 1). Communities are associated with more highly interconnected parts in a social network. Identifying communities is crucial to the understanding of the structural properties of networks [18].

Network measures for ego networks consist of size, density, clustering coefficient etc., while network measures for whole networks mainly include betweenness, degree, average path length and degree distribution [13]. Community network measurements consist of size, density, the link ratio in and out community to measure the tightness of his community membership.

3 Social Relationships Enhanced Credit Risk Assessment Framework

Figure 1 illustrates the framework of our proposed system. There are four phases in the framework: data acquisition, data cleaning and processing, feature engineering and prediction models.

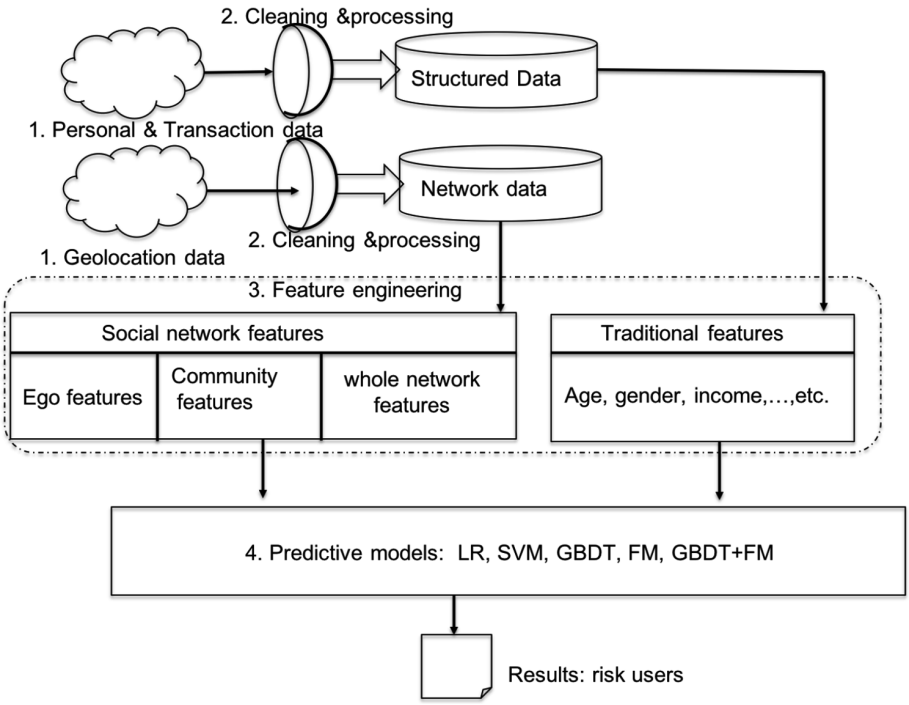

Fig. 1. Proposed credit risk assessment framework

3.1 Data Acquisition

The first phase is data acquisition. We picked a real-world credit data set, which contains two types of data: users' basic features and users' geolocation log data. The former consists of age, gender, years of working, the numbers of credit cards, marital status, which are the conventional features for credit risk prediction. The latter, users' geolocation data are gained by an app software. Via this app, users report the information about their geolocations and time which contain wifi name they connected, wifi address, check-in time and check-in location.

3.2 Data Cleaning and Processing

In the second phase, we clean and process the acquired data. Since basic features in our case are kinds of structured data, we just extract them into the structured database. We use the users' geolocation data to build social relationships among users. Inspire by [15–17], friendship network structure could be inferred by using mobile phone data and physical location. Hence, we use this spatial-temporal location information reported by users to presume users' social connections in real life. As friends tend to visit same places, we consider the number of co-occurrences and the number of locations that two people cooccurred as indicators of friendship. In this paper, if two users connect the same wifi at the same period of time (for example in half an hour, or 1 day), a tie is built between these two users. For each day we built a weighted network, and we aggregate all these networks during a period of time (in our case three months) into a whole network which is used to measure the social relationships among users. Although this network data is not real social relationships, it presumes some relationship among users. Here, we used this network to simulate the social relationships among users.

3.3 Feature Engineering

In the third phase, we mainly build social network features at three different levels: ego-level, community-level and the whole network level. Where ego-level features capture the individual network characteristics, the community level features capture the positions and roles in his social groups, and the whole network level features capture the importance in a society. To obtain community level features, we apply a community detection algorithm known as Clique Percolation Method (CPM) [18] to find all communities in the network. Table 1 lists the social network features we composed in three different levels. We use three measurements to identify network features at ego level: degree, clustering coefficient and the ratio of high-risk friends. Degree measures the local importance of a user in the network, and clustering coefficient measures the ratio of structural embeddedness, i.e. the ratio of his friends are also friends. The ratio of high-risk friends presumes the risk of the user from his social relationships. We use seven features to measure the characteristics of a user at community level. They are how many communities the user located in, the degree he sunk in his communities (his ties in or out of communities), how risk his communities are, and community features like community size, community density and degree dispersion. Finally, we use betweenness centrality to measure the user whole network characteristics. Betweenness centrality is to measure how importance a user is in controlling network communication aspect.

Table 1. Social network features.

Dimensions		Network features
Ego level (individual level)		Degree
		Clustering coefficient
		Ratio of high-risk friends
Community level	Individual	Number of communities he is in
		Number of neighbors in his communities
		Number of neighbors not in his communities
	Community	Ratio of high-risk users in communities
		Community size
		Community density
		Degree dispersion in community
Whole network level		Betweenness Centrality

Table 2. Basic features

Features	Descriptions
Age	User's age
Gender	Female: 0, Male: 1
Income	Annual income
Marital Status	Married: 1, Other: 0
The years he/she worked	Total years the user worked
Number of credit cards	How many credit cards the user has

Moreover, the basic features we extracted are listed in Table 2. They consist of users' age, gender, income, marital status, the years he/she worked, and the number of credit cards.

3.4 Predictive Models

In the fourth phase, we apply both social network and basic features into the predictive models and evaluate their performance. In our framework, five state-of-the-art models are used to test the effectiveness of our proposed social network features, i.e., LR, SVM, GBDT, FM and GBDT+FM.

We choose six measurements to evaluate the effectiveness of the predictive models: Accuracy, Recall, Precision, F-score, AUC and Logloss. The former four are measurements of model accuracy. AUC plots the true positive rate against the false positive rate. An AUC of 1 means a perfect classification whereas 0.5 refers to a random guess. Being more robust against prior distributions, AUC is considered by many researchers to be one of the best indicators of a classifier's performance. Logloss is a measure defined as the negative log-likelihood of the true labels given a probabilistic classifier's predictions, the smaller the value is, the better the prediction.

4 Experiment Analysis

In this section, we conduct several experiments to validate the effects of social network features based on LR, SVM, GBDT, FM and GBDT+FM models. We do the experiments by adding social network features or not to demonstrate the effectiveness of our proposed method.

4.1 Data Description and Basic Statistics

Two data sources provided by a company are used. One is the personal information of applicants for credit, which contains user ID, six basic features listed in Table 2 and one class label (1 as risk user and 0 as normal user). The other data is a user log data recording users' spatial-temporal location information reported by a mobile phone app software from February 1st to April 30th in an area of Guangzhou, China. Totally the users' geolocation data have 6361 users, and 2890096 log records containing information about user ID, wifi name, wifi address and connecting time, check-in time and check-in location. Among 6361 users, there are 993 high risk users are identified. As mentioned in Sect. 3.2, these geolocation data are used to build a social network among users. By deleting isolated users, we obtained a network with 2666 nodes (each node is an individual user), and 34683 links. There are 672 high-risk users. The average degree is around 26.

By applying CPM algorithm [18] to social network data (we set k = 6), we obtain 126 overlapped communities. There are 1212 users not belong to any community. The largest community contains 1272 users. One user locates in 90 communities. 442 of 672 high-risk users are located in at least one community. Three communities have no high-risk users and one community has 85.7% high-risk users which could be labelled as high-risk community.

Furthermore, we compute social network features mentioned in Table 1 and get basic features listed in Table 2 (Sect. 3.3). All the social network features and basic features are as our input features, and class label is 1 or 0 (1: high-risk users and 0: normal users). Now our credit risk assessment could be taken as a binary classification problem. We randomly choose 80% data as training set and the other 20% as testing data.

4.2 Evaluation Criteria

Any item in the prediction can be described with 4 types: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). In credit risk assessment, a high FP is a serious problem for the prediction model, because it would lead to a high risk of capital loss for banks when lending money to a person who would actually default on the loan. In this paper, six commonly employed measures are applied to evaluate the model: Accuracy, Precision, Recall, F-score, AUC and Logloss. These evaluation criteria are introduced as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{2}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{3}$$

$$F - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) is a primary indicator when measuring the classifier without the influence of class distribution. The ROC curve is plotted to reveal the relation between the sensitivity and specificity, with sensitivity on the x-axis and specificity on the y-axis. AUC is the area under the ROC curve, ranging from 0 (no discrimination ability) to 1 (perfect discrimination ability).

Logloss is defined as $-\frac{1}{N}\sum_{i=1}^{N}(y_i \log p_i + (1 - y_i)\log(1 - p_i))$, where y_i is real class of data x_i, and p_i is the probability that sample x_i belongs to positive class. The smaller the value is, the better the prediction.

4.3 Experimental Design and Analysis

4.3.1 Experimental Design

We classified the model input as three parts: basic features, social network features (ego level + whole network level) and community level network features. To test the effectiveness of social network features, we set basic features input as baseline model, and then introduce social network features (ego level + whole network level) and community level network features respectively to explore the impacts of these network features impact on credit risk assessment. Table 3 summarizes our designed experiments. Baseline model only contains basic features, Model 1 inputs basic features with ego and whole network features, Model 2 inputs basic features and community network features, and Model 3 contains all the features. For each experiment in Table 3, we apply them into 5 state-of-the-art prediction models.

4.3.2 Experimental Results

Accuracy

Table 4 gives out the accuracy results of our experiments. We can see that except SVM, the accuracy of other 4 algorithms are improved by adding social network features. The accuracy performance of these 5 predictive models demonstrates SVM>GBDT+FM>GBDT>FM>LR. For further investigation, we found that SVM classify all the test data into normal users (class labeled 0), with both Recall and Precision equal to 0. The AUC of SVM is 0.5. So, the high accuracy of SVM is meaningless. we remove SVM algorithm from our further evaluation and discussion.

Table 3. Experimental design

Experiments	Input Features
Baseline	Basic features
Model 1	Basic features + Social network features (ego+whole)
Model 2	Basic features + Social network features (Community)
Model 3	Basic features + All Social network Features

Table 4. Accuracy metric

	LR	SVM	GBDT	FM	GBDT+FM
Baseline	0.533	0.7598	0.665	0.568	0.731
Model 1	0.578	0.7598	0.683	0.638	0.740
Model 2	0.598	0.7598	0.728	0.623	0.754
Model 3	0.602	0.7598	0.704	0.630	0.746

Recall, Precision and F-score

Figure 2 shows the Recall, Precision and F-score performance of LR, GBDT, FM and GBDT+FM in different 4 experiment settings. In LR, by adding social network features, Recall has almost no change, but Precision and F-Score performance improve a little. In GBDT and GBDT+FM, by comparison of baseline model, Recall, Precision and F-score improve by adding social network features. But in FM, there is no improvement in Recall, Precision and F-score which are need to be investigated further. LR has the highest Recall and F-score value, and GBDT+FM has the highest Precision and Accuracy value.

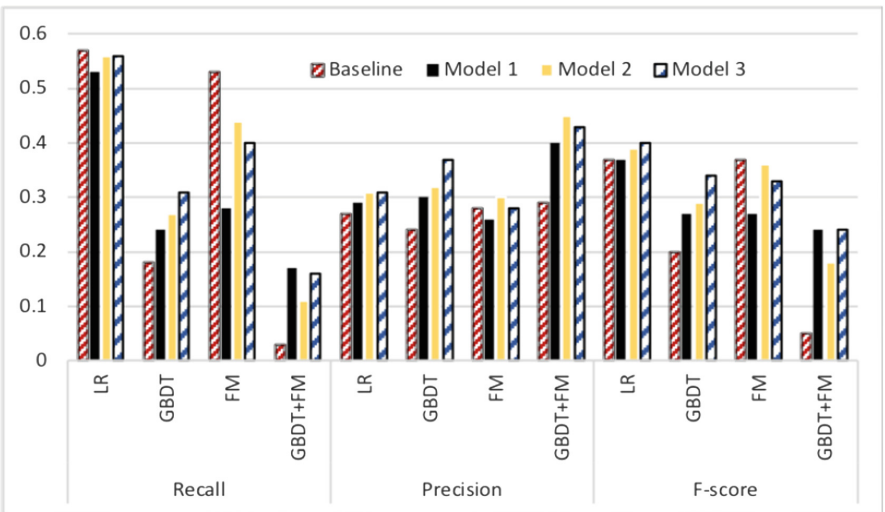

Fig. 2. Recall, precision and F-score metrics

AUC and Logloss

AUC and Logloss are two indicators for the model quality. Figure 3 shows that in the same algorithm, social network features could improve AUC and Logloss values. Among the four algorithms, GBDT+FM demonstrates the higher value of AUC. Logloss has no change in LR algorithm by adding social network features, but in GBDT, FM and GBDT+FM, logloss improve a little. Among the four algorithms, LR has the lowest logloss and accuracy.

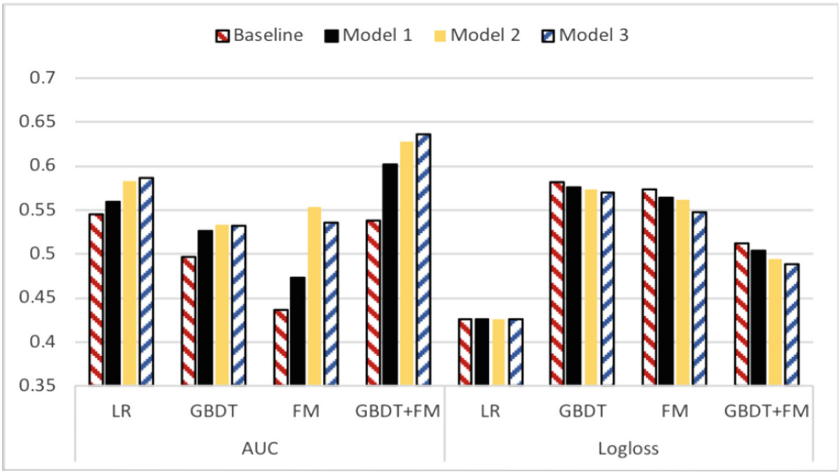

Fig. 3. AUC and logloss metrics

In sum, our experiment results demonstrate that by introducing social network features, except SVM, which cannot identify the high-risk users, all other four predictive models outperform the baseline model (without social network features) by examining six evaluation metrics. Social network features have impact on credit risk prediction.

5 Conclusions, Limitations and Future Work

In this study, we proposed a social relationship enhanced credit risk prediction framework by collecting data from different sources. By building a social network from users' geolocation data, we extract social network features in dimensions at ego, community and the whole network level. Credit risk prediction improves by introducing the novel social network features. Our study demonstrates that social network features are valuable to credit risk prediction and gives out a practical way to acquire social network data and employ feature engineering. Our work gives an insight on obtaining soft information for credit risk assessment performance improvement, such as how to build social relationships from users' mobile phone check in data, and how to construct network features at different levels, especially in community level.

But from the experiment results, the improvement by adding social network features is not large. The method we used to build social network is naïve, further we can apply the algorithms mentioned in [15–17] to rebuild the social network.

Future work can be developed in the following directions. Firstly, although this work has verified the significance of social network features on credit risk assessment, further study needs to done in extracting network features automatically. For instance, we plan to test our data on graph-based deep learning methods, such as Graph Neural Network (GNN), and Deep FM for their powers of feature engineering. Moreover, we will apply more datasets on our proposed framework to demonstrate its effectiveness.

Acknowledgement. We would like to acknowledge the partial financial support from Beijing Social Science Foundation (Project No. 17GLC056) and National Natural Science Foundation of China (Project No. 91546125).

References

1. Rendle, S.: Factorization machines. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), Sydney, Australia, pp. 995–1000. IEEE (2010)
2. Li, Z.: GBDT-SVM credit risk assessment model and empirical analysis of peer-to-peer borrowers under consideration of audit information. Open J. Bus. Manag. **06**(2), 362–372 (2018)
3. Freedman, S., Jin, G.Z.: Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper.com. Working Paper, 2008.11 (2008)
4. Avery, R.B., Calem, P.S., Canner, G.B.: Consumer credit scoring: do situational circumstances matter. J. Banking Finance **28**(4), 835–856 (2004)
5. Harris, T.: Quantitative credit risk assessment using support vector machines: broad versus narrow default definitions. Expert Syst. Appl. **40**(11), 4404–4413 (2013)
6. Liberati, C., Camillo, F.: Personal values and credit scoring: new insights in the financial prediction. J. Oper. Res. Soc. **69**(12), 1994–2005 (2018)
7. Sinnha, A.P., Zhao, H.: Incorporating domain knowledge into data mining classifiers: an application in indirect lending. Decis. Support Syst. **46**(1), 287–299 (2008)
8. Susterstic, M., Mramor, D., Zupan, J.: Consumer credit scoring models with limited data. Expert Syst. Appl. **36**(3), 4736–4744 (2008)
9. Zhang, T., Zhang, W., Xu, W., Hao, H.: Multiple instance learning for credit risk assessment with transaction data. Knowl.-Based Syst. **161**, 65–77 (2018)
10. Granovette, M.: Economic action and social structure: the problem of embeddedness. Am. J. Sociol. **91**(3), 481–510 (1985)
11. Lin, M.F.: Peer-to-peer lending: an empirical study. In: The 15th Americas Conference on Information Systems, AIS eLibrary, P8, San Francisco, USA (2009)
12. Greiner, M.E., Wang, H.: The role of social capital in people-to-people lending marketplaces. In: International Conference on Information Systems, DBLP (2009)
13. Lin, M.F., Prabhala, N., Viswanathan, S.: Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending. Soc. Sci. Electron. Publishing **59**(1), 17–35 (2013)
14. Hanneman, R., Riddle, M.: Introduction to Social Network Methods. University of California, Publisher (2005)

15. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proc. Nat. Acad. Sci. **106**(36), 15274–15278 (2009)
16. Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N.: Bridging the gap between physical location and online social networks. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, pp. 119–128. ACM (2010)
17. Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.: Inferring Social Ties from Geographic Coincidences. Proc. Nat. Acad. Sci. **107**(52), 22436–22441 (2010)
18. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005)