P2P Lending Fraud Detection: A Big Data Approach

Jennifer J. Xu[1(✉)], Yong Lu[2], and Michael Chau[3]

[1]Bentley University, Waltham, MA, USA
jxu@bentley.edu
[2]Pennsylvania State University, Hazleton, PA, USA
Yu114@psu.edu
[3]The University of Hong Kong, Hong Kong, China
mchau@business.hku.hk

Abstract. P2P lending directly connects borrowers and lenders without a financial institution as the intermediary. This new form of crowdfunding brings lenders more investment opportunities, but also poses unprecedented risks of default and fraud. This research-in-progress paper focuses on a specific type of fraud, loan request fraud, which may be unique to lenders on Chinese P2P lending sites due to the lack of nationwide credit rating systems in China. We propose research questions surrounding the problem of loan request fraud (its types, features, and detection methods) and present our research methodology and project plans. Specifically, we plan to develop data mining based methods and employ a big data approach to address our research questions. With the help of large volumes of data from a variety of sources, we will be able to find ways to leverage rich datasets about user behaviors and transaction histories to detect loan request fraud more effectively and efficiently.

Keywords: P2P lending · Loan request fraud · Financial fraud detection · Big data approach

1 Introduction

Recent years have seen the rapid development of crowdfunding around the world due to the advance of Web and information technologies. Since 2013, which is often called the first year of China's Internet finance era, many crowdfunding platforms and websites have quickly emerged in China. In this strong tide of web-based financial innovations, peer-to-peer (P2P) lending has become one of the most significant and important form of crowdfunding.

P2P lending is quite different from the traditional practice of financial loans. Traditionally, a financial institution (e.g., a bank) collects deposits from many individual customers (or lenders) and pays them savings interests. It then lends funds to borrowers (e.g., firms) and charges a higher interest rate. In this process, the lenders and borrowers are only indirectly connected through the financial institution. As a credit intermediary, the financial institution can evaluate the credibility of the borrowers and assess their repaying capabilities while making lending decisions, reducing the default risks to a large extent. In contrast, P2P lending is characterized by the absence of financial

M. Chau et al. (Eds.): PAISI 2015, LNCS 9074, pp. 71–81, 2015.
DOI: 10.1007/978-3-319-18455-5_5

intermediaries. P2P lending websites are simply marketplaces where people sell or buy loans. That is, except for offering the platform to accommodate loan transactions and providing the necessary information about borrowers and their loans, P2P lending websites do not evaluate the borrowers' trustworthiness, assess the financial risks associated with loans, or help lenders make investment decisions. As a result, lenders must bear higher risks than they used to when using traditional financial institutions.

More seriously, lenders may become victims of a new form of financial fraud on these crowdfunding platforms – loan request fraud. For example, a borrower may request a medical emergency loan while the real purpose of borrowing is to purchase luxury goods. A business owner may create a fraudulent loan request seeking funds for a nonexistent project, cheating about the financial performance of the company, exaggerating the company's profit prospects and repaying capabilities. After securing sufficient funds for their loans, these borrowers may default, causing financial losses to lenders. This type of fraud not only harms lenders financially (and even emotionally) but also causes great damages to the P2P websites, destroying their reputations. In 2014, four paper companies in Guangzhou, China, committed fraudulent loans, causing a P2P website, My089, a total of ￥100,000,000 bad debt risk [29].

The existence of loan request fraud may be attributed to several reasons. First, since all users (borrowers and lenders) on P2P sites are anonymous, lenders cannot get complete and accurate information about the identities, incomes, and credit histories of borrowers. This information asymmetry makes lenders very vulnerable to loan request fraud. Second, China does not yet have a nationwide credit rating system to report on individuals' credit histories. As a result, from a borrower's perspective, the consequence of committing fraud or defaulting in loan payments is not devastating; and the negative impact on their chances of getting future loans is minimal. Third, although most P2P lending sites in China have made great efforts to enforce risk management by carefully verifying borrowers' identifies, occupations, incomes, and home ownership, this process is largely manual and error-prone. It is still possible for malevolent borrowers to falsify their records and use misleading or untrue information to lure lenders to fund their loans.

Therefore, it is highly desirable to develop efficient and effective strategies and techniques for detecting loan request fraud. This research is intended to address the problem of fraud detection in the context of P2P lending. Since P2P lending is a relatively new phenomenon, there has been little research on the detection of loan request fraud, making it a challenging task. On the other hand, however, the recent advance in business analytics, cloud computing, and big data areas has also brought new opportunities for developing innovative big data approaches to financial fraud detection. With the increasing amount of data about users' online behaviors, purchasing records, and social networks, it becomes possible to mine the large volumes of multidimensional data about users and their digital footage, cross-validate P2P borrowers' credibility, and detect fraudulent loan requests.

Specifically, our research aims at addressing four research questions:

- What are the different *types* of loan request fraud in P2P lending?
- What *features and variables* can be used to detect loan request fraud?

- Which *methods and techniques* are more effective and efficient in detecting loan request fraud?
- How to leverage the "*big data*" to improve the performance of loan request fraud detection?

The remainder of this paper is organized as follows. In the next section we review the literature on P2P lending and financial fraud detection. Section 3 will describe our research methodology and lay out the project plans. The last section summarizes potential contributions of the research and concludes the paper.

2 Literature Review

In this section will first briefly introduce P2P lending and then review its current status in China, the state-of-the-art research on P2P lending, and the prior work on financial fraud detection.

2.1 P2P Lending

P2P (peer-to-peer) lending is a type of online auction platform that allows individuals to acquire loans from other people without an intermediary financial institution. P2P lending is a type of crowdfunding, which is broadly defined as "the practice of funding a project or venture by raising monetary contributions from a large number of people, typically via the Internet" [31]. Other types of crowdfunding include donation-based, rewards-based, and equity-based [6].

Since its inception, P2P lending has attracted considerable attention from the public, media, and academia. The first P2P lending site, Zopa.com, launched in England in 2005. Since then many P2P lending sites and platforms have emerged in many countries. In the United States, the two largest P2P lending sites are Prosper and LendingClub. For example, launched in 2006, Proposer has had more than two million members and funded over $2,000 million loans [27]. As the first P2P lending site in China, PPDai has attracted more than five million users, becoming one of the leading P2P platforms [26]. Other large P2P lending sites in China include CreditEase, My089, eDai365, etc. The recent two years have seen a surge of new P2P lending sites in China.

On a typical P2P site, an individual user, called borrower, can create a loan request, called listing, with a certain amount for auction. The borrower must also specify the maximum interest rate she is willing to pay. Other individuals, called lenders, can bid on the loan request by making a pledge to contribute a certain amount to the listing and specifying the minimum interest rate she is willing to accept. Auctions typically remain active for several days. Figure 1 displays a sample loan request on the website of PPDai.com. It shows critical information about this listing: title, borrowing amount, asking interest, maturity, and a brief text about the purpose of the loan. Important information about the borrower is also displayed along with the listing including the user name,[1] credit grade (assigned by PPDai), borrowing history, and the optional image of the borrower. A table on the same page records the lenders who have already bidden on the listing, their contribution amounts, and the interest rates they are willing to accept.

The borrower can select either the open or closed auction format. With the closed format, a loan request is closed immediately if enough funding has been received. With the open format, if a loan request is fully funded before the end of its specified duration, lenders can compete and bid down the final interest rate. Regardless of the auction format, a loan request that receives sufficient funding will be materialized into a loan. Each loan has a maturity of 6-36 months and each monthly payment made by the borrower is distributed among the winning lenders based on their contribution proportions. The status of a loan that has been repaid up to the payment schedule is "current." A loan's status can also be "one month late" or "two-month late." A loan that is late for three or more months is considered default.

Fig. 1. A sample loan request on PPDai.com

Although P2P lending has brought lenders new investment opportunities and channels, it also poses unprecedented challenges and risks, among which default is the greatest risk that lenders must bear when lending money to strangers. To protect

[1] To protect the privacy of users, all user names in Figure 1 are fake.

lenders' investment and minimize default risks, each P2P site has employed various risk management measures such as identity verification and credit grade assessment. Although the true identities of users are never publicly revealed on the websites, P2P platforms always verify the users' identities using their social security numbers, driver's license numbers, bank account numbers, or other valid identifications before allowing them to engage in any transactions. P2P sites often also collect and make accessible additional information regarding borrowers. PPDai, for example, assigns each borrower a letter credit grade ranging from AA (high quality) to HR (low quality) based on the borrower's income, occupation, education level, and credit history within PPDai [26]. The borrower's home ownership, past and present delinquencies, current credit lines, and bankcard utilization may also be disclosed on the website.

However, despite these verification and prevention measures, the information that lenders can access regarding borrowers are still very limited; and the fundamental problem of information asymmetry remains. Information asymmetry refers to a situation in which one party in a transaction has more or better information than the other [2]. In the traditional lending context, as the intermediary between borrowers and lenders, financial institutions have access to detailed information of borrowers and possess sophisticated risk-assessment instruments. Thus, they are able to mitigate information asymmetry and reduce investment risks effectively. However, unlike banks, P2P lenders do not have access to the comprehensive information about borrowers. Consequently, information asymmetry between borrowers and lenders is significantly elevated.

2.2 Prior Work on P2P Lending

The literature on P2P lending is mainly focused on lenders' lending decision making processes under the condition of information asymmetry. In general, two categories of research questions have been studied: (a) what type of information affects lenders' lending decisions? and (b) which factors are related to or can be used to predict payment default?

For the first question, research has found that two types of information, hard and soft, play a role in lender' decisions about whether to bid on a loan request or not [18]. Hard credit information refers to a set of quantitative parameters about a borrower (e.g., the credit score or grade, debt-income ratio, homeownership, and number of credit cards). Naturally, it is easier for borrowers with high credit scores and low debt-income ratios to get funded for their loan requests.

However, studies have also reported that loan requests made by borrowers with poor credit grades may also successfully get funded. In this case, it often is the soft information that makes lenders trust the borrowers and believe that they can repay loans on time. Soft information refers to borrower characteristics that are not reflected by hard credit information, such as a borrower's personal capital (e.g., gender, age, race, appearance) [18]. Pope and Sydnor found that loan requests made by females, white or Asian borrowers are more likely to attract funding [25]. Similarly, Duarte et al. found that lenders tend to perceive borrowers with attractive appearances to be trustworthy [11]. Another important type of soft information is related to a borrower's social capital, which can be measured by the borrower's number of friends and the number of endorsements made to the borrower's loan requests by his/her friends [3, 18, 19]. In this case, the social capital serves as a signal of loan quality or trustworthiness of the borrower.

For the second research question, most studies have reported that loans with better credit grades tend to have lower default rates. However, borrowers with good personal or social capital may not necessarily perform well in their ex-post loan payments and tend to default more often than those with good hard scores [4, 10, 12, 14-17].

It must be noted that default is not necessarily caused by a fraudulent loan request. A valid, legitimate loan request made by an honest, trustable borrower can end up with default because of various reasons such as the borrower's unexpected financial hardship, loss of jobs, or health issues. However, on the other hand, default usually is the predestined outcome of a fraudulent loan request, which is intended to deceive lenders for financial gain.

2.3 Fraud Detection

The Merriam-Webster's Dictionary of Law defines fraud as "a misrepresentation or concealment with reference to some fact material to a transaction that is made with knowledge of its falsity or in reckless disregard of its truth or falsity and with the intent to deceive another and that is reasonably relied on by the other who is injured thereby." Loan request fraud can be seen as a type of financial fraud. Phua et al. categorized financial fraud into four types [24]: internal fraud (i.e., fraudulent financial reports of firms), insurance fraud (i.e., fraudulent insurance claims), credit card fraud (i.e., fraudulent purchases), and telecommunication fraud.

Statistical methods and data mining techniques have been widely used for detecting financial fraud. In the data mining approach, regular patterns or models are first extracted from large volumes of historical data about transactions or user behaviors. It then compares any new data against the regular patterns and raises flags for transactions or behaviors that significantly deviate from the regular patterns and models. For example, a credit card company can build a "behavioral profile" for individual customers by finding regular patterns in their purchasing histories, and classifying transactions based on their average monthly spending and the places where they usually shop. A sudden, sharp increase in a customer's monthly spending in unusual places can signal fraudulent transactions.

The detection of financial fraud, especially the detection of internal and credit card fraud, has been studied extensively in literature using various classification, clustering, and outlier analysis methods. For example, Cecchini et al. used Support Vector Machine (SVM) to analyze 23 financial measures of 122 firms and found 132 frauds in a 13-year period [7]. Dechow et al. focused on seven main financial variables and used logistic regression to detect 293 frauds in 896 firms [9]. Abbasi et al. proposed a meta-learning framework and combined several classification methods including SVM, J48, Bayes Net, logit regression, and neural networks [1]. In addition to the 12 common financial variables (e.g., asset quality index, cash flow earnings difference, inventory growth, etc.) extracted from firms' quarterly and annual financial reports, they also leveraged organizational level and industry level data and significantly improved the performance of internal fraud detection. Similarly, credit card fraud detection also often relies on classification and outlier analysis techniques, including Bayesian classifiers [22], hidden Markov models [30], and association rules [28]. Paasch used neural networks and genetic algorithms to detect fraudulent credit card transactions in a real dataset that contained 13-month worth of 50 million credit card transactions [20]. Using the same dataset, Bhattacharyya et al. combined support vector machines and logistic regression models for fraud detection [5].

Online auction fraud has attracted the attention of researchers in recent years. This kind of fraud often occurs on e-commerce platforms (e.g., eBay). Behaviors such as shilling, bid shielding, misrepresentation, selling counterfeits, and triangulation are all considered fraud. Many methods and approaches have been proposed to detect online auction fraud. Pandit et al. extracted suspicious patterns for fraudsters by modeling online auction transactions and users as a Markov Random Field, and then employed a Belief Propagation method to detect likely fraudsters [21]. In addition to these technological approaches, researchers have also proposed many constructive strategies for users, auction platforms, and regulation institutions. Chua et al. advocated for collaboration between online auction user communities and auction platforms to effectively manage the problem of auction fraud [8]. Gavish and Tucci suggested that fraud be reduced by increasing information sharing among auction platforms and users, employing legitimate escrow services, enforcing auction insurance policies, and encouraging self-protection by buyers [13]. Pavlou and Gefen proposed to leverage three IT-enabled institutional mechanisms (feedback, third-party escrow services, and credit card guarantees) to engender buyer trust and reduce auction fraud [23].

Because P2P lending is a relatively new financial phenomenon, there has been little research on fraud detection in this context. Especially, it has not been much report about loan request fraud in developed countries (e.g., the United States), in which the credit rating systems are rather comprehensive and can provide accurate and reliable information about individuals' credit histories. Moreover, because there are many regulations and laws targeting financial fraud in those developed countries, it is relatively difficult for individuals to commit frauds by lying about credit histories, using fake or stolen identities, and providing misleading or untrue information. In contrast, there has not been a nationwide credit rating system in China that monitors and reports the credit histories of individuals. P2P sites can only assess borrowers' credit grades based on limited credit history records within their own platforms. There is also no information sharing between different P2P sites and traditional financial institutions. This makes it possible for a malevolent borrower to commit fraud on one platform and continue to do so on other platforms and sites without being noticed or caught. Our research is intended to develop methods and techniques for detecting such fraudulent online loan requests on P2P platforms in China.

3 Research Methodology

Our research consists of several phases using various research methods.

The first phase is literature review. We have reviewed the literature regarding three aspects related to our research: P2P lending, financial fraud detection using data mining and big data approaches, and the current state of P2P lending fraud in China. We have searched and found published research in a number of finance, business, and information systems journals and conferences in order to gain a complete and comprehensive understanding of the state-of-the-art research on P2P lending. We have also studied the prior work on fraud detection methods and techniques, finding the pros and cons of each type of methods in detecting fraudulent financial transactions or behaviors. A big challenge has been to examine the state of P2P lending fraud. Only a limited number of research studies have been found to investigate Internet financial fraud in China. Alternatively, some news articles have reported real cases of fraud found in crowdfunding platforms and websites. We will continue to look for these types of studies and seek deeper understanding of P2P lending fraud.

We will conduct field studies in the second phase. We will select PPDai.com, one of the largest P2P site in China, as the field organization for our study. Through previous research collaborations, we have built a strong relationship with PPDai, making it possible to enter the site and collect real data. We will interview PPDai's key domain experts who are in charge of risk prevention and management and learn their knowledge and experience about how they have detected fraudulent loan requests in their identity verification and loan approval processes. We will also study the characteristics of fraudulent cases to prepare for the development of large-scale models for loan request fraud.

The task in the third phase is to develop detection methods and prototype systems. Like in most classification applications, one of the important steps in this phase will be feature selection. We will select and construct a set of features that are useful for detecting fraud based on our review of literature on fraud detection, as well as our findings from our interviews with domain experts in the field studies. We will adopt the meta-learning framework proposed by Abbasi et al. [1] and incorporate multiple classification algorithms (e.g., decision trees, SVM, and neural networks) in this framework.

More importantly, we will explore and experiment with the big data approach to improve the detection performance. This approach is characterized by its reliance on large volumes of data from multiple sources with multiple dimensions in multiple formats. Using this approach may help discover novel knowledge about financial fraud and develop innovative detection strategies and techniques. For example, a considerable percentage of PPDai users are retailers on Alibaba.com, the largest e-commerce site in China. These users often choose to display links to their Alibaba online stores on their loan request pages. This can become a unique external source for a rich set of data containing these users' Alibaba seller ratings, their business types, product categories, and sales volumes. Such data can be used to verify the purpose of business loans and to assess borrowers' repaying capabilities.

In addition, advanced techniques will be employed to mine data in different formats (e.g., structured and unstructured). For instance, PPDai borrowers are allowed to write a brief text describing the purpose of a loan request and show evidence for repaying capabilities. Users (borrowers and lenders) can engage in private email conversations within the site regarding a specific loan request. There also is a user community where users can post messages to ask questions, seek investment advice, share lessons and experience, discuss about specific loans, and make friends. This set of unstructured data can be mined to build profiles of users' online behaviors, social relations, and even sketches of personal background, assisting the detection of fraudulent loan requests.

The last stage will be performance testing and evaluation. We will ask domain experts from PPDai to provide a dataset with confirmed fraudulent loan requests and use it as our training and testing data. We will compare different feature sets and methods based on a number of performance metrics (e.g., accuracy, precision, recall, and F-measure).

At present we have finished the first phase and are preparing for the interviews with domain experts in PPDai. We have secured a preliminary dataset containing all loan requests and payment records between 2009 and 2011. We plan to program a Web crawler soon to collect data about PPDai's registered users who own online stores on Alibaba.com.

4 Potential Contributions and Conclusions

The potential contribution of our research is two-fold. First, our research will shed light on a new type of financial fraud, P2P loan request fraud, associated with crowdfunding. This problem must be addressed because it adds risks to lenders' investment on P2P platforms and may cause great financial losses to lenders. To the best of our knowledge, there has not been any study focusing on this new type of fraud. Second, our research pioneers the use of the big data approach to tackling the problem of fraud detection. With the help of large volumes of data from a variety of sources, we will be able to find ways to leverage rich datasets about user behaviors and transaction histories to detect fraud more effectively and efficiently.

Acknowledgements. The authors would like to thank PPDai.com for their support for this project, and Reno Ha and Leo Lu of the University of Hong Kong for their assistance.

References

1. Abbasi, A., Albrecht, C., Vance, A., Hansen, J.: MetaFraud: A meta-learning framework for detecting financial fraud. MIS Quarterly 36, 1293–1327 (2012)
2. Akerlof, G.A.: The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84, 488–500 (1970)
3. Bastelaer, T., Leathers, H.: Trust in lending: Social capital and joint liability seed loans in southern zambia. World Development 34, 1788–1807 (2006)
4. Berger, S.C., Gleisner, F.: Emergence of financial intermediaries in electronic markets: The case of online P2P lending. Business Research 2, 39–65 (2009)
5. Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C.: Data mining for credit card fraud: A comparative study. Decision Support Systems 50, 602–613 (2011)
6. Bradford, C.S.: Crowdfunding and the federal securities laws. Columbia Business Law Review 1, 14–27 (2012)

7. Cecchini, M., Aytug, H., Koehler, G., Pathak, P.: Detecting management fraud in public companies. Management Science 56, 1146–1160 (2010)
8. Chua, C.E.H., Wareham, J., Robey, D.: The role of online trading communities in managing internet auction fraud. MIS Quarterly 31, 759–781 (2007)
9. Dechow, P., Ge, W., Larson, C., Sloan, R.: Predicting material accounting misstatements. Contemporary Accounting Research 28, 17–82 (2011)
10. Do, H., Jeon, S., Banker, R., Lee, B., Yoo, B.: Is the leaderboard information useful to investors? The leaderboard effect in P2P lending. In: The 33rd International Conference on Information Systems, Orlando, FL (2012)
11. Duarte, J., Siegel, S., Young, L.: Trust and credit: The role of appearance in peer-to-peer lending. The Review of Financial Studies 25, 2455–2484 (2012)
12. Freedman, S., Jin, G.Z.: Learning by doing with asymmetric information: Evidence from Prosper.com (2010), sitemaker.umich.edu/sfreedm/files/freedman_jin_prosper.pdf
13. Gavish, B., Tucci, C.L.: Reducing internet auction fraud. Communications of the ACM 51, 89–97 (2008)
14. Herrero-Lopez, S.: Social interactions in P2P lending. In: The 3rd Workshop on Social Network Mining and Analysis (2009)
15. Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., Gorski, S.: Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. Journal of Applied Communication Research 39, 19–37 (2011)
16. Lee, E., Lee, B.: Herding behavior in online P2P lending: An empirical investigation. Electronic Commerce Research and Applications 11, 495–503 (2012)
17. Lin, M.: Peer-to-peer lending: An empirical study. In: AMCIS 2009 Doctoral Consortium, San Francisco, California, pp. 1–7 (2009)
18. Lin, M., Prabhala, N.R., Viswanathan, S.: Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. Management Science 59, 17–35 (2013)
19. Lu, Y., Gu, B., Ye, Q., Sheng, Z.: Social influence and defaults in peer-to-peer lending networks. In: The 33rd International Conference on Information Systems, Orlando, FL (2012)
20. Paasch, C.: Credit card fraud detection using artificial neural networks tuned by genetic algorithms, Technical Report, Hong Kong University of Science and Technology (2007)
21. Pandit, S., Chau, D.H., Wang, S., Faloutsos, C.: NetProbe: A fast and scalable system for fraud detection in online auction networks. In: The 16th International World Wide Web Conference, Banff, Alberta, Canada (2007)
22. Panigrahi, S., Kundu, A., Sural, S., Majumdar, A.K.: Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. Information Fusion 10, 354–363 (2009)
23. Pavlou, P.A., Gefen, D.: Building effective online marketplaces with institution-based trust. Information Systems Research 15, 37–59 (2004)
24. Phua, C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research, http://arxiv.org/pdf/1009.6119v1.pdf
25. Pope, D.G., Sydnor, J.R.: What's in a picture? Evidence of discrimination from Prosper.com. The Journal of Human Resources 46, 53–92 (2011)
26. PPDai. PPDai: About us (2015), http://www.ppdai.com (cited March 21, 2015)
27. Prosper. Prosper.com: About us (2014), https://www.prosper.com/about/ (cited December 30, 2014)
28. Sánchez, D., Vila, M.A., Cerda, L., Serrano, J.M.: Association rules applied to credit card fraud detection. Expert Systems with Applications 36, 3630–3640 (2009)

29. Sina Tech. The biggest bad debt in the P2P lending history: Paper industry loans poses immense risks to My089 (P2P爆最大坏账 红岭创投垫付1亿纸业借款) (2014), http://tech.sina.com.cn/i/2014-08-29/01599583049.shtml (cited August 29, 2014)
30. Srivastava, A., Kundu, A., Sural, S., Majumdar, A.K.: Credit card fraud detection using hidden Markov model. IEEE Transactions on Dependable and Secure Computing 5, 37–48 (2008)
31. Wikipedia. Crowdfunding (2015), http://en.wikipedia.org/wiki/Crowdfunding (cited March 21, 2015)