# Understandig User Behavior in Online Banking System

Yuan Wang[1,2], Liming Wang[1](✉), Zhen Xu[1], and Wei An[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
wangyuan,wangliming,xuzhen,anwei@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

**Abstract.** Currently, online banking has become extremely popular all over the world and plays a significant role in people's daily lives. However, the user behaviors have yet to be studied carefully in existing works. In this paper, we provide a large-scale, comprehensive measurement study of online banking users based on a two-week long dataset consisting of transactions conducted by personal users in one of the top banks in China. We demonstrate the customer behaviors mostly comply with the heavy-tail distribution which implies abnormal activities. In further analysis of those activities, we figure out that most of them are generated by two types of accounts, i.e., corporate accounts paying salaries and dishonest bank employees plastering the achievement. We extract a set of features to classify the two types of abnormal accounts from the benign ones. The experimental result illustrates that our system can accurately detect them with only 0.5% false positive rate.

**Keywords:** Online bank; User behavior; Abnormal detection

## 1 Introduction

Currently, online banking has become extremely popular all over the world and plays a significant role in people's daily lives. Beneficial to online banking, customers can conduct their financial activities like payments, money transfers or investment in a convenient and efficient manner at anytime and anywhere.

The popularity of Internet banking has led to an increase of frauds, perpetrated through cyber attacks, phishing scams and malware campaigns, resulting in substantial financial losses [15]. A lot of researchers dedicate in mitigating those threats [9, 11, 15, 1–3, 14, 4]. The most challenging reason for bank fraud detection is its dynamic behavioral characteristics. To outwit online banking defenses, fraudulent behavior is dynamic, rare, and dispersed in very large and highly imbalanced datasets. In addition, different customer habits vary widely, making it more difficult to distinguish fraudulent transactions from normal ones.

Therefore, in-depth understanding the customer behaviors is urgent because (1) It can make our understanding of customer behavior more clear and profound; (2) It can help us portray customer behavior more appropriately; (3) It provides reliable knowledge to distinguish between suspicious behavior and good behavior. However, few studies have been carried out on understanding customer behaviors due to existing barriers. Privacy is the first one that leads to the unavailability of the data for researchers, and another one is the competition issue. Therefore, most works on online banking only provide coarse results without detailed analysis. Due to the cooperation with a large Chinese bank, we have the opportunity to investigate online banking customer behaviors based on the anonymized ground-truth dataset that will be addressed in detail in Section 2.

In this paper, we systematically studied the customer behavior of online banking. Using the transaction data, we analyzed customer behavior from the access pattern and transaction pattern to characterize how customers access the online banking service and conduct their transaction activities. To deeply understand customer behavior, we compared sessions from thirdparty websites with online banking websites in access pattern analysis and examined the differences between non-transaction and transaction sessions through transaction pattern analysis. Our main contributions are stated as follows:

1) We first present analysis results of customer behavior from a session perspective. We find that session length (of requests) is a heavy-tailed(also power-law) behavior and that sessions from online banking website have a marked bimodal pattern in working hours, while sessions from thirdparty websites do not perform as well.
2) We provide insights into the details of transaction behavior. We find that transaction amount follows a lognormal distribution. Our analysis of the number of transactions and payees per session shows that the dishonest internal employees and corporate customers are among the personal accounts.
3) Based on the analysis, we propose CatchAbs, a supervised method for abnormal activites detection. We show that it can accurately catch these two types of abnormal behaviors.

The remainder of this paper is organized as follows. Section 2 describes the data used in this paper and Section 3 estimates the access patterns of online banking customer, especially characteristics of the session perspective. In Section 4, we investigate the transaction behavior and summarize the findings we derived and implications. In Section 5, we present the detection system and analyze the results. Section 6 discusses the related work, while Section 7 concludes this paper.

## 2 Data Description

With the online banking system, customers log in to access the online banking service through a browser, and they can perform various financial activities, inquiries, money transfers, fee payments, and investments, by using a personal account or corporate account. Each request of a customer will be stored as a record in the data. Our data is collected from one of the top banks in China, whose online banking system provides online services for millions of customers every day. It includes all the records in the duration of 12 days from July 7, 2014 to July 18, 2014. The record contains various information about the customer's behavior, such as time stamp, account ID, payer ID, payee ID, operation (e.g. login, money transfer), amount, login IP, login area, operation status (success or failure).

Personal information was anonymized to meet personal privacy policy. Three types of sensitive information about customer identities are anonymized: (1) customer login ID (or user ID) usually with a unique ID for one customer in the online banking website, (2) customer transaction IDs (payer IDs or query IDs) in the online banking site, one for a card, where sometimes one customer login ID may correspond to a few transaction IDs, (3) payee IDs of the beneficiary account in a transaction. After filtering some error records, we finally have 23,212,800 sessions and 91,002,483 records in total.

The characteristics of the authenticated session are that the session retains the user ID of the successfully logged-in user, while the unauthenticated session does not have the user ID. A total of 4,983,518 authenticated sessions include 3,412,869 customers and 23,863,321 total records in our data. Unauthenticated sessions may be created by the crawler and have less value than authenticated sessions, so our analysis in the following sections will focus on authenticated sessions. Operation describes a customer's action like 'login' or 'logout' in the online banking system. We call a session as a transaction session if there exists at least one transaction (money-moving operation) in a session; otherwise call it as a non-transaction session.

**Table 1.** Summary of the on-line banking data

| | | Records | Sessions | Customers |
|---|---|---|---|---|
| Unauthenticated | | 18,162,843 | 12,467,088 | 0 |
| Authenticated | Website | 56,249,672 | 6,457,533 | 2,707,737 |
| | Thirdparty | 16,589,968 | 4,288,179 | 1,375,557 |
| Total | | 91,002,483 | 23,212,800 | 3,411,486 |

Customers can also access online banking services through a thirdparty website. For example, a customer surfs an online shopping website, finds a product of interest, and then logins to an online banking account for payment. To the online banking system, these sessions from thirdparty websites are given a special login API ('PayGateLogin'). In our paper, we use 'thirdparty' to denote these sessions and 'website' denotes those sessions from online banking homepage.

## 3 Access Patterns

Login frequency is the key metric for the online banking system. We first study the login frequency of customers in term of the number of logins per hour in the duration of 12 days. As shown in Figure 1, both logins from the online banking website and the thirdparty websites reveal a clear diurnal pattern, i.e.

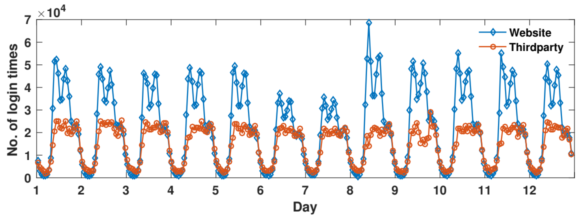
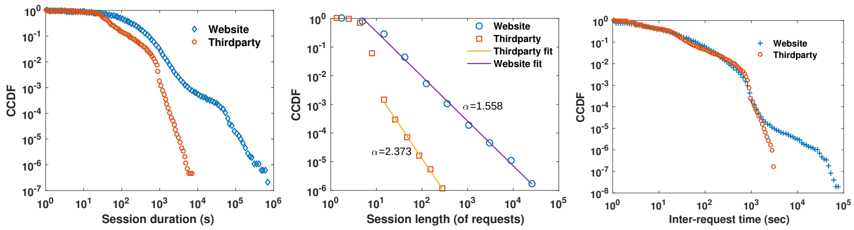


**Fig. 1.** No. of customer logins over time (interval of 1 hour).

the heavy logins in the daytime and the light logins at night. The logins from online banking homepage (website) are significantly different from those via the thirdparty websites (thirdparty). It is observable that the number of website logins on every day is bimodal. Specifically, one peak appears before lunch time (10 am∼11 am), and the other appears at afternoon (3 pm∼4 pm), which implies that more people like to access the online banking service at work time.

Unlike the counter service, customers can access the online banking system at anytime via the online banking system and thereby increase the service capability for the bank. As shown in Figure 1, the online banking system responses to tens of customers every second in the busy hour, while the inter-arrival time even achieves 325 seconds at night for the online banking system. Also, logins of online banking at weekend are remarkably less than that on weekdays. However, the logins from the thirdparty websites do not reveal the bimodal pattern and the weekend pattern.



(a)Session duration (b)Fitting for session length (c)Inter-request time

**Fig. 2.** Characteristic of intra-session.

Figure 2(a) shows CCDF as session duration varies. Here, the complementary cumulative distribution(CCDF) is defined as

$$F(x) = P(X > x) \tag{1}$$

which presents more information on the tail-end extreme events. As shown in this figure, customers spent less time from the thirdparty websites (thirdparty) than that from online banking (website). Specifically, the mean and median of session duration of thirdparty sessions are 71.4s and 34s, 209.6s and 87s for website sessions, which double the time that is spent on the thirdparty sessions on average. Most of thirdparty sessions last for a short time, while some website sessions last for a longtime, even for a few hours. As shown in Figure 2(a), 85.47% of thirdparty sessions are less than 100s, while only 58.62% of website sessions last less than 100s.

In Figure 2(b), we observe a well fitted power-law distribution of session length of online banking sessions (website) in double-logarithmic plot of CCDF and a power-law tail for the distribution of thirdparty sessions. Here, the power-law distribution has a probability density function (PDF)

$$f(x) \sim x^{-\alpha-1}, \quad x \to \infty. \tag{2}$$

where the complementary cumulative distribution (CCDF) is

$$P(X \geq x) \sim x^{-\alpha}, \quad x \to \infty. \tag{3}$$

The power-law-type distribution is called heavy-tailed or fat-tailed if $\alpha < 2$. In this case, the variance of the random variable is infinite. Furthermore, when $\alpha \leq 1$, the mean of random variable is also infinite [5]. The fitting result of website sessions is shown in Eq.(3) with $\alpha = 1.558$. It reveals that the distribution of website sessions is a heavy-tailed distribution. This implies that the number of long sessions is more than expectation as exponential distribution or normal distribution, e.g., the longest session has 44,187 requests.

We also characterize the inter-request time within a session. The CCDF distribution is shown in Figure 3(c). Due to the timeout settled by online banking, the distributions present a large change around 900 seconds(15m). The inter-requests more than 900s happen when customer return to the online banking system after timeout, the online banking will still record this request with the session ID kept on in customer client cache. However, the online banking system will not respond to this request and will require customers to login again.

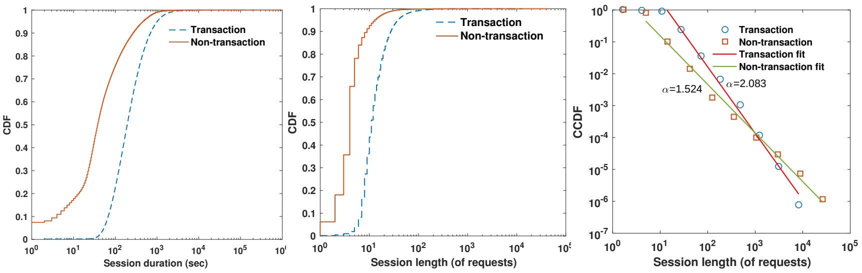
## 4 Transaction Patterns

### 4.1 Transaction & Non-transaction

We first examine the characteristics of non-transaction activities and transaction activities from the session view. Our data has 1,253,652 transaction sessions in total and most transactions are performed through website sessions.

As shown in Figure 5(a), non-transaction sessions take less time than that of transaction sessions. 58.58% of non-transaction sessions are performed within the range of [10$s$, 100$s$], while 72.66% of transaction sessions with falling into the range of [100$s$, 1000$s$]. The mean and median of session duration of non-transaction session are 119.6$s$, and 38$s$, respectively, while they are 329$s$ and 193$s$ for transaction sessions.

Also, non-transaction sessions have less requests as shown in Figure 5(b), 91.25% non-transaction sessions have less than 10 requests, while there are only



(a)Session duration (b)Session length (c)Session length

**Fig. 3.** Characteristics of non-transaction and transaction sessions.

36.14% transaction sessions with less than 10 requests. The mean and median of session length are 16.3 and 11 for transaction sessions, respectively, and they are 5.5 and 4 for non-transaction sessions. Most of the non-transaction sessions are query tasks, which usually take less time than that of transaction tasks, as explained in Section 4. Figure 5(c) shows that the distribution of session length of non-transaction sessions is fitted to a power-law distribution with $\alpha = 1.524$, which means that it also follows the heavy-tailed distribution. It can be explained that customers with transaction tasks have more clear goals than those with non-transaction tasks.

### 4.2 Transaction Amount

In order to understand transaction behavior, we first examine the amount per transaction. As shown in Figure 4, the transaction amount clearly follows a log-normal distribution. The probability distribution function for the lognormal distribution is given by:

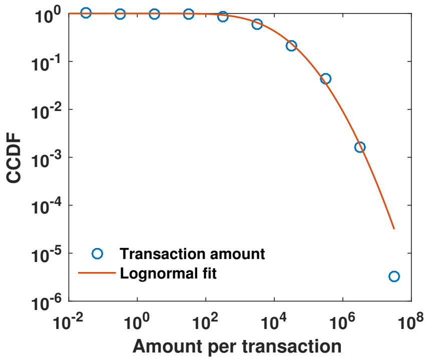
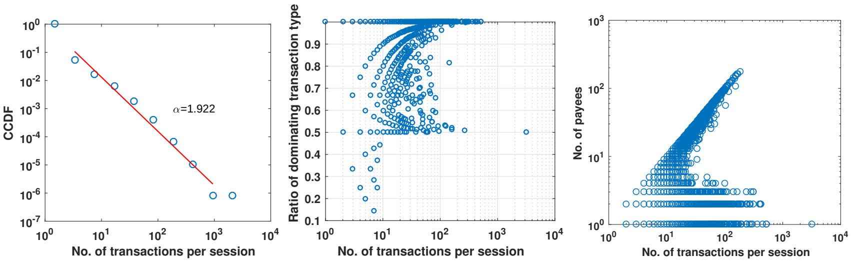


**Fig. 4.** CCDF of amount per transaction, which follows a lognormal distribution.

With this equation, we fit the log-normal distribution of Figure 6 as Eq.(4) with $\sigma = 2.11$ and $\mu = 8.83$. The mean, and median of transaction amounts are 22,786 and 1900(CNY), respectively. Among the transactions, the largest 10% contribute 81.35% of the total transaction amount, which largely follows the 80-20 rule [10].

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}} \tag{4}$$

## 4.3 Transaction Times



(a)No. of transactions in one session

(b)The ratio of dominating transaction type

(c) The No. of transactions and payees

**Fig. 5.** Characteristic of transaction sessions.

The distribution of the number of transactions per session is shown in Figure 7(a), which follows a power-law distribution with $\alpha = 1.922$. 85.46% of transaction sessions have only one transaction in one session. However, the maximum is 3240 transactions in a session.

Then we investigate the dominating transaction type, which contributes the most transactions in the online banking system. Figure 7(b) shows the ratio of the dominating transaction type in a session. We can find that the more transactions there are in a session, the higher the proportion of its dominating transaction type. When a session has more than 10 transactions, the dominating transaction type contributes most of the transactions: the proportion of dominating transaction type for these sessions is all greater than or equal to 0.5; the proportion of dominating transaction type for 90% of sessions exceeds 0.969.

To further examine the transaction behavior, Figure 7(c) shows the number of transactions and payees per session. We find that sessions follow two trends: One trend is that a large number of transactions accompany a small number of payees; Another trend is that the number of payees increases with the number of transactions. 1) We choose the group of sessions with more than 100 transactions and more than 100 payees. After examining the transaction type and amount for every session, we confirm these sessions are wage payment activities by corporate customers. 2) We choose the group of sessions with more than 100 transactions and less than 10 payees to analyze. These sessions are mainly divided into two transaction types. One type is fee payment transaction, which refers to business behavior. The other type has proved to be fraud behavior of bank employee to obtain better job performance.

## 5 Abnormal Detection

In this section, we develop a detection system CatchAbs that distinguishes the abnormal accounts described in 4.3. Since manual tagging takes a lot of time and effort, we only mark sessions with session length greater than 500. In the end, we found 152 anomalous sessions in these 376 sessions, 22 of which were bank employee activities and 130 were corporate events.

### 5.1 Feature Extraction

Intuitively, we should select features which help spot the type of anomalies we are interested in. As mentioned in Section 4.3, we want to find the anomalies with large operations in transaction activities. To accommodate detection of all of these anomalies, we extract 5 features from session-level information.

F1 The number of transactions per session.
F2 The ratio of transactions(F1) to session length per session.
F3 The ratio of payees to transactions(F1) per session.
F4 The average transaction amount per session.
F5 The entropy of transaction amount per session.

### 5.2 Detection and Analysis

After we extract the sessions, we want to predict whether they are abnormal or normal.We use a random forest, a supervised classifier, using the features described above. The major advantage of using random forest lies in the unexcelled accuracy and efficiency.

We computed values for all 5 features for all sessions in our dataset, input the the data to a random forest classifier, and ran 10 fold cross-validation(CV). In 10-fold CV, the data is randomly splitted into 10 folds, where the classifier is trained on 9-folds and tested on the remaining fold. The classifier repeats this process 10 times, each time a different fold is used for testing.

The resulting classification accuracy was 96.9%, with 0.5% false positives (i.e. classify normal users as abnormal) and 0.0% false negatives (i.e. classify abnormal users as normal). Further, we dig into the false positives (FPs) obtained from the classifier to gain a better understanding into possible classes of abnormal activities that we may have missed in our ground truth data. We manually checked each session incorrectly classified as abnormal. We find that most of them stem from a pay fee service with a very little amount. It is not expected to see these service provider activities among normal personal accounts.

## 6 Related Work

A few works have been carried out on customer behavior analysis in online banking because of the privacy, secrecy and commercial interest concerns. Their works focused on two aspects: service quality improving and the online banking fraud detecting.

*Service quality improving:* Some works attempted to improve the service quality of online banking services [12, 7, 13, 6]. These works focused on the attitude of customers who use the online banking services and investigated the factors contributing to customer satisfaction [6]. They usually adopted the way of asking the participants and bank customers with a questionnaire. Different to their work, our analysis is based on the transaction data, which comprehensively describes customer behaviors during transaction activities in server side, with the objective to in-depth understand customer behaviors and to improve the quality of online banking services.

*Online banking fraud detecting:* Some works detected the online banking fraud based on customer behavior analysis [9, 11, 15, 1, 2]. Their research usually models the behavior of each customer and monitors whether it deviates from normal behavior [8]. However, these works do not systematically analyze customer behavior based on the real data. Wei *et al.* [15] introduced a systematic online banking fraud detection method using transaction data from a large Australian bank, but they did not provide any analysis for the online banking customer behavior. Carminati *et al.* [2] developed a semi-supervised and unsupervised fraud and anomaly detection method based on a real-world dataset of a large Italian national bank. His system design is guided by data analysis, but his work only describes the distribution of the amount and transaction frequency. Compared to these studies, our research relies on the dataset that includes more details about the transaction, and more online banking patterns are revealed in this paper. Moreover, because of the lack of publicly available and real-world frauds, most these works resort to synthetically generated frauds. Our work based on ground truth data reveals some of the abnormal behavior that is happening.

## 7 Conclusion

In this paper, we have analyzed the characteristics of online banking customer behaviors based on personal transaction data collected from a large bank in China. To the best of our knowledge, this work is the first attempt to comprehensively understand the usage patterns of the online banking.

We first analyzed the statistical and distribution properties of the important variables of the access patterns from the session level. The analysis showed that most customer behaviors follow a power-law distribution. Then, we investigated the details of the transaction behaviors, e.g., the number of transactions, the transaction amount, and the transaction account. Our analysis also revealed some special accounts, e.g., corporate accounts and dishonest internal employees. Finally, we developed CatchAbs, a supervised method for detection of these two abnormal behaviors. In a word, our work will be helpful to improve the quality and security of the online banking service.

## References

1. G. Cabanes, Y. Bennani, and N. Grozavu. Unsupervised learning for analyzing the dynamic behavior of online banking fraud. In W. Ding, T. Washio, H. Xiong, G. Karypis, B. M. Thuraisingham, D. J. Cook, and X. Wu, editors, *ICDM Workshops*, pages 513–520. IEEE Computer Society, 2013.
2. M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero. Banksealer: A decision support system for online banking fraud analysis and investigation. *Computers & Security*, 53:175–186, 2015.
3. M. Carminati, L. Valentini, and S. Zanero. A supervised auto-tuning approach for a banking fraud detection system. In *International Conference on Cyber Security Cryptography and Machine Learning*, pages 215–233. Springer, 2017.
4. S. Dhankhad, E. Mohammed, and B. Far. Supervised machine learning algorithms for credit card fraudulent transaction detection: A comparative study. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 122–125. IEEE, 2018.
5. J. Gao, Y. Cao, W.-w. Tung, and J. Hu. *Multiscale analysis of complex time series: integration of chaos and random fractal theory, and beyond*. John Wiley & Sons, 2007.
6. P. Hanafizadeh, B. W. Keating, and H. R. Khedmatgozar. A systematic review of internet banking adoption. *Telematics and Informatics*, 31(3):492–510, 2014.
7. C. Herington and S. Weaven. E-retailing by banks: e-service quality and its importance to customer satisfaction. *European Journal of Marketing*, 43(9/10):1220–1231, 2009.
8. V. Jyothsna, V. R. Prasad, and K. M. Prasad. A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, 28(7):26–35, 2011.
9. K. N. Karlsen and T. Killingberg. Profile based intrusion detection for internet banking systems. 2008.
10. R. Kock. 80-20 principle: The secret to success by achieving more with less. *Crown Business*, 1999.
11. S. Kovach and W. V. Ruggiero. Online banking fraud detection based on local and global behavior. In *Proc. of the Fifth International Conference on Digital Society, Guadeloupe, France*, pages 166–171, 2011.
12. K. Pikkarainen, T. Pikkarainen, H. Karjaluoto, and S. Pahnila. The measurement of end-user computing satisfaction of online banking services: empirical evidence from finland. *International Journal of Bank Marketing*, 24(3):158–172, 2006.
13. L. F. Rodrigues, C. J. Costa, and A. Oliveira. How does the web game design influence the behavior of e-banking users? *Computers in Human Behavior*, 74:163–174, 2017.
14. U. Spagnolini and S. Zanero. Fraudbuster: Temporal analysis and detection of advanced financial frauds. *Detection of Intrusions and Malware, and Vulnerability Assessment*, page 211.
15. W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475, 2013.