Received November 30, 2016, accepted December 21, 2016, date of publication January 17, 2017,
date of current version March 13, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2654272

ProGuard: Detecting Malicious Accounts in Social-Network-Based Online Promotions

YADONG ZHOU[1], DAE WOOK KIM[2], JUNJIE ZHANG[2], (Member, IEEE), LILI LIU[1], HUAN JIN[3], HONGBO JIN[3], AND TING LIU[1]

[1]Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, 710049, China
[2]Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435 USA
[3]Tencent Technology (Shenzhen) Company Ltd., Shenzhen, 518000, China

Corresponding author: T. Liu (tingliu@xjtu.edu.cn)

This work was supported in part by Tencent Technology (Shenzhen) Company Ltd., in part by the National Natural Science Foundation under Grant 61572397, Grant 61502383, Grant 91418205, and Grant 61571360, in part by the National Key Research and Development Program of China under Grant 2016YFB0800202, in part by the Fundamental Research Project of Natural Science in Shaanxi Province under Grant 2015JM6298 and Grant 2015JM6299, and in part by The Ministry of Education Innovation Research Team under Grant IRT13035.

ABSTRACT Online social networks (OSNs) gradually integrate financial capabilities by enabling the usage of real and virtual currency. They serve as new platforms to host a variety of business activities, such as online promotion events, where users can possibly get virtual currency as rewards by participating in such events. Both OSNs and business partners are significantly concerned when attackers instrument a set of accounts to collect virtual currency from these events, which make these events ineffective and result in significant financial loss. It becomes of great importance to proactively detecting these malicious accounts before the online promotion activities and subsequently decreases their priority to be rewarded. In this paper, we propose a novel system, namely *ProGuard*, to accomplish this objective by systematically integrating features that characterize accounts from three perspectives including their general behaviors, their recharging patterns, and the usage of their currency. We have performed extensive experiments based on data collected from the Tencent QQ, a global leading OSN with built-in financial management activities. Experimental results have demonstrated that our system can accomplish a high detection rate of 96.67% at a very low false positive rate of 0.3%.

INDEX TERMS Online social networks, virtual currency, malicious accounts, intrusion detection, network security.

I. INTRODUCTION

Online social networks (OSNs) that integrate virtual currency serve as an appealing platform for various business activities, where online, interactive promotion is among the most active ones. Specifically, a user, who is commonly represented by her OSN account, can possibly get reward in the form of virtual currency by participating online promotion activities organized by business entities. She can then use such reward in various ways such as online shopping, transferring it to others, and even exchanging it for real currency [1]. Such virtual-currency-enabled online promotion model enables enormous outreach, offers direct financial stimuli to end users, and meanwhile minimizes the interactions between business entities and financial institutions. As a result, this model has shown great promise and gained huge prevalence rapidly. However, it faces a significant threat: attackers can control a large number of accounts, either by registering new accounts or compromising existing accounts, to participate in the online promotion events for virtual currency. Such malicious activities will fundamentally undermine the effectiveness of the promotion activities, immediately voiding the effectiveness of the promotion investment from business entities and meanwhile damaging ONSs' reputation. Moreover, a large volume of virtual currency, when controlled by attackers, could also become a potential challenge against virtual currency regulation [2].

It therefore becomes of essential importance to detect accounts controlled by attackers in online promotion activities. In the following discussions, we refer to such accounts as malicious accounts. The effective detection of malicious accounts enables both OSNs and business entities to take mitigation actions such as banning these accounts or decreasing the possibility to reward these accounts. However, designing an effective detection method is faced with a few significant challenges. First, attackers do not need to generate malicious content (e.g., phishing URLs and malicious executables) to launch successful attacks. Comparatively, attackers can effectively perform attacks by simply clicking links offered by business entities or sharing the benign content that is originally distributed by business partners. These actions themselves do not perceivably differentiate from benign accounts. Second, successful attacks do not need to depend on social structures (e.g., "following" or "friend" relationship in popular social networks). To be more specific, maintaining active social structures does not benefit to attackers, which is fundamentally different from popular attacks such as spammers in online social networks. These two challenges make the detection of such malicious OSN accounts fundamentally different from the detection of traditional attacks such as spamming and phishing. As a consequence, it is extremely hard to adopt existing methods to detect spamming and phishing accounts.

In order to effectively detect malicious accounts in online promotion activities by overcoming the aforementioned challenges, we have designed a novel system, namely *ProGuard*. *ProGuard* employs a collection of behavioral features to profile an account that participates in an online promotion event. These features aim to characterize an account from three aspects including i) its general usage profile, ii) how an account collects virtual currency, and iii) how the virtual currency is spent. *ProGuard* further integrates these features using a statistical classifier so that they can be collectively used to discriminate between those accounts controlled by attackers and benign ones. To the best of our knowledge, this work represents the first effort to systematically detect malicious accounts used for online promotion activity participation. We have evaluated our system using data collected from Tencent QQ, a leading Chinese online social network that uses a widely-accepted virtual currency (i.e., Q coin), to support online financial activities for a giant body of 899 million active accounts. Our experimental results have demonstrated that *ProGuard* can achieve a high detection rate of 96.67% with a very low false positive rate of 0.3%.

The rest of this paper is organized as follows. Section II introduces the related work. Section III briefly discusses the background of virtual-currency-enabled OSNs. Section IV describes how data was collected and labeled. We present the system design in Section V and evaluation results in Section VI. The discussion is provided in Section VII and Section VIII concludes.

II. RELATED WORK

Since online social networks play an increasing important role in both cyber and business world, detecting malicious users in OSNs becomes of great importance. Many detection methods have been consequently proposed [3]–[10]. Considering the popularity of spammers in OSNs, these methods almost exclusively focus on detecting accounts that send malicious content. A spamming attack can be considered as an information flow initiated from an attacker, through a series of malicious accounts, and finally to a victim account. Despite the diversity of these methods, they generally leverage partial or all of three sources for detection including i) the content of the spam message, ii) the network infrastructure that hosts the malicious information (e.g., phishing content or exploits), and iii) the social structure among malicious accounts and victim accounts. For example, Gao et al. [11] designed a method to reveal campaigns of malicious accounts by clustering accounts that send messages with similar content. Lee and Kim [12] devised a method to first track HTTP redirection chains initiated from URLs embedded in an OSN message, then grouped messages that led to webpages hosted in the same server, and finally used the server reputation to identify malicious accounts. Yang et al. [13] extracted a graph from the "following" relationship of twitter accounts and then propagated maliciousness score using the derived graph; Wu et al. [9] proposed a social spammer and spam message co-detection method based on the posting relations between users and messages, and utilized the relationship among user and message to improve the performance of both social spammer detection.

Compared to existing methods on detecting spamming accounts in OSNs, it is faced with new challenges to detect malicious accounts that participate in online promotion activities. First, different from spamming accounts, these accounts neither rely on spamming messages nor need malicious network infrastructures to launch attacks. Second, social structures are not necessary. Therefore, none of existing methods is applicable to detecting malicious accounts in online promotion activities. To solve the new challenges, our method detects malicious accounts by investigating both regular activities of an account and its financial activities.

Detecting fraudulent activities in financial transactions has also attracted significant research efforts [14], [15]. For example, Olszewski [16] represented the user account records in 2-dimensional space of the Self-Organizing Map grid, and proposed a detection method based on threshold-type binary classification algorithm to solve problems of credit card fraud and telecommunications fraud. Lin et al. [17] ranked the importance of fraud factors used in financial statement fraud detection, and investigated the correct classification rates of three algorithms including Logistic Regression, Decision Trees, and Artificial Neural Networks. Throckmorton et al. [18] proposed a corporate financial fraud detection method based on combined features of financial numbers, linguistic behavior, and non-verbal vocal. Compared to the studied financial fraud detection problems, account behaviors of collecting and using the virtual currency in online promotion activities are almost completely different with traditional financial systems since they do not only involve financial activities but also networking and online promotion activities.

To summarize, our work aims to address a new problem caused by the new trend of integrating online social networks and financial activities. *ProGuard* features new capability of fusing features from both networking and financial aspects for detection. Nevertheless, we believe our method and existing approaches can complement each other to improve the security of online social networks.

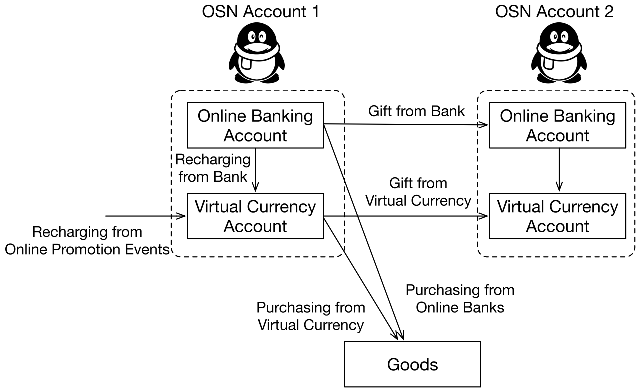


FIGURE 1. The integration of OSN accounts and financial accounts.

III. BACKGROUND

In an OSN that integrates financial activities, an OSN account is commonly associated with accounts for both online banking and virtual currency. Figure 1 presents such an example, where a QQ account, the most popular OSN account of Tencent, is associated with an online banking account for real currency and an account for virtual currency (i.e., Q coin). A user usually directly deposits real currency into her online banking account; she can recharge her virtual currency account from her banking account. By participating online promotion events, a user can also recharge her virtual currency account by collecting rewards from the promotion events. A user can expend from his accounts in two typical ways. First, she can use real or virtual currency to purchase both real and virtual goods (i.e., online shopping). Second, she can transfer both real and virtual currency to another user by sending out gifts.

Figure 2 presents the typical virtual currency flow when malicious accounts participate in online promotion events. The flow is composed of three phases including i) collecting, ii) multi-layer transferring, and iii) laundering the virtual currency. In first phase, an attacker controls a set of accounts to participate in online business promotion activities and each account possibly gets a certain amount of virtual currency as return. In the second phase, the attacker will instrument these currency-collection accounts to transfer the virtual currency to other accounts. Multiple layers of transferring activities might be involved to obfuscate the identities of malicious accounts used for participating online promotion activities. At the end of the second phase, a large amount of virtual currency will be aggregated into a few laundering accounts. In the third phase, the attacker will control the laundering accounts to trade the virtual currency into real cash by selling it to individual buyers. Attackers usually employ two methods to solicit individual buyers including sending spams and advertising through major e-commerce websites such as *www.taobao.com* and *www.tmall.com*. In order to compete with regulated sources for virtual currency (i.e., purchasing virtual currency using real currency), attackers usually offer a considerable discount.

Our objective is to design a detection system capable of identifying malicious accounts that participate in online promotion events for virtual currency collection (at the collection phase) before rewards are committed. Detecting malicious accounts at this specific time point (i.e., before the commitment of rewards and at the collection phase) results in unique advantages. First, as a simple heuristic to prevent freshly registered accounts that are likely to be bots, business entities usually require the participating accounts to be registered for a certain amount of time (e.g., a few weeks). Therefore, the detected and mitigated malicious accounts cannot be immediately replaced by the newly registered accounts, thereby drastically limiting attackers' capabilities. In contrast, no constraint is applied for accounts used for virtual currency transferring and laundering. This implies such accounts can be easily replaced by attackers if detected, resulting negligible impact to attackers' capabilities. Second, our detection system will label whether an account is malicious when it participates in an online promotion event; this enables business entities to make actionable decisions such as de-prioritize this account from being rewarded in this event. Therefore, it can proactively mitigate the financial loss faced by business entities.

IV. DATA

We have collected labelled data from Tencent QQ, a leading Chinese online social network that offers a variety of services such as instant message, voice chat, online games, online shopping, and e-commerce. All these services support the usage of the Q coin, the virtual currency distributed and managed by Tencent QQ. Tencent QQ has a giant body of 899 million active QQ accounts with a reportedly peak of 176.4 million simultaneous online QQ users. Tencent QQ is one of the global leading OSNs that are actively involved in virtual-currency-based online promotion activities.

Our data set is composed of 28,000 malicious accounts and 28,000 benign accounts, where all of these accounts are randomly sampled from the accounts that participated in Tencent QQ online promotion activities in August 2015. The labeling process starts from identifying laundering accounts (i.e., accounts that are associated with virtual currency spams and accounts that sell virtual currency in major e-commerce websites). Specifically, if an account transfers virtual currency to any account that engages in virtual-money laundering activities, this account will be labeled as malicious. Such "trace-back" process may involve multiple layers of transferring, which is visualized at the bottom in Figure 2. *It is worth noting that although both malicious and benign accounts are labelled based on their activities in Phase-2 (i.e., currency transferring) and Phase-3 (i.e., laundering), the data used for building the detection system are collected before the launch of the online promotion event.* The reason is that the objective of our detection system is to identify malicious accounts *before the rewards are committed.*

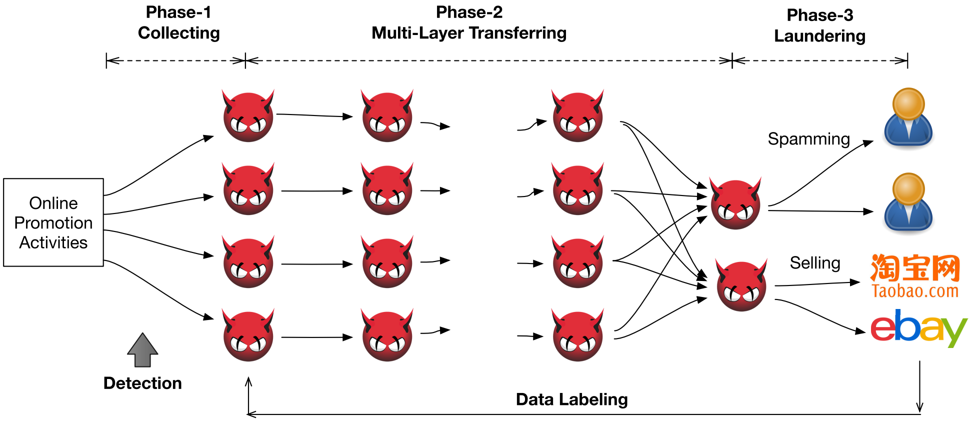


FIGURE 2. Virtual currency flow for malicious OSN accounts.

The top of Figure 3 presents the temporal relationship among the data collection process, online promotion events, and the account labeling process. Therefore, it is worth noting that an account may not have any historical financial activities (even for virtual currency collection activities) since it participates in the online promotion for the first time.

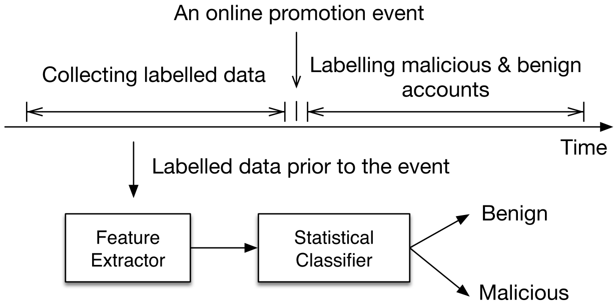

An online promotion event
Collecting labelled data
Labelling malicious & benign accounts
Time
Labelled data prior to the event
Feature Extractor
Statistical Classifier
Benign
Malicious

FIGURE 3. The architectural overview of the system.

Although the aforementioned "trace-back" method is effective in manually labeling malicious accounts, using it as a detection method is impractical. First, it requires a tremendous amount of manual efforts for forensic analysis such as identifying suspicious virtual-currency dealers in external e-commerce websites, correlating spamming content with user accounts, and correlating sellers' profiles with user accounts. In addition, evidence for such forensic analysis will be only available *after* malicious accounts participate in online promotion events. Therefore, this data labeling process, if used as detection method, cannot guide business entities to mitigate their financial loss proactively. In contrast, our method is designed to detect malicious accounts prior to the reward commitment. For each account, we collect a variety of information including 1) login activities, 2) a list of anonymized accounts that this account has sent instant messages to, 3) service purchase activities, 4) the recharging activities, and 5) the expenditure activities.

V. SYSTEM DESIGN

ProGuard is composed of two phases, namely the training phase and the detection phase. In the training phase, a statistical classifier is learnt from a set of pre-labelled malicious and benign accounts. In the detection phase, an unknown account will first be converted to a feature vector and then analyzed by the statistical classifier to assess its maliciousness. The bottom of Figure 3 presents the architectural overview of *ProGuard*. As a variety of statistical classifiers have been developed and widely used, designing features capable of discriminating between malicious accounts and benign accounts becomes of central focus. In this section, we will introduce various features and demonstrate their effectiveness on differentiating malicious accounts from benign ones. We propose three general guidelines to steer the feature design.

- **General Behaviors**: Benign accounts are usually used by regular users for variety of activities such as chatting, photo sharing, and financial activities. In contrast, malicious accounts are more likely to be driven by online promotion events. Therefore, the benign accounts tend to be more socially active compared to malicious accounts.
- **Currency Collection**: The malicious accounts under investigation focus on using online promotion activities to collect virtual currency. In contrast, benign users are likely to obtain virtual currency from multiple resources.
- **Currency Usage**: Attackers' ultimate objective is to monetize the virtual currency. In contrast, benign users use their virtual currency in much more diversified ways.

A. GENERAL-BEHAVIOR FEATURES

Malicious accounts tend to be less active compared to benign accounts with respect to the non-financial usage. Attackers usually control their accounts to only participate in online promotion activities. In contrast, benign accounts are more likely to engage in active interaction with other users.

- ***Feature 1: The Ratio of Active Days.*** This feature represents the ratio of the number of active days of an account for the passed one year. Specifically, if an account is logged in at least once for a day, this day will be labeled as "active" for this account.

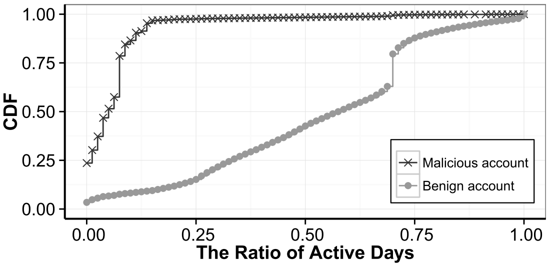


FIGURE 4. **Feature 1 - the ratio of active days.**

Attackers usually login malicious accounts for participating in online promotion activities that involve virtual currency. Therefore, malicious accounts tend to be silent in the absence of online promotion activities. The availability of promotion activities is significantly influenced by timing and spatial factors. For example, promotion activities are intensive over holiday seasons, special dates, and regional events while occasionally available for other time periods. As a consequence, malicious accounts tend to be inactive generally. Comparatively, benign accounts are used by regular users and their logins are driven by the daily usage such as chatting and photo sharing. Many users configure their applications to automatically login upon the bootstrap of the underlying system (e.g., a smartphone), which further facilitates volatility of benign accounts. Figure 4 presents the distribution of feature values for both malicious accounts and benign accounts. As illustrated in the figure, the vast majority of malicious accounts (i.e., approximately 98% of malicious accounts) are active for less than 20% of total days whereas only a small percentage of benign accounts (i.e., less than 20%) experience the same active level (i.e., being active for less than 20% of one year).

- ***Feature 2: The Number of Friends.*** This feature summarizes the number of friends for each account.

As a common feature for almost all online social networks, each OSN account has a list of friends. It usually implies a considerable amount of user-user interaction for one user to add another one as her friend. It is common for a benign user to maintain a relatively lengthy friend list for various social activities such as chatting and photo sharing. In contrast, an attacker usually lacks the motivation to maintain a friend list since it contributes little to promotion participation but costs significant efforts such as solving captcha challenges. Figure 5 presents the distribution of values for this feature, where malicious accounts tend to have much less friends compared to benign accounts. Specifically, approximately 80% of malicious accounts have less than 40 friends while about 70% benign accounts have more than 200 friends.

- ***Feature 3: The Number of Services Purchased By An Account.*** This feature represents the total number of types of upgraded membership that an account has paid for through all possible methods.

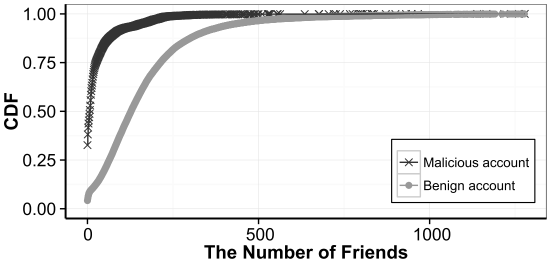


FIGURE 5. **Feature 2 - the number of friends for an account.**

It is a common feature in many online social networks that an user can upgrade his/her account by making a certain amount of payment through various ways such as credit card, wire transfer, and virtual currency. In the Tencent dataset, we consider 8 types of most popular upgraded membership including QQ VIP, Qzone, SVIP, QQ Music, Hollywood VIP, QQ Games, QQ books, Tencent Sports. An upgraded account can a wide range of paid benefits such as advanced capabilities for an online game avatar, enriched decoration for the account appearance, and expanded visibility of visitors. While a certain amount of benign users are inclined to be motivated to upgrade their accounts, accounts controlled by attackers are extremely unlikely to participate in such paid upgrade since the upgraded membership contributes nothing to their collection of virtual currency. Figure 6 presents the distribution for this feature: while approximately 37% of benign users purchased at least one type of upgraded membership, the vast majority of malicious accounts do not make any purchase.

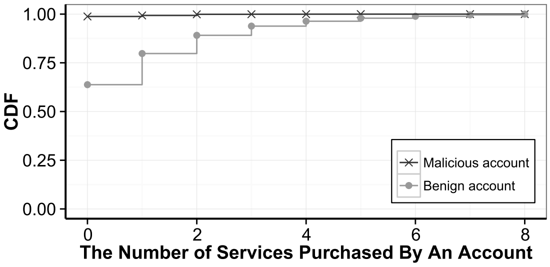


FIGURE 6. **Feature 3 - the number of services purchased by an account.**

B. CURRENCY COLLECTION FEATURES

In addition to collecting virtual currency by participating in online promotion activities, an OSN user can recharge her account with virtual currency through various ways such as wire transfer, selling virtual goods, and transferring from other accounts. Generally, benign users should be more active with respect to recharging their accounts. We propose two features to characterize this trend from two aspects including the amount of recharging and the important sources for recharging.

- ***Feature 4 - The Average Recharge Amount of Virtual Currency.*** This feature represents the average amount of virtual currency for each recharge regardless of the sources for recharging.

Benign users who participate in online promotion activities are usually also interested in other online financial activities. Therefore, these benign users tend to actively recharge their accounts. The recharge amount for each time by a benign user is commonly considerably large since users tend to decrease the hassle of recharging. In contrast, if a malicious account has been recharged, the amount of virtual currency for each recharge is usually bounded by a relatively small volume offered by the online promotion activity. Figure 7 presents the distribution of this feature for benign and malicious accounts, respectively. Specifically, the average recharge amount is higher than 1100 Chinese cents[1] for more than 50% of benign users, where only a small percentage (i.e., approximately 15%) of malicious users has an average amount that is higher than 140 Chinese cents.

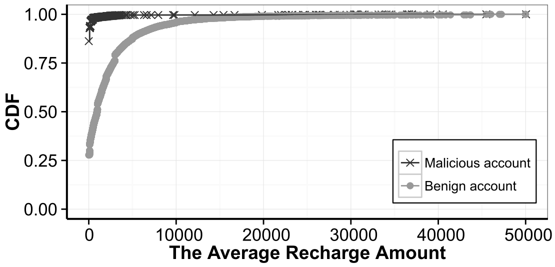


FIGURE 7. **Feature 4 - the average recharge amount.**

We then consider the sources for recharge. Despite a variety of possibilities, we focus on one source that is rewards from promotion activities, and accordingly design one self-explanatory features as follows:

- ***Feature 5: The Percentage of Recharge From Promotion Activities.***

The feature intuitively profiles how significantly online promotion activities contribute to the wealth of an account. Benign users are inclined to employ a variety of sources for recharge. Comparatively, malicious accounts usually exclusively rely on online promotion activities to collect virtual currency. Figure 8 presents the distribution of this feature. Both the majority (approximate 88%) of benign and malicious accounts have not collected any virtual currency from promotion activities. In spite of the similarity, malicious accounts differentiate themselves from benign ones by experiencing a strong bipolar distribution: approximate 88% of malicious accounts do not collect virtual currency from online promotion activities at all where as the remaining (i.e., approximate 12%) accumulate their wealth exclusively from online promotion activities; the number of malicious accounts, whose currency is partially collected from online promotion activities, is negligible. This implies that it is the first time for about 88% of malicious accounts to participate in online promotion events where 12% are reused accounts.

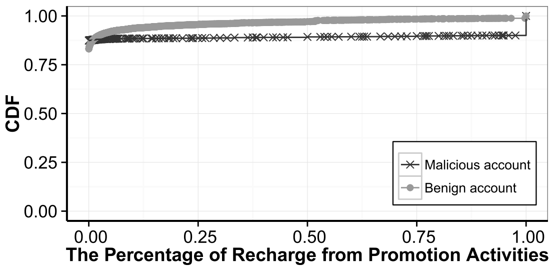


FIGURE 8. **Feature 5 - the percentage of recharge from promotion activities.**

C. FEATURES OF USAGE ACTIVITIES

As an increasing number of business capabilities are integrated into social networks, users conduct a variety of activities such as shopping and gifting. Features in this category characterize how users spend their wealth. As a means towards this end, we propose three features.

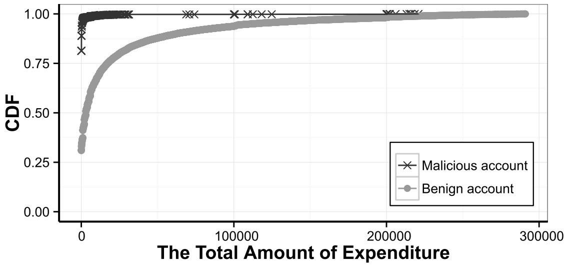


FIGURE 9. **Feature 6 - the total amount of expenditure.**

- ***Feature 6: Total Amount of Expenditure.***

This feature characterizes the total amount of expenditure of an account regardless of the possible sources such as the associated bank accounts, the virtual currency, and other online social network platforms. As the popular online social networks are integrated into almost all mainstream e-business infrastructures, shopping and gifting through these accounts becomes prevalent. Users keep recharging their accounts, persistently associate their bank accounts with OSN accounts, and actively engage in shopping and gifting. Therefore, we expect that benign accounts accumulate a high amount of expenditure. Comparatively, the total amount of currency controlled by each malicious account is constrained by the total number of virtual currency collected from online promotions, which is expected to be relatively small. Figure 9 presents the distribution of this feature. At least 50% of benign accounts have spent more than 4000 Chinese cents. Comparatively, only a tiny percentage of 0.9% malicious accounts spent more than 4000 Chinese cent; the vast majority of malicious accounts never commit any spending.

- ***Feature 7: The Percentage of Expenditure From Banks.***

As we have introduced, a user can associate her bank account with the OSN account. This bank account can be directly used for shopping and gifting in addition to recharging the OSN account with virtual currency. Such association may greatly facilitate financial activities but result in

[1] Each Chinese cent is approximately $\frac{1}{6}$ of a U.S. cent

exposure of users' bank identities in case of law enforcement. Figure 10 presents the evaluation of this feature based on the real-world data. A very tiny percentage of malicious accounts expended their currency from bank accounts. Comparatively, this percentage is considerably high for benign users (i.e., around 45%).

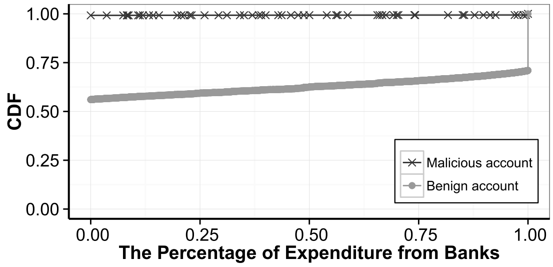


FIGURE 10. **Feature 7 - the percentage of expenditure from banks.**

- ***Feature 8: The Percentage of Expenditure as Gifts.***

After malicious accounts collect virtual currency from the online promotion activities, they will transfer it to malicious accounts used for trading. Sending online gift cards becomes the best option for malicious accounts to transfer currency for two reasons. First, sending online gift cards inside an OSN usually does not incur any cost. Second, such transfer is independent to any bank, thereby requiring no personal information and consequently minimizing the exposure of attackers. We therefore design this feature to quantify the percentage of all expenditure that is used for gifts. Figure 11 presents the empirical analysis of this feature. Specifically, malicious accounts show a strong pattern of bipolar distribution. Particularly, approximate 81% of malicious accounts never sent any gifts. Most likely, these accounts have never successfully participate any promotion activities and therefore have nothing to transfer. The rest of them (i.e., about 19%) spend all of their expenditure on gifts, which implies that they transfer all of their wealth to other accounts. Similar to malicious accounts, most of benign accounts (i.e., about 80%) never engaged in any gifting activities. However, the remaining benign accounts (i.e., about 20%) seem reluctant to use all of their wealth as gift.

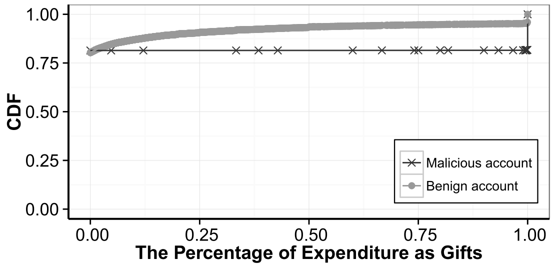


FIGURE 11. **Feature 8 - the percentage of expenditure as gifts.**

VI. EVALUATION

We performed extensive evaluation of *ProGuard*, which focuses on the overall detection accuracy, the importance of each feature, and the correlation among these features. For this evaluation, we used totally 56,000 accounts whose entire dataset is divided into 28,000 malicious accounts and 28,000 benign accounts. Such data serve as a well-balanced dataset for training a statistical classifier [19].

A. DETECTION ACCURACY

We have used the normalized Random Forest (RF) as the statistical classifier for *ProGuard* and evaluated its detection accuracy. RF classifier [20] is an ensemble of unpruned classification trees, which is trained over bootstrapped samples of the original data and the prediction is made by aggregating majority vote of the ensemble. In order to avoid the bias caused by the selection of specific training set, we also performed 10-fold cross-validation. Specifically, the entire dataset is partitioned to 10 equal-size sets (i.e., 10-folds); then iteratively 9-folds are used for training and the remaining 1-fold is adopted for testing. The RF classifier was trained with 3000 trees and randomly sampled 4 features for each of tree splitting [21]. The receiver operating characteristic (ROC) that characterizes the overall detection performance of *ProGuard* is presented in Fig. 12. The experimental results have shown that *ProGuard* can achieve high detection accuracy. For example, given the false positive rate of 0.3%, *ProGuard* can accomplish a high detection rate of 96.67%.

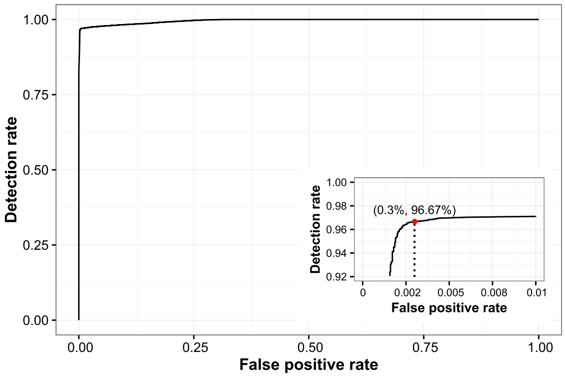


FIGURE 12. **ROC curve on 8 features.**

In practice, alternative statistical classifiers might be adopted to render new performance benefits such as scalability. Therefore, we also evaluate how *ProGuard* performs when alternative classifiers are used. As a means towards this end, we used Support Vector Machine (SVM) [22] and Gradient-Boosted Tree [23] to repeat our experiments. Specifically, we used 10-fold cross validation for each of classifiers and calculated the area under the ROC curve (AUC) [24], a widely used measure of quality of supervised classification models, which is equal to the probability that a randomly chosen sample of malicious accounts will have a higher estimated probability of belonging to malicious accounts than a randomly chosen sample of benign accounts. Since AUC is cutoff-independent and values of AUC range from 0.5 (no predictive ability) to 1.0 (perfect predictive ability), a higher AUC of a classifier indicates the better prediction performance, irrespective of the cutoff selection.

Table 1 lists the AUC values for all three classifiers used in the experiments. Both SVM and Gradient-Boosted Tree accomplished high detection results, comparable with the Random Forest which has the best performance on AUC. The experimental results imply that our proposed features are not sensitive to the selection of statistical classifiers.

TABLE 1. AUCs for three classifiers.

Classifier	AUC
Random Forest	0.9959
SVM	0.9753
Gradient-Boosted Tree	0.9781

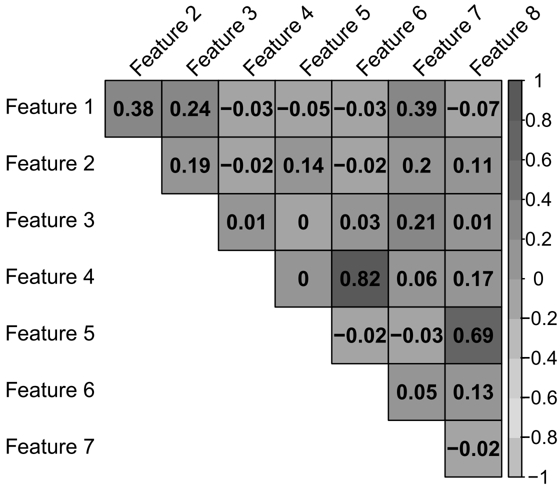
B. FEATURE IMPORTANCE AND CORRELATION

We investigated the relative importance of the proposed features in the context of Random Forest classifier, which has accomplished the best detection accuracy according to our experiments. We employed the variable importance of each feature to the Random Forest classification model using permutation test [21]. The variable importance for each feature is computed by mean decrease in accuracy, which is defined as a prediction error rate after permuting an each feature [21]. The rank of features based on the variable importance is shown in Table 2. Specifically, the ratio of active days (Feature 1), the average recharge amount of virtual currency (Feature 4), and the percentage of expenditure from banks (Feature 7) represent the most significantly for detection. It is worth noting that these top three features cover three complementary aspects including the general behaviors, currency collection, and currency usage that guide the feature design.

TABLE 2. Feature importance rank of *ProGuard* by random forest.

Rank	Variable importance
Feature 1	465.4
Feature 4	349.9
Feature 7	246.6
Feature 2	61.31
Feature 5	56.91
Feature 8	52.17
Feature 6	46.44
Feature 3	35.63

We also performed the correlation among various features, where the correlation implies the extent to which a feature might be redundant given other features. Two widely-adopted methods have been used in our experiments. First, the upper triangular of correlation matrix is carried out for discovering if a pair of strongly correlated features appear within the features, where each column in the upper triangular matrix represents the Pearson's r correlation coefficient [25] of a pair of two distinct features. The Pearson's correlation coefficient



	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
Feature 1	0.38	0.24	−0.03	−0.05	−0.03	0.39	−0.07
Feature 2		0.19	−0.02	0.14	−0.02	0.2	0.11
Feature 3			0.01	0	0.03	0.21	0.01
Feature 4				0	0.82	0.06	0.17
Feature 5					−0.02	−0.03	0.69
Feature 6						0.05	0.13
Feature 7							−0.02

FIGURE 13. Upper triangular matrix.

$r \in [-1, 1]$ of two features X and Y can be defined as

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

where $\bar{X}$ and $\bar{Y}$ denote the means of the two features. Fig. 13 shows that the most of features are not strongly correlated one to each other (i.e, Pearson's correlation coefficient $|r| \geq 0.9$). For example, a pair of two features, Feature 1 (*The Ratio of Active Days*) and Feature 8 (*The Percentage of Expenditure as Gifts*) represents that the highest negative correlation score is 0.07 and the highest positive correlation between Feature 4 (*The Average Recharge Amount of Virtual Currency*) and Feature 6 (*The Total Amount of Expenditure*) is 0.82.

Next, we analyzed Principal Component Analysis (PCA), which can be used to evaluate variable correlation in regard to the variance of the data [26]. Figure 14 shows the experimental result on PCA variables factor map [27]. In the variable factor map, each of features is expressed as an arrow and the angle between the two arrows of features implies the correlation among the respective features on the third and fourth principal components (PC). For example, given the angle between the two arrows of different two features goes near 90 degrees, they might not be correlated. As can be seen in Figure 14, the angles between the most of features are found proximate to 90 degrees (e.g., Feature 3 (*The Number of Services Purchased By An Account*) and Feature 5 (*The Percentage of Recharge from Promotion Activities*) onto the 3rd and 4th PCs), implying a weak correlation between features. According to the correlation matrix and PCA variable factor map, which show little correlation with each other, we conclude that majority of the features complement each other given their tendency towards linearly independence.

VII. DISCUSSION

Attackers may attempt to evade our detection after they know the design of *ProGuard*. This represents a general challenge for all detection systems rather than a specific design flaw of the proposed system. Specifically, attackers can instrument their accounts so that their behaviors are

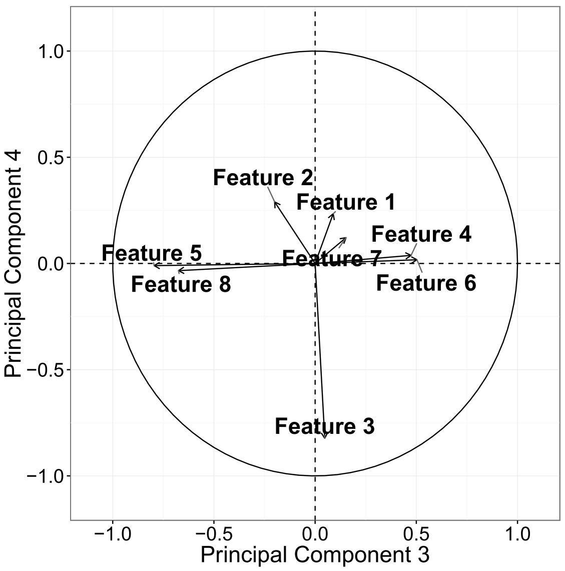


FIGURE 14. The variables factor map (PCA).

indistinguishable from benign accounts. However, since *ProGuard* detection features characterize elements of malicious accounts that are critical to their success of attacks and stealthiness against other detection systems, the successful evasion may fundamentally constrain attackers' capabilities. For example, attackers can significantly increase the number of active days of malicious accounts. However, it may expose malicious accounts to existing bot-account detection systems that leverage frequent login patterns of malicious accounts [28]. Attackers can also increase the number of friends by adding malicious accounts as friends. Nevertheless, this may qualify the applicability of many detection systems that take advantage of social structures such as [13], [29], and [30]. Attackers can also increase the diversity for recharging sources, the amount of recharging, and the expenditure from bank accounts. However, these solutions directly increase the financial cost for launching the attacks, which could make attacks themselves meaningless. Attackers might also attempt to decrease the percentage of expenditure as gifts, which, however, fundamentally limits the bandwidth to launder the collected virtual currency.

It exists the possibility that an attacker may hack some benign accounts and use them to participate online promotion events. However, hacking a considerable number of benign accounts is not a trivial task, which usually implies significant cost. In addition, mainstream social networks have usually enforced effective means to assist victim users to recover their hacked accounts. On the contrary, it is free for any user, including the attacker, to register a large number of accounts, which are dedicated to persistent malicious activities. In summary, attackers have extremely limited motivation to use hacked accounts for this type of attacks. Nevertheless, if a hacked account is indeed used by an attacker for such attacks, this account will experience mixed benign and malicious behavior. If the malicious behavior dominates (i.e., the benign online financial activities are negligible), then we expect our method can still detect this account; unfortunately, if the benign activities dominates (i.e., this account is very active at online financial activities), this account is likely to introduce a false negative. Addressing false negatives in this case is definitely an important issue and seeking effective solutions falls into our future work.

Considering the active trend of integrating OSNs with financial capabilities, detecting malicious accounts that engage in suspicious financial activities becomes of central importance. Although the design and evaluation of *ProGuard* are based on real-world data collected from Tencent QQ, a leading OSN with 899 million active accounts, the features and the detection framework can be easily applied to other OSNs that integrate financial activities. Specifically, all the proposed features are based on essential financial functions such as recharging and gifting. In addition, all current features rely on coarse-grained information that minimizes privacy concerns, which may foster the deployment of the proposed system in a detection-as-service model. Despite the fact that *ProGuard* can effectively detect malicious accounts used for collecting virtual currency from online promotion activities, it is not designed for detecting malicious accounts used for transferring and laundering virtual currency. Extending *ProGuard* to include such detection capabilities falls into our future work.

VIII. CONCLUSION

This paper presents a novel system, *ProGuard*, to automatically detect malicious OSN accounts that participate in online promotion events. *ProGuard* leverages three categories of features including general behavior, virtual-currency collection, and virtual-currency usage. Experimental results based on labelled data collected from Tencent QQ, a global leading OSN company, have demonstrated the detection accuracy of *ProGuard*, which has achieved a high detection rate of 96.67% given an extremely low false positive rate of 0.3%.

REFERENCES

[1] Y. Wang and S. D. Mainwaring, "Human-currency interaction: Learning from virtual currency use in China," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 25–28.

[2] J. S. Gans and H. Halaburda, "Some economics of private digital currency," Rotman School Manage., Toronto, ON, Canada, Tech. Rep. 2297296, 2013.

[3] X. Hu, J. Tang, and H. Liu, "Online social spammer detection," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 59–65.

[4] X. Hu, J. Tang, and H. Liu, "Leveraging knowledge across media for spammer detection in microblogging," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2014, pp. 547–556.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, Bot, or cyborg?" *IEEE Trans. Depend. Sec. Comput.*, vol. 9, no. 6, pp. 811–824, Nov. 2012.

[6] Z. Chu, S. Gianvecchio, A. Koehl, H. Wang, and S. Jajodia, "Blog or block: Detecting blog bots through behavioral biometrics," *Comput. Netw.*, vol. 57, no. 3, pp. 634–646, 2013.

[7] S. Fakhraei, J. Foulds, M. Shashanka, and L. Getoor, "Collective spammer detection in evolving multi-relational social networks," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1769–1778.

[8] Y.-R. Chen and H.-H. Chen, "Opinion spammer detection in Web forum," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 759–762.

[9] F. Wu, J. Shu, Y. Huang, and Z. Yuan, "Social spammer and spam message co-detection in microblogging with social context regularization," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag.*, 2015, pp. 1601–1610.

[10] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Sep. 2014.

[11] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 35–47.

[12] S. Lee and J. Kim, "Warningbird: Detecting suspicious URLS in twitter stream," in *Proc. NDSS*, vol. 12. 2012, pp. 1–13.

[13] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers," in *Proc. Int. Workshop Recent Adv. Intrusion Detection*, 2011, pp. 318–337.

[14] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, Oct. 2016.

[15] J. West and M. Bhattacharya, "Intelligent financial fraud detection: A comprehensive review," *Comput. Secur.*, vol. 57, pp. 47–66, Jun. 2016.

[16] D. Olszewski, "Fraud detection using self-organizing map visualizing the user profiles," *Knowl.-Based Syst.*, vol. 70, pp. 324–334, Jan. 2014.

[17] C.-C. Lin, A.-A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," *Knowl.-Based Syst.*, vol. 89, pp. 459–470, Sep. 2015.

[18] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decision Support Syst.*, vol. 74, pp. 78–87, Jun. 2015.

[19] Z. Afzal, M. J. Schuemie, J. C. van Blijderveen, E. F. Sen, M. C. Sturkenboom, and J. A. Kors, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," *BMC Med. Informat. Decision Making*, vol. 13, no. 1, p. 1, 2013.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[21] A. Liaw, M. Wiener, *Package 'Randomforest'*, accessed On Jun. 15, 2016. [Online] Available: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[22] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[23] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[24] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[25] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statist.*, vol. 42, no. 1, pp. 59–66, 1988.

[26] I. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2005.

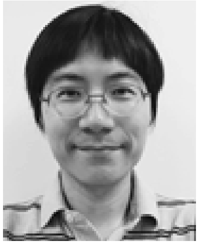
[27] (2014). *R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing Vienna, Austria*, accessed on Jun. 15, 2016. [Online]. Available: http://www.R-project.org/

[28] Y. Zhao *et al.*, "Botgraph: Large scale spamming botnet detection," in *Proc. NSDI*, vol. 9, pp. 321–334, Jan. 2009.

[29] J. Song, S. Lee, and J. Kim, "Spam filtering in twitter using sender-receiver relationship," in *Proc. Int. Workshop Recent Adv. Intrusion Detection*, 2011, pp. 301–317.

[30] T.-S. Moh and A. J. Murmann, "Can you judge a man by his friends?—Enhancing spammer detection on the twitter microblogging platform using friends and followers," in *Proc. Int. Conf. Inf. Syst. Technol. Manage.*, 2010, pp. 210–220.

YADONG ZHOU received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University, China, in 2004 and 2011, respectively. He was a Post-Doctoral Researcher with The Chinese University of Hong Kong in 2014. He is currently an Assistant Professor with the Department of Automation, Xi'an Jiaotong University. His research focuses on data analysis and mining, network science and its applications, software defined networks.



DAE WOOK KIM received the B.S. degree in computer science and engineering from Michigan State University, USA, in 2003, the M.S. degree in computer science and information from Syracuse University, USA, in 2007. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Wright State University. His research interest lies primarily in network security.

JUNJIE ZHANG received the B.S. degree in computer science and the M.S. degree in system engineering from Xi'an Jiaotong University, China, in 2003 and 2006, respectively, and the Ph.D. degree in computer science from the Georgia Institute of Technology in 2012. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Wright State University. His current research focuses on network security and Cyber-physical system security.

LILI LIU received the B.S. degree in control science and engineering from Xi'an Jiaotong University, China, in 2014, where she is currently pursuing the master's degree with the Department of Automation. Her research focuses on data analysis and data mining.

HUAN JIN received the B.S. and M.S. degrees in computer science and technology from the Xi'an University of Technology, China, in 2006 and 2009, respectively. He is currently a Senior Engineer with Tencent Technology (Shenzhen) Company Ltd. His current research focuses on online payment and risk control.

HONGBO JIN received the B.S. degree in control theory and engineering from the Chongqing University of Posts and Telecommunications, China, in 2003, and the M.S. degree in control theory and engineering from Chongqing University, China, in 2006. He is currently a Senior Engineer with Tencent Technology (Shenzhen) Company Ltd. His current research focuses on online payment and risk control.

TING LIU received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University, China, in 2003 and 2010, respectively. He is currently an Associate Professor with the Department of Automation, Xi'an Jiaotong University. His research focuses on security and reliability in computer network, Cyber-physical system, and software system.

• • •