FrauDetector: A Graph-Mining-based Framework for Fraudulent Phone Call Detection

Vincent S. Tseng[1]*, Josh Jia-Ching Ying[2], Che-Wei Huang[2], Yimin Kao[3], and Kuan-Ta Chen[4]

[1] Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, ROC

[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, ROC

[3] Gogolook Co. Ltd., Taiwan, ROC

[4] Institute of Information Science, Academia Sinica, Taiwan, ROC

*correspondence: vtseng@cs.nctu.edu.tw,

{jashying, weiiboy}@gmail.com, yiminkao@gogolook.com, swc@iis.sinica.edu.tw

ABSTRACT

In recent years, fraud is increasing rapidly with the development of modern technology and global communication. Although many literatures have address the fraud detection problem, these existing works focus only on formulating the fraud detection problem as a binary classification problem. Due to limitation of information provided by telecommunication records, such classifier-based approaches for fraudulent phone call detection do not work well. In this paper, we develop a graph-mining-based fraudulent phone call detection framework for mobile application to automatically annotate fraudulent phone numbers with a "fraud" tag, which is a crucial prerequisite for distinguishing fraudulent phone calls from normal phone calls. Our detection approach performs a weighted HITS algorithm to learn the trust value of each remote phone number. Based on the telecommunication records, we build two kinds of directed bipartite graph: i) contact_book-remote phone number graph (CPG) and ii) user-remote_phone_number graph (UPG) to represent telecommunication behavior of users. To weight the edges of CPG and UPG, we extract features for each pair of user and remote phone number in two different yet complementary aspects: 1) duration relatedness (DR) between user and phone number; and 2) frequency relatedness (FR) between user and phone number. Upon weighted CPG and UPG, we determine the trust value for each remote phone number. Finally, we conduct a comprehensive experimental study based on a real dataset collected through an anti-fraud mobile application, *Whoscall*. The results demonstrate the effectiveness of our weighted HITS-based approach and show the strength of taking both DR and FR into account in feature extraction.

Keywords

Telecommunication Fraud, Trust Value Mining, Fraudulent Phone Call Detection, Weighted HITS algorithm.

1. INTRODUCTION

In our daily life, fraudulent activities can be seen on different communication channels, such as telecommunication networks, on-line banking, and e-commerce. Fraud is increasing rapidly with the development of modern technology and global communication. As the result, millions of people suffer terribly with these fraudulent activities. To protect people against the damage caused by fraudulent activities, many anti-fraud mechanisms have been proposed, where such mechanism for telecommunication are still in its infancy stage. Actually, all of the anti-fraud mechanisms for telecommunications more or less rely on the annotations made by the crowd. A person would annotate a phone number as fraud while he receives a phone call from the phone number and realizes that the phone call is fraudulent. In other words, frauders still can approach innocent people via the remote phone number before that particular number is blacklisted as fraud. Because not everyone is willing to annotate a fraudulent phone number, it is possible that the frauders can make fraudulent calls through the phone number for days, weeks, or even months. We denote this phenomenon as the *time lag* in the fraud detection problem. Consequentially, dealing with time lag of fraud detection problem for fraudulent phone calls detection has become an important issue to be explored.

Figure 1. A screenshot of Whoscall, an anti-fraud mobile app.

In recent years, a new breed of smartphone applications for anti-fraud such as *Whoscall*[1], *Pindrop*[2] and *Truecaller*[3], have emerged. Such smartphone applications identify background information of incoming unknown calls in seconds through tags reported by other users, Internet search results, and yellow/white pages. Obviously, this kind of applications is still based on the principle of annotation by crowd to detect the fraudulent remote phone number and broadcast the list of detected fraudulent remote phone numbers to each anti-fraud application user. As mentioned earlier, the principle of annotation by crowd would cause the time lag of fraud detection problem. Fortunately, such anti-fraud applications make collecting records of telecommunication much easier. Figure 1 shows a screenshot of an anti-fraud application, *Whoscall*. We can observe that the phone number would be recorded as well when a person makes a phone call through the phone number. Furthermore, the timestamps of receiving phone call request, starting communication and ending communication can also be recorded. While receiving a phone call, the user can

[1] http://Whoscall.com/

[2] http://www.pindropsecurity.com/

[3] http://www.truecaller.com/

annotate the remote phone number whether it is fraudulent. Therefore, we can discover the characteristics of fraudulent remote phone numbers through mining these communication records. Based on the discovered characteristics of fraudulent remote phone numbers, we can build the model for detecting fraudulent remote phone numbers.

Intuitively, as shown in formula (1), given a set of remote phone numbers P, the fraudulent remote phone number detection can be formulated as predicting trust value of a given remote phone number.

$$f(p) \to \tau, \text{where } p \in P \text{ and } \tau \in [0, 1] \quad (1)$$

In daily life, a person may receive a lot of normal phone calls relatively compared with fraudulent phone calls. In other words, the number of remote phone numbers tagged with "fraud" may rarely occur in telecommunication records. Hence, fraudulent remote phone number detection from telecommunication records may be addressed as an *imbalance-label prediction problem* [5] [21]. While imbalance-label prediction techniques have been developed for many fraud detections, such as credit card fraud detection [6], telecommunications fraud detection [7] and insurance fraud detection [19], the problem has not been well explored under the context of communication records, where we can only operate over telecommunication activities for certain phone calls.

Although the issues of telecommunication fraud detection have been discussed in many literatures, existing studies mostly consider only on the *classifier-based fraud detection* [7] [9] [10] [28] in which the fraud detection problem is formulated as a binary classification problem. Notice that a fundamental issue of classifier-based fraudulent phone call detection is to identify and extract a number of descriptive features for each remote phone number in the system. Selecting the right features is important because the selected features have a direct impact on the effectiveness of the classification task. As a result, the behaviors of users who receive a phone call are aggregated into a single value of a feature for the remote phone number. For example, as Figure 2 shows, $user_1$ and $user_2$ directly hung up the phone, $user_3$ answers the call for a while, and $user_4$ and $user_5$ miss the call and call back later. Thus, the values of some features that reflect the call duration of users to the known phone should be increased by the behaviors of $user_4$ and $user_5$. Accordingly, the features might not distinguish the fraudulent remote phone numbers from normal remote phone numbers because both two kinds of remote phone numbers have similar call durations with users. Additionally, frauders tend to make fraud telecommunication to the same group

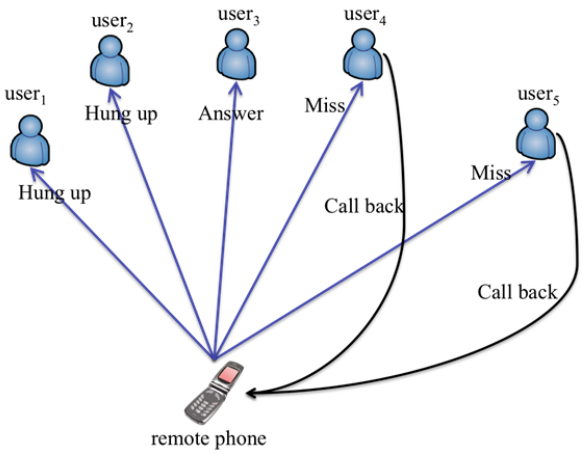


Figure 2. An example of the behaviors of users who receive a phone call.

of users because the frauders have those users' profiles. Therefore, the classifier-based fraud detection requires the analysis of users' profiles to produce useful features. However, the analysis of users' profiles costs a lot of effort and may even violate the privacy policy.

To support fraudulent phone call detection based only on the telecommunication records, we propose a novel graph-mining-based fraudulent phone call detection framework, called *FrauDetector,* to evaluate the trust value of an unknown remote phone number. The framework consists of two major modules: 1) offline learning module, and 2) on-line detection module. In the offline mining module, we adopt the notion of *weighted Hyperlink-Induced Topic Search* (*HITS*) *model* to learn the users' experience values and trust values of remote phone numbers. The notion of *Hyperlink-Induced Topic Search* (*HITS*) has been proposed by Kleinberg *et al.* [12] [13]. Basically, the HITS algorithm is a link analysis algorithm that rates the importance of web pages. The idea behind Hubs and Authorities is originally based on a particular insight into the creation of web pages when the Internet was originally forming. Meanwhile, certain web pages, known as hubs, do not actually provide authoritative information; rather, they provide a number of links which point users to other authoritative pages. In other words, a good hub represents a page that points to many other pages, where a good authority represents a page that is pointed by many hubs. In the context of telecommunication networks, the *hubs* can represent experienced users in the telecommunication network, who mostly accept normal phone calls and reject suspicious and fraudulent calls. Take Figure 2 as an example. Suppose the phone calls made by the "remote phone" are fraudulent, $user_1$ and $user_2$ directly hung up the phone, $user_3$ answers the call for a while, implying that $user_1$ and $user_2$ are more experienced than $user_3$. Similarly, a good *authority* in the communication network represents a phone number that was answered by many experienced users (i.e., hubs).

Accordingly, we can easily realize that the fraudulent phone call detection inherently could be formulated as a problem of learning hubs and authorities. Here, we treat the authority value of a remote phone number as "trust value" of itself and the hub value of a user as his "experience value". When an unknown remote phone number is stored in the telecommunication record, the trust value of the unknown remote phone number is contributed by experience values of the user who answered the phone. Although the HITS algorithm has widely been applied in many research areas, such as social network analysis [31], to our best knowledge, this is the first work that explores the HITS-based learning for fraud detection in telecommunications. To support our proposed weighted HITS model, we propose two kinds of directed bipartite graphs, *contact_book-remote phone number graph* (*CPG*) and *user-remote_phone_number graph* (*UPG*). To weight the edges of CPG and UPG, we extract features for each pair of a user and a remote phone number in two different but complementary aspects: 1) duration relatedness (DR) between a user and a remote phone number; and 2) frequency relatedness (FR) between a user and a remote phone number. In the on-line prediction module, based on discovered experience values of users, we evaluate the trust value of the unknown remote phone number to determine whether the unknown remote phone number is fraudulent. To our best knowledge, this is the first work on detecting fraudulent phone call by exploiting weighted HITS model. Through an experimental evaluation, we show that the proposed fraudulent phone calls detection delivers excellent performance.

The contributions of our research are six-fold.

- We propose the *FrauDetector* framework, a novel approach for fraudulent phone call detection. The problems and ideas in *FrauDetector* have not been explored previously in the research community.

- We develop the weighted HITS algorithm to learn experience values for individual users and trust values for remote phone numbers.
- We develop two kinds of structures to support the weighted HITS algorithm. One is contact_book-remote phone number graph (CPG) which reflects whether the remote phone number occurs in the user's contact book. The other is user-remote_phone_number graph (UPG) which reflects whether the remote phone number has had telecommunications with the user's.
- We propose two kinds of features to weight CPG and UPG. The proposed features can represent each pair of user and remote phone number in two different but complementary aspects: 1) duration relatedness (DR) between a user and a remote phone number; and 2) frequency relatedness (FR) between a user and a remote phone number.
- Based on the learned experience values for individual users and trust values for remote phone numbers, we propose a fraud detection strategy to detect fraudulent phone calls.
- We use a real dataset from *Whoscall* in a series of experiments to evaluate the performance of our proposal. The results show superior performance over other classifier-based fraud detection techniques in terms of precision, recall, F-measure and AUC.

The rest of this paper is organized as follows. We briefly review the related work in Section 2 and provide an overview of our graph-mining-based fraudulent phone call detection framework in Section 3. We detail the proposed trust value mining in section 4 and describe our fraudulent phone call detection in Section 5. Finally, we present the evaluation result of our empirical performance study in Section 6 and discuss our conclusions and future work in Section 7.

2. RELATED WORK

In this section, we briefly introduce three popular kinds of fraud detection, Telecommunications Fraud Detection, Insurance Fraud Detection, and Credit Card Detection. As mentioned earlier, most fraud detection techniques formulate the detection problem as a binary classification problem. Such classifier-based approach requires sufficient features to build classification model. Therefore, very limited numbers of literatures have address the telecommunications fraud detection because the phone users usually provide limited profile for applying a phone number such that we can not compute sufficient features only from telecommunication records.

2.1 Fraud Detection in Telecommunications

The histories of fraud detection at AT&T have been described in [1], one of the companies to address the fraud detection and discuss some techniques used to address them. Pal *et al.* [20] describe how data mining techniques help the telecommunication in many applications. Weatherford *et al.* [28] focus on neural networks, which utilize current user profiles that store long-term information to define normal patterns of use. After training the model, the behavior of fraud is probably different than the normal one. To detect frauders' phone numbers, Onderwater [18] adopts outlier detection techniques to identify unusual user profiles. Yusoff [32] proposed an approach using Gaussian mixed model (GMM), a probalistic model which is successful applying on speech recognition problems where the model can apply on it as well. Based on LDA, [17] detects the fraud using a threshold-type classification algorithm. Cahill *et al.* [7] builds upon the adaptive fraud detection framework [10] by using an event-driven approach of assigning fraud scores to detect fraud as it happens, and weighting recent mobile phone calls more heavily than earlier ones. The new framework [7] can also detect types of fraud using rules, in addition to detecting fraud in each individual account, from large databases. This framework has been applied to both wireless and wire line fraud detection systems with over two million customers. The adaptive fraud detection framework presents rule-learning fraud detectors based on account-specific thresholds that are automatically generated for profiling the fraud in an individual account. The system, based on the framework, has been applied by combining the most relevant rules, to uncover fraudulent usage that is added to the legitimate use of a mobile phone account [9] [10].

2.2 Fraud Detection in Other Field

Insurance Fraud Detection. Pocard [22] provide a survey of recent development on economic analysis of insurance fraud. Ngai *et al.* [16] proposed a series of classification schemes used for financial fraud detection, and 49 journals are analysed and 6 classification approaches are utilized. Ormerod *et al.* [19] proposed a Mass Detection Tool (MDT) for early detection of insurance fraud. The MDT uses a dynamic Bayesian Belief Network (BBN) of fraud indicators, whose weights are extracted from the rule generator's outcomes and claims handlers to keep up with the evolving fraudulent behaviors. The approach of fraudulent types evolves when the fraudulent behavior changes while capturing unexpected anomalies detected by claims handlers. Williams *et al.* demonstrated insurance and fraud applications [29], which have been applied, to health care for the Australian Health Insurance Commission. This methodology first performs clustering, *k-means* for cluster detect, and applies *C4.5* algorithm to build the model and evaluate the rules from the domain knowledge, statistical summaries and visualization. Williams [30] proposed an approach to hot spot data mining in which the measure of interestingness and the description of groups evolved according to a user guide the system towards significant discovery. Brockett *et al.* [4] utilized the similar methodology of self-organizing Map to uncover automobile bodily injury claims fraud. [11] used the method of two-class neural network classification for medical practice developed at the Health Insurance Commission. Like the hot spot methodology, this approach can be applied on instances of the Australian Health Insurance Commission health practitioners' profiles. Šubelj *et al.* [26] described an expert system for insurance fraudsters and the experiments shows the better effectiveness.

Credit Card Detection. Credit card fraud detection is of great importance to financial institutions. Raj *et al.* [24] provided a complete survey of techniques for credit card fraud detection and evaluate the effectiveness of each approach. Bhattacharyya *et al.* [2] focus on two advanced approaches: 1) support vector machine 2) random forests 3) regression to achieve better performance on credit fraud detection. Maes *et al.* [14] applied two machine learning techniques: Artificial Neural Network (ANN) and Bayesian Belief Networks (BBN) to the problem and show the results on the real-world financial dataset. Comparative results show that BBNs were more accurate and much faster while BBNs are slower when applied to new instances. Chan *et al.* [6] proposed a method of combining multiple fraud detectors under cost model to evaluate C4.5, CART, Ripper and NB classification models. The result shows that partitioning the large dataset into small subsets to utilizing different classification algorithms used for detecting fraud improves the cost of savings by using stacking to combine multiple models and the method was applied to credit card transactions from two major US banks, Chase Bank and First Union Bank. Weatherford [28] describes the FairIsaac, which produces software for detecting credit card fraud, and utilizing the BP neural network to detect fraudulent activities.

According to above-mentioned three kinds of fraud, we can observe that all existing work formulate the fraud detection as a binary classification problem and try to extract features from users' profile. This strategy is reasonable for detecting insurance and

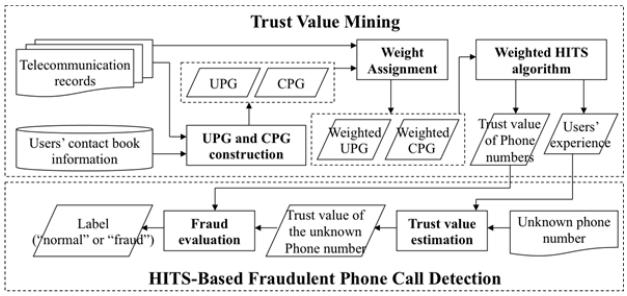


Figure 3. The *FrauDetector* framework for fraudulent phone call detection.

credit card fraud because people should provide details of their profile when they apply insurance or credit card. However, the telecommunication is much different from insurance or credit card. We only can collect very simple profile, which is insufficient to build a precise classifier. As the result, in the next section, we propose a novel fraudulent phone call framework, based only on telecommunication records.

3. OVERVIEW OF OUR PROPOSED *FrauDetector* FRAMEWORK

With the notion of weighted HITS algorithm, we propose a novel fraudulent phone call framework, namely, *FrauDetector,* based only on telecommunication records. The *FrauDetector* framework consists of 1) a trust value mining module, and 2) a HITS-based fraudulent phone call detection module. Figure 3 shows the framework and its flow of data processing. The idea is to explore the telecommunication behaviors of mobile users, captured in telecommunication records, to reveal trust value of a remote phone number. As shown in Figure 3, the trust value mining module is an offline module which includes three steps. The first step, called *UPG and CPG construction*, transforms users' telecommunication as directed graph. The second step, called *weights assignment*, extracts features to weight the edges of CPG and UPG. Meanwhile, two different but complementary aspects, 1) duration relatedness (DR) between user and remote phone number and 2) frequency relatedness (FR) for each pair of user and remote phone number are proposed. The third step, called *trust value discovery*, perform weighted HITS algorithm on the weighted CPG and UPG to learn experience value for individual users and trust value for remote phone numbers. The HITS-based fraudulent phone call detection module is an on-line module in which we propose a scoring function to evaluate the probability for a remote phone number to be fraudulent. We then details the value mining module in next section and the HITS-based fraudulent phone call detection module in the section 5.

4. TRUST VALUE MINING

To learn the trust value for each remote phone number and the experience value for each user without users' profile, we perform a weighted HITS algorithm based on the telecommunication records. To support weighted HITS algorithm, we proposed two kinds of directed graph, CPG and UPG, to represent behavior of users' telecommunications. To weight our proposed CPG and UPG, we propose two different but complementary aspects, 1) duration relatedness (DR) between user and remote phone number and 2) frequency relatedness (FR) for each pair of user and remote phone number.

4.1 CPG and UPG Construction

Since original idea of HITS is that a good hub represented a web page that pointed to many other pages, and a good authority represented a web page that was linked by many different hubs. Therefore, constructing a directed graph to represent the relation between users and remote phone numbers is required. Intuitively, it inherently represents the relation between a user and a remote phone number to see if the user has a call from/to the remote phone number. Based on this idea, we propose a directed bipartite graph, *user-remote_phone_number graph* (*UPG*), of which edges represent call directions.

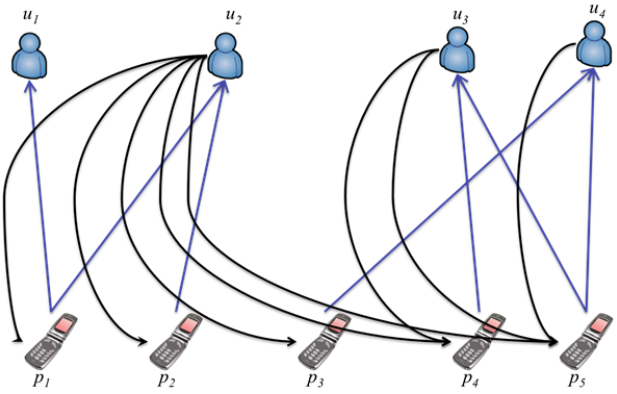


Figure 4. An example of user-remote_phone_number graph (UPG).

Definition 1. User-remote_phone_number graph (UPG). Given a telecommunication record set T, a user set U and a remote phone number set P, the *User-remote_phone_number graph* is a directed bi-partite graph $UPG = (V, E)$ with the vertex set $V = U \cup P$ the edge set $E = \{(u, p) \mid u \in U, p \in P$, and $\exists$ a call from u to p in $T\} \cup \{(p, u) \mid u \in U, p \in P$, and $\exists$ a call from p to u in $T\}$.

Example 1. Table 1 shows an example of telecommunication record set. The vertex set should be $\{u_1, u_2, u_3, u_4, p_1, p_2, p_3, p_4, p_5\}$, and the edge set should be $\{(p_1, u_1), (p_1, u_2), (p_2, u_2), (u_2, p_1), (u_2, p_2), (u_2, p_5), (u_2, p_3), (u_2, p_4), (p_4, u_3), (u_3, p_5), (u_3, p_4), (p_5, u_3), (p_3, u_4), (p_5, u_4), (u_4, p_5)\}$. The illustration of user-remote_phone_number graph is shown in Figure 4.

Table 1. An example of telecommunication record set.

Users	*Remote Phone Numbers*	*Receiving /Dialing Time*	*Starting Time*	*Ending Time*	*Call Direction*
u_1	p_1	17:01:48	17:02:08	17:06:08	in
u_2	p_1	17:07:48	-	-	in
	p_2	17:02:08	17:02:28	17:04:08	in
	p_1	18:05:21	18:05:30	18:06:01	out
	p_2	18:07:08	18:07:17	18:10:31	out
	P_5	19:00:08	-	-	out
	p_3	19:10:08	-	-	out
	p_4	19:12:32	19:12:38	19:22:01	out
	p_5	19:22:55	19:23:01	19:23:58	in
u_3	p_4	16:22:01	16:22:21	16:25:40	in
	p_5	16:25:41	-	-	out
	p_4	16:25:45	16:26:02	16:30:10	out
	p_5	16:32:01	16:32:13	16:33:50	in
u_4	p_3	17:09:01	17:09:08	17:12:02	in
	p_5	17:12:18	17:12:33	17:30:08	in
	p_5	19:01:05	19:01:16	19:02:28	out
	p_5	20:02:17	20:02:28	20:40:08	in

Actually, the telecommunication activities can be classified into two categories. One is telecommunication with someone who is well known, and the other is telecommunication with stranger. Accordingly, if the users provide the contact book information for our model building, we could utilize this information to construct the directed graph for HITS algorithm. Based on this idea, we formally define a directed graph, *contact_book-remote phone number graph* (*CPG*), as follows:

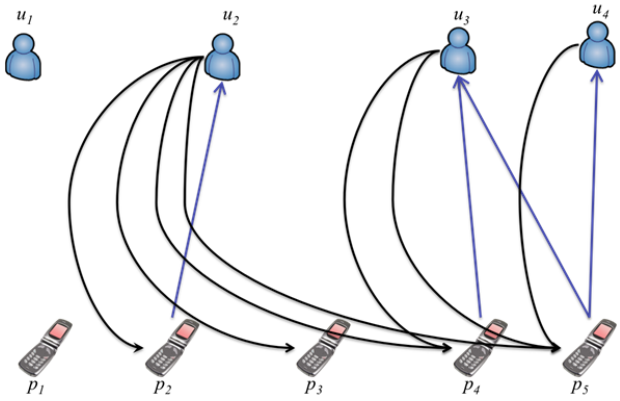


Figure 5. An example of contact_book-remote phone number graph (CPG).

Definition 2. Contact book phone number (CP). Given a user u, the *Contact book phone number* of u, formally denoted as $CP(u)$, is set list of phone numbers that are recorded in u's contact book.

Example 2. Table 2 shows an example of contact book of each user. The contact book phone number of u_2 should be $\{p_2, p_3, p_4, p_5\}$, i.e., $CP(u_2) = \{p_2, p_3, p_4, p_5\}$

Table 2. An example of contact book of each user.

Users	*Contact Book Phone Numbers*
u_1	p_2, p_4
u_2	p_2, p_3, p_4, p_5
u_3	p_4, p_5
u_4	p_2, p_5

Definition 3. Contact_book-remote phone number graph (CPG). Given a telecommunication record set T, a user set U and a remote phone number set P, the *Contact_book-remote phone number graph* is a directed bi-partite graph $CPG = (V, E)$ with the vertex set $V = U \cup P$ the edge set $E = \{(u, p) \mid u \in U, p \in CP(u),$ and $\exists$ a call from u to p in $T\} \cup \{(p, u) \mid u \in U, p \in CP(u),$ and $\exists$ a call from p to u in $T\}$.

Example 3. Table 1 shows an example of telecommunication record set, and Table 2 shows an example of contact book of each user. The vertex set should be $\{u_1, u_2, u_3, u_4, p_1, p_2, p_3, p_4, p_5\}$, and the edge set should be $\{(p_2, u_2), (u_2, p_2), (u_2, p_5), (u_2, p_3), (u_2, p_4), (p_4, u_3), (u_3, p_5), (u_3, p_4), (p_5, u_3), (p_5, u_4), (u_4, p_5)\}$. The illustration of user-remote_phone_number graph is shown as Figure 5.

Here, we can observe that CPG can precisely represent relation between users and remote phone numbers. However, some remote phone numbers would easily be isolated in a CPG, such as p_1 in Figure 5. This phenomenon might lead the HITS algorithm does not work well. As the result, both UPG and CPG are potentially useful for leaning trust values of remote phone numbers.

4.2 Weights Assignment

Based on our observation, users' telecommunication behaviors could be categorized into two types. One is to make phone call to someone familiar, and the other is to make phone call to someone who is a colleague. Usually, people would chatter to someone familiar for a while and frequently talk to colleagues. That means user usually has call to/from normal remote phone number either frequently or for a while. Therefore, to weight a specific user-remote_phone_number pair, we explore the telecommunication behaviors in two different but complementary aspects: 1) duration relatedness (DR) between user and remote phone number; and 2) frequency relatedness (FR) between user and remote phone number. Before introducing the duration relatedness features, we first describe the formal definitions for illustrating the duration relatedness features:

Definition 3. Conditional Telecommunication Record Set (CT). Given a telecommunication record set T, a user u, a remote phone number p and call direction d, the *Conditional Telecommunication Record Set*, denoted as $CT(T, u, p, d)$, is a subset of T in which user is u, remote phone number is p, and call direction is d.

Example 3. Consider Table 1 as an example of telecommunication record set T, the conditional telecommunication record set $CT(T, u_4, p_5, in)$ is shown as Table 3.

Table 3. An example of conditional telecommunication record set.

Users	*Remote Phone Numbers*	*Receiving /Dialing Time*	*Starting Time*	*Ending Time*	*Call Direction*
u_4	p_5	17:12:18	17:12:33	17:30:08	in
	p_5	20:02:17	20:02:28	20:40:08	in

4.2.1 Duration Relatedness

Our goal is to extract duration relatedness features for pairs of users and remote phone numbers. Intuitively, users have different telecommunication duration for different people due to the nature of relationship between the users and the remote phone numbers. As a result, different call durations, naturally formed in aggregated behaviors of users to various kinds of remote phone numbers, are embedded in the users' telecommunication activities which are logged in telecommunication records. In a telecommunication record, the most important information is remote phone number and duration, besides the user himself. In the following, we propose to extract two duration features to depict relation between users and remote phone numbers as below.

- **Total Call Duration (TCD).** Intuitively, the user is likely to share some interests with his or her friends or someone familiar. Thus, if a user spends much time for a phone call to/from to a remote phone number, he may contribute his trust to the remote phone number. Accordingly, aggregating user's call duration of a remote phone number can be used to infer the probability that a user trusts that remote phone number. Formally, given a telecommunication record set T, a user u and a remote phone number p, total call duration from u to p and that from p to u are respectively formulated as:

$$\begin{cases} TCD(u \to p \mid T) = \sum_{\forall r \in CT(T,\, u,\, p,\, out)} (r.ending\ time - r.starting\ time) \\ TCD(p \to u \mid T) = \sum_{\forall r \in CT(T,\, u,\, p,\, in)} (r.ending\ time - r.starting\ time) \end{cases} \quad (2)$$

- **Average Call Duration (TCD).** It is trivial that the total call duration includes two aspects, duration of each phone call and phone call frequency. That might weaken the significance of duration relatedness. For example, if a user makes short duration calls to a remote phone number many times, the total call duration might be pretty long. Accordingly, averaging user's call duration of a remote phone number can adjust this bias evaluation. Formally, given a telecommunication record set T, a user u and a remote phone number p, average call duration from u to p and that from p to u are respectively formulated as:

$$\begin{cases} ACD(u \to p \mid T) = \dfrac{TCD(u \to p \mid T)}{CT(T,\, u,\, p,\, out)} \\ ACD(p \to u \mid T) = \dfrac{TCD(p \to u \mid T)}{CT(T,\, u,\, p,\, in)} \end{cases} \quad (3)$$

Take Table 1 as telecommunication record set T. The total call duration from p_5 to u_4 is (17:30:08 − 17:12:33) + (20:40:08 − 20:02:28) = (00:17:35) + (00:37:40) = 3315 (sec.), and the average call duration from p_5 to u_4 is 3315/2 = 1657.5 (sec.).

4.2.2 Frequency Relatedness

As mentioned earlier, user usually has call to/from normal remote phone number either frequently or for a while. The above-proposed duration relatedness features have completely represented the aspect, having a call for a while. In this subsection, our goal is to extract frequency relatedness feature for pairs of users and remote phone numbers. Intuitively, users have different telecommunication frequencies for different people due to the needs of communication between the users and the remote phone numbers. Thus, we also can realize that different call frequency are embedded in the users' telecommunication activities that can be utilized to distinguish fraudulent remote phone numbers from normal remote phone numbers. In the following, we propose to extract a call frequency feature to depict relation between users and remote phone numbers.

Based on the observation, each person's call frequency is different from other person's call frequency. Some active person might frequently make phone calls to other people. Therefore, we should not directly utilize call frequency as the feature to represent relation between users and remote phone numbers. Accordingly, we propose a Normalized Call Frequency (NCF) to capture relation between users and remote phone numbers. Formally, given a telecommunication record set T, a user u and a remote phone number p, normalized call frequency from u to p and that from p to u are respectively formulated as:

$$\begin{cases} NCF(u \to p \mid T) = \dfrac{CT(T,\ u,\ p,\ out)}{\max\limits_{p^* \in P}\{CT(T,\ u,\ p^*,\ out)\}} \\ NCF(p \to u \mid T) = \dfrac{CT(T,\ u,\ p,\ in)}{\max\limits_{p^* \in P}\{CT(T,\ u,\ p^*,\ in)\}} \end{cases} \quad (4)$$

Take Table 1 as telecommunication record set T. The normalized call frequency from p_5 to u_4 is 2/4.

4.3 Trust Value Learning

Based on the weighted CPG and UPG, we can perform a weighted HITS algorithm to learn trust value for each remote phone number and experience value for each user. Given m users and n remote phone numbers, we build an $m \times n$ weighted adjacency matrix X and an $n \times m$ weighted adjacency matrix Y for the UPG or CPG. Formally, $X = [v_{ij}]$ and $Y = [w_{ji}]$,$0 \le i <$m; $0 \le j <$ n, where v_{ij} represents the weight of the edge linking the from ith user to the jth remote phone number, and v_{ij} represents the weight of the edge linking from the jth remote phone number to the ith user. Formally, the random walk model applied to the UPG or the CPG can be described as follows:

$$\begin{aligned} u_{user}^{k+1} &= u_{user}^{k} + Y^T \times \frac{1}{norm(u_{phone}^{k})} u_{phone}^{k} \\ u_{phone}^{k+1} &= u_{phone}^{k} + X^T \times \frac{1}{norm(u_{user}^{k+1})} u_{user}^{k+1} \end{aligned} \quad (5)$$

where k indicates the number of iterations; Y indicates the phone-to-user stochastic matrix (computed by UPG or CPG); X indicates the user-to-phone stochastic matrix (computed by UPG or CPG); and norm() indicates the "Euclidean norm[4] of a vector," which represents the length of a trust value vector (respectively experience value vector).

As we described earlier, there are three types of weight,

[4] http://en.wikipedia.org/wiki/Euclidean_distance

```
Input:  Weighted directed bipartite graph UPG (or CPG)
Output: Trust value vector of remote phone number τ
        Experience value vector of user u
1   X ← transform UPG to user-by-phone matrix
2   Y ← transform UPG to phone-by-user matrix
3   for each user i //initial experience value
4     u[i] ← 0.
5   End for
6   for each remote phone number j //initial trust value
7     if jth remote phone number is fraudulent
8       τ[j]← -1
9     otherwise
10      τ[j]← 1
11    End if
12  End for
13  for iteration < N
14    τ ← τ × 1/norm(τ) // normalize by norm
15    u ← u + X^T×τ
16    u ← u × 1/norm(u) // normalize by norm
17    τ ← τ + Y^T×u
18  End for
19  Return τ and u
```

Figure 6. Weighted HITS algorithm.

normalized call frequency (denoted as FR), total call duration (denoted as DR_T) and average call duration (denoted as DR_A), and there are two kinds of graph, CPG and UPG. Therefore, there are six kinds of weighted graph (3×2 = 6). Based on one of the six kinds of weighted graph, we perform the weighted HITS algorithm, as shown in Figure 6, to calculate the trust value for each remote phone number and experience value for each user. First, the weighted adjacency matrix X and Y are generated from input UPG or CPG (see lines 1 through 2 of Figure 6). Then, the initial state of each user and each remote phone is obtained (see lines 3 through 12 of Figure 6). For each iteration, we aggregate trust value (τ) into experience value (u) and experience value (u) into trust value (τ) (see lines 13 through 18 of Figure 6). Finally, we can obtain the trust value (τ) and experience value (u) (see lines 9 through 11 of Figure 6). As the result, we can use the learned trust value for each remote phone number and experience value for each user to build the fraudulent phone call.

5. HITS-BASED FRAUDULENT REMOTE PHONE NUMBER DETECTION

After learning trust value and experience value, we must estimate the trust value of unknown remote phone number based on the learned trust value and experience value. Intuitively, we can insert the unknown remote phone number in UPG or CPG and use the learned users' experience values to estimate the trust value of the unknown remote phone number. Since frauders make phone call first before they receive any phone call, there is no in-direct edge point to the vertex, which represents the unknown remote phone number in UPG or CPG. However, the fraudulent phone call detection should detect the fraud as soon as possible. Thus, we assume one of user might call back. Accordingly, we directly use the weighted average experience value to estimate the possible

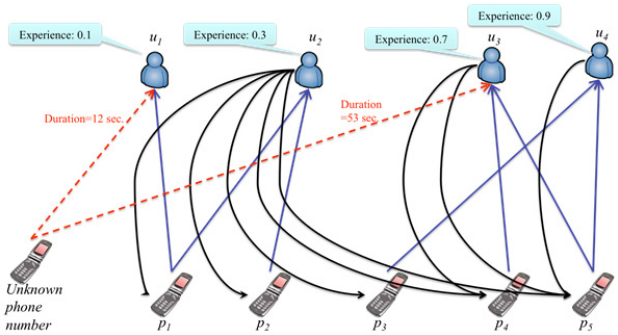


Figure 7. An example of estimation of trust value for an unknown remote phone number.

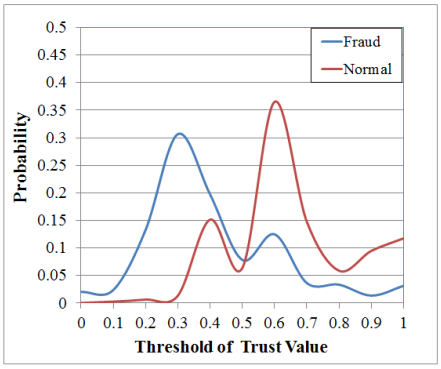


Figure 8. Distributions of learned trust values.

trust value of the unknown remote phone number. Given an unknown phone number p^*, a telecommunication record set T and user set U, the estimated trust value of p^* is formally defined as follows:

$$\frac{\sum_{u\in U}\left[\left(w(p^*\to u\,|\,T)\times u.experience\right)+\left(w(u\to p^*\,|\,T)\times u.experience\right)\right]}{\sum_{u\in U}\left(w(p^*\to u\,|\,T)+w(u\to p^*\,|\,T)\right)}, \quad (6)$$

where $w()$ is the weighting function utilized for UPG or CPG. Take Figure 7 as an example. The unknown phone number has have a phone call to the user u_1 and u_2, and the graph is weight total call duration. The estimated trust value of the unknown phone number should be

$$\frac{12\times 0.1+53\times 0.7}{12+53}=0.59\,.$$

After estimating possible trust value of the unknown remote phone number, we then determine whether the unknown remote phone number is fraudulent according to the possible trust value. Accordingly, a suitable threshold is required for evaluate whether the possible trust value provide significant evidence to prove the unknown remote phone number is fraudulent. Intuitively, we can use the distribution of learned trust value of phone numbers, utilized for HITS, to compute the threshold. Figure 8 shows the distributions of learned trust values of normal remote phone numbers and that of fraudulent remote phone numbers, learned from *Whoscall* telecommunication record set. We can observe that Although the distribution of learned trust values of normal remote phone numbers is significantly different from that of fraudulent remote phone numbers, there is a quite portion of fraudulent remote phone numbers get high trust value. To avoid the effect of these outliers, we utilize the k percentile[5] of distribution of learned trust values of fraudulent remote phone numbers as the threshold. Based on the threshold, we classify the unknown remote phone number into the "fraud" category if the possible trust value of it is lower than the threshold.

6. EXPERIMENTAL EVALUATIONS

In this section, we present the results from a series of experiments to evaluate the performance of UPOI-Walk using Gowalla and EveryTrail datasets. All the experiments are implemented in Java JDK 1.6 on an Intel Xeon CPU W3520 2.67 GHz machine with 48 GB of memory running Microsoft Windows 7. We first describe the preparation of the datasets, and then introduce the evaluation methodology. Finally, we present and discuss our experimental results.

6.1 Dataset Description

We used the data provided by *Whoscall*, which is a powerfulsmartphone application that can detect the incoming call which is tell-marketing, harassment, call centers from the known information such as Yellow Pages, remote phone numbers on the network, etc. The dataset collected for a month in August 2014 in Taiwan (see Table 4) consists of 218,060,640 call logs, where there are 1,324,217 *Whoscall* users and 14,573,937 remote phone numbers. Each call record logs time, user id, remote phone number, duration of calls, ringtone type, whether the incoming call is in contact book, the country code of the incoming call, whether the call is missed, etc. In Table 5, we show the normal remote phone number and fraudulent remote number in the training data and testing data, where the fraudulent remote phone numbers are marked by the *Whoscall* users. According to the call records, we extract the total volumes of each remote phone number, the call duration of each remote phone number, whether the incoming call is in the user's contact books, which we take as the attributes used for classification. Note that, we focus on the fraudulent remote phone numbers, but in our dataset, most of the remote phone numbers are normal. Therefore, the dataset is an imbalance-label dataset, which the proportion of fraudulent remote phone numbers and normal remote phone numbers is about 1:400. Thus, in our experiment, with baseline methods, we solve the imbalance-label problem with our proposed method mentioned before. Otherwise, we came up with the approach to fraudulent detection using the web search engine algorithm Hyperlink Induced Topic Search (HITS), which is compared with in the later section.

Table 4. Basic statistics of Whoscall dataset.

	# of Calls in August,2014, Taiwan
# of call records	218,060,640
# of unique numbers	14,573,937
# of unique users	1,324,217

Table 5. Distributions of training and testing dataset.

	Training Data	***Testing Data***
# of fraudulent number	26,807	10,030
# of normal number	10,905,721	3,631,379

6.2 Performance Metrics

To compare with the performance of each approach, we can evaluate the effectiveness by the following standards that we focus on the fraudulent remote phone numbers: (1) Precision: the number of correct positive results divided by the number of all positive results (i.e., of all the classified fraudulent remote phone numbers, the rate of truely classified as fraudulent remote phone numbers). (2) Recall: the number of correct positive results divided by the number of positive results that should have been returned (i.e., of all the real fraudulent remote phone numbers, the rate of truely classified as fraudulent remote phone numbers). (3) F1-Measure: the F1-Measure can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. (4) ROC curve: A graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the *True Positive* rate against the *False Positive* rate. When making decision, ROC analysis can not be cost impact, and give the objective and neutral advices. (5) AUC: the area under the ROC curve, representing to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In the next section, we will evaluate the performance using the standards.

6.3 Evaluation on Baseline Approaches

Due to the imbalance-label problem, first we use chi-square to select the features that are useful; second, take each remote phone number as a raw data, cluster the remote phone numbers using *k-means* into groups which we prune the normal remote phone numbers that are prone to fraudulent remote phone numbers; third,

[5] http://en.wikipedia.org/wiki/Percentile

use sampling approaches [5] (e.g., *over-sampling, down-sampling, SMOTE*) to eliminate the imbalance. After these steps, we use five existing approaches as baseline approaches that are listed below.

- **Naïve Bayes.** Naïve Bayes classifiers are a family of probabilistic classifier based on Bayes' theorem with naïve independence assumptions between the features [17]. In some applications, the effectiveness of classification is better than ANNs and decision trees. Naïve Bayes classifier is sometimes used for large dataset and it sometimes obtains predictive and effective classified results [15] [25].
- **Random Forests.** Random Forests first proposed by Amit and Geman in 1997, after that it was completely integrated by Leo Breiman and Adele Cutler [3]. This algorithm combines the thought of "Bootstrap aggregating" and the method of "random subspace" to construct a multitude of decision trees at training time and output the class with the most popular class.
- **C4.5 decision tree.** An algorithm used to generate a decision tree by Ross Quinlan [23]. C4.5 is an extension of Quinlan's earlier ID3 algorithm, and the decision trees that C4.5 create is not only used for classification, but for statistical classifier.
- **Support Vector Machines (SVM).** SVMs are supervised learning models with associated learning algorithms which analyze data and recognize patterns, and used for classification and regression analysis. SVMs are classification algorithms that can apply in linear data and nonlinear data, and can transform the initial data into higher dimension for more accurate classification [8].
- **Artificial Neural Networks(ANNs).** ANNs are one of statistical algorithms that inspired by biological neural networks [27]. In most cases, ANNs are a self-adaptive system that can change infer-structure based on the input data, and ANNs are a nonlinear statistical tool, which usually building models for complicated relationship between input data and output data. Like other machine learning methodologies, ANNs have been solved a wide variety of tasks including computer vision and speech recognition.

6.4 Internal Experimental Results

Table 6. Abbreviation of our proposed weighted HITS.

Directed Graph	*Weighting fuction*	*Abbreviation*
UPG	none	*UPG*
	TCD (see section 4.2.1)	*UPG_TCD*
	ACD(see section 4.2.1)	*UPG_ACD*
	NCF(see section 4.2.2)	*UPG_FR*
CPG	none	*CPG*
	TCD (see section 4.2.1)	*CPG_TCD*
	ACD(see section 4.2.1)	*CPG_ACD*
	NCF(see section 4.2.2)	*CPG_FR*

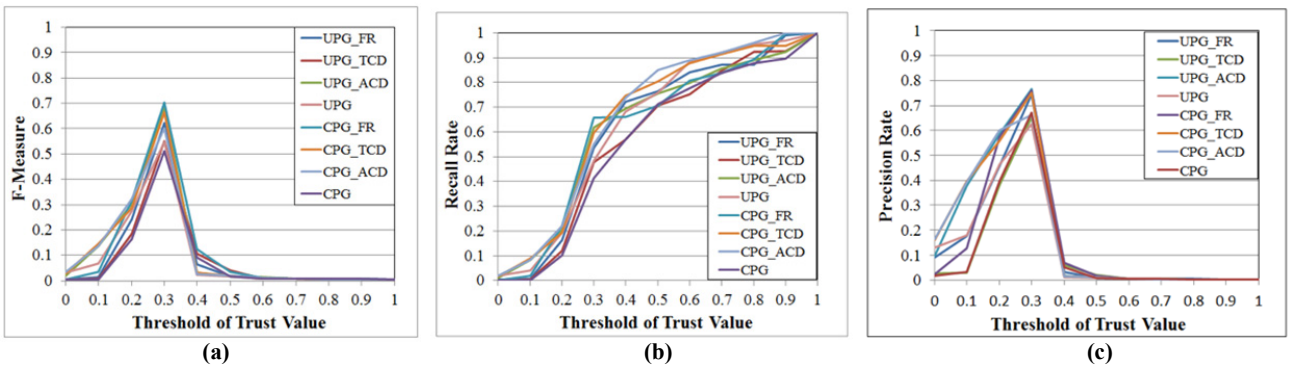
As mentioned in the Section 4, we have various weighted directed graphs, which can capture some aspects of users' telecommunication activities. In this section, we compare them and seek the best parameter settings. To conveniently present our experimental results, we give the abbreviations of our proposed HIST-based model under various weighted directed graph as shown in Table 6.

Figure 9 (a) indicates the precision of Graph-mining-based models with different weighting strategies under various threshold setting. Here, the x-axis represents *k* percentile of distribution of learned trust values of the fraudulent remote phone number. In other words, 0.2 means we set threshold as the learned trust value of the fraudulent remote phone number, which is higher than the learned trust values of 20% the fraudulent remote phone numbers. As shown in Figure 9 (a), we can recognize that the precision of most of the lines increase stably and reach the highest peak when threshold is set as 30 percentile. Besides, the precision decrease rapidly down to zero when threshold is set as greater than 0.4. The reason is that a large volumes of normal remote phone numbers are determined as fraud (False Alarm) when threshold is set as greater than 40 percentile so that the false alarm occupy in majority. Furthermore, we can figure out that the UPG_ACD has the highest precision 0.765 when threshold is set as 30 percentile.

Figure 9 (b) shows the recall of Graph-mining-based models with different weighting strategies under various threshold setting. Unlike the precision shown in Figure 9 (a), the recall increases rapidly until threshold is 30 percentile and increases smoothly when threshold is set as greater than 30 percentile. The reason for the former phenomenon is a large number of fraudulent remote phone numbers which are determined as fraudulent remote phone numbers (True Positive), and the latter phenomenon is the number of decreasing increased rate. Besides, we can see that the trends of the all models are quite similar, and when threshold is set as greater than 30 percentile, the CPG_ACD performs higher recall than all the other features, and the recall of CPG shows lowest recall.

Figure 9 (c) indicates the F1-Measure of Graph-mining-based models with different weighting strategies under various threshold setting. Taking into account of precision and recall, F1-Measure provides more comprehensive effectiveness evaluation. As shown in Figure 9 (c), the highest peak occurs when threshold is set as about 30 percentile, and the F1-Measure decrease down to zero when threshold is set as greater than 30 percentile. It is reasonable because the precision is very close to zero when threshold is set as greater than 30 percentile. We also can observe that the F1-Measure of all models show the highest peak when threshold is set as 30 percentile. Meanwhile, the F1-Measure of the CPG_FR is up to 0.71 and the F1-Measure of the CPG is 0.51, which means that weighting strategy has played an important role on the weighted HITS model.

Figure 10 shows the ROC curve of Graph-mining-based models with different weighting strategies. As shown in Figure 10,



(a) **(b)** **(c)**

Figure 9. Precision, Recall and F1-Measure of weighted HITS models.

the true positive rates of all models are increased rapidly while the fale positive rates of all models are low. It means that all HITS-based approaches we propose have excellent ability for dealing with imbalance-label problem. We also can observe that the curve of UPG lies below than other models. This phenomenon shows that all weighting functions are useful. Figure 11 shows the AUC of Graph-mining-based models with different weighting strategies. We can see that the AUC of all Graph-mining-based models with different weighting strategies are greater than 0.8 except for the CPG, of which is 0.791, and the UPG, of which is 0.746.

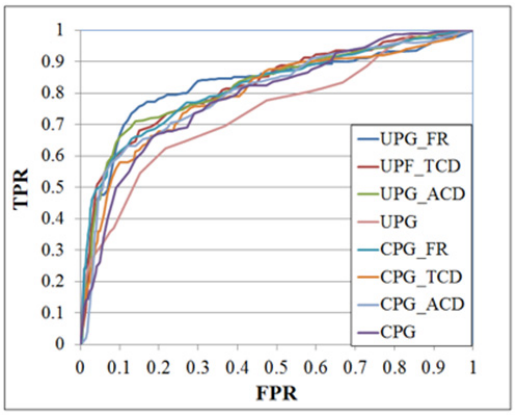


Figure 10. ROC curve of weighted HITS models.

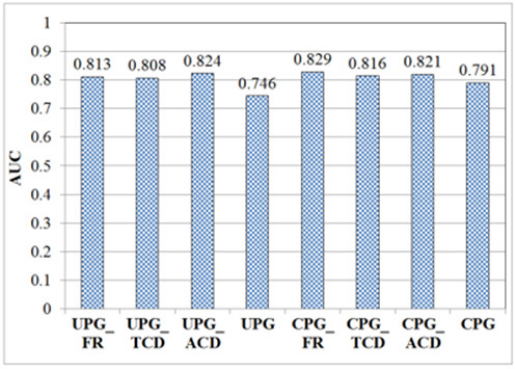


Figure 11. AUC of weighted HITS models.

In this internal experiment, we can conclude that 1) CPG_FR has the optimal effectiveness and 2) all models have best effectiveness when threshold is set as 30 percentile. As the result, we select the CPG_FR with threshold is set as 30 percentile to represent our proposed method in following experiments.

6.5 Comparison with Classifier-based Fraud Detection

Figure 12 indicates the ROC curve of comparison of our proposed HITS-based approach with classifier-based fraud detection approaches. As shown in Figure 12, our proposed HITS-based approach outperforms than the classifier-based fraud detection approaches. Besides, we can observe that SVM is significantly worse than other classifier-based fraud detection approaches. The reason is that SVM requires sufficient features to achieve quite high effectiveness, but our proposed HITS-based approach does not. Figure 13 indicates the AUC of comparison with classifier-based fraud detection approaches. As t shown in Figure 13, the AUC of our proposed HITS-based approach (CPG_FR) is 0.829, and the highest AUC of the classifier-based fraud detection approach (NaiveBayes) is 0.656. Compared with the best of classifier-based fraud detection approaches, our proposed HITS-based approach achieve 26.4% improvement rate (i.e., (0.829 − 0.656)/ 0.656).

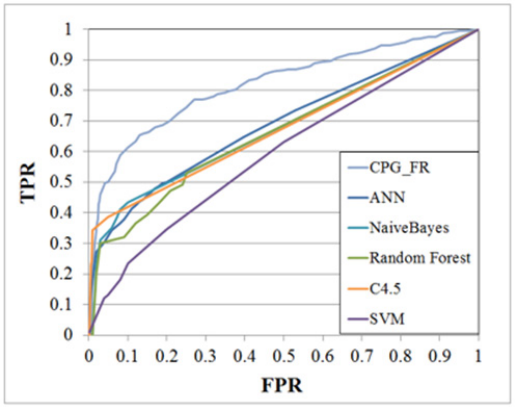


Figure 12. ROC curve of comparison with classifier-based fraud detection approaches.

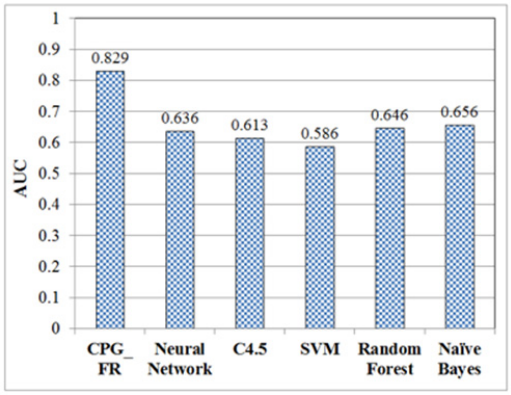


Figure 13. AUC of comparison with classifier-based fraud detection approaches.

6.6 Discussions

Some lessons are learned from the above experimental results:

1. The experiments results show that our proposed approaches outperform the classifier-based methods significantly. One observation is that most classifiers require quite sufficient features which might not be provided in this application. Another observation leads to that HITS algorithm is a good solution for this application since the data which could be modeled as a network when the information shown in the data is very limited.
2. CPG_FR shows the best performance in our proposed approaches. It means that contact book and call frequency are the two most important kinds of information for identifying fraudulent phone calls.
3. In *Whoscall* dataset, the 30 percentile of trust values of fraudulent remote phone numbers is a good criteria for identifying fraudulent phone calls. It means there still are many fraudulent remote phone numbers that can get high trust value. Accordingly, although our proposed HITS-based approach is not exactly perfect, the weakness can easily be fixed through a series internal experiments.

7. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel framework named *HITS-based fraudulent phone call detection (FrauDetector)* for detecting fraudulent phone calls by mining users' telecommunication records. We also tackled the problem of mining trust value from telecommunication activity, which is a crucial prerequisite for fraud detection. The core task of fraudulent phone call detection is conveniently transformed to the problem of trust value prediction. We evaluated the trust value of each phone number by training a weighted HITS model. In the proposed *FrauDetector,* we consider two kinds of

telecommunication activities to build the directed graphs, CPG and UPG. Furthermore, we have explored 1) duration relatedness (DR) and 2) frequency relatedness (FR) by exploiting telecommunication to extract descriptive features to weight the CPG and UPG. To the best of our knowledge, this is the first work on fraudulent phone call detection that considers mining trust value of remote phone number by performing weighted HITS algorithm on telecommunication records. Through a series of experiments using a real dataset, provided by *Whoscall*, we have validated the proposed *FrauDetector* and shown that it has excellent performance compared with state-of-the-art classifier-based fraud detection methods under various conditions. In future work, we plan to design more sophisticated methods to enhance the quality of our proposed *FrauDetector* for various urban computing service applications.

REFERENCES

[1] R. A. Becker, C. Volinsky, and A. R. Wilks. Fraud Detection in Telecommunications: History and Lessons Learnd. *Technometrics*, vol. 52, No. 1, pp. 20–33, February 2010.

[2] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. Westland. Data mining for credit card fraud: a comparative study. *Decision Support Systems*, 50 (3), pp. 602–613, 2011

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] P. Brockett, X. Xia and R. Derrig. Using Kohonen's Self Organising Feature Map to Uncover Automobile Bodily Injury Claims Fraud. *Journal of Risk and Insurance*, USA, 1998

[5] J. Burez, D. Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 2009

[6] P. Chan, W. Fan, A. Prodromidis and S. Stolfo. Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems,* **14,** pp67-74, 1999.

[7] M. Cahill, D. Lambert, J. Pinheiro and D. Sun. Detecting Fraud In The Real World. *The Handbook of Massive Data Sets*, Kluwer, pp911-930, 2002.

[8] C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*. 20, 273–297, 1995.

[9] T. Fawcett and F. Provost. Combining Data Mining and Machine Learning for Effective User Profiling. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Oregon, USA, 1996.

[10] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery,* Kluwer, 1**,** pp291-316, 1997

[11] H. He, J. Wang, W. Graco and S. Hawkins. Application of Neural Networks to Detection of Medical Fraud. *Expert Systems with Applications*, 13, pp329-336, 1997.

[12] J. Kleinberg. Hubs, Authorities, and Communities. *Cornell University*, December 1999.

[13] J. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM Computing Surveys*, 46 (5): 604–632, 1999.

[14] S. Maes, K. Tuyls, B. Vanschoenwinkel and B. Manderick. Credit Card Fraud Detection Using Bayesian and Neural Networks. In *Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies,* Havana, Cuba, 2002.

[15] K. P. Murphy. Naive Bayes classifiers. *http://www.cs.ubc.ca/murphyk/Teaching/CS340-Fall06/reading/NB.pdf*

[16] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, X. Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50 (3) (2011), pp. 559 – 569

[17] D. Olszewski. A probabilistic approach to fraud detection in telecommunications. *Knowledge Based Systems,* Volume 26, pp.246-258, 2012

[18] M. Onderwater. Detecting unusual user profiles with outlier detection techniques. VU University Amsterdam, 2010.

[19] T. Ormerod, N. Morley, L. Ball, C. Langley and C. Spenser. Using Ethnography To Design a Mass Detection Tool (MDT) For The Early Discovery of Insurance Fraud. In *Proceedings of ACM CHI Conference,* Florida, USA, 2003.

[20] S. H. Pal, J. N. Patel. Data Mining in Telecommunication: A Review. *International Journal of Innovative Research in Technology*, 2014

[21] C. Phua, D. Alahakoon and V. Lee. Minority Report in Fraud Detection: Classification of Skewed Data. *ACM SIGKDD Explorations Newsletter,* vol. 6, no. 1, pp. 50-59, 2004.

[22] P. Picard. Economic analysis of insurance fraud. G. Dionne (Ed.), Handbook of Insurance, Kluwer Academic Press, Boston (2000), pp. 315 – 362

[23] J. R. Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufman*, 1993.

[24] S. B. E. Raj and A. A. Portia. Analysis on Credit Card Fraud Detection Methods. In *Proceedings of International Conference on Computer, Communication and Electrical Technology – ICCCET2011*, 18th & 19th March, 2011

[25] I. Rish. An empirical study of the naive bayes classifier. In *Proceedings of IJCAI-01 workshop on Empirical Methods in AI, International Joint Conference on Artificial Intelligence*, 2001, pages 41– 46.

[26] L. Šubelj, S. Furlan and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. Expert Systems with Applications, 38(1), 1039-1052, 2011.

[27] S.C. Wang. Interdisciplinary Computing in Java Programming Language. *Dordrecht the Netherlands, Kluwer Academic Publishers*, 2003.

[28] M. Weatherford. Mining for Fraud. *IEEE Intelligent Systems*, July/August Issue, pp4-6, 2002.

[29] G. Williams and Z. Huang. Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases. In *Proceedings of the 10th Australian Joint Conference on Artificial Intelligence*, Perth, Australia, 1997.

[30] G. Williams. Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries. In *Proceedings of the 3rd Pacific-Asia Conference in Knowledge Discovery and Data Mining*, Beijing, China, 1999.

[31] J. J.-C. Ying, E. H.-C. Lu, W.-N. Kuo and V. S. Tseng. Mining User Check-in Behavior with a Random Walk for Urban Point-of-interest Recommendations. *ACM Transactions on Intelligent Systems and Technology (TIST)* Volume 5 Issue 3, September 2014, Article No. 40

[32] M. I. M. Yusoff. Fraud detection in telecommunication industry using Gaussian mixed model. *Mathematical Problems in Engineering*, Volume 2013