Socioscope: I Know Who You Are, A Robo, Human Caller or Service Number

Muahammad Ajmal Azad[a], Mamoun Alazab[b], Farhan Riaz[c], Junaid Arshad[d], Tariq Abullah[a]

[a]*Department of Computer Science, University of Derby, United Kingdom*
[b]*College of Engineering, IT and Environment, Charles Darwin University, Australia*
[c]*Department of Computer and Software Engineering, National University of Science and Technology, Islamabad, Pakistan*
[d]*School of Computing and Engineering, University of West London, United Kingdom*

Abstract

Telephony technologies (mobile, VoIP, and fixed) have potentially improved the way we communicate in our daily life and have been widely adopted for business and personal communications. At the same time, scammers, criminals, and fraudsters have also find the telephony network an attractive and affordable medium to target end-users with the advertisement, marketing of legal and illegal products, and bombard them with the huge volume of unwanted calls. These calls would not only trick call recipients into disclosing their private information such as credit card numbers, PIN code which can be used for financial fraud but also causes a lot of displeasure because of continuous ringing. The fraudsters, political campaigners can also use telephony systems to spread malicious information (hate political or religious messages) in real-time through audio or text messages, which have serious political and social consequences if malicious callers are not mitigated in a quick time. In this context, the identification of malicious callers would not only minimize telephony fraud but would also bring peace to the lives of individuals. One way to classifies users as a spammer or legitimate is to get feedback from the call recipients about their recent interactions with the caller, but these systems not only bring inconvenience to callees but also require changes in the system design. The call detail records extensively log the activities of users and can be used to categorize them as the spammer and non-spammer. In this paper, we utilize the information from the call detailed records and proposed a spam detection framework for the telephone network that identifies malicious callers by utilizing the social behavioral features of users within the network. To this extent, we first model the behavior of the users as the directed social graph and then analyze different features of the social graph i.e. the *Relationship Network* and *Call patterns* of users towards their peers. We then used these features along with the decision tree to classify callers into three classes i.e. human, spammer and call center. We analyzed the call record data-set consisting of more than 2 million users. We have conducted a detailed evaluation of our framework which demonstrates its effectiveness by achieving acceptable detection accuracy and extremely low false-positive rate. The performance results show that the spammers and call center numbers not only have a large number of non-repetitive calls but also have a large number of short duration calls. Similarly, on the other hand, the legitimate callers have a good number of repetitive calls and most of them interacted for a relatively long duration.

Keywords: Social Network Analysis, Telephone Spam Detection; Robocalls; Telephone Call Records; Telemarketers; User Characterization

1. Introduction

In recent years, social networks like Facebook, Twitter, WhatsApp as well as traditional telecommunications networks (mobile, wired, and VoIP etc) have become an integral part of one's life for the interactive communications. Such networks have also been used to disseminate information such as ads, product marketing, and other social-political information. The customer-base and growth of these technologies has also attracted fraudsters, criminals scammers and digital marketers to misuse the medium for social engineering attack to convince users to disclose their private information to be used for executing financial frauds, and marketing of legal and illegal products. The unwanted calls and instant messaging over telephony networks have a greater adverse impact than the traditional email or message spamming because a telephone call (whether a legitimate or non-legitimate) alerts the call recipient in a real-time and requires an immediate response from the call receiver. Furthermore, the exceptional penetration of technology, ease, cheap and real-time nature of attack mechanisms have also convinced scammers to use this medium for the unsolicited communication [6]. Besides committing financial frauds, spammers can also bombard user's voice-mail box with the unwanted content which makes this resource unavailable for legitimate messages whilst making it difficult to remove such content. From the perspective of the service provider, such calls could affect the reputation of the service provider and might also bring financial loss. Therefore, it is of utmost importance for the telecommunication operators to block unsolicited calls in order to provide trustworthy services to their customers.

In telecommunication networks, caller[1] and callee identities are used for initiating and receiving a call. The telecommunication service providers (TSP) record call transactions between a caller and the callee using these identities in a database referred as the call detailed record (CDR). The user's identity has been widely used for blocking the malicious users in the social networks, email networks [24, 35, 49, 12] and telecommunication networks [18, 26, 14] by characterizing the social behavior of user in the networks. In these scenarios, the service provider manages the list of black-listed and white-listed users using identity (Telephone number, IP address or email address) of the users. In telephony networks, users normally develop a social relationship with other users over the time and exhibit different calling behavior across their social groups. These groups can be categories in to three types: *normal users*, *telemarketers* or *robo-callers,* and *call centers* or *service numbers.* The normal or legitimate users typically obey the signed agreement of the service providers, the call centers provide value added services to their users (help number or organization service numbers), and spammers who do not follow the signed agreement and are usually involved in massive unsolicited calling. The spammers can be further grouped into two groups: the robo-callers (automated calling) – pre-configured machines used to generate the large volume of calls on the random numbers and plays the recorded message after the connection also called *pre-recorded calls*, and human-generated calls – affordable human operators are hired for the massive marketing calls.

The above three categories of callers have different call patterns. For example, it is shown that 90% of the legitimate callers usually have only five strongly connected friends whom they mostly talk for 80% of their total talk time [20]. On the other hand, spammers normally call a large number of people and develop disconnected social connections with many of callees with small duration calls. In contrast, call centers have completely different calling behavior i.e. they receive a large number of calls with only a few or zero outgoing calls. There are two types of calls in call centers, inbound and outbound but the phone numbers for either purpose are different. It is also believed that spammers normally have a disconnected network because they initiate interaction with a large number of people and only a few people respond them back [24], [35], [49]. To the best of our knowledge, there is no such study available that analyzed the variety of social network features of different classes of users in a telephone network.

In this paper, we present "Socioscope", a framework for identifying different types of callers in a large-scale telecommunication network. The Socioscope framework consists of two main components: 1) a social network analysis module – that models the raw CDR as the graph network and extracts the social behavior and call patterns of callers, and 2) a decision tree module that classifies callers into three main classes. To model the CDR as the call graph network, We analyzed the anonymized CDRs of around more than 2 million users gathered over one day by a anonymous telecommunication. To this extent, first, we represent the communication network of users as the social weighted directed call graph and secondly, we analyze the connectivity and call patterns of different groups of users namely: the legitimate users, call centers, and the telemarketers or spammers. Specifically, we used the known behavior of three classes for the following features: in-degree and out-degree distribution of users in a respective class, connectivity behavior of users, talk time of users in each class, and centrality measure of users. The evaluation results show the following major findings: spammers exhibit unbalanced communication behavior, for example, they have a very high out-degree and a very small in-degree as confirmed from spammer's behavior in online social networks [30],[49]; call center users have a very high in-degree and a very small or zero out-degree, and the legitimate users have balanced in-degree and out-degree. Moreover, the behavior of legitimate users is also different from call centers and spammers in terms of social network connectivity, clustering coefficient, centrality, and average call duration. Finally, we applied the decision tree method to understand the effectiveness of feature selection for the classification of caller in one of the mentioned class. The major contributions of this paper are as follows:

- We propose Socioscope, a framework for classifying telecommunication users into different classes. The framework specifically utilizes the semantics of graph features, call and connectivity patterns of users to label them as the telemarketers, call center user or the legitimate user.

- We model the CDRs provided by the anonymous telecommunication operator consisting of 2 million users and more than 10 million call records for a single day.

- We developed a prototype to evaluate the performance of our proposed features and the decision tree method. The results show that the proposed features achieve acceptable detection accuracy with a very small false positives.

The rest of the paper is organized as follows. Section 2 introduces types of telecommunication users, the robocall problem and its significance in telecommunication networks. Section 3 reviews existing efforts characterizing the behavior of users in online social and telecommunication networks. Section 4 presents background and the social network features used in this study. Section 5 defines multi-class classification problem addressed by this paper. Section 6 provides details of our proposed approach with discussion on the feature set and machine learning model.

[1] Caller is the person who initiates the call, and callee is the person who receives the call.

Section 7 evaluates the performance of the framework in classifying caller into their respective calls. The Section 8 concludes the paper.

2. Types Telecommunication Users

This section briefly describes the types of telecommunication users.

2.1. Robo and Telemarketing Calls

A *Robocall* or the telemarketing call is the phone call that utilizes telecommunication medium and the computerized autodialer to deliver a pre-recorded telemarketing message to a call receiver. These calls are typically unsolicited and often made for political campaigns, offering holiday packages, promoting political and religious thoughts and selling legal and illegal products. These calls can be made at any hour of the day and require an immediate response from the recipient, thereby a nuisance to call recipients while at work, disturb them in their family times, and can even interrupt their sleep in late hours at night. Recent statistics on telephony spam have revealed that answering a spam call would result in an estimated loss of 20 million man-hours for a small business enterprise in the United States with the estimated loss of about $475 million annually [3]. It is estimated that by the end of 2019 44.6% of phone calls in the US will be spam calls, reflecting the exponential growth and severity of the problem [4, 9]. Another study estimated that in 2018, 1 in every 10 US citizen lost money to the fraudsters who are using telephone channel to attack the users and many of them have become a victim of scam more than once [8]. It is also estimated that telephone scammers managed to have the benefit of $357 per victim, with aggregated overall loss approximately $8.9 billion in total losses [8]. The YouMail (anti-robocalling company) estimates that there were around 3 billion robocalls in February 2019 which are increasing day by day [5]. Further, it is also estimated that half of the calls a user received are originated from unsolicited sources [2]. Consequently, every year service providers, regulators, and law enforcement agencies receive thousands of complaints from consumers for unsolicited, unauthorized, and fraudulent callers trying to abuse them. Furthermore, these calls can also be the first step towards serious frauds such as identity theft and financial scams. Federal Trade Communication (FTC) has estimated that every year scammers and spammers cause a loss of $8.6 billion annually to citizens of USA due to frauds with the majority of them initiated through the telephone [32].

Among other telephony scams, *Wangiri* (literally, ring and disconnect) is a form of fraud call that was first originated in Japan back in 2002 [7]. In a Wangiri call attack, the attacker attempts to convince call recipients to make a premium rate national or international call. The attacker typically makes a *missed call* to the victim identity whereby curious recipients lead to the assumption that they have missed a call from a legitimate caller and call the callee back. The victim is then charged at the premium call rate for this call and only notice upon receiving the monthly bill from their service provider.

2.2. The Call Centers

A call center is an office, which is responsible for managing incoming and outgoing telephone calls from the new and existing customers. The call centers normally operate in two settings, the public call centers that are normally set up for the emergency services or the privately held call centers owned by companies for the advertisement of products or solving problems of the customers. The telemarketers also set up call centers in countries such as India, Pakistan or China where labor cost is extremely low. The legitimate call centers normally obey the *do-not-call registry* and do not attempt to call those number present in the do-not-call registry due to deterrence through detection and penalties for malpractice. However, there may be exceptions where call centers might be involved in massive and advertisement calls.

2.3. The Legitimate Callers

The legitimate callers are normally end-users or human beings legitimately using the telecommunication services for social interaction. Normally, these users develop a relationship over time with others which may be either weak or strong relationship depending upon the call patterns.

3. Related Work and Motivation

In telecommunication networks, Social Network Analysis (SNA) has been used for several purposes including churn analysis, fraud detection and classifying callers as the spammer or non-spammer. Specifically, call records can be analyzed for different behavioral features i.e. the number of phone calls made by the user, the average talk time and the number of unique callees of caller [47], can be analyzed for business analytics informing the operator to improve the quality of service [41], and analyse the call logs for the churn prediction and user retention by creating custom offers through the use of machine learning models over the call logs [48], [44].

In terms of detecting the malicious callers in the network, a range of efforts have been made which utilize the semantics of social network to block caller before they affect a large number of users. Several social graph features have been proposed to characterize the behavior of caller as the spammer and non-spammer. For example, average call duration is proposed in the [18] to estimate the reputation of caller, the explicit feedback from the callee is used to compute the reputation in [26], and call duration feature and out-degree distribution are collectively used in [14] to compute the direct trust and global reputation of caller. The computed reputation is then used along with machine learning models to block the spam caller.

The social network features have also been used to connect the identities that belong to the same physical user in the network [15]. The identity linking is essential for effective detection of spammers and fraudsters as attackers are frequently changing their identities using identity spoofing techniques[13]. In [22], the authors analyzed data from a very large telecommunication network and cluster the users into different clusters based on the communication behavior of users. In [47] authors analyzed the call duration and out-degree distribution of users in a mobile network and suggested that these features could be used for anomaly detection and predicting a wealth of the users. In[11], author detected the anomalous users in a telecommunication call detailed records using weighted reciprocity feature. In [28], authors use the call graph extracted from the CDR to identify the gender of the caller.

Several other approaches have also been proposed for detecting the spammers in the network. Yu-Sung et al. [51] use the extended K-mean clustering algorithm based on the call parameters (messages exchanged during call setup, and termination) along with the callee feedback about the behavior of the subscriber. Azad et al. [17] utilize the K-mean clustering algorithm to mark the caller as a spammer or a non-spammer. Liu et al. [38] discovered the telephone numbers involved in spam campaigns by using the unsupervised and supervised machine learning methods along with the known spam phone numbers to find out new spammers. Sharbani et al. [43] estimate the effectiveness of spam blacklists by measuring their ability to block future unwanted phone calls. Li et al. [36] use 29 features along with the machine learning algorithms to predict whether the subscriber is a legitimate user or a spammer. Chiappetta et al. [23] used an unsupervised clustering algorithm i.e., the K-Means algorithm to group users based on the behavioral model. Collaborative approaches have been proposed [10, 16] where the number of telecommunication operators collaborates for improving the detection time and the detection rate.

Social network analysis has also been used to block the spammers in online social networks and countering web and email spamming [46, 24, 33]. In [46], authors used the structural properties of social graph extracted from the communication logs of spammers and non-spammers on Web and Twitter network. In [30], authors present an approach to characterize the behavior of the spammer on a Facebook social network which uses a structure of connectivity graph of users and their wall post to identify the spammers. The authors identify that Web and Twitter graph of a spammer and a non-spammer is similar which shows that users normally exhibit similar networking behavior across different social networks. In [25], the authors characterize the behavior of bots, legitimate human users, and cyborgs on the twitter network. Authors identified that the bots are more active in posting the message than legitimate users, for example, bots post messages throughout the week whereas the human has a small posting on the weekends, nights and early hours in the morning. In [52], authors proposed graph-based social network features like clustering coefficient, closeness and betweenness centrality of users to identify spammers. In [45], the authors analyzed several social behavioral features for characterizing the behavior of users in online social networks and then used these features for detecting compromised account in the network. In [19], the authors analyzed the social structure of users posting videos on the Youtube network and identified possible self-promotors on the network.

Our proposed approach is based on the following methodology. Firstly, the interactions between caller and callee are represented as a graph network; secondly, different social network features are computed from the weighted and non-weighted graph network, and thirdly a fixed threshold decision tree is applied for classifying caller into one of the three groups: legitimate caller, telemarketer or robocaller, and the service number or helpline. To the best of our knowledge, this is the first study towards modeling user behavior and grouping them into different groups in the telecommunication network to study their behavioral patterns. The approach presented here is also effective in timely identification of scammers and fraudsters in a telecommunication network.

4. Background

This section presents details of call records, call graph semantics and social network features extracted from the call detail records. In this section, we also discuss how different users exhibit different social behavior in a telecommunication network.

4.1. Call Detail Records

Telecommunication service providers record call transactions of customers in a *Call Detail Record* that are used for billing purposes and network management. Service providers can also utilize these records for characterizing call patterns of users for other purposes such as marketing, the personalized offering of new products and identification of malicious users involved in massive calling. CDR is a meta-data of call transactions between a caller and the callee and does not contain recorded speech contents. A typical CDR mainly consists of many fields however only a few features are used to model the call graph network and characterize the behavior of a user. These fields typically are: caller and callee identity, time of call when the caller initiates the call, time when the call is disconnected, duration of the call, who disconnected the call, call type (voice, SMS, MMS) and status of the call (successful or failed), channel activity time, and billing amount. In this paper, we consider five important variables for the construction of a weighted social graph of a user i.e. anonymized identity of caller and callee (it can be telephone number or the IP address assigned to the user), call direction (who initiated the call), call duration (talk time), time of the call when call is initiated, and the frequency of calls. We developed a social call graph over the massive anonymized CDR made available by a service provider consisting of 10 million calls made by 2 million subscribers. The general statistics of the call record is as follows: the average calling rate per second is around 10 to 280 calls during midnight and mid-day, and an average number of calls made by a subscriber to different subscribers is 2.8 with the average call duration of around 60 seconds.

4.2. Call Graph Semantics

The CDR logs can be modeled as a social graph which can be directed and weighted. The call graph is represented as GV, E, W: where V is the set of vertices and is represent the identity of caller or the callee, E a set of edges i.e. the link between caller and the callee if they have been involved in communication at-least-once, and W is the edge weight. In this study, we modeled the CDRs as the direct call graph, which is drawn by creating a separate edge between the identities for both directions (incoming and outgoing), and weights on the edges are defined through interaction frequency and total time a caller and callee are engaged with each other. For an unweighted graph, the edge weight is assigned with constant 1 whereas in the weighted graph the edge value is replaced with weight. Specifically, the weights on the edges represent the strength of the relationship between caller and the callee. The call graph G can be represented as the adjacency matrix. An example of the call graph among users of the network is shown in a Figure 1 and is represented as a sparse adjacency matrix, where 1 represents that caller S has interacted with callee R and 0 represents no interaction between the caller and the callee had happened. A nxn adjacency matrix A is represented by elements as:

$$A_{SR} = \begin{cases} Interacted; & \text{if } S \text{ interacted } R \\ not-interacted; & \text{Otherwise} \end{cases} \tag{1}$$

Where $S, R \in A$ In the case of the weighted call graph, A_{ij} is replaced by the weights determined from the frequency of interaction and sum of call duration between caller and the callee. In this paper, the weighted call graph is constructed by extracting the following three parameters from the call records.

Talk Time: The talk time or call duration represents the time two users talked to each other. Specifically, talk time of caller S with callee R is the time difference when callee accepts the call and the time when a call is disconnected either by callee or caller. The total talk time between caller and the callee is the sum of the duration of all calls made by caller S to callee R, and overall talk time of S is the sum of the duration of all calls made by S to their callees. The long duration calls between S and R represents sign of a strong relationship between caller and the callee.

Call-Rate: Call-Rate represents the interaction rate between caller and the callee over the defined period. Specifically, the call rate between caller S and callee R is the sum of all calls made from caller S to callee R. The call-rate can be grouped as the incoming call-rate (calls received by the S) and the out-going call-rate (call made by the S). The aggregated call-rate, therefore, is the sum of all calls made and received by the user A i.e. $S, R \in A$. The more the call-rate caller S with callee R the stronger the relationship exists between caller and the callee

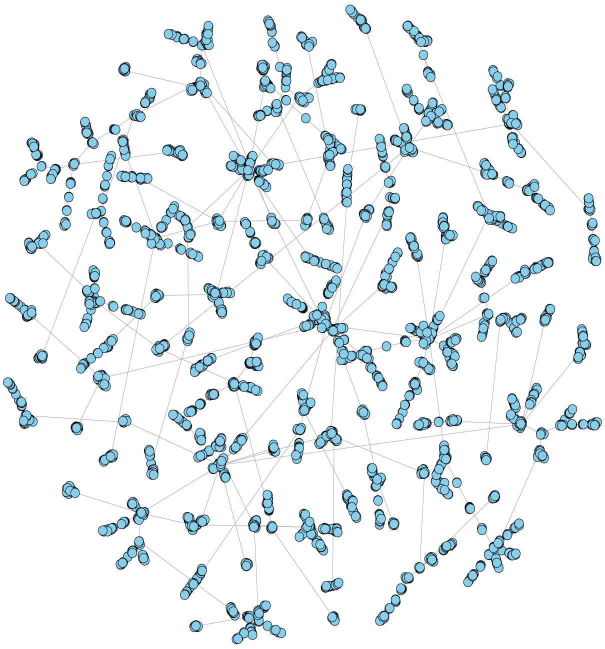


Figure 1: Representation of Users Call Records as a Call Graph.

Partners: Partner is the total number of unique callee certain user-initiated calls to or received calls from and can be grouped into incoming and outgoing partners. The incoming partners of caller S are represented as IDC_S and out-going partners of caller S is represented as ODC_S. The out-going Interactions represents that user is more important to the certain user than those he did not initiate any call.

5. Problem Definition

The telecommunication service providers record each interaction of their customers in a call detailed record. Given a set of N users in the telecommunication network represented as $A = a_1, a_2, \ldots, a_n$, a caller $S \subseteq A$ is the set of users that have originated at least one call, and the call recipient R is the set of users that have received at least one call from the $S \subseteq A$. The call detailed logs between the users are available in the form of different features that includes the time of call, the identity of caller and the call recipient, and the duration, respectively. The identity of the user is the phone number assigned to the user at the time of registration with the service provider. In the data set we have set of users labeled as the 0,1,2; 1 if caller is legitimate, 0 if the caller is labeled as spammer and 2

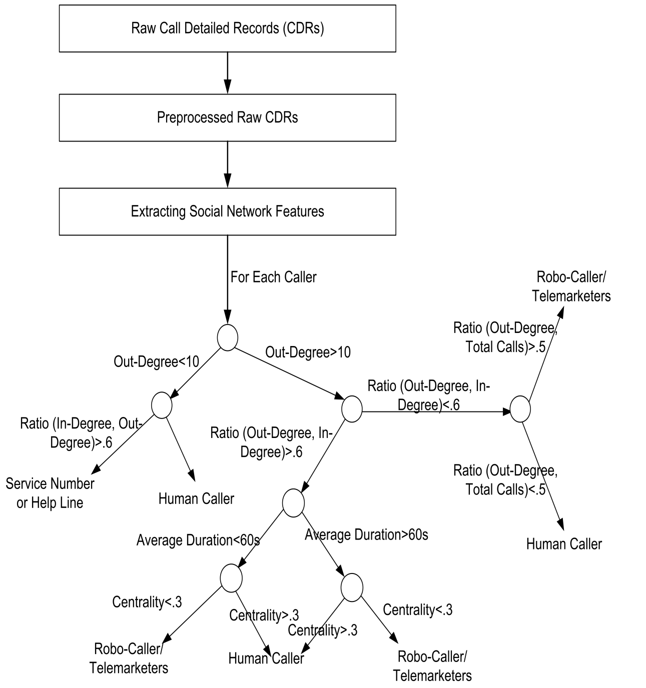

Raw Call Detailed Records (CDRs)
Preprocessed Raw CDRs
Extracting Social Network Features
For Each Caller
Out-Degree<10
Out-Degree>10
Ratio (In-Degree, Out-Degree)>.6
Service Number or Help Line
Human Caller
Ratio (Out-Degree, In-Degree)>.6
Ratio (Out-Degree, In-Degree)<.6
Robo-Caller/ Telemarketers
Ratio (Out-Degree, Total Calls)>.5
Ratio (Out-Degree, Total Calls)<.5
Human Caller
Average Duration<60s
Average Duration>60s
Centrality<.3
Centrality>.3
Centrality>.3
Centrality<.3
Robo-Caller/ Telemarketers
Human Caller
Robo-Caller/ Telemarketers

Figure 2: Example Decision for the Daily Representation of Users in a Telecommunication Network.

if caller is labeled as the call center number. The goal of this work is twofold: first, understand the call and social behavior of all three respective classes, and secondly automatically classify the remaining unlabelled users with the score of 0,1,2. The social connectivity score is computed using different social network features and can also represent the legitimacy of caller.

6. Proposed Socioscope Framework

In this section, we present a discussion on the Socioscope framework and outline its working towards classifying caller.

6.1. Description of Solution

Figure 2 presents the building block of the proposed solution for categorizing caller and identifying spam caller. Firstly, a raw CDRs are processed and social call graph is constructed from the processed CDRs. A social call graph can be represented as the direct graph where caller is represented as the nodes and call transaction between nodes is represented as an edge. Secondly, social network features and call features are computed, and finally, a machine learning method such as the decision trees is used to classify caller as the legitimate, spammer and call center. The remainder of this section details the steps involved. Each part of the system will be described in detail in the following sections.

6.2. Pre-Processing and Construction of Call Graph Network

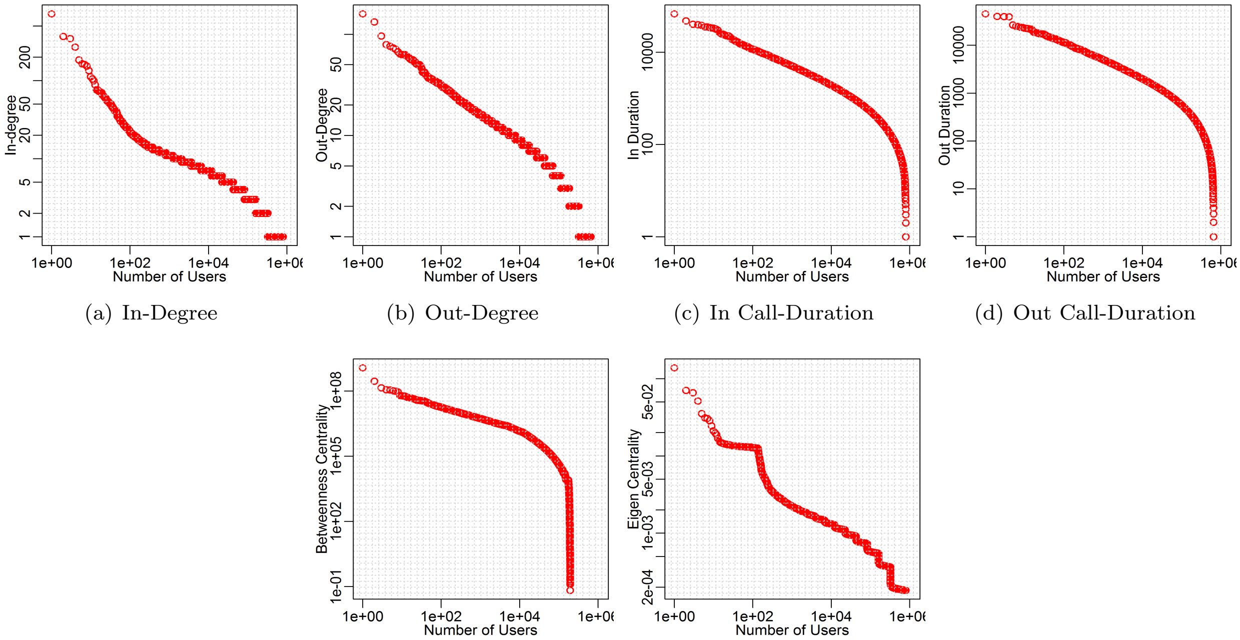
Almost all telecommunication operators maintain a database that keeps logs of the call transaction of users of the network. The logs keep record of both failed and successful calls but do not record any speech content. The call log has the identities of the parties involved in communication, time of call, call duration, and price of the call. The social call is constructed using the data for the particular time window by parsing the call logs through the processing engine. We used only the successful calls to construct the social graph, as we are interested in understanding the social connectivity of the users in the call logs. The information about the call transaction is extracted from the call logs.

Let: $CallerID(a_i, a_j)$ = number of calls made by the user a_i to a_j , where $ai \in S$, $aj \in R$, and $S \cup R = A$. A few sets of the users are labeled as the spammers, legitimate users, or the call center as C = 0, 1,2. In our social graph, some of the users a_i have different roles i.e. act as caller S as well as the callee R. To address this, we model the social graph as directed with a separate link between users for the incoming and outgoing calls. The edge between the caller or the callee only exist once but is given weights if multiple transactions have happened between the users.

6.3. Social Behavior Features

People use the telephony network to interact with friends, family members, and other unknown users. The users normally develop a social relationship (can be classified as weak or strong) with each other over time depending on the frequency of calls and the duration of calls. On the other hand, spammers or telemarketers also try to exploit the telephony network for financial benefits (e.g marketing of products, advertising, Phishing, and frauds, etc.) thus also develop a strong and weak social network with many users. However, the social behavior of a legitimate user is different from that of a spammer and therefore can reveal insights which can be used to block unwanted callers in the network. This section presents social and calls characteristics of the user that can help segregate spammers and non-spammers. We outline the calling behavior of spammers and non-spammers for different call features such as several callees the user calls, the number of calls the user is receiving from others, the call duration of user's incoming and outgoing calls, and incoming and outgoing call-rate of the user. We also extracted the social network features such as *closeness centrality*, *betweenness centrality* and *Eigenvector centrality* measures of the user.

Figure 3 presents distribution for different social network features extracted from the data of 2 million users for a single day. Specifically, Figure 3.a and b present the Indegree and Out-degree patterns as observed in the data set respectively whereas Figure 3.e presents the betweenness Centrality for users in the data set.



(a) In-Degree

(b) Out-Degree

(c) In Call-Duration

(d) Out Call-Duration

(e) Betweenness Centrality

(f) Eigen Centrality

Figure 3: Social Network Features for Users Extracted from 2 Million Users.

6.3.1. Degree Centrality

One of the most important structural features of a user in a social call graph is the degree centrality of the user. The degree centrality of a user in a social call graph represents the number of unique connections the user has with others. In a directed social call graph, the user has two different measures for degree centrality: the out-degree and the in-degree. The out-degree of a user i in an adjacency matrix A is the sum of all outgoing calls made by the user i to other unique identities. The in-degree is the sum of unique users calling the user i. The out-degree of the user i can be represented as:

$$ODCi = \sum_{i=1}^{n} A_{ij} \tag{2}$$

$$IDC_i = \sum_{j=1}^{n} A_{ji} \tag{3}$$

Where $A_{ij} = 1$ if there is an outgoing link from user i to user j, and zero otherwise. Similarly, $A_{ji} = 1$ if user j has outgoing link to user i and zero otherwise. In the case of a weighted network, the out-degree and in-degree are simply summed of weights of rows and columns of the user i. The degree of the user can also be represented as the degree distribution which is the probability distribution of the user's degree of the whole network. Many real networks such as World Wide Web[42], phone call graph, [41], network of autonomous IP systems [29] and online social networks [39] exhibit a power-law degree distribution. It might be possible that inclusion of a large number of spammers in a network would divert the degree distribution from a power-law degree distribution to some other distribution [40].

6.3.2. Betweenness Centrality

Betweenness centrality of a user in a network is computed by quantifying a number of times it is present in the shortest paths between other users of the network. Betweenness centrality is an important metric in social networks to identify the control of a human on communication between other humans. Specifically, a node with a high betweenness centrality has greater influence and control over the flow of information through a network thereby also representing the importance of the user within a network. In the context of the social call graph, the betweenness of a user i is computed as

$$CB(i) = \sum_{j<k} \frac{SP_{jk}(i)}{SP_{jk}} \tag{4}$$

In equation 4, SP_{jk} is a total number of shortest paths from node j to node k and $SP_{jk}(i)$ is the number of those paths that pass through i. Users with a high betweenness centrality score are considered as the important players in the network and are considered as a pivot point to control the flow of information between other users. Within the context of this research, spammers are envisaged to have high betweenness centrality as they are involved in massive calling to a large number of users. On the other hand, a legitimate caller may have small betweenness centrality measures because of the existence of only a few links to other nodes in the network.

6.3.3. Closeness Centrality

Closeness centrality is the measure to estimate how close a user is to all other users in the social call graph. It is computed by taking the reciprocal of the sum of the shortest distances from a given user to every other user in the network. The user having high closeness score the lower would be its total distance from all other users. The closeness measure identifies the users which could reach others more quickly. The closeness centrality of the user i can be defined as:

$$CC(i) = \sum_{j=1}^{N} \frac{1}{d(i,j)}. \tag{5}$$

In 5, $d(i,j)$ is the shortest distance between node i and j, and N is the total number of nodes in the network.Spammers normally have a high closeness centrality since they exist very near to ever another user in the network, on the other hands, the legitimate users normally have a small closeness centrality because of their few callees and callers.

6.3.4. Eigen Centrality

Eigenvector centrality estimates the centrality of a user in a social call graph by computing the Eigenvector of the largest positive eigenvalues of the user. Specifically, Eigen Centrality considers connectivity information of a user's neighbors. Eigenvector provides information about the connectivity of user with the other users. Users have a connection with the users having high centrality would have high centrality score than users connected to the low centrality subscribers. The Eigen centrality of a user from an adjacency matrix is defined as:

$$EC(A) = \sum A_{ij} * x_j \tag{6}$$

Where A_{ij} is the adjacency matrix of a graph G and x_j is the initial centrality score of subscriber j. The users connected to high scoring users would typically have high eigenvector centrality. Spammers normally choose a large number of reputed recipients, thus probably would have high centrality score, whereas legitimate user only has connections with few reputed users thus resulting in a small centrality score. Considering, high centrality score as a sign of legitimacy would result in blocking many legitimate subscribers and allowing many spammers. However, we believe that Eigenvector centrality would not only be limited to the structure of the network but also need to consider the relationship strength between users. Computing centrality by considering the trust weights would probably result in a small centrality score for the spammers and high centrality score for the non-spammer.

6.3.5. Average Call Duration

The call duration or the talk time represents the length of duration the users talked to each other. The average call duration between users is the sum of the duration of calls to the total number of calls between users. The average call duration of user S with the user R is defined as follow:

$$Avg.Duration(SR) = \frac{\sum Duration(SR)}{\sum Interactions(SR)} \tag{7}$$

In this paper, we assign edge weights by computing the average call duration between caller and the callee. The average call duration also represents the importance of the link between caller and the callee. The high duration is the sign that users are strongly connected and small duration calls represent a weak relationship. In the context of spamming, spammers in telecommunication typically have a large number of small duration calls thus develop weak relationship network with a large number of users, whereas legitimate callers develop a strong relationship over time.

6.4. Call Graph Processing and Feature Vector

For caller classification, we propose to know that whether the particular caller is spammer, a legitimate caller or the call center by using the machine learning approaches applied to the social graph feature vector. This section describes applying a decision tree algorithm to the given social feature vector. The feature vector of the user who made at least one call is computed by using the features construction method discussed in the previous section. We modeled 5 features to construct the feature vector of the user. In some circumstances using only one feature would be enough for the classification but it might not have the trade-off between true positive and false positive. For example, if outdegree is used as the sign of spamming then the system would be flagged all callers having a high degree as the spammer despite several long duration calls. It is important to have a trade-off between false positive and true positive by combining a set of features. In our representation a set of 5 features $f_i(i = 1, \dots, 5)$ are used to construct the feature row for the user a_i as $x_i = (f_1, f_2, \dots f_i, \dots, f_n)$ for each of the user $a_i \in S$. We normalized the values of features so to consider the same scale for all the features. We normalized the value of the feature by simply dividing the feature value of the user by the sum score of all features of the user $a_i \in S$. The feature vector can be easily extended to include other features such as reciprocity index of the user and Katz centrality measure.

6.5. Decision Tree for Classification

Decision Tree (DT) refers to a supervised classification technique that segregates various attributes in the data to logically accommodate them into their respective classes. DTs are used to build logical models of data with higher accuracy. The features that we have used in the proposed framework are not continuous (e.g., data acquired from a sensor, image data, etc.). Appropriately, the use of various feature space classifiers such as neural networks [37, 50], support vector machines [21] etc. is not useful in our scenario. Given this, we have used DT to provide rule-based classification framework [34], that tries to perform prediction on the data based on the maximization of entropy (inter-feature information gain). In our implementation, the input to the decision tree classifier has the following structure for the user a_i:

$$(X, Y) = (F_1, F_2, F_3 \dots F_j, Y) \tag{8}$$

Where $X(F_1, F_2, F_3 \dots F_j)$ are the feature values that are used for the classification task, and Y is the class of the user (spammer, legitimate, call center). Let S be a group of samples and $N_r(C_i, S)$ be the frequency of occurrence of samples in S that belong to the class C_i. Now, assuming that we have k number of classes, and the total number of samples in S are $|S|$, the entropy of S can be calculated as follows:

$$E(S) = -\sum_{i=1}^{k} \Big(\frac{N_r(C_i, S)}{|S|} . \log_2 \Big(\frac{N_r(C_i, S)}{|S|} \Big) \Big)$$

Once $E(S)$ is calculated, S is partitioned into n number of outcomes with respect to a feature F_j. Thus, $E(S)$ with respect to F_j becomes the weighted sum of entropies of all individual samples. The final entropy and information gain are calculated as follows:

$$E_{F_j}(S) = -\sum_{i=1}^{n} \Big(\frac{|S_i|}{|S|} E(S_i) \Big)$$

$$IG(F_j) = E(S) - E_{F_j}(S)$$

Where IG refers to the information gain. The IG is calculated for all the attributes and an attribute with a maximum IG is selected to partition S. An example of such a decision tree is presented in Figure 2. We specifically used around 5 social networks and call features to place the user in a specific group. At the end of the classification process, each caller in the row represented by the feature vector along with the class type. It is believed that callers having similar calling behavior would have almost similar values for their social network features vector and lies close to each other.

7. Evaluation

In this section, we evaluate the performance of the proposed Socioscope framework in classifying caller into three different classes and identify spammers. All the experiments are carried out on real CDRs.

7.1. Data Set

To evaluate our proposed framework, we used the data set provided by the telecommunication service provider. The CDRs are completely anonymized i.e. the original identity of the user is modified with a random string. The identity of caller, callee and the time stamp of call records have been anonymized by the operator. The region of the provided data is not disclosed, thus cannot be deanonymized by the data handler. T timing information of the call is also rounded-off to minimize any linkability in condition if adversary managed to have limited background information. Further, data is seen only by one author based in Pakistan and not moved outside jurisdiction of a research organization. We analyzed the records over one day. The average call rate of users per second is 10 to 280 calls during midnight and mid-day. An average number of calls made by a user to other user is around 2.8 calls per day. We also have small sample labeled data set for each class.

7.2. Performance Metric

The evaluation of the machine learning methods and classification is based on the evaluation measures widely used in the machine learning and information retrieval domain. Given a classification method, the confusion matrix is created representing the true values for classification results achieved by the classification method. The confusion matrix is represented as follows: In a two-class classification problem (spammer or non-spammer), TP represents the number of users spam class that were correctly classified as the spammer, FN represents the spam users that were incorrectly classified as the non-spammer examples that were falsely classified as non-spam, FP represents the number of non-spam users that were incorrectly classified as the spammer, and TN represents the number of non-spam users that were correctly classified as the non-spam users. The confusion matrix can also be used to estimate other performance metrics such as precision, recall, and F-measure. The confusion matrix can be easily extended for the three-class problem.

		Predicted	
		Spammer	non-Spammer
Actual	Spammer	TP	FN
	non-Spammer	FP	TN

Table 1: Confusion Matrix.

7.3. Result Analysis

The first behavior attribute of the spammers in any type of network is to spread the information to comparatively large footprint to gain more profit. For this purpose they operate in two phases, crawling of address space, and sending the recorded core message to the selected random number. However, the behavior of call recipients varies, some call recipients answer the call from an unknown number, and others do not like to answer the call from unknown numbers. In our analysis, it has been seen that only a short fraction of target recipients accept the call from the unknown numbers, and the majority of those who accepted the call disconnected the call within first 20 seconds of duration. Only a small fraction of recipients who accepted the call has talked for the long duration. Our analysis also showed that a large number of users who received a call from an unknown number did not reply to the call. The high out-degree is the sign that caller is spammers but it might have some false positives as some users also have high out-degree but their behavior tells another way around, for example, they have a large number of good duration calls. Thats why in our analysis we combine the call duration and out-degree feature to analyze the behavior of caller. The normal callers, on the other hand, not only have a small number of friends but also have a large number of good duration calls with them over time, thus developing a strong social circle. Moreover, the legitimate callers also have a connected social network that results in a comparatively high Eigen centrality measure and high clustering scores for the legitimate callers.

Our analysis over the real data-set shows that the Socioscope has successful identifies around 250 identities as the spammer, 50 identities as the call center, and remaining are identifies as the legitimate users. The distribution for the different social network features is presented in a figure 3. The results revealed that the identified spammers have exceptional high out-degree with an average of 32 unique victims and for some callers, it goes to more than 100 unique targets. On a further investigation, our analysis also shows that the identified spammers have average call duration of around 90 seconds with each of his target victims, with a large number of calls having call duration less than 50 seconds. This behavior of the spammer is almost similar to the one characterized by [27] and [31]. On aggregate, the identified number of spammers have some unexpected behavior in terms of aggregate call duration, for example, many identified spammers have a total call duration of more than 6 hours on a single day with an average of 50 callees, which might not characterize the behavior of legitimate human callers. On further investigation of these callers, we have also found that these callers have zero-incoming calls. Spammers normally target users only once from his single identity and this behavior has also been seen for the classified spammers.

The second group of caller that our analysis has identified is the call center or help numbers. These callers normally have a huge number of incoming calls and a very small number of outgoing calls. Our analysis results show that call centers have an average of 81 unique users calling. Some of the call centers even received more than 1500 calls on a single day. The analysis show that the call centers hardly made calls to the users (average out-going number of users is 0.81), they have average incoming call duration of 136 seconds which is slightly greater than the average call duration of 911 calls in USA [1]. Moreover, these callers have high clustering coefficient greater than that of spammers.

The identified normal callers not only have outgoing calls but also have incoming calls. These callers also have good duration calls in both directions. The analysis revealed that the out-degree and in-degree of legitimate caller normal lies close to each other that shows that legitimate callers have a reciprocal communication behavior.

As the data-set in hand is not labeled by the service provider, we do not have a way to evaluate the classification accuracy of the proposed approach other than further analyzing the behavior of the user in the respective class. On further analyzing the spammer class, we identified that only a few legitimate callers exhibit spamming behavior because they have high out-degree. This corresponds to a very small false positive rate which is acceptable in the large telecommunication networks. On the other hands, we have not found any high out-degree callers in the non-spamming class which shows the high true positive rate of the proposed framework.

8. Discussion and Conclusions

Spamming in a telecommunication network is not only intrusive to users of the technology but scammers also convince call recipients to disclose their private information that leads to the financial frauds. The telecommunication users can be grouped into three main categories depending on the behavior of caller and require a framework that automates the classification process by widely utilizing several social and call features. To address this challenge, the proposed framework facilitate classification of each of the user into one of the three categories. To this extent, we first analyzed the social network features and call patterns of the user, and then apply the decision tree to label the user as the spammer, legitimate or the call center. The classification process consists of three steps: first raw CDRs are processed to represent users interactions as a weighted directed call graph, secondly statistics for several social network features the user are computed and analyzed from the weighted call graph and finally, users are classified into one of three classes using fixed threshold-based decision trees. Once users are placed in a classification group, we then analyzed the behavioral patterns of the user in each class to get a better insight into the behavior of users in each class. The experimental results show that spammers have non-repetitive calling behavior which is similar to the behavior of spammers in an online social and email networks. The experimental results show that spammers have non-repetitive calling behavior which is similar to the behavior of spammers in online social and email networks. This is because spammers normally target a victim only once, do not repeat calling same user again and again, thus develop a non-repetitive communication behavior. On the other hands, normal callers not only have repetitive calling behavior but also have a strong social network.

The challenge while deploying such a system in a real setup is the use of the identity for modeling the behavior of the user. In reality, spammers largely spoofed the identity of other legitimate entities which might leads to poor classification results as the system see random identity as the new. This would also lead to false-positive rates if the

CDR were recorded against the real identity. The solution to such problem is to incorporate the strong authentication mechanism that ensures that caller owns the identity he is claiming.

Currently, we have only used the five features for the classification, in future we are planning to extended the system in two directions i.e. inclusion of more features, using incremental machine learning and performing experiments on the labeled data set.

References

[1] Dauphin county call statistics for 911 service.

[2] FCC: Nearly half the calls you receive this year will be spam, url = https://edition.cnn.com/2019/02/14/politics/fcc-robocalls-report/index.html, owner = ajmal, timestamp = 2015.09.14.

[3] Spam Phone Calls Cost U.S. Small Businesses Half-Billion Dollars in Lost Productivity, Marchex Study Finds.

[4] The Fight to Eliminate Unwanted Robocalls, url = https://www.aarp.org/content/dam/aarp/ppi/2017/10/the-fight-to-eliminate-unwanted-robocalls.pdf, owner = ajmal, timestamp = 2015.09.14.

[5] Whats The Point Of All Those Robocalls, Anyways?, url = https://blog.youmail.com/2018/02/robocalls/, owner = ajmal, timestamp = 2015.09.14.

[6] Symantec Intelligence Report (Reterived September 2015), 2015.

[7] Mobile phone scam kills curious cats, 01, December 2016.

[8] Estimated 24.9m americans lost $8.9b in phone scams, 01, December 2018.

[9] Phone spam stats: The big picture, 01, December 2019.

[10] M. Ajmal, S. Bag, S. Tabassum, and F. Hao. privy: Privacy preserving collaboration across multiple service providers to combat telecoms spam. *IEEE Transactions on Emerging Topics in Computing*, pages 1–1, 2017.

[11] L. Akoglu, M. McGlohon, and C. Faloutsos. *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part II*, chapter oddball: Spotting Anomalies in Weighted Graphs, pages 410–421. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[12] Mamoun Alazab. Profiling and classifying the behavior of malicious codes. *J. Syst. Softw.*, 100(C):91–102, February 2015.

[13] M. A. Azad, S. Bag, C. Perera, M. Barhamgi, and F. Hao. Authentic-caller: Self-enforcing authentication in a next generation network. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2019.

[14] Muhammad Ajmal Azad and Ricardo Morla. Caller-Rep: Detecting unwanted calls with caller social strength. *Computers & Security*, 39, Part B:219–236, 2013.

[15] Muhammad Ajmal Azad and Ricardo Morla. Early identification of spammers through identity linking, social network and call features. *Journal of Computational Science*, pages –, 2016.

[16] Muhammad Ajmal Azad and Ricardo Morla. Rapid detection of spammers through collaborative information sharing across multiple service providers. *Future Generation Computer Systems*, 95:841 – 854, 2019.

[17] Muhammad Ajmal Azad, Ricardo Morla, Junaid Arshad, and Khaled Salah. Clustering voip caller for spit identification. *Security and Communication Networks*, 9(18):4827–4838.

[18] V.A Balasubramaniyan, M Ahamad, and H Park. CallRank: Combating SPIT Using Call Duration, Social Networks and Global Reputation. In *Proceedings of Fourth CEAS2007.*, 2007.

[19] Fabricio Benevenuto, Fernando Duarte, Tiago Rodrigues, Virgilio A.F. Almeida, Jussara M. Almeida, and Keith W. Ross. Understanding video interactions in youtube. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 761–764, New York, NY, USA, 2008. ACM.

[20] H.K. Bokharaei, A. Sahraei, Y. Ganjali, R. Keralapura, and A. Nucci. You can SPIT, but You can't hide: Spammer Identification in Telephony Networks. In *Proceedings of 2011 IEEE INFOCOM*, pages 41–45, 2011.

[21] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[22] S. Chiappetta, C. Mazzariello, R. Presta, and Simon P. Romano. An Anomaly-based Approach to the Analysis of the Social Behavior of VoIP Users. *Computer Networks*, 57:1545 – 1559, 2013.

[23] S. Chiappetta, C. Mazzariello, R. Presta, and S.P. Romano. An anomaly-based approach to the analysis of the social behavior of voip users. *Computer Networks*, 57(6):1545 – 1559, 2013.

[24] Paul-Alexandru Chirita, Jörg Diederich, and Wolfgang Nejdl. MailRank: Using Ranking for Spam Detections. In *Proceedings of 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 373–380, 2005.

[25] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, New York, NY, USA, 2010. ACM.

[26] R Dantu and P Kolan. Detecting Spam in VoIP Networks. In *Proceedings of the Steps to Reducing Unwanted Traffic on the Internet ,Berkeley, CA, USA*, pages 31–37. USENIX, 2005.

[27] Nico d?Heureuse, Sandra Tartarelli, and Saverio Niccolini. Analyzing Telemarketer Behavior in Massive Telecom Data Records. In *Springer Trustworthy Internet*, pages 261–271. 2011.

[28] Zheng-Bin Dong, Guo-Jie Song, Kun-Qing Xie, and Jing-Yao Wang. An experimental study of large-scale mobile social network. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1175–1176, New York, NY, USA, 2009. ACM.

[29] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 251–262, 1999.

[30] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 35–47, New York, NY, USA, 2010. ACM.

[31] P. Gupta, B. Srinivasan, V. Balasubramaniyan, and M. Ahamad. Phoneypot: Data-driven Understanding of Telephony Threats. In *20th NDSS*, 2015.

[32] Payas Gupta, Roberto Perdisci, and Mustaque Ahamad. Towards measuring the role of phone numbers in twitter-advertised spam. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, pages 285–296, New York, NY, USA, 2018. ACM.

[33] M. Jiang, P. Cui, and C. Faloutsos. Suspicious behavior detection: Current trends and future directions. *IEEE Intelligent Systems*, 31(1):31–39, Jan 2016.

[34] Anish Jindal, Amit Dua, Kuljeet Kaur, Mukesh Singh, Neeraj Kumar, and S Mishra. Decision tree and svm-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, 12(3):1005–1016, 2016.

[35] HoYu Lam and DitYan Yeung. A Learning Approach to Spam Detection Based on Social Networks. In *Proceedings of Fourth Conference on Email and Anti-Spam (CEAS2007)*, 2007.

[36] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, W. Zhang, Y. Yu, and D. Song. A machine learning approach to prevent malicious calls over telephony networks. In *Proceedings of 2018 IEEE Symposium on Security and Privacy (SP)*, pages 53–69, May 2018.

[37] W. Li and A. W. Moore. A machine learning approach for efficient traffic classification. In *2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 310–317, Oct 2007.

[38] Jienan Liu, Babak Rahbarinia, Roberto Perdisci, Haitao Du,

and Li Su. Augmenting telephone spam blacklists by mining large cdr datasets. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, pages 273–284, 2018.

[39] A. Mislove, M. Marcon, P. Gummadi, Krishna, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42, 2007.

[40] F. Moradi, T. Olovsson, and P. Tsigas. Towards Modeling Legitimate and Unsolicited Email Traffic Using Social Network Properties. In *5th Workshop on Social Network Systems*, pages 1–6, 2012.

[41] A.A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications. In *15th ACM CIKM '06*, pages 435–444, 2006.

[42] M. E. J. Newman. Power laws, Pareto Distributions and Zipf?s Law. In *Contemporary Physics*, pages 323–351. 2005.

[43] Sharbani Pandit, Roberto Perdisci, Mustaque Ahamad, and Payas Gupta. Towards measuring the effectiveness of telephony blacklists. In *Proceedings of 23rd NDSS*, 2015.

[44] C. Phadke, H. Uzunalioglu, V. B. Mendiratta, D. Kushnir, and D. Doran. Prediction of subscriber churn using social network analysis. *Bell Labs Technical Journal*, 17(4):63–76, March 2013.

[45] X. Ruan, Z. Wu, H. Wang, and S. Jajodia. Profiling online social behaviors for compromised account detection. *IEEE Transactions on Information Forensics and Security*, 11(1):176–187, Jan 2016.

[46] Yardi Sarita, M. Romero Daniel, Schoenebeck Grant, and boyd danah. Detecting spam in a twitter network. *First Monday*, 2010.

[47] Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskove. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 596–604, New York, NY, USA, 2008. ACM.

[48] Wouter Verbeke, David Martens, and Bart Baesens. Social network analysis for customer churn prediction. *Applied Soft Computing*, 14, Part C:431 – 446, 2014.

[49] Alex Hai Wang. Don't Follow me: Spam Detection in Twitter. In *2010 International Conference on Security and Cryptography (SECRYPT),*, pages 1 –10, 2010.

[50] Chih-Hung Wu. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications*, 36(3, Part 1):4321 – 4330, 2009.

[51] Yu-Sung Wu, Saurabh Bagchi, Navjot Singh, and Ratsameetip Wita. Spam Detection in Voice- Over-IP Calls through Semi-Supervised Clustering. In *Proceedings of 39th Annual IEEE/IFIP DSN, Portugal*, pages 307–316, 2009.

[52] Chao Yang, Robert Chandler Harkreader, and Guofei Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection*, RAID'11, pages 318–337, Berlin, Heidelberg, 2011. Springer-Verlag.