Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090407

Clustering of Telecommunications User Profiles for Fraud Detection and Security Enhancement in Large Corporate Networks: A case Study

Constantinos S. Hilas[1,*], *Paris A. Mastorocostas*[1] *and Ioannis T. Rekanos*[2]

[1] Dept. of Informatics Engineering, Technological Educational Institute of Central Macedonia, GR-62124, Serres, Greece
[2] Dept of Electrical and Electronics Engineering, School of Engineering, Aristotle University of Thessaloniki, GR-54124, Thessaloniki, Greece

Received: 8 Oct. 2014, Revised: 8 Jan. 2015, Accepted: 9 Jan. 2015
Published online: 1 Jul. 2015

Abstract: A user's transactions with modern networks and services produce a vast amount of user related data. The byproduct of every phone call a person makes or every web page one visits is translated into a log record with usage data. By studying these log records, the user's behavior is revealed and one may come up with clues about user preferences, identify security issues, or discover fraudulent use of the network or the service one provides. Thus, the modeling of network users' behavior may serve as an invaluable tool for the IT manager. In this paper, many of these issues are discussed and emphasis is given on the construction of appropriate user behavior representation in telecommunications. As an example, the application of two clustering techniques is presented, with the task to identify appropriate user behavior representations (profiles) inside a large organization's telecommunications network, in order to spot fraudulent usage. Through this study a researcher and/or the organization's network manager may gain more insight into the problems of user profiling and fraud detection.

Keywords: User profiling, clustering applications, data mining, fraud detection, telecommunications

1 Introduction

A user profile is a collection of personal data associated to a specific user. Ideally a user's profile would be an explicit digital representation of that person. However, one may have several different profiles depending on the system with which he interacts. For example, from a personal computer's perspective a user profile is a collection of settings that make the computer look and work the way you want it to. For an e-commerce system, e.g. an electronic book store, a user's profile may include identity information, like login credentials; billing information, like credit card data; and more importantly, a user's preferences as regards literature genres he prefers, recent purchases, etc. Some of this knowledge was explicitly entered to the system by the user, while other was implied by the system after analyzing past transactions. In this sense a profile emerges on the basis of monitoring ones usage patterns, so it is only relevant to a user's attitude against a specific service. Of course, these data may be correlated with data from other databases and yield a more specific inference about one's preferences. User profiles may be found in operating systems, e-commerce applications, social networking sites, recommender systems, intelligent tutoring systems, etc.

The main idea behind user profiling is that the past behavior of a user can be accumulated in order to construct a profile, or a "user dictionary", or a "user signature" of what might be the expected values of a user's behavior. In its simplest form this profile is a vector that contains single numerical summaries of some aspect of behavior or some kind of multivariate behavioral pattern. A profile may also contain categorical, censored, or other non-numeric data.

The profile follows the logic of the recordable data, with the constraints inherent in computer technology. This issue, the inherently reductive character of a profile, is important because profiles may impact privacy and identity in the strong sense (concerning our sense of self).

* Corresponding author e-mail: chilas@teicm.gr

Since profiles will often affect our lives (providing or prohibiting access, enabling selection, inclusion and exclusion) it is of utmost importance to clarify in what ways and on what basis they affect our lives, without conflating profile and profiled person [1]. This remark should always be kept in mind when constructing user profiles whether this is done in order to help or protect the user from other dangers.

After a profile is constructed, it can be used appropriately. An operating system stores user profiles that contain one's settings for desktop backgrounds, screen savers, pointer preferences, sound settings, and other features. A recommender system uses past user preferences in order to predict items that the user has not yet considered, e.g. a book, a song, or even a friend [2]. Traditionally, in computer security user profiles are constructed based on any basic usage characteristic such as resources consumed, login location, typing rate and counts of particular commands. Future behavior of the user can then be compared with his profile in order to examine the consistency with it (normal behavior) or any deviation from his profile, which may imply a breach in security or some fraudulent activity.

In particular, fraud detection is important to the telecommunications industry because companies and suppliers of telecommunications services lose a significant proportion of their revenue as a result. Moreover, the modeling and characterization of users' behavior in telecommunications can be used to improve network security, improve services, provide personalized applications, and optimize the operation of electronic equipment and/or communication protocols.

Several categories of telecommunications fraud have been reported in the literature. The most prominant are the technical fraud, the contractual fraud, the procedural fraud, and the hacking fraud [3]. The first three usually burden the economics of the service provider, while hacking fraud also harms the subscriber. Hacking fraud is usually met in the form of the superimposed fraud where the fraudster (hacker) uses a service concurrently with the subscriber and burdens his account. The present paper focuses on superimposed fraud identification.

The Communications Fraud Control Association (CFCA) recently announced the results of a global survey carried out in 2013. Experts estimate 2013 fraud losses at $46.3 billion (USD), up 15% from 2011. As a percent of global telecom revenues, fraud losses are approximately 2.09%, a 0.21% increase from 2011. The main reason for the relative increase in fraud is due to more fraudulent activity targeting the wireless industry [4]. It should also be noted that the relative decrease in fraud, that was apparent in previous surveys, was not an actual decline in absolute values but it had been attributed to the fact that the growth in global telecom revenues had outpaced the growth in fraud losses in the past (e.g. the CFCA 2011 survey).

Exchange of ideas in fraud detection is limited by the fact that it makes no sense to describe the methods in detail, as it gives fraudsters the information they require to evade detection. Moreover, companies and organizations that have been defrauded refrain from revealing the situation due to reputation concerns. Adding to this, fraud detection problems involve huge data sets, which are constantly evolving. Data sets can be as large as tenths of thousands of calls per weekday for a large organization with 3 or 4 thousand employees, to hundreds of millions of calls for national carriers. One should also consider the size of the related metadata.

Another difficulty with fraud detection is the fact that, nowadays, the term telecommunications is wider than ever. It includes both wired and wireless systems, mobile and cellular systems, legacy systems (PSTN, ISDN), terrestrial and satellite networks, and a plethora of Internet–based communication applications. The diversity of network types and applications, along with the deregulation of the market and the relocation of services to the cloud, makes fraud detection a complex task.

Research in telecommunications fraud detection is mainly motivated by fraudulent activities in mobile technologies [3,5]. Recent research also focuses on VoIP technologies [6]. Fraud detection methods can be based on statistical or machine learning techniques and may be supervised or unsupervised [7,8,9]. Fawcett and Phua [10] have ensembled the bibliography on the use of data mining and machine learning methods for automatic fraud detection up to 2005.

This paper proceeds as follows: In the next Section the user modeling procedure and the proposed user profiles are presented. A brief presentation of the clustering techniques and the clustering quality statistics that are used is given in the third Section. The user data and the outcome of the analysis are described in the fourth Section. In the last Section conclusions are discussed.

2 Behavior Modeling and Profiling in Telecommunications

The data that can be used to construct the basic profile vector for a telecommunications user are contained in the Call Detail Record (CDR) of any Private Branch Exchange (PBX) or any VoIP switch. The format of the CDR varies among providers or programs. Some programs allow CDRs to be configured by the user. In most cases a CDR contains at least data such as: the caller ID, the chargeable duration of the call, the called party ID, the date and the time of the call, etc [11]. In mobile telephone systems, such as GSM, the data records that contain details of every mobile phone attempt are the Toll Tickets. Location data may also be useful. In computer security, user profiles may be constructed based on any basic usage characteristic such as resources consumed, login location, typing rate and counts of particular commands.

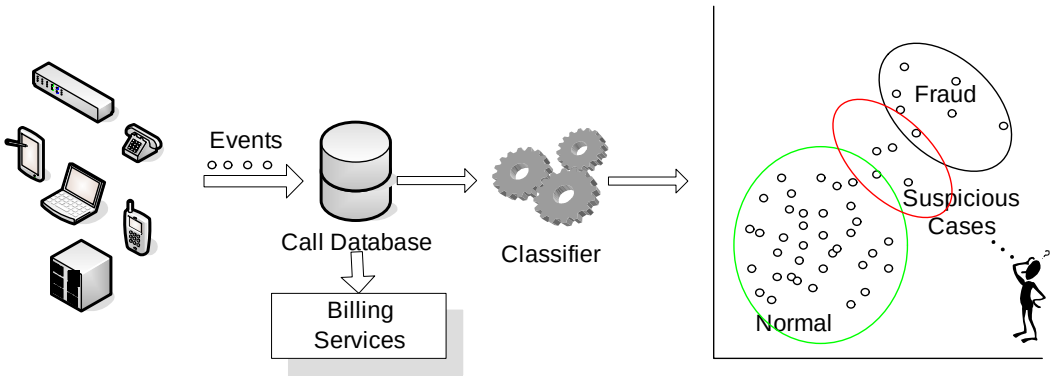


Fig. 1: Both the billing and the security systems draw usage related data from a common database.

CDR data are primarily collected in order to charge the user for the service he gets, but may additionally be used to manage the security of the system. The study of CDRs may reveal cases of unauthorized use of the system which may burden both the provider and the subscriber (Fig.1). The system administrator may investigate unusual call patterns, multiple calls to the same number, calls outside normal working hours, calls with long duration, phone calls to suspicious destinations such as international or premium-rate services (e.g. 090, 1-900, 900). Also, in the case of private PBXs, where access to outgoing destinations is done by means of personal authorization codes (PAC), one may check for frequent or simultaneous use of a PAC, attempts to import nonexistent system codes (efforts to find codes), or using PACs that are not associated with a user [12].

In order to develop models that can be applied to distinguish legitimate from fraudulent usage, one needs examples of both cases. Finding data of normal usage is easy. In fact, most of the data generated during the operation of a system are of this type. The data relating to fraud are relatively rare and any effort to characterize them in detail is difficult, time consuming, and requires specialized knowledge. Moreover, the processing and storage of user data is subject to restrictions by the legislation that protects privacy [13, 14].

Several methods appear in the literature for collecting and classifying data usage. An approach that involves the user in the process, and thus may overcome privacy concerns, is *block crediting*, proposed by Fawcett and Provost [15]. According to this method both the provider and the subscriber are involved in characterizing the data. If the subscriber has any objection on the amount of his bill, he may contact the provider and collaborate in order to characterize the calls in his detailed bill and then pay for the corresponding price. Typically, the account is divided into two periods, one with the calls of the legitimate user and one that also contains calls from the usurper. However, this kind of separation does not provide high accuracy in the characterization of cases. On the other hand, a detailed per phone call characterization of the account is by nature expensive and requires too much human intervention.

In general, fraud detection focuses on the analysis of users' activity and the related approaches are divided into two main subcategories. The absolute one that searches for limits between legal and fraudulent behavior, and the differential approach that tries to detect extreme changes in a user's behavior. All cases of telecom fraud can actually be viewed as fraud scenarios which are related to the way the access to the network was acquired.

One of the most interesting aspects of the problem is the evaluation of different user representations (profiles) and their effect towards the proper discrimination between legitimate and fraudulent activity.

In the present analysis, for each user, three different profile types are constructed and tested. The first profile (Profile1) is build up from the accumulated weekly behavior of the user. The profile consists of seven fields which are the mean and the standard deviation of the number of calls per week (calls), the mean and the standard deviation of the duration (dur) of calls per week, the maximum number of calls, the maximum duration of one call and the maximum cost of one call (Fig.2). All maxima are computed within a week's period.

The second profile (Profile2) is a detailed daily behavior of a user which is constructed by separating the number of calls per day and their corresponding duration per day according to the called destination, i.e., national (nat), international (int), and mobile (mob) calls, and the time of the day, i.e., working hours (w), afternoon hours (a), and night (n) (Fig.3).

Last, the third profile (Profile3) is an accumulated per day behavior (Fig.4). It consists of the number of calls and their corresponding duration separated only according to the called destination, that is, national, international and mobile calls.

The last two profiles were also accumulated per week to give Profile2w and Profile3w. So, overall five different user profile representations are evaluated.

mean(calls)	std(calls)	mean(dur)	std(dur)	max(calls)	max(dur)	max(cost)

Fig. 2: Profile1 of telephone calls

nat_calls_w	nat_dur_w	nat_calls_a	nat_dur_a	nat_calls_n	nat_dur_n
mob_calls_w	mob_dur_w	mob_calls_a	mob_dur_a	mob_calls_n	mob_dur_n
int_calls_w	int_dur_w	int_calls_a	int_dur_a	int_calls_n	int_dur_n

Fig. 3: Profile2 of telephone calls

nat_calls	nat_dur	mob_calls	mob_dur	int_calls	int_dur

Fig. 4: Profile3 of telephone calls

In the past, feed-forward neural networks have applied to classify cases of user behavior [16]. No matter how well a neural network classifier may have performed, there is no clue about the features it actually used in order to achieve its performance. So, it is difficult to identify important characteristics that led to a successful classification.

In order to further investigate the problem of appropriate user modeling towards fraud detection, we also apply clustering techniques on the data. The aim is to test whether cases from the same class tend to form clusters and under which condition.

An important characteristic of most clustering methods is that they are actually unsupervised learning approaches. Thus, one does not have to provide the corresponding algorithms with class examples. On the contrary, the algorithm is left to decide on case similarities by means of a pre-selected distance (similarity) measure [17].

It is hoped that clustering will unveil important information on the nature of user data and on the key features that may be used to distinguish between legitimate and fraudulent usage.

3 Clustering Techniques and Clustering Quality Measures

Clustering is one of the most important sets of unsupervised learning tools among the machine learning techniques and algorithms. It deals with finding a structure or an intrinsic grouping in a collection of unlabeled data. During the clustering procedure some objects are labeled “similar” to each other and “dissimilar” to objects belonging to other clusters. So, clustering aims on organizing objects into groups whose members are similar in some way. The researcher needs to decide on the similarity criterion that will be used during the process as well as on the clustering quality criterion that will be applied to decide what constitutes a good clustering. Clustering has been applied to a wide range of topics, such as pattern recognition, compression, and classification, as well as in diverse disciplines like biology, marketing, psychology and business. For a detailed introduction of clustering techniques the reader is referred to Kaufmann and Rousseeuw [17].

Two of the most common clustering techniques are used in the present work, namely the partitioning and the hierarchical clustering. As a main representative of the partitioning techniques we will apply the k-means algorithm. The hierarchical clustering technique that will be used is the agglomerative clustering. Their main difference is that the first needs the user’s input on the expected number of clusters while the latter needs no user intervention.

3.1 K-means clustering

The main idea behind the k-means algorithm is that in order to obtain k clusters, the algorithm selects k objects from the data sets. The remaining objects are then assigned to the nearest representative object, as the algorithm attempts to minimize the average squared distance between objects. Other distance measures can be used as well.

A graphical representation of the clustering is provided by displaying the silhouettes introduced by Kaufmann and Rousseeuw [17]. Silhouettes are constructed in the following way. Given the number of clusters in the problem, the value $s(i)$ is defined for each object i and these numbers are presented in a plot. The value $s(i)$ is defined by:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{1}$$

where $a(i)$ is the average dissimilarity of i to all other objects in A, where A is the cluster in which the object i is first assigned, with $d(i,C)$ being the average dissimilarity of i to all objects of any other cluster C different from A. Object i is assigned to the cluster B that satisfies $b(i) = d(i,B)$.

3.2 Agglomerative Clustering

During hierarchical agglomerative clustering the user does not specify the expected number of clusters k. Instead, the algorithm constructs a tree-like hierarchy, a dendrogram, which implicitly contains all values of k. The root of the tree structure defines a cluster that contains all data, while its leafs represent n clusters, each one containing one of the n objects. The agglomerative clustering algorithm starts with each object representing a cluster, called a singleton, and proceeds by fusing the

closest ones until a single cluster is obtained. Therefore, a measure of dissimilarity between two clusters must be defined [17].

Two different distance measures, namely the Euclidean distance and the correlation between objects, are used for both clustering algorithms.

3.3 Clustering Quality Measures

In order to judge on the quality of the clustering structure, we apply three quality statistics. These are the *silhouette coefficient (SC)*, the *agglomerative coefficient (AC)* and the *cophenetic coefficient (CC)*. Visual inspection of the dendrograms and the *consistency* of their structure may also reveal certain characteristics.

The *silhouette coefficient* (*SC*) is given by:

$$SC = \max_k \bar{s}(k), \tag{2}$$

where $s(k)$ is derived from (1) and the maximum is taken over all k for which the silhouettes can be constructed. The *SC* is a useful measure of the amount of clustering structure that has been discovered by the classification algorithm.

Another measure, that allows us to decide whether the clustering reveals a clear separation between groups, is the comparison of the within–cluster dissimilarity with the between–cluster dissimilarity. This measure in a tree structure is equivalent to the tendency that the tree branches get taller. If for each object i one measures its distance, $l(i)$, from the tree root, normalize the value in [0-1] and compute the average value, then he has a measure of the average tree height. This average, whose larger values imply better clustering, is named *agglomerative coefficient* (*AC*), and it is given by:

$$AC = \frac{1}{n}\sum_{i=1}^{n} l(i), \tag{3}$$

The *cophenetic coefficient*, *CC*, is a measure of how faithfully a dendrogram preserves the pair-wise distances between the original un-modeled data points. Given the matrix Z with the distances between clusters and the object dissimilarity matrix Y, then *CC* can be used as a measure of correlation between the object distances and their distance from the tree root. The value of *CC* is given by:

$$CC = \frac{\sum\limits_{\substack{i,j \\ i<j}} \left(Y_{ij} - \hat{Y}\right)\left(Z_{ij} - \hat{Z}\right)}{\sqrt{\sum\limits_{\substack{i,j \\ i<j}} \left(Y_{ij} - \hat{Y}\right)^2 \sum\limits_{\substack{i,j \\ i<j}} \left(Z_{ij} - \hat{Z}\right)^2}}, \tag{4}$$

where Y_{ij} is the distance of object i from j. *CC* values closer to 1 reveal better clustering structures [18].

One last criterion that may help choose in favor of one clustering structure over another is the *consistency* of the structure. If a branch in a dendrogram is at the same height with its neighbors, then this reveals similarity between the corresponding objects at that level of the structure. This behavior can be considered consistent. Whenever a branch height significantly differs from that of its neighbors, this behavior reveals high values of dissimilarity between the corresponding objects and can be named inconsistent. Inconsistent links in a dendrogram may reveal cluster separation points. For a more detailed analysis on cluster selection criteria the reader is referred to [19].

4 An Application of User Profile Clustering

Our experiments are based on real data extracted from a database that holds the CDR for a period of eight years from an organization's PBX. According to the organization's charging policy, only calls to national, international and mobile destinations are charged. Calls to local destinations are not charged so they are not included in the examples. In order to properly charge users, for the calls they place, a system of Personal Identification Numbers (PIN) is used. Each user owns a unique PIN which "unlocks" the organization's telephone sets in order to place costly outgoing calls. If anyone (e.g., a fraudster) finds a PIN he can use it to place his own calls from any telephone set within the organization.

Several user accounts, which have been defrauded, have been identified. The detailed daily accounts were examined by a field expert and each phone call was marked as either normal or defrauded. If during a day no fraudulent activity was present, then the whole day was marked as normal. If at least one call from the fraudster was present, then the whole day was marked as fraud. This separation of the classes is named *detailed characterization*. Adding to this, each day was also marked, according to the first time that fraudulent activity appeared. Thus, each user's account is split into two sets, one pre- and one post-fraud. This separation of the classes is named *coarse characterization*.

In this work we use the examples of 6 users (it has been stressed earlier that it is difficult to isolate fraud cases), 5 profile representations, and 2 different ways to characterize the user accounts as normal of fraudulent. The above give 60 different data sets.

The user daily profiles will be used as an input for the two algorithms. We will ask the k-means algorithm to partition the input space in two distinct groups. If the legitimate and the fraudulent behavior cases are sufficiently different from each other, then the k-means algorithm will provide us with two distinct clusters of data. If this is not the case, then the division will not be good. Then the same input data will be fed into the agglomerative clustering algorithm and we will also check whether there is an output with distinct cases separation.

Table 1: Clustering outcome for a defrauded user account using the five profiles and two similarity measures, Euclidean distance (Eucl) and Correlation (Correl)

Profile	Distance Measure	Correct Clustering (%)	*SC*	*AC*	*CC*
Profile1	Eucl	77.9310	0.7734	0.9632	0.9143
Profile1	Correl	**84.8276**	**0.8729**	0.9881	0.9299
Profile2	Eucl	78.0374	0.5303	0.9738	0.9514
Profile2	Correl	67.2897	0.2624	0.9952	0.9698
Profile3	Eucl	67.7570	0.8104	0.9824	0.9160
Profile3	Correl	76.1682	0.7668	**0.9984**	**0.9725**
Profile2w	Eucl	78.0374	0.5264	0.9738	0.9514
Profile2w	Correl	77.1028	0.3429	0.9952	0.9698
Profile3w	Eucl	71.2329	0.7533	0.9748	0.9338
Profile3w	Correl	77.3973	0.7296	0.9956	0.9489

Table 2: Clustering statistics for six defrauded user accounts and two similarity measures – Profile1 has been used in all cases

Profile	Distance Measure	Correct Clustering (%)	*SC*	*AC*	*CC*
User1	euclidean	77.9310	0.7734	0.9632	0.9143
	correlation	**84.8276**	0.8729	0.9881	0.9299
User2	euclidean	65.2582	0.9466	0.9770	0.9118
	correlation	**78.8732**	0.7152	0.9975	0.9232
User3	euclidean	**73.3154**	0.9989	0.9944	0.9910
	correlation	67.1159	0.5847	0.9945	0.6782
User4	euclidean	79.4979	0.8570	0.9706	0.8857
	correlation	**92.4686**	0.6865	0.9949	0.8302
User5	euclidean	62.1367	0.5784	0.9544	0.6278
	correlation	**79.4149**	0.8750	0.9706	0.8785
User6	euclidean	76.9031	0.7632	0.9612	0.8949
	correlation	**86.4668**	0.7445	0.9903	0.8302

In Table 1 the outcome of the clustering procedure for a defrauded user's account is listed for the five profiles and two similarity measures. These are the results for User1. In general, the best results, i.e. the higher percentage of correct clustering, were achieved with the use of Profile1 combined with correlation as the similarity measure.

In Table 2 the values of the statistics of Section 3.3 are presented for the cases of four defrauded user accounts. For each user the similarity measure and the correct clustering percentage are also given. Bold numbers are used to stress best statistics values. In general, correlation gave better results. The case of User3 (Table 2) is an exception that will be dealt with shortly. Due to space limitations only the findings for Profile1 will be depicted in the figures that follow.

First, the clustering ability of each profile is tested by means of the *k*-means algorithm using the Euclidean distance as a similarity measure. Comparison of the clustering outcome with the original class separation reveals that the clustering is correct by 77.9%. However, visual observation of the silhouette diagrams (Fig.5(a)) shows the existence of negative silhouette values at the bottom of the figure which imply wrong clustering. The *SC* for the clustering in this figure is 0.7734. If the procedure is repeated but with the correlation of the input vectors as the similarity measure, then the outcome is depicted in (Fig.5(b)). Now, the percentage of correct clustering is 84.8% and the *SC* value is 0.8729.

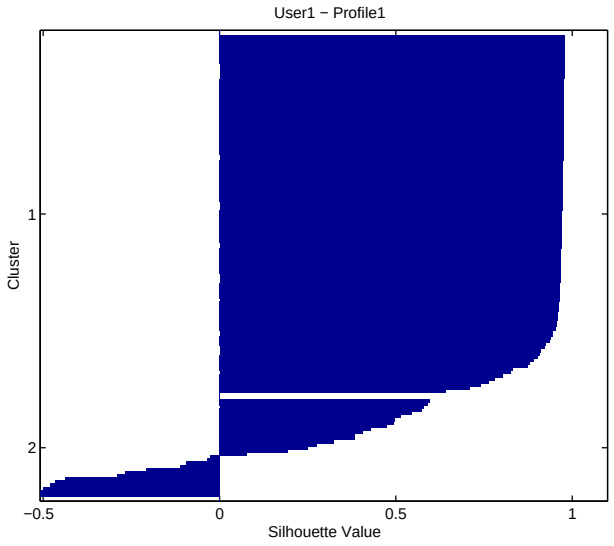
In general, the *k*-means clustering showed that better separation between classes is achieved when user behavior is modeled with Profile1, where the detailed characterization of the cases is used and correlation is employed as the similarity measure.

Examples of the agglomerative clustering of the same user's behavior and the same profile (Profile1) are shown in Fig.6. The former plot (Fig.6(a)) is a case where the Euclidean distance is used as measure of the similarity between objects, while in the latter (Fig.6(b)) the objects' correlation is used instead. Fig.6 is reprinted here from [8] with permission from Elsevier.

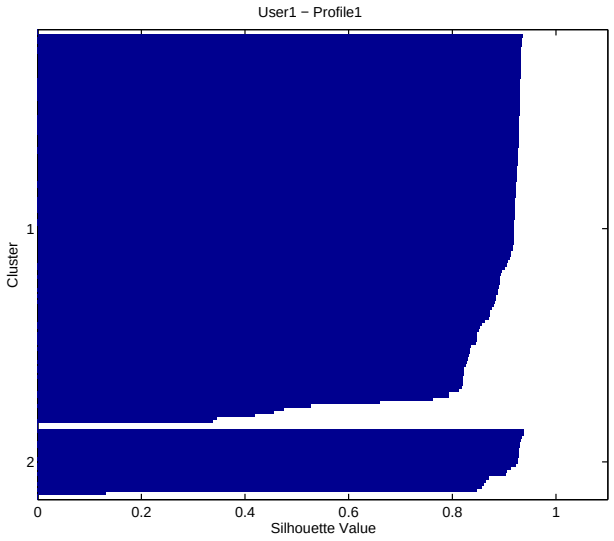
In all cases the analysis reveals that whenever the correlation is used as the distance measure then one gets the highest percentage of correct clustering. Visual inspection of the dendrograms is also important. The difference of correct clustering between the dendrograms in Fig.6(a) and Fig.6(b) is only 7%. However, each one reveals two completely different views of the data. In the first one, one distinct cluster is formed that contains all the outliers regardless of their class membership. In the second, clusters are formed which include only pure fraud cases. In this case there are three clusters that contain the fraud cases [8]. This implies a type of mixed behavior of the fraudster. Actually, a dendrogram structure implicitly contains all possible cluster separations and is up to the analyst to decide the correct number of clusters and the point of separation.

It is interesting to study the silhouette diagrams in conjunction with their agglomerative clustering counterparts. Especially in the case where the Euclidean distance is used, one may observe that the silhouette diagram shows some kind of "bad" clustering that may mislead the researcher, but its corresponding dendrogram conveys a behavior that needs to be further explored. In fact, it was an expert's intervention, who examined all the cases in the separate cluster and revealed that these belong to outliers, i.e. rare cases of extreme network usage by the legal user or the fraudster.

Fig. 7, which is reprinted here from [8] with permission from Elsevier, shows the dendrogram for User3 of Table 2 when the correlation is used as the similarity measure. From Table 2 one may observe that for the case of User3 the clustering statistics for correlation are worst than the ones for the Euclidean distance similarity measure. However, a distinct cluster with pure legal behavior is formed ((Fig. 7)). On the contrary, the silhouette diagram for the same user with the Euclidean distance, gives two clusters where the second one has only one member. Then the first one includes all legitimate usage cases and gives the high (73.31) correct clustering percentage.

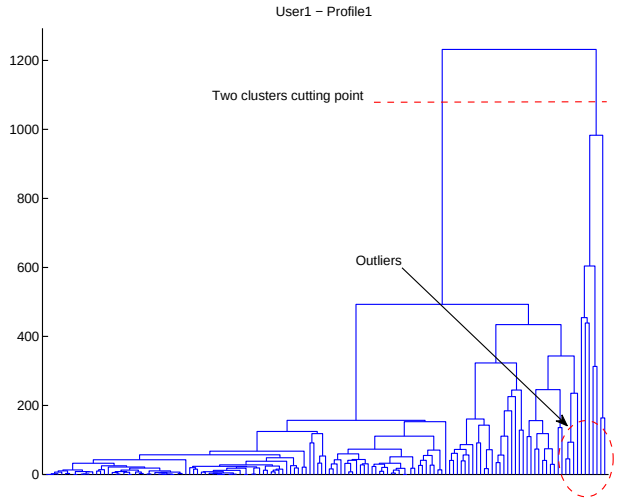


(a)

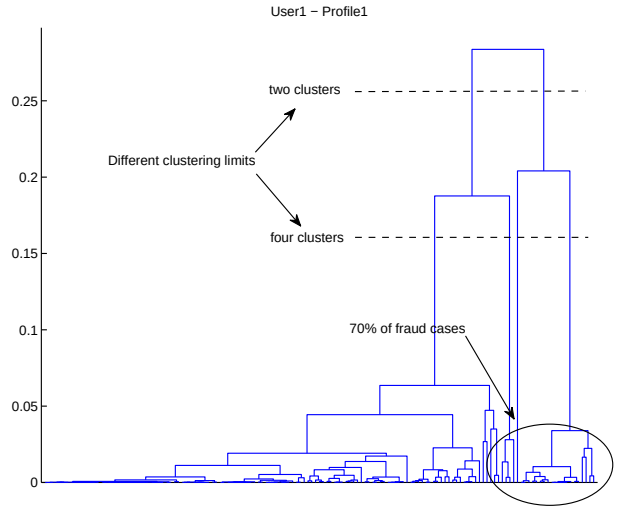


(b)

Fig. 5: Silhouette diagram for the clustering of a user's profile into two clusters using (a) the Euclidean distance as a similarity measure – the negative silhouette values at the bottom of the figure imply wrong clustering of the user behaviour, and (b) correlation as a similarity measure – here two well separated clusters are formed



(a)



(b)

Fig. 6: Hierarchical agglomerative clustering of a user's profile using as similarity measure (a) the Euclidean distance between objects – revealing outliers, and (b) the correlation between objects – revealing distinct case groups

5 Conclusions and Discussion

As large companies, organizations or institutions seek to grow their customer base and revenue, they must increasingly combat sophisticated security threats while navigating the growing challenge of compliance risk related to adhering with national or international laws. Business needs drive an unprecedented demand for complex IT networks and applications that call for the establishment of authentication and fraud detection solutions throughout the enterprise. To effectively improve security and compliance controls, total cost of ownership and the end-customer experience, companies are adopting identity-based security and fraud detection platforms that span enterprise-wide.

In particular, for the telecommunications industry, fraud detection is important because companies and suppliers of telecommunications services lose a significant proportion of their revenue as a result. As regards large organizations and companies, problems with internal fraudulent activities, i.e. misuse of the corporate

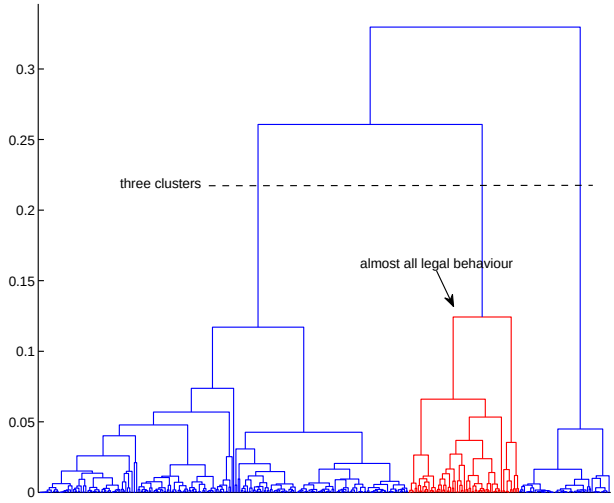


Fig. 7: The dendrogram for User3 of Table 2 when the correlation is used as the similarity measure. A distinct cluster with all legal behavior is formed (reprinted from [8] with permission from Elsevier)

network or usurpation of colleagues' identities burden the company's budget or degrade the management reputation among employees. Adding to these, the modeling and characterization of users' behavior in telecommunications can be used to improve network security, improve services, provide personalized applications, and optimize the operation of electronic equipment and/or communication protocols.

In this context, we present how some well established unsupervised learning techniques may be applied on telecommunications data in order to provide us with more insight on both a legitimate user's and a fraudster's behavior. Prior to this, raw usage data must be transformed into appropriate user profiles. The profile construction and selection is also a challenge.

From the analysis it is concluded that, as regards user profile building, accumulated characteristics of a user yield better discrimination results. However, aggregating user's behavior for larger periods than a week was avoided in order to preserve some level of on-line detection ability.

Clustering reveals that misclassification occurs due to mixed types of behavior. That is, there are cases for which the legitimate user acts like a fraudster, e.g. making long or expensive calls, while at the same time a fraudster may act like a normal user, e.g. he is more refrained and tries to mimic legitimate user behavior. So, if one forces the clustering procedure to produce two distinct groups of data (one with the legitimate and one with the fraudulent usage) it will fail to do it correctly.

It was evident from the analysis that there was also some kind of "noise" in the data. This is due to the fact that even the field expert, that manually characterized the cases in the first place, had a hard time distinguishing between the two classes. This observation reveals the complexity in a fraud detection problem. This "grey" area in the classification problem is depicted as the suspicious cases in Fig.1; a space between pure legal and fraudulent activity.

Taking into account the aforementioned observations, i.e. the mixed types of behavior and the noisy data, it is concluded that clustering alone can not be used as the only decision technique on whether a user account is defrauded or not. Input from other methods is also needed, and a more complex framework should be used. Nonetheless, clustering may provide the researcher with invaluable insight on the nature of data and on the nature of user behavior.

The user profiles that were presented, in this work, have the benefit of respecting users' privacy. That is, except from some coarse user behavior characteristics, all private data (e.g. called number, calling location, etc) are hidden from the analyst. Private data would definitely enhance the accuracy of fraud detection. In fact, the expert who characterized the data sets, in the first place, used rules based on private data and his domain specific knowledge. However, our aim is to test the ability to detect fraud given the minimum possible information about the user.

A fundamental aspect of any predictive problem in data analysis is the choice of an appropriate criterion for estimation and performance assessment. In the case of fraud, one needs, in particular, to combine both classification accuracy and timeliness of classification. Moreover, the granularity of the timeline of events plays an important role in the predictive ability of models. This, also, means that standard measures of classification performance, such as the error rate, AUC, KS statistic, information value, etc, may not be sufficient. This is evident in the present work by the fact that it is difficult to judge on the most appropriate clustering outcome just by evaluating similarity statistics. As can be seen in Table 1 the percentage of correct clustering is a vital criterion, while visual observation of the corresponding graphs is also needed. One of the main tasks in future research should be, apart from the selection of appropriate user models (profiles), the effort to design suitable measures and performance curves which combine these aspects.

The development of generalized frameworks that would be capable of adapting to different environments is highly desired. The presented methodologies may, also, be applied to study similar problems in mobile or data networks. The first step should be the appropriate user modeling, i.e. the selection of the appropriate attributes for profile building. Towards this step, an expert's intervention is highly needed. Adding to this, due to privacy concerns and restrictions, the user's collaboration may also be asked for. The user may give his consent for the analysis of his account which helps overcome some legislation problems regarding privacy. His collaboration may also help the analyst to identify fraudulent activity more precisely.

Nonetheless, one should keep in mind that security problems often depend on the specific nature of each problem and there is a risk when directly generalizing the findings even to similar environments. Among other risks, there underlies the danger of information leakage between interconnected IT systems which may burden users' privacy.

Future research includes the application of other clustering methods like the subtractive clustering, or other unsupervised techniques like the SOM, on the problem. Social network analysis is also of great interest and seems like an appealing alternative to statistical approaches or computational intelligence ones. What is intriguing about its application on the problem is that we expect to approach user behavior not by means of some usage statistic but by detecting transactions between the persons behind the statistics. Recent evidence suggests that technical controls only detect one third of fraud cases with zero time exposure and loss. More complex fraud is detected with a range of technical and sociotechnical controls from inside and outside the firm [20].

It is hoped that further research will give important findings on how to distinguish between legitimate and fraudulent network usage. Moreover, the departure from a strict technical approach and the use of social and behavioral controls used in the organizational environment may provide more clues about the problem.

Acknowledgement

The authors would like to thank the members of the Telecommunications Centre of the Aristotle University of Thessaloniki for their contribution of anonymised user data. The authors wish to acknowledge financial support provided by the Research Committee of the Technological Education Institute of Central Macedonia (Serres) under grant SAT/IC/181212-157/7

References

[1] Hildebrandt, Mireille and James Backhouse (eds.), D7.2: Descriptive analysis and inventory of profiling practices. FIDIS Deliverable, available online at: http://www.fidis.net/resources/deliverables/profiling/ (2005).

[2] Adomavicius, G., Huang, Z., and Tuzhilin, A., Personalization and Recommender Systems. In Z.-L. C. a. S. Raghavan (Ed.), State-of-the-Art Decision Making Tools in the Information-Intensive Age (pp. 55-100): Tutorials in Operations Research (2008).

[3] P. Gosset and M. Hyland, Classification, detection and prosecution of fraud in mobile networks, Proceedings of ACTS Mobile Summit, Sorrento, Italy, June (1999).

[4] Communications Fraud Control Association. 2013 Global Fraud Loss Survey, Press Release, Roseland, NJ (CFCA) October 10, 2013, available online: http://www.cfca.org/press.php (2013).

[5] Moreau Y., Preneel B., Burge P., Shawe-Taylor J., Stoermann C., Cooke C., Novel Techniques for Fraud Detection in Mobile Telecommunication Networks, ACTS Mobile Summit, Granada Spain (1997).

[6] Kapourniotis, T. Dagiuklas, T., Polyzos, G., Alefragkis, P., Scam and fraud detection in VoIP Networks: Analysis and countermeasures using user profiling, 50th FITCE Congress (FITCE), Palermo Italy, doi: 10.1109/FITCE.2011.6133427, (2011).

[7] R. J. Bolton and D. J Hand, Statistical fraud detection: a review, Statistical Science, **17.3**, 235255, (2002).

[8] Constantinos S. Hilas and Paris As. Mastorocostas, An Application of Supervised and Unsupervised Learning Approaches to Telecommunications Fraud Detection, Knowledge-Based Systems, **21**, 721 726, doi:10.1016/j.knosys.2008.03.026, (2008).

[9] Olszewski, Dominik, A probabilistic approach to fraud detection in telecommunications, Knowledge-Based Systems, **26**, 246-258, (2012).

[10] Fawcett, T. and Phua, C., Fraud Detection Bibliography, Accessed From: http://liinwww.ira.uka.de/bibliography/Ai/fraud.detection.html (2005).

[11] S. F. Hinde, Call Record Analysis, Making Life Easier - Network Design and Management Tools (Digest No: 1996/217), IEE Colloquium on, 8/1 8/4, (1996).

[12] DEFINITY Enterprise Communication Server (ECS): Technical Articles and Technical Tips, available online: http://esearch.avaya.com/

[13] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal L 281/31 of 23.11.95, pp 31-50, available online at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX: 31995L0046:EN:NOT (1995).

[14] Directive 2002/58/EC. Directive on privacy and electronic communications. Official Journal L 201, 31.7.2002, 3747, available online at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do? uri=CELEX: 32002L0058:EN:NOT (2002)

[15] Fawcett, T. and F. Provost, Adaptive fraud detection, Journal of Data Mining and Knowledge Discovery, **1.3**, 291316 (1997).

[16] Constantinos S. Hilas, and John N. Sahalos, Testing the fraud detection ability of different user profiles by means of FFNN classifiers, in Collias St. et al ed.. Lecture Notes in Computer Science, **4132**, Part II, Berlin-Heidelberg: Springer Verlag, 872-883 (2006).

[17] Kaufman L., and Rousseeuw P.J., Finding groups in data: an introduction to cluster analysis, New York: John Wiley & Sons, Inc. (1990).

[18] Rohlf, F. J. and David L. Fisher, Test for hierarchical structure in random data sets, Systematic Zool., **17**, 407-412 (1968).

[19] Witten I. H., and Frank E., Data mining: practical machine learning tools and techniques with Java implementations, London: Academic Press, (2000).

[20] Goode, S., and Lacey, Detecting complex account fraud in the enterprise: The role of technical and non-technical controls, Decision Support Systems, **50.4**, 702-714 (2011).

Constantinos S. Hilas received the B.S. degree in Physics, in 1989, the M.S. degree in Electronic Physics - Radioelectrology, in 1996, and the Ph.D. degree in Physics, in 2007, from the Aristotle University of Thessaloniki (AUTH), Greece. He also received the M.S. degree in Information Systems from the University of Macedonia (UM), Thessaloniki, Greece, in 2000. From 1993 to 2002, he was Computer Administrator and the Administrator of the Telecommunications Network of AUTH. Since 2002, he has been with the Technological Educational Institute of Central Macedonia (Serres, Greece), where he is currently an Associate Professor in the Dept. of Informatics Engineering. His basic research interests include user characterization, service development, and data mining techniques in telecommunications and networking.

Paris A. Mastorocostas received the Diploma and Ph.D. degrees in Electrical & Computer Engineering from Aristotle University of Thessaloniki, Greece. Presently he serves as Professor at the Dept. of Informatics Engineering, Technological Education Institute of Central Macedonia, Serres, Greece. His research interests include fuzzy and neural systems with applications to identification and classification processes, data mining and scientific programming.

Ioannis T. Rekanos received the M.Sc. and the Ph.D. degree in electrical and computer engineering, in 1993 and in 1998, respectively, both from the Aristotle University of Thessaloniki (AUTH), Greece. From 2000 to 2002, he was a Postdoctoral Senior Researcher in the Radio Laboratory at the Helsinki University of Technology, Finland. From 2002 to 2006, he served as a faculty member of the Department of Informatics and Communications, TEI of Serres, Greece. Since 2006, he has been with the AUTH, where he is currently an Associate Professor in the Department of Electrical and Computer Engineering. His research interests include electromagnetic and acoustic wave propagation, inverse scattering, computational electromagnetics, and digital signal processing.