CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2015; **00**:1–11
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe

A Novel Machine Learning Approach to the Detection of Identity Theft in Social Networks based on Emulated Attack Instances and Support Vector Machines

E. Villar-Rodríguez[1], J. Del Ser[1,*], A. I. Torre-Bastida[1], M. N. Bilbao[2] and S. Salcedo-Sanz[3]

[1]*TECNALIA RESEARCH & INNOVATION, P. Tecnológico Bizkaia, Ed. 700, 48160 Derio, Spain*
[2]*Department of Communications Engineering, University of the Basque Country UPV/EHU, 48013 Bilbao, Spain*
[3]*Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain*

SUMMARY

The proliferation of social networks and their usage by a wide spectrum of user profiles has been specially notable in the last decade. A Social Network is frequently conceived as a strongly interlinked community of users, each featuring a compact neighborhood tightly and actively connected through different communication flows. This realm unleashes a rich substrate for a myriad of malicious activities aimed at unauthorizedly profiting from the user itself or from his/her social circle. This manuscript elaborates on a practical approach for the detection of identity theft in social networks, by which the credentials of a certain user are stolen and used without permission by the attacker for its own benefit. The proposed scheme detects identity thefts by exclusively analyzing connection time traces of the account being tested in a non intrusive manner. The manuscript formulates the detection of this attack as a binary classification problem, which is tackled by means of a Support Vector classifier applied over features inferred from the original connection time traces of the user. Simulation results are discussed in depth towards elucidating the potentiality of the proposed system as the first step of a more involved impersonation detection framework, also relying on connectivity patterns and elements from language processing. Copyright © 2015 John Wiley & Sons, Ltd.

Received . . .

KEY WORDS: Identity theft; Social Networks; Machine Learning; Support Vector Machines

1. INTRODUCTION

In the last years social networks (SN) have become a widespread relational tool at both professional and personal levels, with usage statistics reportedly increasing at unprecedented rates in the history of Internet. Unfortunately, so has grown the interest of cyber-attackers in these platforms as means to increase the diversity and effectiveness of their malicious activities [1]. Facts speak by themselves: according to Digital Insights, over 500 million tweets are posted every day collaboratively amongst its more than 255 million active users [2]. As evinced by the unauthorized access to the details of approximately 250.000 Twitter users in early 2013 [3], such amount of data motivates cyber criminals to discover new procedures towards taking advantage and eventually exploiting the lack of knowledge and/or negligence of potential victims regarding good practices and policies in terms of information security.

Goals pursued by attacks in SN may reside not only in the economic profitability of the attacker, but also in other interests achievable by unauthorizedly accessing the information of the victim

*Correspondence to: TECNALIA RESEARCH & INNOVATION, P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain. E-mail: javier.delser@tecnalia.com.

(e.g. bullying or intimidation, particularly frequent within the teenage community). It is often the case that sensitive information items are carelessly posted in social networks, whose revelation may trigger dramatic consequences, security breaches and eventually fatal circumstances for the victim. Although the need for detection schemes specially tailored to attacks in social networks has been noted by the research community, contributions in this matter are relatively scarce. Furthermore, they hinge mostly on ad-hoc designed detectors for a certain attack class approach based mainly on analyzing private features from the user account (e.g. content of the messages or contact list).

From a more general point of view, motivations and goals for cyber-crimes may vary within a wide spectrum of possibilities that unchain an equally diverse portfolio of detection methods. In particular, phishing refers to those procedures used for broadcasting messages from apparently reputable sources succinctly devoted to capturing sensitive information such as account credentials or credit card details [4, 5, 6, 7]. Research on this class of attacks has gravitated on the application of textual phishing indicators [8, 9] and information retrieval algorithms such as Hidden Markov Models (HMM), Latent Dirichlet Allocation (LDA) or naïve bag-of-words procedures [10]. Other features used for the detection of phishing attacks have been found in the Internet search engines, which help finding inconsistencies between the fake and the authentic identity [11].

This work focuses rather on identity thefts, which are especially recurrent in communities of young users and teenagers. More generally, in a masquerade attack the criminal pretends to be an authorized user of a system in order to gain access or to be granted with more extended privileges than he/she is authorized for [12]. Identity theft is another related term coined to name the same sort of masquerading acts when the attack is performed through legitimate access identification in order to accomplish illegitimate activities. Those aforementioned attacks have drawn the attention of the research community, which has so far reported several fully dedicated systems that analyze specific characteristics of messages sent/left by the impostor.

This manuscript focuses on those scenarios in social networks with an unauthorized user making use of the victim's account by stealing his/her credentials, which results in the victim being impersonated. Nonetheless, an implicit assumption in all previous systems dealing with identity thefts is the interaction of the criminal with the compromised system/account, which does not match the entire range of privacy invasions intended by identity thefts. For instance, when held within communities of student teenagers one of the most reported purposes of identity thefts is to sneak the information of the victim in his/her profile, which does not involve any interaction of the attacker with the victim's account beyond the unauthorized usage of his/her credentials. On the contrary to former approaches, preliminary work recently published by the authors in [13] proposed to disengage from the particular purpose of the identity theft attack and instead focus on characterizing the usage profile of the potential victim in terms of privacy-aware features not controllable by the attacker him/herself. This methodological approach is supported by a premise: any behavioral deviation of the account usage with respect to its regular use may eventually correspond to an attacker using the stolen account in a different hence detectable fashion.

Features that could be used for the detection of identity thefts in social networks can be found in different domains depending on their level of privacy awareness. Scarce contributions can been found in the literature regarding the detection of compromised social network accounts [14, 15], where several features inferred from the social graph of the user are investigated as inputs of a detector for this class of attacks. However, as already anticipated some identity theft attacks are targeted at merely gossiping personal or sensitive information about the attacked user, i.e. without any proactive interaction of the attacker with the network. In this case, connection time statistics such as frequency or periodicity could be fairly discriminative so as to discern a regular connection time behavior from an unconventional usage schedule of the account, always subject to the erraticism of the user himself when accessing the network. Nevertheless, this early detection stage based on connection time statistics could serve as a trigger for alternate preventive mechanisms aimed at verifying the identity of the user logged in the social network.

In this context, this paper introduces a practical scheme for the aforementioned early identity theft detector, which is framed within a more involved 3-stage system each operating on gradually more intrusive feature sets. The goal of this first detector is to trigger an initial alarm of a potential

DOI: 10.1002/cpe

identity theft attack, alarm that could be subsequently fed to the other two detection stages. These secondary detectors would leverage the social graph of the user (via e.g. dynamic link grouping or centrality metrics) or the content itself (by turning to e.g. semantic and natural language processing procedures). The early-warning detection stage proposed in this paper 1) transforms the connection time information to a feature space yielding more condensed multidimensional profiles for the user under consideration; 2) trains a binary support vector machine (SVM [16]) classifier with the available feature history and synthetically generated yet realistic connection traces of potential attackers; and 3) estimates the false alarm and detection probabilities that reflect its performance. This work builds upon [13] and extends it by 1) describing the proposed detector in detail; 2) providing rationale on the need and subsequent construction of a synthetic validation dataset; and 3) discussing extensive simulation results aimed at verifying the performance behavior of the proposed scheme under different parameters of the attack itself and the set of emulated identity theft attacks.

The manuscript is structured as follows: first Section 2 formally poses the detection of identity thefts as a mathematical hypothesis testing problem. Next Section 3 and subsections therein delve into the user profiling scheme that lies at the core of the proposed detector, including the extraction of an alternate set of features from the connection time traces, the generation of emulated attack instances and the selection of the classifier. Section 4 describes the obtained simulation results over over the aforementioned validation dataset and, finally, Section 5 concludes the paper and outlines lines of future research.

2. PROBLEM FORMULATION

The detection of identity theft attacks in social networks can be conceived as a binary hypothesis testing problem where session times for user A are denoted by the time-variant vector $\mathbf{w}_t^A \triangleq \{w_{t,1}^A, \ldots, w_{t,N}^A\} = \{w_{t,n}^A\}_{n=1}^N$. In this definition, $w_{t,n}^A$ stands for the total duration of the session for user A sampled at a certain granularity given by parameters t and n. It should be emphasized that this nomenclature accommodates any granularity in the recording of the connection statistics on which the testing process is performed. The hypotheses being tested refer explicitly to the detection of a potential identity theft at time T^* given the recorded connectivity time information captured up to that given moment. Mathematically speaking, the detector should determine which of the hypotheses

$$\mathcal{H}_0\text{: user } A \text{ has NOT undergone any impersonation attack,} \tag{1}$$

$$\mathcal{H}_1\text{: user } A \text{ has undergone an impersonation attack,} \tag{2}$$

holds at time T^* by exploiting the connection time information stored in the matrix $\mathbf{W}_{T^*}^A \triangleq \{\mathbf{w}_t^A\}_{t=1}^{T^*}$. The fact that the hypothesis testing is performed at time T^* yields an implicit dependence of the two hypotheses on the information used for their testing, i.e.

$$\mathcal{H}_0^{T^*}\text{: user } A \text{ has NOT undergone any impersonation attack at time } T^*, \tag{3}$$

$$\mathcal{H}_1^{T^*}\text{: user } A \text{ has undergone an impersonation attack at time } T^*. \tag{4}$$

As mentioned before, this test can be approached from an algorithmic perspective as a binary classification problem where a predictive model is trained with the entries in $\mathbf{W}_{T^*}^A$ so as to classify correctly a new connection time trace as one between two types of classes: 1 (the new trace $\mathbf{w}_{T^*+1}^A$ can be regarded as an indicator of a potential identity theft attack) and 0 ($\mathbf{w}_{T^*+1}^A$ does not suggest any identity theft being performed on A's social network account). The model that relates connection time traces to labels 0 and 1 can be built from the record $\mathbf{W}_{T^*}^A$ of connection time traces under the assumption that no identity theft has been committed during its time frame. Unfortunately, when dealing with this kind of subtle attacks counter-examples for possible identity thefts that would be needed for training a balanced supervised classifier are difficult to register and infer in practice. As a workaround two strategies can be adopted:

DOI: 10.1002/cpe

A. To train the model exclusively with $\mathbf{W}^{A}_{T^*}$, yielding a particular instance of the so-called one-class classifier.
B. To synthetically construct connection time traces representing events of identity thefts over A's account driven by the casuistry of this kind of attacks.

The work presented in this paper follows the strategy B since, when lacking of samples corresponding to identity thefts, one-class classifiers can overfit the feature space spanned by user A under analysis, yielding a potentially low probability of detection due to the misclassification of true identity thefts as no attacks. This being said, the set of features extracted from $\mathbf{W}^{A}_{T^*}$ and the classifier itself should render a high rate of *true positives* (i.e. confirmed attacks should be detected as resiliently as possible) and a low rate of *false positives* (correspondingly, wrongly detected attack events or *false alarms*). In what follows these detection performance indicators will be expressed as $P^{T^*}_{d} \triangleq Pr\{\mathcal{H}^{T^*}_{1}|\mathcal{H}^{T^*}_{1}\}$ and $P^{T^*}_{fa} \triangleq Pr\{\mathcal{H}^{T^*}_{1}|\mathcal{H}^{T^*}_{0}\}$, respectively.

3. PROPOSED USER PROFILING APPROACH

User profiling refers to those processes aimed at inferring properties of a certain user-generated dataset towards developing a user model well-suited for subsequent classification, prediction or clustering stages. In this context *habits* are those regular patterns within the properties best describing a normal or expected behavior of the user based on the information contained in the dataset. When used as properties for the detection problem tackled in this paper, connection time traces also convey valuable information of the behavioral patterns of the user under analysis. In other words, connection time statistics of his/her social network account may reflect daily activities which may constitute peculiarities worthy of being established as early indicators of behavioral changes, from e.g. users who never connect within their working hours to those who generate short connections scattered throughout daytime.

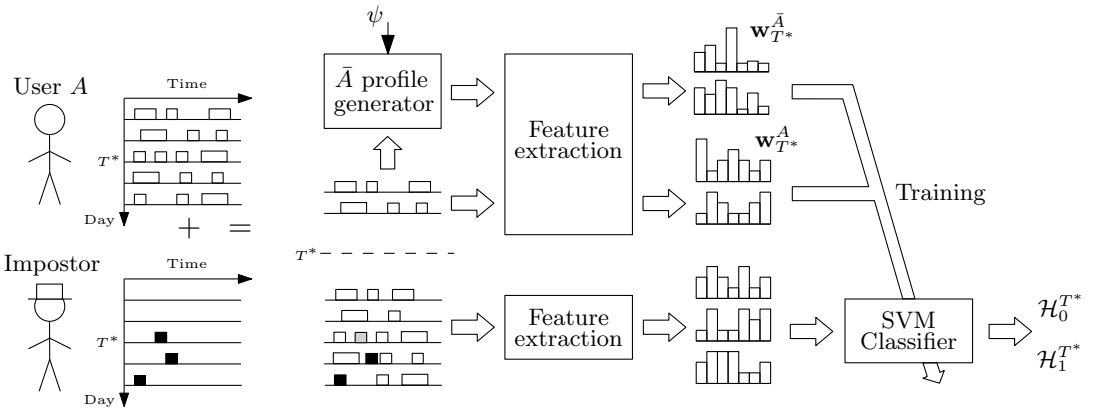

Figure 1. Proposed connection time based detector of impersonation attacks.

Under this assumption, a connection-based system for impersonation detection has been built according to the scheme depicted in Figure 1. Connection time information for user A is recorded at a sampling granularity driven by t and n which needs a proper adjusting in an attempt to find a balance between losing relevant data and not reaching an over-fitted model that triggers an alarm under small yet not necessarily meaningful behavioral outliers. A behavior is originally modeled as the hours at which such connections have taken place and their particular duration. In order to determine if a behavior with regard to connections is suspicious according to the learned patterns, a precise sampling granularity is required to fairly represent the connection behavior of the user. However, it should not be set as stringent as to consider every minor change or perturbation as a potential impersonation attack.

DOI: 10.1002/cpe

Taking into account this trade-off the proposed user profiling scheme generates a feature set based on connection information aggregated on an hourly basis. This permits collecting general measures over connection frequencies and durations avoiding that an one-hour deviation in the connection habits could produce misclassified instances. This feature space characterizes the essence of the behavioral patterns not assuming that a user should be regular enough to keep a severe and rigid daily connection schedule with less-than-one-hour deviations. Therefore, the meaningfulness of the original connection time traces is captured by a feature transformation with a two-fold aim: 1) to narrow down the high-dimensional feature space resulting into a lower computational complexity for the classifier and 2) to abstain from feeding the SVM with very large datasets due to the so-called course of dimensionality, which forces the number of input examples to grow exponentially so as to obtain statistically reliable results.

This devised feature space $\mathbf{w}_{T^*}^A$ aggregates the captured time statistics of the user and group it into periods corresponding to morning, noon, evening and night, which are the typical intervals in which it is hypothesized that a user shows certain regularity in his/her connection habits. Therefore, following the notation in Section 2 the new feature space is defined by $\mathbf{w}_t^A \triangleq \{w_{t,n}^A\}_{n=1}^N$ with $N = 10$ and the following entries:

- Overall duration of connections in the morning (07:01-13:00).
- Overall duration of connections during lunchtime (13:01-17:00).
- Overall duration of connections in the evening (17:01-0:00).
- Overall duration of connections at night (0:01-07:00).
- Number of hours with at least one connection in the morning.
- Number of hours with at least one connection during lunchtime.
- Number of hours with at least one connection in the evening.
- Number of hours with at least one connection at the night.
- Mean duration averaged over the longest eight daily connections.
- Median of the duration of all connections within the day.

This alternative feature space is postulated to embed the generalities of the behavior of any user of the social network, but sampled at a granularity that permits discovering strange connections and distinguish him/her from an impostor. The inferred model must discern behavioral patterns in a concise representation that allows discerning stealthy identity thefts. This new space is expected to delimit the user feature space in such a way that the triggering of false alarms is set to a minimum while better discriminating true attacks than when using hourly statistics as mentioned before.

3.1. Complementary Space

Once the features $\mathbf{w}_{T^*}^A$ have been extracted from the raw connection time traces, synthetic attacks are added to the dataset in order to include negative instances of both classes and subsequently allow for a balanced supervised classifier. This set of synthetic attack examples, hereafter referred to as *complementary space*, will be comprised by connection time traces not representing any of the patterns of the user, hence aimed at embodying traces generated by potential impersonation attacks. This is accomplished by means of an interspersing parameter which denotes the percentage of the set of real connection time traces of the user at hand that is replicated in the synthetic connection time trace, hence yielding a mixture of values of samples belonging to the trace of the legitimate user and other generated for representing the effect of an identity theft. As such, when the interspersing parameter ψ is set to 0, features $\mathbf{w}_{T^*}^{\bar{A}}$ corresponding to the complementary space are generated from progressively upscaled connection time traces from the set of positive examples. This simple procedure finds its rationale on the assumption that connection time statistics for a social network user follow a multi-variable non-uniform statistical distribution of some kind (i.e. they are regular to a lesser or greater extent). As such, the mean and standard deviation of the real user connection time records at every sampled time n establish the statistical boundary beyond which *detectable* negative instances must lie. By upscaling the average original connection time traces beyond these limits, features extracted therefrom should resemble theft profiles. As shown in Figure 2.a, when $\psi = 0$ distant traces (light solid lines) from the user ones (light dashed lines) are produced,

DOI: 10.1002/cpe

hence representing better detectable identity theft attacks. However, as ψ increases (Figure 2.b) the complementary space is generated by a mixture of real samples of the connection time traces of the user and values outlying beyond their statistical boundary (bold solid line), the latter modeling an eventual, sporadic session of the identity thief to the account of the user. Asymptotically when $\psi \rightarrow 1$, the traces would equal those of the user under analysis.

In summary: this modeling procedure allows quantifying the effect of randomly placed, small increases in the connection time habits of the user under analysis. However, it is important to emphasize that the value of the interspersing parameter ψ must be tuned so as to yield a classification model with high prediction indicators (high $P_d^{T^*}$ and low $P_{fa}^{T^*}$), which can be done via balanced training and testing sets by virtue of the proposed synthetic attack model.

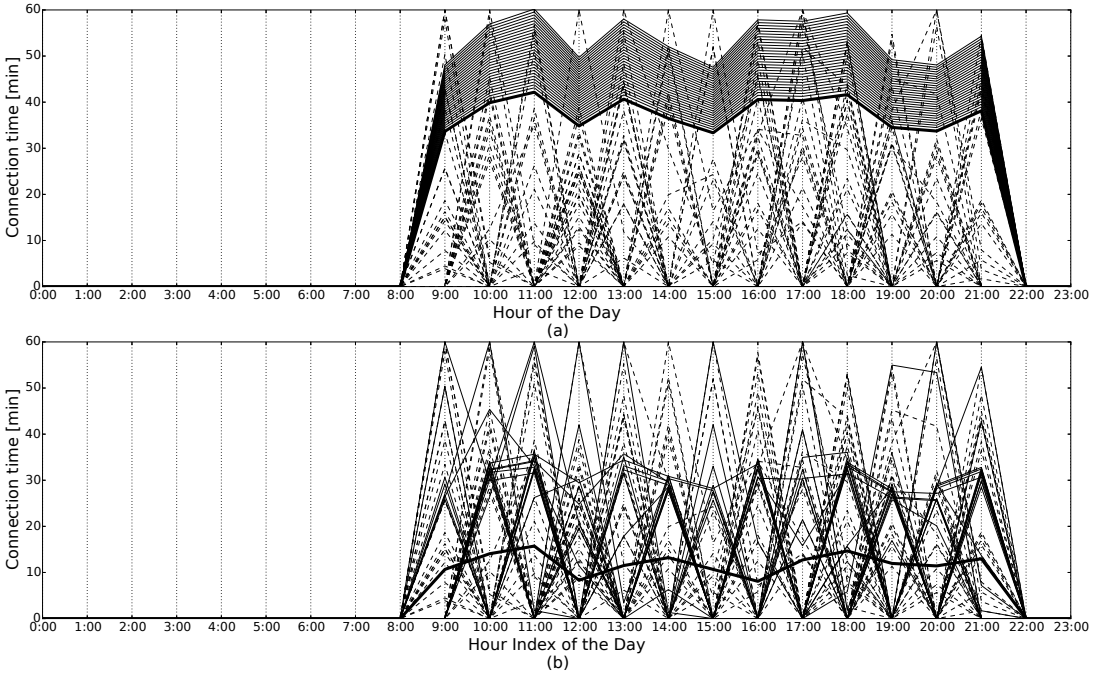

Figure 2. Exemplifying realization of user connection time traces (light dashed lines) overlapped with a complementary space (light solid lines) generated with a) $\psi = 0.0$; b) $\psi = 0.5$. The bold solid line represents the upper statistical limit of the trace region spanned by the traces of the user (given by the sum of their mean and standard deviation).

The complementary space in Figure 2.a is merely a linear increase over the connection time traces of the real user. However, the attacker might access the account just in a few specific hours; this subtle attack might be easily overlooked if the model is not fitted properly to the space of connection time traces of the user in a balanced manner, i.e. by taking simultaneously into account the sought detection score for subtle attacks and the need for keeping the alarm rate low enough not to annoy the legitimate user with useless warnings. Likewise, the value of ψ must be adjusted so as not to overfit the user space and eventually increase the rate of false alarms (i.e. examples of the user wrongly classified as potential attacks). Thus, the design goal of the detector is to produce synthetic attack instances by setting an interspersing value ψ that meets a trade-off between the detection and false alarm rates.

3.2. Selection of the Classifier

Methodologically speaking, throughout most of the related literature cyberattacks are identified by virtue of textual or context features, which are subsequently examined and reasoned by a machine learning technique or statistical methods. Support Vector Machines (SVM) have been empirically

DOI: 10.1002/cpe

proved to be one of the most effective method as evinced by their generally superior performance in comparison with other classifiers, even though Artificial Neural Networks (ANN), Self Organizing Maps (SOMs) and other machine learning schemes have been applied with similar satisfactory results [17]. This paper joins this research trend by considering a SVM classifier with a radial basis kernel, which permits not to assume any statistical distributions on the input set. The precise adjustment of the parameters C (penalty of the error term) and γ (kernel coefficient) controlling the SVM classifier eventually leads to a fine-delimited region with negative areas in-between where traces for a subtle attack could be located. Having said this, the SVM classifier has been fed with the transformed feature set $\mathbf{w}_t^A$ corresponding to the connection time statistics of the user under consideration. Similarly, as shown in Figure 1 and discussed in Section 2 and 3.1, the training process also includes a second set of synthetically generated connection traces as the negative category designed based on – and balanced in number with – the original set of connection time traces of the user.

4. UTILIZED DATASET AND EXPERIMENTAL RESULTS

Currently available privacy configuration options in social network platforms vary within a wide spectrum of levels, which let users decide who will access their posts or their profile as a security measure devised to protect oneself from sexual predators, stalkers, identity thieves or other potential dangers. Unfortunately, such security levels are stringently bound to the messages or multimedia content and do not allow for any chance to retrieve connection statistics of the account at hand. Although this information is systematically stored by the social network platform itself, to the best of the authors' knowledge it is not made available to third parties, nor is the chance to authorize the access to this information by the owner of the account. It is indeed this information concealment what has jeopardized the data acquisition process towards testing the proposed identity theft detector in a practical scenario with real connection time traces.

As a workaround, missing information has been synthetically generated by resorting to different statistical distributions under the realistic assumption that connection time habits are systematically regular for a number of real user profiles. This formulated hypothesis finds its roots in a survey performed over the social circles of the authors, from which several connection time profiles have been concluded to hold consistently in practice:

A) users connecting after their work schedule with no Internet access from his/her mobile device, which corresponds to a regular usage pattern (session start and duration) at evening hours (e.g. half an hour on average sometime between 20:00 and 21:00);
B) users whose accounts are used as a communication channel for business related matters (e.g. marketing campaigns led by managers of the corporate network account), with a connection time usage schedule restricted to regular working hours;
C) users using mobile devices with multiple, short connections during office hours and shifting to web interfaces in the evening (as done by e.g. teenagers);
D) users with connections held in the evening (representing, for instance, shift workers with mobile Internet access); and
E) users who may establish long connections via web interfaces all day long, not as regularly on a specific time period as user A (e.g. retirees, unemployed people or users with less strict, organized and rigid habits).

Bearing these connection patterns in mind, different Poisson and Gaussian distributions have been utilized for synthetically yet realistically generating connection time traces based on non-uniformly distributed random connections over certain hours determined by each of the above usage profiles. The major benefit obtained from these profiles has been the chance of producing as many profile instances as needed for experimentation. As mentioned earlier, it is important to note that this approach does not incur any loss of generality for the designed scheme since this step is conceived as the first of a more complex impersonation detection system which will take into account a broader set of features at distinct levels, i.e. connectivity and content of the exchanged messages.

DOI: 10.1002/cpe

The emphasis in this initial detection phase is placed on the regularity in terms of users connection habits: we henceforth postulate that by using statistical distributions to model connection time traces the resulting dataset meets the real behavior of social network users with regular connection habits.

For the sake of brevity in foregoing discussions we have selected 3 out of the 5 aforementioned user profiles (namely, A, B and E) whose frequencies (normalized in relation to the total number of days) are depicted in Figure 3. User A corresponds to a person who usually connects in the evening after work, whereas profile B gathers all users whose interaction with the social network falls within the work time slot in an attempt at representing corporate social network accounts. User E may represent, instead, young users with mobile phones whose activity levels are intermittent albeit continuous along the day. Despite the relative simplicity of the models, the aim is to demonstrate that the existence of patterns behind the usage of social networks (subject to external factors such as culture, socioeconomic status, contextual facts or even technology development of the area/country of the user) can be exploited to reveal potential identity theft attacks without the need for accessing private information of the user, often assumed by the related state of the art.

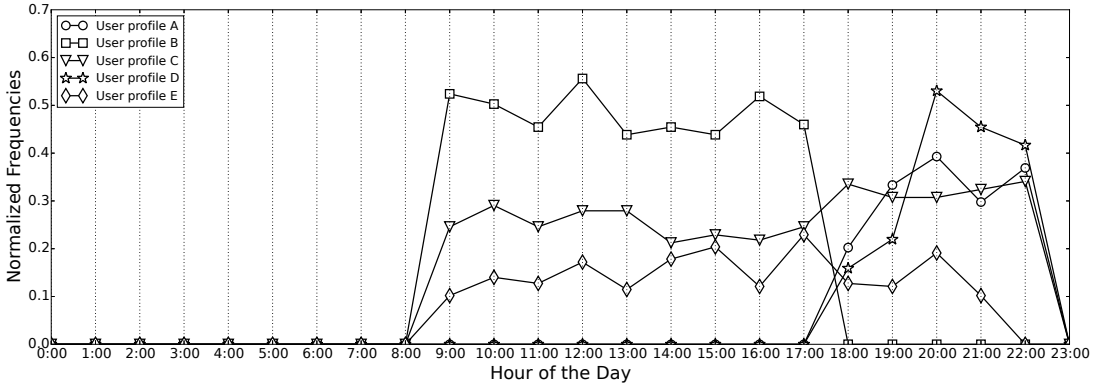

Figure 3. Normalized frequencies of the user profiles considered in the experiments. For instance, 50 % of the total number of days a user following profile A would connect to the social network at 9:00.

In order to verify the hypothesized connection time behavioral regularity an analysis of the Facebook wall posts published by 46,952 users and compiled in the so-called WOSN dataset [18] has been performed. To this end, usage frequencies per user have been inferred from the post messages contained in the WOSN dataset (assuming that a message sent by a certain user implies his/her connection to the social network during 10 minutes) and aggregated aiming at finding an overall behavioral pattern matching that of the above profiles. So do the obtained results when ordering their quartiles in terms of their relative frequencies, as shown in Figure 4: if the WOSN dataset featured no overall usage pattern a uniform distribution over the hours of the day would have resulted from this analysis. However, the obtained plot shows a shape that resembles that of the emulated traces when distributed in its quartiles.

4.1. Experimental Results

In order to assess the performance of the developed detection scheme, evaluation samples have been also produced under the hypothesis that any identity theft trace results in an absolute increase of the real user connection records. Samples are obtained from the same profile template than the user under analysis, but with such an attack implemented as an Gaussian distributed increment (with mean T_{attack} and variance 5 minutes) of the connection trace over a randomly picked hour. A total of $T^* = 250$ positive examples are fed the classifier as the training set, which correspond to true connection traces of the user under analysis. Similarly, the training set is complemented with 250 emulated attack instances based on a value of the interspersing parameter ψ, which permits to control the percentage of authentic values copied from the legitimate user onto the trace representing the emulated attack. In other words, traces corresponding to synthetic identity theft attacks comprise

DOI: 10.1002/cpe

values drawn from the set of original connection time traces of the user (with probability ψ) and values above the limits imposed by the mean and variance of the traces of the user at the considered hour (with probability $1 - \psi$).

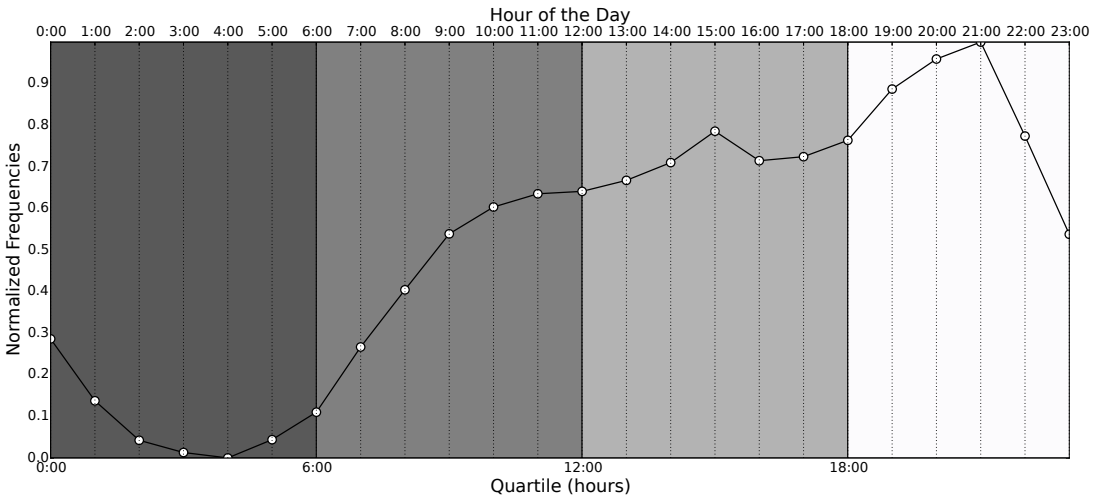

Figure 4. Aggregated normalized frequencies of the connections inferred from the WOSN dataset. The fact that the aggregated frequency profile is not uniformly distributed along the day unveils an overall connection pattern followed by the users within this database: sessions in the social network happen to be more often in the evening and very scarce at night.

The performance of the classifier is assessed by means of $T^* = 250$ samples of each category assembling a total of $T^* = 500$ instances to be evaluated. The detection rate estimates the portion correctly categorized as *attack* out of the whole positive samples set provided for the analysis, whereas the false alarm rate comprises the negative instances misclassified as an attack. As discussed before, an identity theft commits the crime by gaining unauthorized access to the user's account and then eventually adding session time to the real connection trace of the user. This being said, negative instances were generated (providing the same statistical user model) for the testing set by attaching one connection per day/sample at a randomly chosen hour. This illegal connection is parametrized by T_{attack} representing the average duration of the attack, which will be a key parameter to evaluate the detection performance of the detector proposed in this paper.

At this point it should be clarified that the classifier performance is influenced by the interspersing parameter ψ user for producing the complementary training space and the average time T_{attack} during which the attacker utilizes the credentials of the user account to commit the attack. Consequently, detection performance scores must be analyzed in terms of these two parameters and averaged over a number of different Monte Carlo realizations. Figures 2.a to 2.f depict the obtained detection and false alarm rates – averaged over 30 realizations – as a function of both parameters for the three considered user profiles. The value of the parameters C and γ of the SVM classifier have been optimized through a grid search over the whole set of considered (ψ, T_{attack}) combinations.

The interspersing value ψ is involved in the training phase of the classification model, whereas the time consumed in the attack T_{attack} impacts exclusively on the evaluation set. Consequently, the false alarm rate results to be invariant with respect to the attack time since this score is computed over the set of negative samples, i.e. traces where no true attack is being held. Intuitively, the detection rate should increase with T_{attack} as the model should be able to discriminate patterns with longer attacks. This effect becomes evident in Figures 2.a, 2.c and 2.e, where the relevance of the interspersing value is evinced to be essential for the adjustment of the model. This design parameter of the detector provides detection scores at different (ψ, T_{attack}) values depending on the distance between the two classes (*attack* and *no attack*). Hence, the interspersing needed for achieving high detection rates suggests how far attacker and user points are located in the feature space.

DOI: 10.1002/cpe

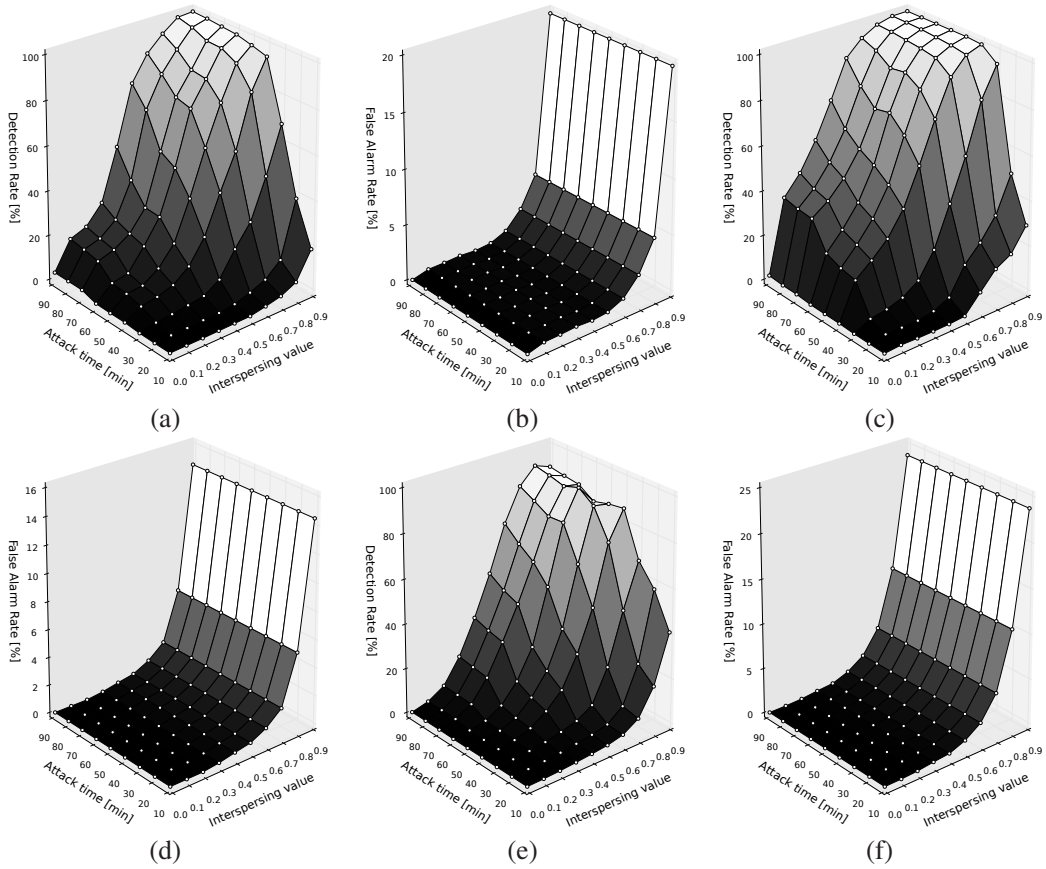

Figure 5. Performance results in terms of average detection ($P_d^{T^*}$) and false alarm ($P_{fa}^{T^*}$) rates for profile A (subplots 2.a and 2.b), profile B (subplots 2.c and 2.d) and profile E (subplots 2.e and 2.f). Results are depicted as a function of the interspersing parameter ψ and attack time T_{attack}.

When implementing a practical detector based on the setup proposed in this paper, the design difficulty is to detect short identity thefts due to the fact that they may fall within the statistical range belonging to the traces of the user and thus may be incorrectly declared as *no attack*. However, in this case increasing the detection rate would increase the false alarm rate as a consequence of decision boundaries being strongly fitted to the user feature space. Indeed, false alarm rates for the three user profiles under analysis exhibit a sharp increase for $\psi > 0.8$, which must be considered as a design threshold to avoid issuing too many alerts to the user.

5. CONCLUSIONS AND FUTURE WORK

This paper finds its motivation in the upsurge of social networks witnessed in the last decade and the wide variety of cyber-crimes that have emerged at the same pace. Social networks allow spreading malicious messages or interacting with personal information through much easier, accessible means. In this context, the manuscript has elaborated on a novel approach for detecting identity theft attacks in social networks based on connection time traces. This particular class of attacks in social networks is often committed for non-interactive purposes, e.g. gossiping. To overcome an eventual lack of content-related traces left by the attacker during the attack, this work takes a step further beyond previous work gravitating on other attack models by proposing to infer a user profile in terms of connection time information. The patterns followed by users when accessing their social network accounts are postulated as crucial when uniquely identifying their degree of dependency,

DOI: 10.1002/cpe

time availability and daily habits with respect to the usage of this technology as a socialization tool. Patterns inferred from the traces of users with regular connection habits (from those owned by individuals to corporate accounts strictly utilized during working hours) are later fed to a SVM classifier jointly with a synthetically generated feature set representing any behavior not observed in the user connection record and emulating a potential impersonation attack. Experiments have been performed and discussed based on a set of synthetic connection time traces that serve as a realistic workaround for the lack of public social network datasets containing session information.

The scheme proposed in this paper must be conceived as an early warning approach that issues an alert regarding a change in the connection behavior of the user currently logged in the account under analysis. This alert can be exploited to send a notification to the owner of the account for his subsequent supervision and/or used for triggering a more complex detection system operating on user-generated content (via natural language processing) and features related to how and with whom the legitimate user interacts. Future research will be devoted towards the design and validation of this overall identity theft detection platform.

ACKNOWLEDGMENTS

The presented work has been possible thanks to the funding support of the Basque Government under the CYBERSID project grant and the computing infrastructure of the i2BASQUE academic network. The authors would also like to thank Dr. Sergio Gil-Lopez from ARIADNA INSTRUMENTS for fruitful technical discussions.

REFERENCES

1. Abdulhamid S. M., Ahmad S., Waziri V. O., Jibril F. N. 2014. *Privacy and National Security Issues in Social Networks: The Challenges*. arXiv preprint arXiv:1402.3301.
2. Social Media Stats in 2014, http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html, retrieved on January 2015.
3. Keeping our users secure, https://blog.twitter.com/2013/keeping-our-users-secure, retrieved on January 2015.
4. Martin A. Anutthamaa N. B., Sathyavathy M., Saint Francois M. M., Venkatesan P. 2011. *A Framework for Predicting Phishing Websites Using Neural Networks*. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, pp. 330–336.
5. Salem M. B., Stolfo S. J. 2011. *Modeling User Search Behavior for Masquerade Detection*. Recent Advances in Intrusion Detection (RAID), pp. 181–200.
6. Fette I., Sadeh N., Tomasic A. 2007. *Learning to Detect Phishing Emails*. Proceedings of the 16th international conference on World Wide Web, pp. 649–656.
7. Dhamija R., Tygar J. D. 2005. *The Battle against Phishing: Dynamic Security Skins*. Proceedings of the 2005 Symposium on Usable Privacy and Security, pp. 77–88.
8. Miyamoto D., Hazeyama H., Kadobayashi Y. 2007. *A proposal of the AdaBoost-based detection of phishing sites*. Proceedings of the Joint Workshop on Information Security.
9. Zhang Y., Hong J., Cranor L. 2007. *Cantina: A content-based approach to detecting phishing web sites*. Proceedings of the International World Wide Web Conference (WWW).
10. Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S. 2007. *A Comparison of Machine Learning Techniques for Phishing Detection*. Proceedings of the Anti-phishing Working Groups of the 2nd Annual eCrime Researchers Summit, pp. 60-69.
11. Xiang G., Hong J. I. 2009. *A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval*. Proceedings of the 18th International Conference on World Wide Web, pp. 571-580.
12. Lane T., Brodley C. E. 1997. *Sequence Matching and Learning in Anomaly Detection for Computer Security*. AAAI Workshop: AI Approaches to Fraud Detection and Risk Management, pp. 43–49.
13. Villar-Rodriguez E., Del Ser J., Salcedo-Sanz S. 2014. *On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks*. International Symposium on Intelligent Distributed Computing, Studies in Computational Intelligence 570: 259-268.
14. Egele M., Stringhini G., Kruegel C., Vigna G. 2013. *COMPA: Detecting Compromised Accounts on Social Networks*. ISOC Network and Distributed System Security Symposium (NDSS).
15. Gao, H., Chen Y., Lee K., Palsetia D., Choudhary A. 2012. *Towards Online Spam Filtering in Social Networks*. Symposium on Network and Distributed System Security (NDSS).
16. Cortes C., Vapnik V. 1995. *Support-Vector Networks*. Machine Learning, Vol. 20, N. 3, pp. 273-297.
17. Liu W., Huang G., Liu X., Zhang M., Deng X. 2005. *Detection of Phishing Web Pages based on Visual Similarity*. Proceedings of the International World Wide Web Conference (WWW), pp. 1060–1061.
18. Viswanath B., Mislove A., Cha M., Gummadi K. P. 2009. *On the Evolution of User Interaction in Facebook*. Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09).