Universidad de Alcalá

Campus Universitario
Dpto. de Teoría de la Señal y Comunicaciones
Ctra. Madrid-Barcelona, Km. 36.6
28805 Alcalá de Henares (Madrid) Telf: +34 91 885 88 99
Fax: +34 91 885 66 99

DR. D. SANCHO SALCEDO SANZ, Profesor Titular de Universidad del Área de Conocimiento de Teoría de la Señal y Comunicaciones de la Universidad de Alcalá, y y DR. D. JAVIER DEL SER LORENTE, Director Tecnológico del Área OPTIMA (Optimización, Modelización y Analítica de Datos) de la Fundación TECNALIA RESEARCH & INNOVATION,

CERTIFICAN

Que la tesis "Advanced Machine Learning Techniques and Meta-Heuristic Optimization for the Detection of Masquerading Attacks in Social Networks", presentada por Esther Villar Rodríguez y realizada en el Departamento de Teoría de la Señal y Comunicaciones bajo nuestra dirección, reúne méritos suficientes para optar al grado de Doctor, por lo que puede procederse a su depósito y lectura.

Alcalá de Henares, Octubre 2015.

Fdo.: Dr. D. Sancho Salcedo Sanz

Fdo.: Dr. D. Javier Del Ser Lorente

Universidad de Alcalá

Campus Universitario
Dpto. de Teoría de la Señal y Comunicaciones
Ctra. Madrid-Barcelona, Km. 36.6
28805 Alcalá de Henares (Madrid) Telf: +34 91 885 88 99
Fax: +34 91 885 66 99

Esther Villar Rodríguez ha realizado en el Departamento de Teoría de la Señal y Comunicaciones y bajo la dirección del Dr. D. Sancho Salcedo Sanz y del Dr. D. Javier Del Ser Lorente, la tesis doctoral titulada "Advanced Machine Learning Techniques and Meta-Heuristic Optimization for the Detection of Masquerading Attacks in Social Networks", cumpliéndose todos los requisitos para la tramitación que conduce a su posterior lectura.

Alcalá de Henares, Octubre 2015.

EL COORDINADOR DEL PROGRAMA DE DOCTORADO

Fdo: Dr. D. Sancho Salcedo Sanz

Escuela Politécnica Superior

Dpto. de Teoría de la Señal y Comunicaciones

Doctorado en Tecnologías de la Información y las Comunicaciones

Memoria de Tesis Doctoral

Advanced Machine Learning Techniques and Meta-Heuristic Optimization for the Detection of Masquerading Attacks in Social Networks

Autor

ESTHER VILLAR RODRIGUEZ

Directores

Dr. Sancho SALCEDO SANZ
Dr. Javier DEL SER LORENTE

Octubre 2015

Para todos los "TI"'s que saben que son ellos

Abstract

According to the report published by the online protection firm Iovation in 2012, cyber fraud ranged from 1 percent of the Internet transactions in North America Africa to a 7 percent in Africa, most of them involving credit card fraud, identity theft, and account takeover or hijacking attempts. This kind of crime is still growing due to the advantages offered by a non face-to-face channel where a increasing number of unsuspecting victims divulges sensitive information. Interpol classifies these illegal activities into 3 types:

- Attacks against computer hardware and software.
- Financial crimes and corruption.
- Abuse, in the form of grooming or "sexploitation".

Most research efforts have been focused on the target of the crime developing different strategies depending on the casuistic. Thus, for the well-known phising, stored blacklist or crime signals through the text are employed eventually designing ad-hoc detectors hardly conveyed to other scenarios even if the background is widely shared. Identity theft or masquerading can be described as a criminal activity oriented towards the misuse of those stolen credentials to obtain goods or services by deception. On March 4, 2005, a million of personal and sensitive information such as credit card and social security numbers was collected by White Hat hackers at Seattle University who just surfed the Web for less than 60 minutes by means of the Google search engine. As a consequence they proved the vulnerability and lack of protection with a mere group of sophisticated search terms typed in the engine whose large data warehouse still allowed showing company or government websites data temporarily cached.

As aforementioned, platforms to connect distant people in which the interaction is undirected pose a forcible entry for unauthorized thirds who impersonate the licit user in a attempt to go unnoticed with some malicious, not necessarily economic, interests. In fact, the last point in the list above regarding abuses has become a major and a terrible risk along with the bullying being both by means of threats, harassment or even self-incrimination likely to drive someone to suicide, depression or helplessness. California Penal Code Section 528.5 states:

> "Notwithstanding any other provision of law, any person who knowingly and without consent credibly impersonates another actual person through or on an Internet Web site or by other electronic means for purposes of harming, intimidating, threatening, or defrauding another person is guilty of a public offense punishable pursuant to subdivision [...]".

Therefore, impersonation consists of any criminal activity in which someone assumes a false identity and acts as his or her assumed character with intent to get a pecuniary benefit or cause some harm. User profiling, in turn, is the process of harvesting user information in order to construct a rich template with all the advantageous attributes in the field at hand and with specific purposes. User profiling is often employed as a mechanism for recommendation of items or useful information which has not yet considered by the client. Nevertheless, deriving user tendency or preferences can be also exploited to define the inherent behavior and address the problem of impersonation by detecting outliers or strange deviations prone to entail a potential attack.

This dissertation is meant to elaborate on impersonation attacks from a profiling perspective, eventually developing a 2-stage environment which consequently embraces 2 levels of privacy intrusion, thus providing the following contributions:

- The inference of behavioral patterns from the connection time traces aiming at avoiding the usurpation of more confidential information. When compared to previous approaches, this procedure abstains from impinging on the user privacy by taking over the messages content, since it only relies on time statistics of the user sessions rather than on their content.
- The application and subsequent discussion of two selected algorithms for the previous point resolution:
  - A commonly employed supervised algorithm executed as a binary classifier which thereafter has forced us to figure out a method to deal with the absence of labeled instances representing an identity theft.
  - And a meta-heuristic algorithm in the search for the most convenient parameters to array the instances within a high dimensional space into properly delimited clusters so as to finally apply an unsupervised clustering algorithm.
- The analysis of message content encroaching on more private information but easing the user identification by mining discriminative features by Natural Language Processing (NLP) techniques. As a consequence, the development of a new feature extraction algorithm based on linguistic theories motivated by the massive quantity of features often gathered when it comes to texts.

In summary, this dissertation means to go beyond typical, ad-hoc approaches adopted by previous identity theft and authorship attribution research. Specifically it proposes tailored solutions to this particular and extensively studied paradigm with the aim at introducing a generic approach from a profiling view, not tightly bound to a unique application field. In addition technical contributions have been made in the course of the solution formulation intending to optimize familiar methods for a better versatility towards the problem at hand. In summary: this Thesis establishes an encouraging research basis towards unveiling subtle impersonation attacks in Social Networks by means of intelligent learning techniques.

Resumen

La empresa Iovation dedicada a la protección ante actividades fraudulentas en la red publicó en 2012 que el fraude informático se sitúa entre el 1 por ciento del total de las transacciones de Internet en América del Norte y un 7 por ciento en África, estando la mayoría de ellas relacionadas con el fraude de tarjetas de crédito, robo de identidad y apropiación indebida de cuentas. Este tipo de delincuencia sigue creciendo debido a las ventajas que ofrece un canal de interacción indirecta donde un número cada vez mayor de víctimas inocentes divulga información confidencial. Interpol clasifica estas actividades ilegales en 3 tipos:

- Ataques contra hardware y software.
- Crímenes económicos y corrupción.
- Abuso, tanto acoso sexual infantil como explotación sexual.

La mayoría de los esfuerzos de investigación se han centrado en el objetivo del crimen en cuestión, desarrollando diferentes estrategias dependiendo de la casuística. Así, para el phising se emplean listas negras almacenadas o señales de crimen en los textos para diseñar detectores ad-hoc que son difícilmente extrapolables a otros escenarios pese a que el trasfondo sea similar. El robo de identidad o el masquerading constituyen una actividad criminal orientada hacia el uso indebido del credenciales robadas para obtener, mediante engaño, algún beneficio. El 4 de Marzo de 2005 una gran cantidad de información sensible como números de tarjetas de crédito y de la seguridad social fue extraída en menos de 60 minutos por White Hat hackers en la Universidad de Seattle haciendo únicamente uso de Google. A consecuencia de este ataque quedó en evidencia la vulnerabilidad y la falta de protección mediante la escritura de un simple conjunto de sofisticados términos en el motor de búsqueda, cuya base de datos aún revelaba información de compañías y del propio gobierno que había sido anteriormente almacenada.

Como se ha mencionado anteriormente, las plataformas para conectar personas en las que la interacción no es directa suponen una atractiva entrada para terceras partes no autorizadas que fingen ser el usuario lícito en un intento de pasar desapercibido con intereses malintencionados y no necesariamente económicos. De hecho, el último punto de la lista anterior en relación con los abusos se ha convertido en un importante y terrible riesgo junto con la intimidación por medio de amenazas, acoso o incluso la auto-incriminación que pueden conducir a alguien al suicidio, la depresión o a su total indefensión. La Sección 528.5 del Código Penal de California dictamina que:

> "Sin perjuicio de cualquier otra disposición de la ley, cualquier persona que suplante a sabiendas y sin el consentimiento de manera creíble a otra persona real a través de o en un sitio Web de Internet o por otros medios

> electrónicos con el objetivo de dañar, intimidar, amenazar o defraudar a otra persona es culpable de un delito público punible conforme a la subdivisión [...]".

Así, la suplantación consiste en cualquier actividad criminal en la que alguien asume una identidad falsa y actúa como tal con la intención de obtener un beneficio pecuniario o causar algún daño. "User profiling", a su vez, es el proceso de recolección de la información del usuario con el fin de construir un modelo orientado al área de interés y con propósitos específicos. Los perfiles de usuario se utilizan a menudo como mecanismo para la recomendación de elementos o información útil que no han sido considerados previamente por el consumidor. Sin embargo, también puede resultar ventajoso extraer las tendencias del usuario o preferencias para definir el comportamiento inherente y abordar el problema de la suplantación mediante la detección de valores atípicos que puedan representar un potencial ataque.

El fin de esta disertación es profundizar en los ataques de impersonación mediante el desarrollo de un entorno de 2 etapas que invaden 2 niveles distintos de privacidad. Las contribuciones principales del trabajo son las siquientes:

- El análisis de patrones de comportamiento en las trazas de conexión con el objetivo de evitar la usurpación de información más confidencial. En comparación con los enfoques anteriores, este procedimiento se abstiene de invadir la privacidad del usuario mediante la obtención del contenido de los mensajes, ya que sólo hace acopio de estadísticas de tiempo de las sesiones de usuario y no accede a ningún contenido.
- La aplicación y posterior discusión de dos algoritmos para la resolución del punto anterior: por un lado, uno de los algoritmos más populares de clasificación binaria que, ante la falta de ejemplos de impersonación, ha forzado la inclusión de un mecanismo para la generación automática de tales ejemplos ausentes. Por otro lado, un algoritmo meta-heurístico empleado para la búsqueda de los parámetros que más convenientemente sitúen los ejemplos en el espacio multi-dimensional de características, con el fin último de facilitar la tarea de clustering encargada a un algoritmo no supervisado.
- El análisis de contenido de los mensajes que implica una intromisión en información más privada, pero que facilita la identificación del usuario mediante la extracción de características discriminatorias gracias a técnicas de Procesamiento del Lenguaje Natural (PLN). En este contexto se propone un nuevo algoritmo de selección de características basado en teorías lingüísticas, motivado por la cantidad masiva de características que gobiernan los enfoques vinculados al tratamiento de textos.

En resumen, esta tesis pretende ir más allá de los enfoques típicos adoptados tanto en la investigación previa realizada sobre robo de identidad como en la correspondiente a la atribución de autoría en textos planteando una solución que, aunque haya sido diseñada a medida de estas ampliamente estudiadas problemáticas, introduzca un enfoque genérico desde una visión de Profiling que permita su extensión a otros campos de aplicación. Además, se han realizado contribuciones técnicas en el trascurso de la formulación del problema en un intento de optimizar métodos típicos para obtener una mayor versatilidad en el tratamiento del ejercicio central de la Tesis.

Acknowledgements

Bad decisions have led me up to this point. And fortunately all those mistakes were committed, much later assumed, now have defined a winding path to walk on. I do not want to be dishonest, not at least in this sheet, I do not want either to make any promise about my own future lines. Anyway, my path is now which needs to be described so as to give some background for my acknowledgments. A road dressed with plenty of rough stones and breathtaking landscapes. My first deserved gratitude is for the former: thank you for being so discouraging, for making me fight harder, for letting me find out what eventually involves a profound lack of passion.

As for the latter, there are many people to include.

All the aforementioned bad decisions were suffered by firstly my family. A family which has given a wholehearted and solid support without asking for anything but a smile or a hug. My parents, a role model of perseverance, will, courage and heroic sacrifice which have brought up two daughters with a deep sense of proud. My mother, the strongest person ever, instilling in us that we should hold our heads up high, learn how to be a woman and how to get the moon by ourselves instead of needing any man to do so. My father, in my heart, in my tears, in my toughest moments, in my affectionate memories, the man of my life: "Sometimes your light shines so bright that it blinds people from seeing who you really are." You have always been able to see the good in me, maybe the good is just you. My sister, my soul mate, sometimes so far, indefinitely so near. Nothing compares to you, no one can fly so high, has such a forthright view of the universe. To my little family members: Xabi, Maialen, Aratz, Uxia and Izei (welcome!). For them, fervently praying for you to find your place in this world, to remember happiness is in you and you mean happiness to me (and so on and on).

It was not originally in the template, perfectly designed by my supervisor, confidant and friend Javier Del Ser (now it is your turn to cry), the chance to introduce any quote here, but this was for sure compelled to me and this would be written down in bold in the upper side of this sheet because all this is as well your work, your persistence:

> *"The whole art of teaching is only the art of awakening the natural curiosity of young minds for the purpose of satisfying it afterwards"*
>
> Anatole France

Nothing left to say. That nowadays seemingly antiquated sentence is your vivid description. That "afterwards" constitutes the conclusive proof that demonstrates that you never left me alone, you walked the whole path by MY SIDE. BY MY SIDE, teaching, awakening the natural curiosity of my young mind. Thank you. In parallel, your Thesis contained in your acknowledgments the following sentence (this is hence

a plagiarism act, matching perfectly as an introduction to the criminal activity later illustrated): “Nevertheless, I eagerly believe that devotion to research reduces to a two-fold rationale: passion for the undiscovered and rewardless satisfaction”. Hopefully, I will shortly be a rewardful satisfaction for you. BY YOUR SIDE.

To Sancho Salcedo Sanz, for opening his arms and feeding my devastated illusion with new opportunities, new intellectual horizons. To the University of Alcalá for cleansing and sweeping away all my first awful memorized steps in this – sometimes corrupted – environment. To both, for putting on a solid and [re|in]novated coat of paint.

I feel also grateful to TECNALIA RESEARCH & INNOVATION, for the confidence placed in me. Particularly, to my closest colleagues at the OPTIMA (Optimization, Modeling and Analytics) area of the ICT/ESI Division. With special reference to my generous comrades: Ana, Maitena, Lara and Cristina. You entail the best coffees, the talks, the ADVICE and the LOYALTY. Today a curious glance for me from the discovering of what I have accomplished: I am a mirror where you all can admire your clear reflection over me.

To Sergio for your constructive discussions finally ended up in arguments, for your humanity, four your passionate and personal commitment without expecting any recompense, for giving me back hope.

To all my missing people on earth.

To all people reading this work, for your patient and the cokes needed to swallow it.

I also apologize in advance for any oversight.

Contents

Contents IX

List of Figures XI

List of Tables XIII

List of Acronyms and Symbols XV

1 Introduction 1

1.1 Motivation . . . 3
1.2 General and Specific Objectives . . . 6
1.3 Structure of the Thesis . . . 7

2 Background Material 9

2.1 Machine Learning . . . 9
2.1.1 Unsupervised Learning . . . 11
2.1.2 Supervised Learning . . . 12
2.1.2.1 Random Forests . . . 13
2.1.2.2 Support Vector Machines . . . 13
2.2 Optimization Problems . . . 16
2.3 Meta-heuristic Optimization . . . 18
2.3.1 Harmony Search . . . 19
2.4 Natural Language Processing . . . 22

3 Mining Connections to Discover the User 27

3.1 Problem Formulation . . . 30
3.2 Proposed User Profiling Approaches . . . 32
3.2.1 SVM as a Supervised Learning Algorithm for the Impersonation Detector . 36
3.2.2 HS as a Meta-heuristic Learning Algorithm for the Impersonation Detector 40
3.3 Comparison between the Proposed Schemes . . . 49

3.4 Conclusions 51

4 Mining Content to Discover the Author 53

4.0.1 Feature Selection 56
4.1 Results and Discussion 61
4.2 Conclusions 66

5 Concluding Remarks and Future Research Lines 67

5.1 Research Outcomes 68
5.2 Future Directions 69

Bibliography 71

List of Figures

1.1 Exponential upsurge of social networks in the last 20 years (source: Merchant). . 2
1.2 Evolution of the age profile of Facebook users through the period between 2011 and 2014 (source: Mashable). 4
1.3 Flow diagram representing the structure of the Thesis. Dashed arrow lines indicate that the reading of the destination chapter is optional. 7

2.1 Soft margin approach utilized by the conventional SVM model to allow for permitted classification errors. 14
2.2 Trade-off between bias and variance in supervised learning models. 15
2.3 Example of a two-dimensional (i.e. $|\mathfrak{X}| = 2$) function $f(\mathbf{X})$ with multiple local optima and isolated global optima. 17
2.4 Block diagram of a generic HS algorithm. 21
2.5 Communication act with tools and resources needed for a proper interpretation. 24

3.1 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. 34
3.2 Real (white bars) and emulated (black lines with different markers per model) connection time frequencies ordered by quartiles. The black bold line corresponds to the overall connection model computed by aggregating patterns of all models. Models A, B, C, D and E as explained in the text are marked with ▼, ■, ★, • and ♦ markers, respectively. 35
3.3 Proposed connection time based detector of impersonation attacks. 37
3.4 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). 38
3.5 Performance results in terms of average detection ($P_d^{T^*}$) and false alarm ($P_{fa}^{T^*}$) rates for profile A (subplots 3.4.a and 3.4.b), profile B (subplots 3.4.c and 3.4.d) and profile E (subplots 3.4.e and 3.4.f). Results are depicted as a function of the interspersing parameter ψ and attack time T_{attack}. 40
3.6 Block diagram of the proposed connection time based detection for impersonation attacks. 41

3.7 Cluster arrangement for the first experimental scenario provided by the combination of the K-Means algorithm and the Elbow method. The area shaded in gray corresponds to the space where connection time patterns are declared as legitimate. 45
3.8 Cluster arrangement for the first experimental scenario provided by the combination of the K-Means algorithm and the HS solver. It is important to notice that axis have been scaled for a better understanding of the plot. 46
3.9 False alarm ($P_{fa}^{T^*}$) and detection loss rates ($P_{loss}^{T^*} = 1 - P_d^{T^*}$) for model A as a function of the duration of the attack T_{attack} (in minutes). Bar amplitudes and whiskers represent the mean and standard deviation of the detection metrics computed over the Monte Carlo simulations. 47
3.10 False alarm and detection rates for model B as a function of T_{attack} (in minutes). . 48
3.11 False alarm and detection rates for model E as a function of T_{attack} (in minutes). . 49

4.1 Feature growth rate for both total and essential features versus the number of senders to be identified. 58
4.2 Diagram showing the two considered sender identification approaches. 59
4.3 Schematic diagram showing the essence extraction procedure for sender i. 61
4.4 Progression of the accumulated number of features per type (3-gram, PoS bigram, essential) for different essence selection threshold schemes and number of considered users. It can be seen that the overall number of features does not increase when considering the last user (user 5), fact that unveils that a single message may not be sufficient for uniquely discriminating among different possible authors, especially when dealing with datasets containing messages of reduced length. . . . 65

List of Tables

3.1 Comparison between the proposed impersonation detectors based on connectivity time information in terms of average probability of detection (P_d) and average probability of false alarm (P_{fa}), both measured in %. 50

4.1 Precision score for different supervised learning techniques and authorship classification approaches. Scores are given as *mean/standard deviation* computed over 10 stratified folds. 63
4.2 Absolute and relative number of features used for each approach (A & B) and threshold selection method. 63
4.3 Normalized confusion matrix corresponding to Approach B, soft voting, self-adjusted $\Psi_{i,j}$. 64
4.4 Precision score (mean /std) for majority voting of successive message and its comparison to the figures of merit of Approaches A and B. 65

List of Acronyms and Symbols

DDoS	Distributed Denial of Service
NLP	Natural language Processing
DBSCAN	Density-based spatial clustering of applications with noise
MDL	Minimum Description Length
BIC	Bayesian information criterion
$h_k(X)$	Collection of k weak classifiers or learners in the bagging ensemble
X	Training vector drawn at random
Y	Category of the training vector X
mg(X,Y)	Margin function for a random vector X belonging to class Y
PE^*	Generalization error
$P_{X,Y}$	Probability over the X, Y space
∂_k	Parameters of the k^{th} classifier of the ensemble
$h(X, \partial_k)$	Classifier configured by the ∂_k parameters
SVM	Support Vector Machines
$\mathbb{N}$	Set of all natural numbers
$\mathbb{N}^+$	Set of natural positive numbers
$\mathbb{N}^N$	Cartesian product of N copies of $\mathbb{N}$
$\langle \phi(\mathbf{x}_i), \mathbf{w} \rangle + b$	Nonlinearly mapped feature space over the linear classifier or hyperplane defined by $\mathbf{x}_i$, w and b
ξ_i	Positive slack variables for the soft margin formulation
C	Parameter defining the penalty over the misclassified instances
$K(\mathbf{x}, \mathbf{z})$	Kernel function applied to the SVM to map the two samples $\mathbf{x}$ and $\mathbf{z}$ into a new feature space
RBF	Radial Basis Function
$\mathbf{X}$	Set of candidate solutions
$\mathfrak{X}$	The search space for the decision variable $\mathbf{X}$
$f : \mathfrak{X} \mapsto \mathbb{R}$	The fitness function for $\mathfrak{X}$ mapped to a Real value
N_c	Number of generic constraints
$\{g_i(\cdot), G_i\}$	Generic constraints over the optimization function
$h_i^\flat(\mathbf{X})$	Hard inequality constraints
$h_i^\sharp(\mathbf{X})$	Hard equality constraints
$N_c^=$	Number of $h_i^\flat(\mathbf{X})$
$N_c^<$	Number of $h_i^\sharp(\mathbf{X})$
$\mathfrak{X}^*$	The collection of optimum solutions

$\mathcal{N}_{\mathbf{X}^{\wp}}$	Immediate neighborhood for a feasible solution $\mathbf{X}$
HS	Harmony Search
HM	Harmony Memory
Ψ	Size of HM or Interspersing Parameter (depending on the context)
HMCR	Harmony Memory Considering Rate
$\wp$	Probability of HMCR
RSR	Random Selection Rate
ε	Probability of RSR
PAR	Pitch Adjusting Rate
ϑ	Probability of PAR
$\mathbf{X}(\psi)$	Harmony vector in the ψ-th position of the HM
$X(\psi, k)$	k-th note of $\mathbf{X}(\psi)$
$\widehat{X}(\psi, k)$	New value of $X(\psi, k)$ after applying the parameter PAR
ω_{HMCR}	Variable in charge of the HMCR
η	Pitch Bandwidth for PAR operator
ω_{PAR}	Auxiliary PAR variable
GHS	Global-best Harmony Search
T^*	Time period
$\mathbf{w}_t^A$	Vector representing connections for user A in the day t
$\mathbf{w}_{T^*}^A$	Set of vectors $\mathbf{w}_t^A$ recorded during time T
$\mathbf{w}_{T^*}^A$	Set of vectors representing the complementary space
$\mathcal{H}_0$	Hypothesis for not undergoing an attack
$\mathcal{H}_1$	Hypothesis for undergoing an attack
$\mathcal{H}_0^{T^*}$	Hypothesis $\mathcal{H}_0$ tested at time T^*
$\mathcal{H}_1^{T^*}$	Hypothesis $\mathcal{H}_1$ tested at time T^*
$P_{fa}^{T^*}$	False Alarms probability at time T^*
$P_d^{T^*}$	Detection probability at time T^*
$P_{loss}^{T^*}$	Loss probability at time T^*
$E^{T^*}(\mathbf{w}, y)$	Error function for vector w according to category y
T_{attack}	Period of time during which the account theft is being committed
$\mathbf{\Lambda}^A$	Multiplicative coefficients for user A connection time traces
λ_n^A	Multiplicative coefficient for the n-th feature
$\mathbf{\Lambda}_m^A$	Multiplicative coefficients belonging to the m-th harmony
$\lambda_{m,n}^A$	Multiplicative coefficient $\mathbf{\Lambda}_m^A$ for the n-th feature
$\widehat{\mathbf{w}}_t^A$	Vector $\mathbf{w}_t^A$ after applying multiplicative coefficients
$\widehat{w}_{t,m,n}^A$	Multiplicative coefficient $\widehat{\mathbf{w}}_t^A$
$\xi(\mathbf{\Lambda}^A)$	Fitness function applied to the set of multiplicative coefficients
$\xi(\mathbf{\Lambda}_m^A)$	Fitness function applied to the m-th harmony
BW	Maximum variational bandwidth parameter
$Z_{m,n}$	Equivalent to ω_{HMCR} but contextualized for the detector in Chapter 3
NLP	Natural Language Processing
POS	Part-Of-Speech
CFG	Context-Free Grammars
G	Grammar in CFG theory
V	Set of non-terminal symbols for G
Σ	Set of terminals for G
R	Set of rules or relations for G
SMS	Short Message Service

SN	Social Network
OvO	One Versus One
FS	Feature Selection
S	Number of senders
FS_i	FS for every sender in the set S
$\mathbf{m}_{i,j}^{k}$	k-th message between sender i and receiver j
$\mathcal{J}_i$	Set of receivers of sender i
$\mathcal{N}$	Set of originally extracted features
$\mathcal{N}_i$	Number of features subset of $\mathcal{N}$ for sender i
$\mathcal{N}_{i,j}$	Number of features for the sender-receiver pair
$\mathcal{N}_{i\cup i'}$	Union of feature subsets $\mathcal{N}_i$ and $\mathcal{N}_i'$ for senders i and i'
$f_{i,j}^{n}$	Frequency of occurrence of the n-th feature between sender i and receiver j
$K_{i,j}$	Number of messages from sender i to receiver j
$\mathbb{I}$	Indicator function
$\Psi_{i,j}$	Minimum frequency threshold for FS imposed over $\mathcal{N}_{i\cup i'}$
$\bigcap$	Intersection of multiple indexed sets
$n_{i,j}^{*}$	Inflexion point in the ordered frequency histogram delimited by $\Psi_{i,j}$
$\lvert\cdot\rvert$	L_2-norm when applied to vectors, cardinality when applied to sets

CHAPTER 1

INTRODUCTION

"The scientist is motivated primarily by curiosity and a desire for truth."

- Irving Langmuir

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. Nevertheless, this is a result of a long walk in which several components have influenced its pace. The earliest forms of the Internet, such as CompuServe (a business-oriented mainframe computer communication solution), were developed in the 1960s along with the primitive emails which were also introduced back then. Bulletin Board Systems were on-line meeting places advanced by Ward Christensen in 1978 which effectively allowed users to communicate with a central system to download files or games (including pirated software) and post messages to other users. Internet relay chats (IRCs) were first employed in 1988 by virtue of the improvement of networking technologies undergone in the previous decade. Geocities was launched in 1994 as the pioneer of social networking sites on the Internet, offering the possibility of developing free home pages within neighborhoods specially designed to allocate and arrange content of specific topics. In 1995, TheGlobe.com went on stage, providing users the ability to interact with people sharing the same interests and publish their own content. However, the first recognizable manifestation of a social media site, as we know it nowadays, was Six Degrees, created in 1997, whose name honors the concept of Six Degrees of Separation and enables users to upload a profile, send messages, post bulletin board items and make friends with other users [1].

As depicted in Figure 1.1, the new millennium brought a plethora of new social networking sites which have raised, evolved and now have become extremely popular worldwide. Friendster in its first three months acquired 3 million users and became the model for posterior web sites such as MySpace, who cloned the former, and was launched after just 10 days of coding. Others popped up consecutively in the few following years: Classmates.com, LinkedIn or Tribe.net as a result of the outstanding success of these preceding entertainment juggernauts, including the most popular social networking website in the Internet history. Facebook.com was founded by Mark Zuckerberg with his college roommates and classmates originally devised as a type of "hot or not" game in its predecessor Facemash, and intended to be limited to the Harvard College students. In its first month, over half of the 19,500 students signed up prompting the gradual

mass movement which ease the open registration to non-college student across the state. In 2008, Facebook surpassed MySpace as the leading social networking website [2].

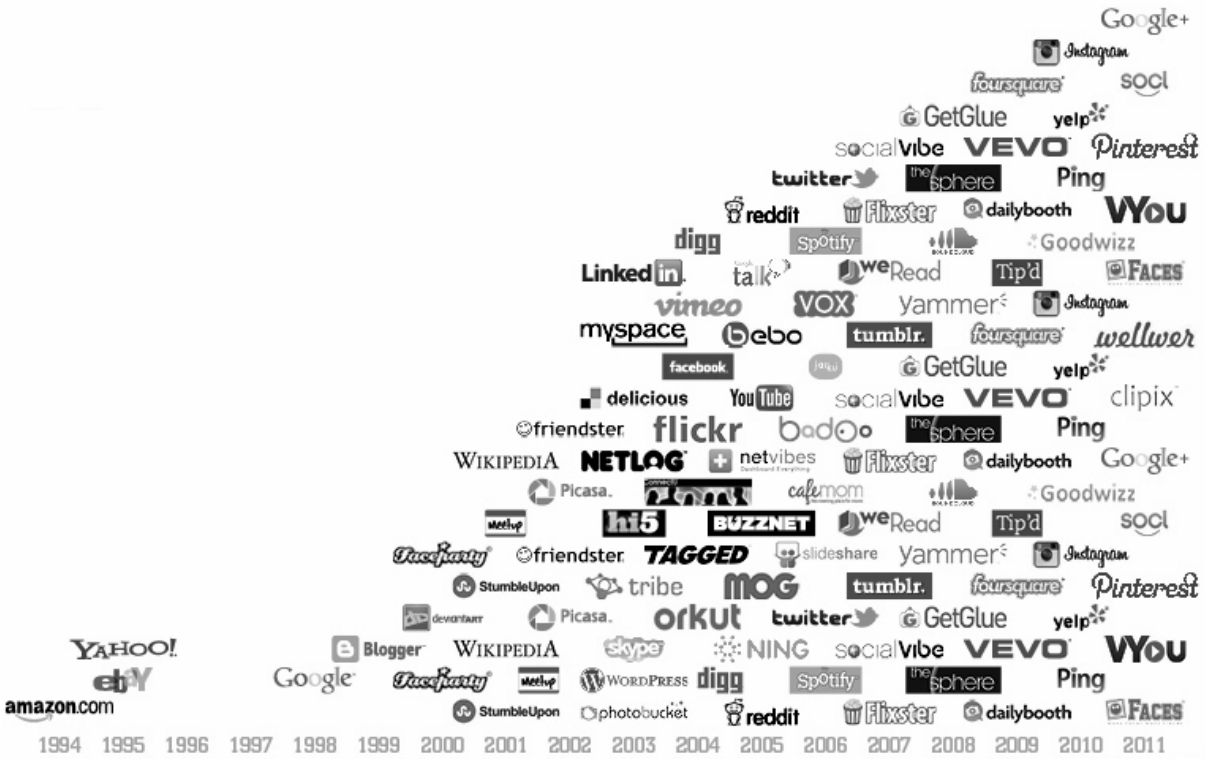

Figure 1.1: Exponential upsurge of social networks in the last 20 years (source: Merchant).

In parallel, Cyber attacks have recently gained momentum in the research community as a sharply concerning phenomenon further ignited by such proliferation of social networks. These attacks have unfolded a variety of ways for cybercriminals to access compromised information of social network users. The general lack of awareness regarding these risks and the consequences of an eventual security breach ends up with large amounts of exposed data susceptible to be stolen and/or exploited with malevolent and fraudulent objectives (e.g. phishing or bullying).

Under the motto "Who needs a gun when you have a keyboard?", the history of cyber attacks became extensive. Also known as Skywiper and Flamer, Flame is a modular computer malware which was discovered on May 28, 2012 by the MAHER Center of Iranian National Computer Emergency Response Team (CERT), the CrySys Lab and Kaspersky Lab as a virus devised for espionage purposes, and used to attack computer systems in Middle Eastern countries that run on Microsoft Windows as their operating system [6]. In July 2009, coordinated attacks against major government, financial websites and news agencies of both the United States and South Korea were deployed bringing about overloads due to the flooding of traffic called DDoS attack [7]. During the 2008 U.S. presidential campaign trail, FBI was forced to retain all the electronic devices due to a attack performed by suspected hackers from China or Russia who had got access to sensitive information used in the campaigns of both Barack Obama and John McCain [8]. In 2007, a malware loaded onto the servers of Hannaford Bros, a grocery retailer, intercepted 4.2 million card numbers stored on the magnetic stripe of payment cards during the four months that the breach was active [9]. Official reports amount to countless cases.

Social Networks have also suffered from this sort of crime. In August 7, 2009 a denial-of-service strike hit Facebook, LiveJournal and Twitter. The government of Russia was informally accused by the victim Cyxymu – the screen name of a Georgian blogger – of being the target in an attempt to silence his criticism over Russia's conduct in the war over the disputed South Ossetia region [3]. In February 2013, Burger King account was hacked and tweeted that the company had been sold to rival hamburger chain McDonald's [4]. On January 12, 2015, hackers claiming loyalty to the Islamic State of Iraq and Syria (ISIS) compromised the Twitter account and YouTube channel of the United States Central Command (CENTCOM) [5]. There are a myriad of malicious activities aimed at unauthorizedly profiting from the user itself or from his/her social circle. For instance, impersonation attacks in account takeovers, whose motivation may go beyond economic interests of the attacker towards getting unauthorized access to information and contacts which can be later used to launch targeted attacks, such as targeted phishing. Criminals are even polishing their tricks by using corporate impersonation aiming at scamming customers, connecting with and phishing employees or slandering the brand, amongst others.

1.1 Motivation

Due to the considerable and noticeable danger which people are exposed to when employing social networks, several alternatives can be embraced. Obviously, awareness campaigns for the general public have been lately deployed, delegating the primary responsibility to the users and their required sense of accountability [10, 11]. A great sum of novel users coming from all different backgrounds and age ranges are increasingly signing in and their lack of knowledge about the underlying implications are leading them to certainly jeopardize their private and sensitive information as well as even their honor and reputation if malicious undesirable intruders find a breach to break into their social through-Internet lives. As Jason Hong states in their published research on phising [13]: "It doesn't matter how many firewalls, encryption software, certificates, or two-factor authentication mechanisms an organization has if the person behind the keyboard falls for a phish". This assumption demonstrates that shifting the whole burden of preventing those harmful activities to the final users is not an advisable policy.

SN sites such as Facebook, Twitter and Gmail, have currently appended to their security registers certain information about connection date, hour and even access point (IP or/and location) from which the user has logged on. Furthermore, some additional security measures are activated in case that SN administrators detect some unusual activity. Gmail sends an informative email to a secondary verified account when the system is suspicious about the browser or device employed with the logged account. More severe restrictions like a double verification and additional security questions increase the security level, but also may deny access to real users. In 2010, Facebook implemented a new security check requiring users to identify friends from tagged photos in order to log in from an unfamiliar computer causing a deep annoyance among the users who were unable to pass the test and thus, were denied to access their legitimate accounts [12].

According to the USA Network commissioned survey entitled "Nation Under A-Hack" [14], if keeping this breach of privacy, 75 percent of young people would be at least "somewhat likely" to deactivate their personal social media accounts, while 23 % would be "highly likely" to do so in the future. These results underline the concern about social media personal privacy which

eventually puts at risk the continuation of the current mainstream use of these platforms.

A combination of factors are encountered here hindering the creation of a stable and safe environment: the absence of incentives to educate users on security, privacy and identity protection; the increased global use of social media and a lack of standards or policing of these standards to prevent from identity theft and fraud. Those factors make Social Media sites have the greatest potential for abuse. Some good practices are strongly recommended such as be wise about what one posts (users feel so comfortable about publishing that they forget the danger involved and announce such hazardous and critical information as they are leaving town), be careful about downloading free applications for use on their profiles, do not give passwords or other account credentials, avoid participating in quizzes which may require the disclosure of personal details, do not click on links that lead to other websites, create strong passwords (mix of upper and lower case letters, characters and numbers not connected to personal information) and change them often or use the highest level privacy settings allowed.

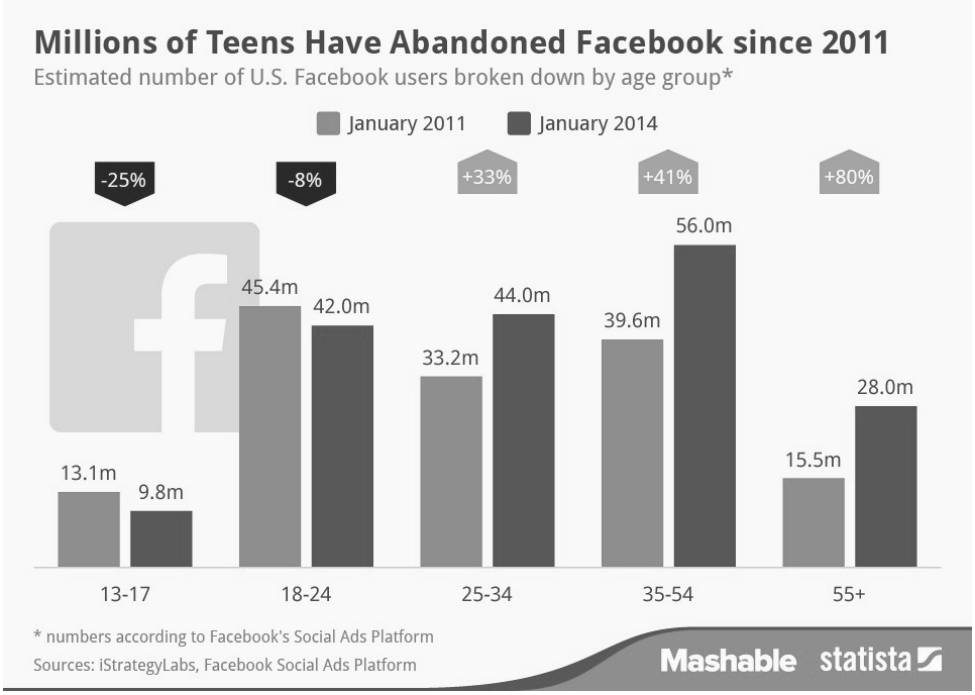

Figure 1.2: Evolution of the age profile of Facebook users through the period between 2011 and 2014 (source: Mashable).

Nevertheless, as aforementioned, an increasing number of people from all generations is now using SN's causing that companies start putting far much time, effort and money into using social media platforms for branding, marketing and engagement to a every segment of the population. According to Mashable (Figure 1.2), between 2011 and 2014 the range of Facebook senior users aged 55 or over has become 80 % larger. Notwithstanding, their unfamiliarity with the wide range of veiled and risky aspects of browsing the web make seniors especially susceptible to such traps. And definitively, awareness campaigns are far from being an appealing and suitable alternative for those least technically prepared who are barely establishing their first contact with such technologies.

Having said this, this Thesis attempts at tackling this matter considering the inherent chal-

lenges and issues arising from finding a compromise between 1) the provision of an technically effective solution; and 2) favoring a safe and agreeable environment compliant with the concept of transparency for the user:

- In foregoing paragraphs, it has been shown that the urgent need of applying dramatic measures to the online crime casuistry, which is fast evolving by discovering new back doors to access private information or deceive users to trigger subliminal and illegal activity by their own. The decrease of hardware and connectivity costs has inevitably leveled the so-called digital divide, hence allowing more and more people to participate in the Internet revolution. Such opening up also involves correlated consequences of diverse nature: people from all ages evincing a growing interest in taking part, an exponential rise in online crime as a result of additional possibilities derived from newborn informational channels and the widespread perception of being spied, controlled and extremely weak and vulnerable as for security and protection issues. The challenge here is how to provide a safe environment without going beyond moral and private own concerns by encroaching on confidential information. Needless to say that annoying and exhaustive controls are as well unwelcome as it was exemplified by the security check implemented on photos by Facebook.

- SN's are also in the spotlight for enterprises which have discovered in this tool a helpful and practical instrument to increase the brand recognition or the brand loyalty. According to Hubspot, 92 % of marketers in 2014 claimed that social media marketing was important for their business [15]. This fact underlines that not every account belongs to a teenager or a person searching for amusement. From a profile perspective, a pattern could be also mined from connection time traces and could become fruitful to depict at least the most sedentary behaviors (i.e. those with marketing purposes). Secondly, this approach poses a non intrusive or invasive exploration which may help defuse any reticence or sensitivities towards the aforesaid perception about being closely monitored on the pretext of keeping the social networking a safer experience.

- Whatever the purpose or the target of the cybercrime, researchers have put all the efforts on generating ad-hoc detectors, especially designed to individually deal with such particular challenges. In turn, authorship attribution has originally devised as a mechanism to determine or verify the signature of anonymous texts. By resorting to those techniques primarily based on computational linguistics, one may provide a feasible and a generic solution dodging, because of pursuing over-ambitious objectives by assuming specific features for the singular problem at hand, the conception of systems tightly adjusted which do not allow for any minor variation in such changeable environments. Profiling would lay a firm foundation for the detection of compromised accounts disregarding the target intended by the attacker at hand.

Bearing all the above evidences in mind and the brief discussion about the matters of interest involved in this context, generating a general-purpose system constitutes the main objective of this doctoral Thesis. By relying on a 2-stage detector regarding the level of privacy usurpation, we propose an environment with the least involvement from the user side compliant with the main purpose aforementioned above: transparency.

1.2 General and Specific Objectives

This Thesis will thoroughly delve into the possibilities to create a 2-stage impersonation detector which accordingly covers different grades of intrusiveness, which in turn are related to potential distinct attacks. Thus, a novel approach is first introduced to deal with gossiping attacks in which no interaction is carried out. Most users have the unfortunate habit of allowing the browsers to negligently save their credentials which becomes a tempting and inviting lure for prying eyes. This behavior is emphasized when it comes to smartphones or tablets. Secondly, a content analyzer is proposed founded on premises from the discipline of authorship attribution. This is a harder task due to our risky assumption that even in those short texts extracted from chats a behavioral profile based on linguistic features is able to distinguish the legitimate user from the rest in a multi-class classification fashion. Although the impersonation detection paradigm by itself can be better formulated as a one-class classification problem, we tackle it as multi-class to evince, from a practical standpoint, that the linguistic essence of the user can be of great help to determine his/her identifiability. Specifically, the following setups will be henceforth devised:

- The first contribution is on the first stage of the detector which exclusively works on analyzing connection time traces of the account being tested in a non intrusive manner to detect evidences of an impersonation attack. The first algorithmic methodology elaborates on a practical approach formulating 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. The second approach consists of a meta-heuristically optimized learning model as the algorithmic core of the detector scheme. This proposed scheme hinges on the K-Means clustering approach, applied to a set of time features specially tailored to characterize the usage of users, which are weighted prior to the clustering under detection performance maximization criteria. The obtained results of the simulations shed light on the potentiality of the proposed methodology for its practical application to real social networks.

- The second scenario considered in this Thesis gravitates on analyzing textual features from messages leveraging Authorship Attribution advances mostly coming from computational intelligence, machine learning and Natural Language Processing (NLP). In this context, this Thesis postulates the identification of the sender of a message as an useful approach to detect impersonation attacks in interactive communication scenarios. In particular conventional yet innovative characteristics of messages will be extracted via NLP techniques which comprise an assembly of linguistic features incorporating not only word-based, grammatical and syntactic features springing from NLP, but also social media and instant messaging based features and a novel indicator of the structural complexity of the message at hand. The feature selection will be handled by means of a newly devised intuitive algorithm based on the dissociation between essential traits of the sender and contextual influences. The proposed selection method yields promising with real SMS data in terms of identification accuracy, and paves the way towards future research lines focused on applying the concept of language typicality in the discourse analysis field. A machine learning classifier is subsequently applied to efficiently exploit the concept of feature essence for authorship attribution at a significantly lower computational complexity than conventional machine learning schemes.

1.3 Structure of the Thesis

This Thesis is organized in two different thematic tracks as remarked (bold line) in Figure 1.3. The first, embodied in Chapter 3, addresses the derivation of trends from connection time traces, whereas Chapter 4 delves into the analysis of textual content under the premises of Authorship Attribution which has proved technically fruitful, respectively. Both chapters are complemented by Chapter 2 which, if needed (dashed line in the Figure), provides the reader with the fundamentals on all the applied algorithms and methodologies necessary to make the document self-contained. Each of these Chapters also includes a literature review of the recent activity around the problem considered therein. Finally, Chapter 5 concludes this Thesis by drawing the main conclusions and outlining several further related lines of research, some of which are being pursued at the time of writing. The scientific publications derived from this work are enumerated and classified within Chapter 5. Acronyms and symbols are collected at the beginning of the Thesis for the interest and perusal of the reader.

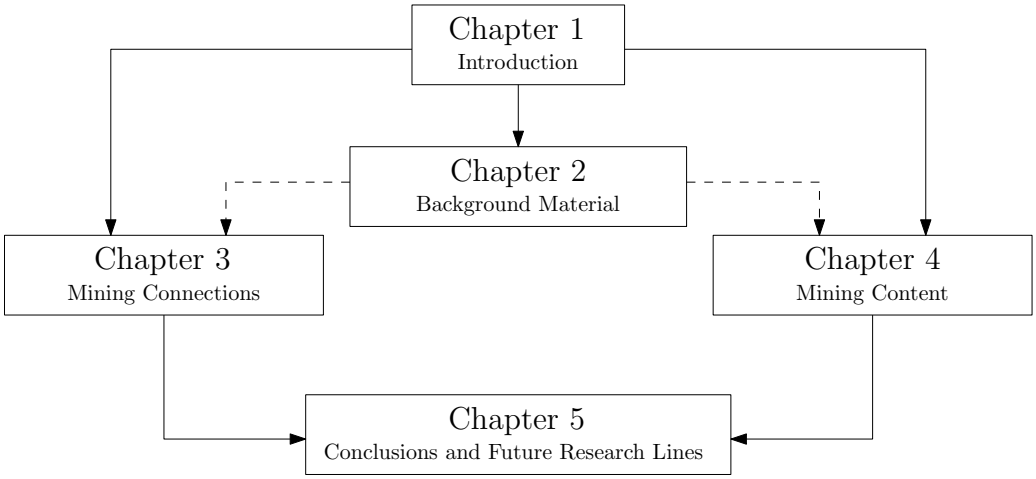

Figure 1.3: Flow diagram representing the structure of the Thesis. Dashed arrow lines indicate that the reading of the destination chapter is optional.

In any case, this Thesis outline is merely a proposal for a coordinated and effective understanding of this report; any other criterium for the reading order can be followed at will.

CHAPTER 2

BACKGROUND MATERIAL

"The important thing is not to stop questioning. Curiosity has its own reason for existing."

- Albert Einstein

This Thesis has been developed under the basis of manifold issues surrounding a specific problem and aiming at isolating those different perspectives under the rationale of giving distinct and precise treatments to every aggregated obstacle. Impersonation is a widely studied field, in this dissertation discovered as a puzzle of numerous components which are far better analyzed deeming each one as a coherent package of determined features regarding the same nature. While skimming through the bibliography, efforts from past research have been focused on confronting the conglomerated constituents of the challenge at hand as a whole, mixing characteristics which are intrinsic to disparate and specialized areas which in turn require exhaustive procedures not to distract oneself from the individual peculiarities.

This chapter formally introduces concepts from the various fields involved – Machine Learning, Optimization and Computational Linguistics – highlighting and elaborating on those techniques which have been employed for the methodological intervals of this research. The purpose lays in providing some background as well as giving brief evidences of the motivations behind the selection of the applied methods encountered in successive chapters.

Nevertheless, it is highly recommended, if interested, to refer to the vast amount of literature concerning Impersonation and Authorship Attribution due to the large number of variants and denominations which have diversified around them. Impersonation is often pointed out as identity theft or even masquerading mainly depending on the environment where is produced and above all the procedural approach implemented by scientists in charge (ranging from the analysis of the remote accesses by means of network protocols up to the study of intrusions gathering content hints). This Thesis builds also upon Authorship attribution in an attempt to match the profiling phase responsible of generating representative and distinct patterns with statistical or computational methods able to infer characteristics of the author from textual features and stylistic character.

2.1 Machine Learning

Machine Learning comprises those algorithms targeted to extract knowledge from data, relying on fundamental concepts in computer science, statistics, probability and optimization. These

techniques are meant to properly represent raw data featuring past experience and rendering it into a model able to gain insights and make either decisions or predictions. Machine Learning is closely related to data mining, although the latter fundamentally concentrates on the exploratory analysis whilst the former draws upon other artificial intelligence disciplines such as computational statistics or pattern recognition. These methods were born from the urgency of delegating tasks up to now manually performed and operated by a human with the purpose of addressing those assignments in a transparent manner, ideally trying to be eventually mistaken for human beings. In that context, ELIZA was created as the first chess game-playing by Arthur Samuel (IBM) and neural networks had their first prototype as a large combination of simple linear classifiers registered as Perceptron by Frank Rosenblatt at the Cornell Aeronautical Laboratory in 1957 [30]. Since then, Machine Learning has been increasingly gaining momentum, initially by the synergy formulated by Statistics and Computer Science which together favored the problem resolution from a mathematical and probabilistic perspective and consecutively at the present by the growing maturity of the Big Data age.

Machine Learning algorithms comprises both descriptive and predictive techniques. Descriptive names those methods focused on describing the data, categorizing or summarizing it whereas predictive analysis is concentrated on drawing conclusions, behaviors or trends which may be useful to anticipate future outcomes. With respect to the learning style applied to the model generation, Machine Learning techniques are typically classified as supervised, unsupervised, semi-supervised and reinforcement learning:

- Supervised Learning: labeled input data feeds the algorithm in the training phase having being compiled granting a balanced number of all distinct categories to be discern. The model or inferred function will be generated under the premise of minimizing an error function or, on the contrary, of maximizing the precision. It is intended to correctly map unseen examples. Mostly addressed problems are those of classification and regression.

- Unsupervised Learning: no label for any input vector is provided since the objective is to find the structure behind the patterns and thus there is no supervisory or reward signal. The model analyzes and deduces peculiarities or common traits in the instances so as to discover similarities and associations amongst the samples. Example problems are clustering and latent variable models.

- Semi-Supervised Learning: labeled and unlabeled instances feed the algorithm thence falling between the previously mentioned categories. The acquisition of labeled data is fairly expensive and requires human skills while unlabeled data can be of great practical value surpassing the performance of any other previous learning approaches. The goal can be oriented towards a transductive learning (deriving the labels of the unlabeled data by searching for analogies) or inductive learning (inferring the mapping from initially labeled vectors to their corresponding categories).

- Reinforcement Learning: the system interacts with its environment by producing actions and receiving either a positive or a negative stimulus from the events in response. These stimuli prompt the translation of that feedback into a learning process aiming at minimizing the punishment or maximizing the gained reward. This sort of learning is typical of robotics and its realistic environments which require algorithms for identifying relevant peripheral events in the stream of sensory inputs to execute proper control actions.

In this research we have leveraged both supervised and unsupervised methods covering different stages of the problem resolution.

2.1.1 Unsupervised Learning

Unsupervised methods have been helpful to recognize the hidden structure of the data. K-Means is one of the most popular clustering method taking as input a set of observations and beforehand the number of desired clusters or groups [26]. The aim is to assign every instance to its closest centroid in terms of certain similarity measure as a result of minimizing the distortion defined as the sum of the squared distances between each observation and its designated centroid. The iterative process reallocates the instances and centroids are again computed until the stop criterion is accomplished (a threshold indicating the required minimum progress measured by the distortion variation or the maximum number of iterations).

Nonetheless, determining K is an ambiguous and intricate issue, dodged by other algorithms such as DBSCAN or hierarchical clustering, since in the case that a penalty over the distortion was imposed the best result would be a cluster for each observation which deviates from the prime purpose. Finding the trade-off is usually a conclusive decision and demands human expertise in the domain under analysis. Several procedures have been introduced to effectively tackle this challenge. Measures such as Bayesian [31], Akaike [32] or Deviation Information Criteria [33] which estimate the maximum likelihood, under the spherical Gaussian distribution assumption, of the different probable models (determined by K) imposing a penalty that discourages the increase of the parameters abstaining from overfitted models. These are measures of fit, typically formulated by deviance statistic, and complexity represented by the number of free parameters in the model. In this context, X-Means starts with one cluster and gives birth to new ones by means of an iterative selection of those more deserving to be split according to the BIC criterion [16]. G-means increases K in a hierarchical fashion until the algorithm is able to state that the data assigned to each centroid follows a Gaussian distribution [18]. On the contrary, there are other approaches in the bibliography that start with a large value for K and keep removing centroids until the description length is no longer reduced. Namely, Minimum Description Length (MDL) supports the adoption of the model which gives the most compact description of the data, including the description of the model itself [17].

Basically, one would expect to find a balance between two factors aiming at creating cohesive clusters: the number K of groups and the average variance contained within the clusters which is a symptom of the cluster compactness. The idea underlying any clustering algorithm is to attempt to find clusters able to minimize the within-group variance, which amounts to maximize the between-group one, since the total variance is fixed. In this context, Silhouette is a metric devised to measure and compare the internal distances of the data (intra-cluster distance) up to their centroids with respect to how loosely it is linked to their neighbouring clusters [25]. Likewise and supporting the same conviction, a number of distinct measures have been developed in the literature in order to select the most suitable or the optimal partition amongst the plausible resulting ones, such as Davies-Bouldin [20], F-statistic [19] or the Dunn's index [21]. Thirty validity indices were summarized by Milligan and Cooper in hierarchical clustering field [23] whereas Dimitriadou et al [24] establish comparisons amongst fifty for binary data sets. On the other hand, conventionally and according to the BIC criterion, the right K is identified as the first decisive local maximum of the aforementioned BIC measure, whose performance can be improved by embracing the "knee" point estimation configured as the point at which increasing

BIC by adding more clusters is no longer fruitful, given the extra computational requirements of a more complex solution [22]. As the number of clusters increases, the intra-cluster distance which measures the average distance from each sample to its dominating centroid is obviously getting lower but instead the inter-cluster distance is amplified. Elbow, another angle-based method, is responsible for determining that curve position where intra-cluster variance and K are lumped together in an abrupt change in function evolution suggesting an optimal solution.

2.1.2 Supervised Learning

On the other hand, classification in this Thesis has also been considered from a supervised learning perspective. Among the supervised learning models found in the literature decision trees [34] are a popular family of classifiers which renders the model in a graph or tree where leaves represent the categories to be predicted and branches stand for the features (or conjunctions of features) that better lead the instance classification, after the training stage, to its class label. The strength of this classifier lies in its ability to divide a complex problem into a collection of smaller and simpler multiple-stage decision tasks yielding a more intelligible solution. In addition, intricate decision regions, especially those derived from high-dimensional spaces, are also approximated at distinct levels of the tree by joining local decision regions. In one-stage classifiers, one subset of features is previously selected based on its discriminatory capability measured by some globally optimal criterion whereas Decision Trees allows for a specific feature selection phase at every node hoping to obtain more discriminating characteristics amongst the subset filtered in the preceding ancestor. This consequent flexibility is likely to furnish a performance improvement.

The main objectives of this algorithm consist of generalizing beyond the training set intending to provide high accuracy once the unseen samples are classified as well as keeping the simple structure aforementioned. However, the tree design is driven by certain optimality criteria being the most often applied those of maximum average mutual information gain, minimum number of nodes, minimum error rate, min-max path length and minimum expected path length. A tree is modeled by splitting the training set into smaller subsets based on an attribute value test. Repetitively, each offspring nodes are similarly split in this recursive partitioning methodology. The recursion is finished by the time the subset of instances within a node belong to the same target variable or other secondary targets/conditions have been accomplished by reaching the maximum value of an imposed threshold over the levels, the number of nodes or the quantity of samples divided by a new split.

The criterion for the selection of the attribute at each node can be guided by a diversity of measures, being the most prevailing some well-known measures such as Impurity, Gini, Information Gain [35] or Entropy whose application at the node become the most time-consuming part of Decision Tree induction. Pruning, removing sub–branches, is the strategy originally suggested in [28] devised to deal with the dilemma of choosing either loose stopping criteria which might produce underfitted models or tight criteria generating the contrary effect of overfitted and complex models with a great sum of nodes and levels which do not contribute to the generalization accuracy.

2.1.2.1 Random Forests

Random Forests is an ensemble approach in which divide-and-conquer strategy is used to improve the performance of the classifier [29]. A number of Decision Trees are computed as *weak learners* following the bagging fashion being all together unexcelled in accuracy among current algorithms and preventing from overfitted models. A subset of the training set is randomly selected for each Decision Tree to be constructed and a random subset of features is analogously picked at each node accordingly to the idea of random subspace selection from Ho [27]. This strategy permits tackling the tough challenge of dealing with thousands of input variables without considering feature deletion. In this case, each tree is grown to the largest extent with possibility of no pruning. Bagging alludes to the process of working independently on equal-sized training subsets, collected by sampling with replacement, which are meant to feed independent classifiers which will cast a unit vote for the individually deduced category. Thence, the ensemble turns robust with respect to noise generating an internal unbiased estimation and gets a generalization error finally converging as forest growing progresses.

The margin function for a K-sized collection of *weak* classifiers $h_k(X)$ measures the extent to which the average number of votes for the right class at a randomly selected training sample X exceeds the average vote for any other class.

$$mg(X,Y) = \frac{\sum_{k=1}^{K} I(h_k(X=Y)}{K} - \max_{j \neq Y} \left[\frac{\sum_{k=1}^{K} I(h_k(X=j)}{K} \right] \tag{2.1}$$

where I(·) is the indicator function which equals 1 or 0 depending on the enclosed condition. Therefore, the larger the margin, the more confidence in the classifiers outcome. On the contrary, a negative margin would represent an error in the classification yielding the following equation of generalization error:

$$PE^{*} = P_{X,Y}(mg(X,Y) < 0). \tag{2.2}$$

In random forests, the ensemble of decision trees $h = \{h_1(X), h_2(X), \ldots h_k(X)\}$ is configured as $\{h(X, \partial_k)\}_{k=1}^{K}$ where ∂_k is taken as the parameters of the k^{th} decision tree (structure of the tree, variables chosen for the splits,...). As shown in [29], as the random forest gets larger ($K \to \infty$) the generalization error following the Strong Law of Large Numbers converges to

$$PE^{*} \to P_{X,Y}\left(P_{\partial}(h(X,\partial) = Y) - \max_{j \neq Y} P_{\partial}(h(X,\partial) = j) < 0 \right). \tag{2.3}$$

This confirms that the generalization error has a limiting value then demonstrating that RF prevents from overfitted models as more trees are added (model complexity does not affect so radically) as it occurs with other algorithms such as the Support Vector Machines, which will be next introduced.

2.1.2.2 Support Vector Machines

In 1963 Vladimir N. Vapnik and Alexey Ya. Chervonenkis invented the original Support Vector Machine (SVM) developed from Statistical Learning Theory. SVM are commonly applied to

classification, regression and outliers detection problems having been proved to be successful in different fields such as bioinformatics, text or image recognition. Formally, given training data, SVM outputs an optimal hyperplane which categorizes new examples aiming at searching for the largest margin in order not to be too sensitive to noise. The strategy consists of finding a function describing a hyperplane such that the expectation of the error on the data be minimum. It is assumed that the larger the perpendicular distance between the nearest two opposite points related to their label, the better is expected to be the unseen instances classification process hence minimizing the occurrence of error.

Notationally, given a nonlinear mapping $\phi(\cdot)$, the SVM method solves as quadratic programming optimization:

$$\min_{\mathbf{w},\xi_i,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \tag{2.4}$$

subject to:

$$y_i(\langle\phi(\mathbf{x}_i),\mathbf{w}\rangle + b) \geq 1 - \xi_i \qquad \forall i = 1,\ldots,n \tag{2.5}$$

$$\xi_i \geq 0 \qquad \forall i = 1,\ldots,n \tag{2.6}$$

where $\mathbf{w}$ and b define a linear classifier in $\mathbb{R}^N$ since $\mathbf{x}_i$ are in $\mathbb{R}^N$ and ξ_i are positive slack variables enabling to deal with permitted errors (Figure 2.1.a) according to the soft margin theory.

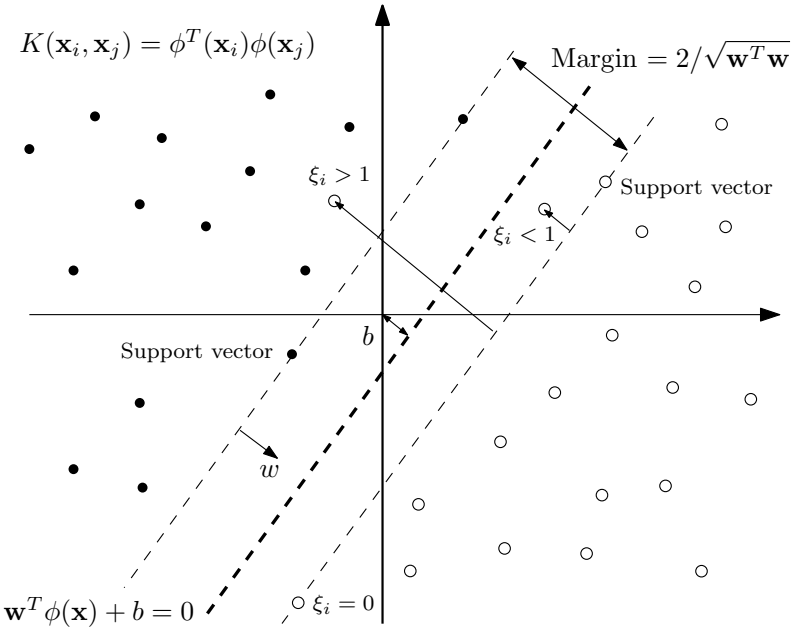

Figure 2.1: Soft margin approach utilized by the conventional SVM model to allow for permitted classification errors.

Those aforementioned slack variables measure the degree of misclassification of the data point which is in consequence penalized by the parameter C. That penalty of cost $C\sum_{i=1}^{n}\xi_i$ is applied to any data point that falls within the margin on the correct side of the separating hyperplane (i.e., when $0 < \xi_i \leq 1$), or on the wrong side of the separating hyperplane (correspondingly, when $\xi_i > 1$) allowing for a functional margin that is eventually less than 1. The regularization term C provides a mechanism to control overfitting: Large C makes the cost of misclassification

high ("hard margin") making unattractive not respecting the data at the cost of reducing the geometric margin (therefore forcing the algorithm to become stricter and potentially generate an overfitted model); whereas conversely with a small value of C the classifier turns flat accounting for some outliers urging the model to search for a larger-margin separating hyperplane.

SVM's strength lies in its ability to effectively deal with high dimensional spaces as well as with few samples in comparison with the number of features. Originally, the proposed solver hyperplane corresponded to a linear classification although not every labeled data collection is this way separable yielding an eventual underfitted model with a high generalization error. The kernel trick was a devise conceived in 1992 by Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik as a response to nonlinear problems (originally proposed by Aizerman et al.[36]). Non-linear patterns correspond to those not straightaway separable in low dimensions being required previous manipulations in order to transform the initial attribute space into a higher (possibly infinite) dimensional space where a linear classifier becomes feasible. Namely, Kernel methods are supported on the concept of similarity considering it from distinct perspectives thus relying on diverse functions. Most commonly used kernel functions include linear $K(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle$, the polynomial $K(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + 1)^d$, $d \in \mathbb{Z}^+$, and the Radial Basis Function (RBF), $K(\mathbf{x}, \mathbf{z}) = \exp\left(-\|\mathbf{x} - \mathbf{z}\|^2 / 2\sigma^2\right)$, $\sigma \in \mathbb{R}^+$.

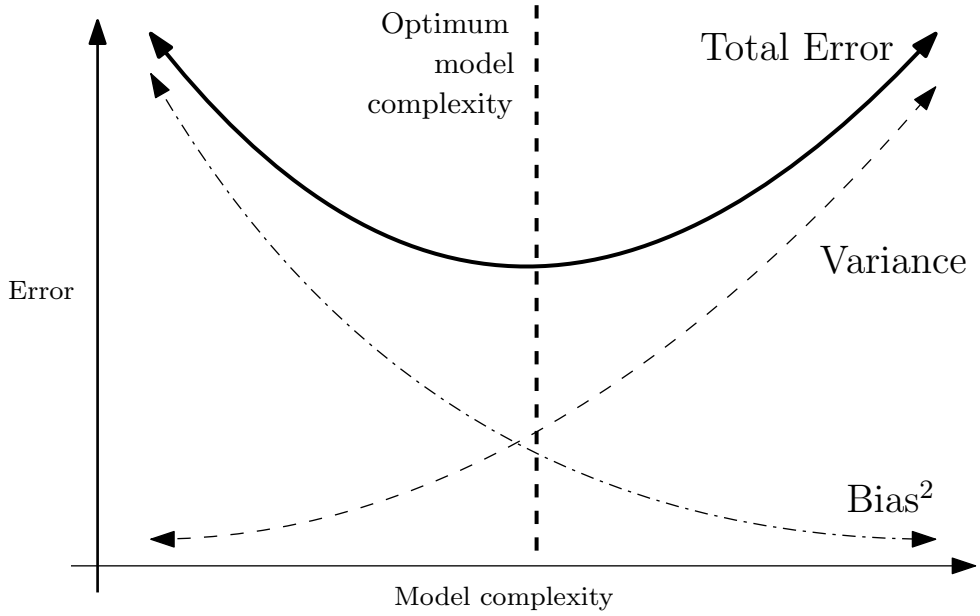

Figure 2.2: Trade-off between bias and variance in supervised learning models.

Such Kernel functions will yield different mapping for the input data in the search of more complex high dimensional spaces to successfully fit the data into these non-linear surfaces. Nevertheless, it cannot be taken for granted that every instance will be placed in the right side according to this kernel function, therefore a cautious approach is to provide a more relaxed constrain over the expected error parametrized by that already introduced parameter C, the penalty constant. This value stands for the amount of error allowed to be ignored and will play a crucial role in the bias variance trade-off during the classification process.

The bias error is deemed as the difference between the expected or average prediction of our model for a certain data point and the correct value which is hoped to be predicted. Bias is

hence considered in terms of accuracy. On the other hand, the variance error accounts for the variability of a model prediction for a given instance implying that for slight changes to the data, the solution changes dramatically. Both factors behave as a function of the model complexity. Consequently, simple models that exhibit small variance and high bias underfit the truth target whereas complex models that exhibit high variance and low bias overfit it.

Transferring those concepts to the world of SVM's and the RBF kernel: a large C will yield a low bias (you penalize up to the extent to expect good predictions) and a high variance (finnally generating such a complex model that little tweaks in the data produce considerable modifications in the outcomes) whereas a small C will result in a high bias and low variance; a small σ will furnish low bias and high variance while a large gamma will give you higher bias and low variance. Basically, the critical point is to reach an agreement between the variables which govern the complexity of the model.

In summary, the ability of the algorithm for properly defining the boundaries will be delimited by the selection of kernel, the kernel's parameters to be settled, and the soft margin C. A grid search strategy is often adopted intending to exhaustively examine all the possible combinations of the parameters which will be checked by a cross-validation methodology. Grid search will compute manifold processes of the algorithm as a result of a cross-product of all the specified values of the required estimators and will return the best configuration in terms of accuracy.

2.2 Optimization Problems

In mathematical terms, an optimization problem is that of searching for the best solution from the set of all feasible ones considering feasible as those alternatives residing within the boundaries delimited by the imposed constraints. Analogously, best refers to the desired solution related to any objective function which is expected to be minimized or maximized. The optimization process consists of inferring a model taking account for 1) the aforementioned objective function depicting the system performance as a quantitative measure 2) the constraints or restrictions to be fulfilled which demarcate the allowable search space 3) the decision variables (quantitative or qualitative) or the parameters which the model is based on. It should be mentioned that multi-objective is a compatible version which is reformulated as single-objective by adding some of them as constrains or by constructing a weighted combination of the multiple initial objectives. The objective functions may be deterministic implying a fixed value for every combination of the decision variables and stochastic when uncontrollable random variables affect the values of the function.

Formally, an optimization problem is defined as to select the best candidate solution

$$\mathbf{X}^* \triangleq \{X(0), \ldots, X(K-1)\} \in \mathcal{X},$$

with $\mathcal{X}$ denoting the search space for $\mathbf{X}$, such that it fulfills

$$\underset{\mathbf{X}\in\mathcal{X}}{\text{Optimize}} \quad f(\mathbf{X}), \tag{2.7}$$

$$\text{subject to} \quad g_i(\mathbf{X}) \gtreqless G_i, \ i = 0, \ldots, N_c - 1, \tag{2.8}$$

where $f : \mathcal{X} \mapsto \mathbb{R}$ is called the objective or *fitness* function which will measure the merit or how close a candidate solution is to achieving the purpose, and $\{g_i(\cdot), G_i\}$ define the N_c constraints

playing the part of the restrictions over the search space $\mathfrak{X}$ to a specific *feasible* subspace. $\mathbf{X}^*$ stands for the best value of $f(\cdot)$ among $\mathfrak{X}$ (the set of all values $\mathbf{X}$ candidates for eventually being the solution of the problem) simultaneously complying with the N_c defined constraints. When $\mathfrak{X}$ are solutions pertaining more than 1 dimension, they are often referred to as *decision variables*. Accordingly, an optimization problem consists of a set of variables, each associated to a certain domain, an objective function mapping variables to real numbers; and an optimality criterion typically intended to search for the best solution which minimizes or maximizes the objective function. In turn, constraints confines the initially possible assignments to those valid turning the problem into the aim of finding the best result which satisfies the restrictions. Restrictions may be classified as soft or hard according to the level that they are imposed at. A soft constraint works as a cost for each value assignment indicating preferences but not compulsory obligations. In this case, the value of the fitness function on a total assignment will be measured as the sum of the costs provided by the soft constraints. Hard constraints instead are restrictions which cannot be violated forcing the system to adhere to them being modeled as a cost of infinity.

Hard constraints may be reorganized in a collection of equalities and inequalities, mathematically describes as:

$$h_i^{\flat}(\mathbf{X}) < 0, \qquad i = 0, \ldots, N_c^{=} - 1, \tag{2.9}$$

$$h_i^{\sharp}(\mathbf{X}) = 0, \qquad i = 0, \ldots, N_c^{<} - 1, \tag{2.10}$$

Consequently, a feasible solution will be any vector $\mathbf{X} \in \mathfrak{X}$ satisfying the $N_c^{=} + N_c^{<}$ constraints. A globally optimal solution is a feasible solution $\mathbf{X}$ among the candidates $\mathfrak{X}$ with the highest achievable value of objective function $f(\cdot)$ (property attributable to linear models). Nevertheless, it should be noted that optimality does not involve uniqueness and many even infinite solutions can be derived resulting in composed solution $\mathfrak{X}^* \subseteq \mathfrak{X}$. When dealing with nonlinear optimization, several solutions may be encountered making gradient based solvers converge to a local optimum where no better solution can be detected in its immediate neighborhood $\mathcal{N}_{\mathbf{X}^p}$.

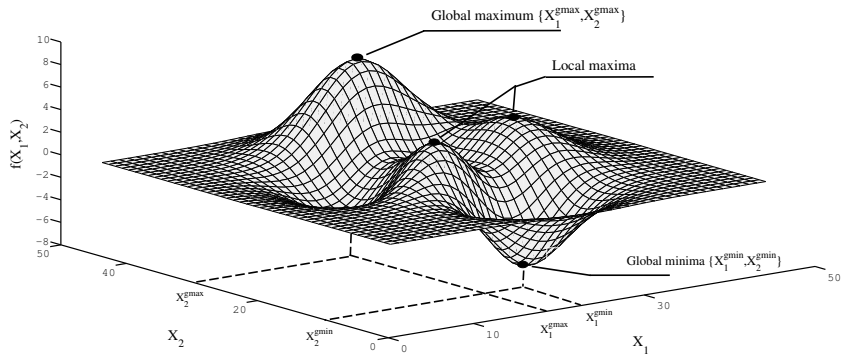

Figure 2.3: Example of a two-dimensional (i.e. $|\mathfrak{X}| = 2$) function $f(\mathbf{X})$ with multiple local optima and isolated global optima $\mathbf{X}_{gmax} = \{X_{gmax}(0), X_{gmax}(1)\}$ and $\mathbf{X}_{gmin} = \{X_{gmin}(0), X_{gmin}(1)\}$.

In the field of approximation algorithms, methods are designed to find near-optimal solutions to hard problems. The concept of near-optimum refers to a feasible solution whose fitness value seems to be in the near surroundings of the optimal solution region $\mathfrak{X}$. These solutions are often

associated with NP-hard problems when acceptable or approximate solutions are sufficient most likely due to the exponential time required to be solved. Thus, some techniques later described are inherently suboptimal by nature.

The taxonomy for optimization problems are diverse but a comprehensive classification may embrace: Unconstrained Optimization versus Constrained Optimization, Deterministic Optimization versus Stochastic Optimization and Continuous Optimization versus Discrete Optimization. Models with discrete variables are discrete optimization problems whereas models with continuous variables $\{X(k)\}_{k=0}^{K-1}$ where $\mathcal{X}$ is a subset with a K-length extension $\mathbb{R}^K$ of the set of real numbers are continuous optimization problems. When dealing with discrete problems the feasible solutions are restricted to be discrete variables although the smoothness of the functions in continuous problems can be helpful in the deduction of information about points in a neighborhood. Notwithstanding, the infinite set of candidate solutions bounded to the dimensionality of the search space in continuous optimization problems may dramatically increase in size and complexity. Typically, Continuous optimization problems leverage the previously generated range of values of the variables, known as iterates, and then move forward by deciding on the potential perturbation in the state of the model taking advantage of the knowledge already gained at previous iterates. This changes aim at representing the sensitivity and are commonly defined by the first and second derivatives of the objective functions. As for single objective deterministic optimization problems, the methods can be grouped into two main types: mathematical programming comprising Linear programming methods, Linear programming methods, Linear integer programming methods, Linear mixed integer programming methods, Nonlinear programming, Dynamic programming and Heuristic Optimization methods classified as Meta-heuristic Methods and Heuristic methods with limited applicability.

2.3 Meta-heuristic Optimization

In this Thesis we focus on meta-heuristic solvers for optimization problems, which are methods able to ignore the closed-form formula of the objective function most of aforementioned techniques rely on. Meta-heuristics have become a widely applied mechanism to tackle optimization problems with large feasible solution spaces. Their ancestor is the stochastic optimization which employs randomness to find optimal (or as optimal as possible) solutions to hard problems. *Derivative-free* or *black-box* approaches received this nomination due to the fact that they draw on numerical values of the objective function without resorting to their derivatives. Meta-heuristics means beyond heuristics since heuristic entails to discover by trial and error and meta-heuristic solvers work as master iterative procedures governing the implicit heuristic and therefore generally surpassing the performance of simple heuristics in terms of affordable time and cost. Hill-climbing is a simple meta-heuristic algorithm which relies on the belief that iteratively small and well-behaved modifications, thus testing new candidates, once near the optimal permit “climb the hill” of quality up to the desired solution. Meta-heuristics complies with this heuristic belief.

The word "meta-heuristic" was first coined by Fred Glover in his seminal paper [37], and a meta-heuristic can be deemed as a "master strategy that guides and modifies other heuristics to produce solutions beyond those that are normally generated in a quest for local optimality" (Glover and Laguna 1997). They embraces those algorithms used to solve problems when there is a lack of information (there is no idea about how optimal solution looks like and we cannot

make such strong assumption imposed in the Gradient Ascent Algorithm about knowing the first derivative) that prevent from applying other more formal methods and brute-force search mechanisms are useless or not very fruitful regarding time and cost because the space is far too large. For instance, the aforementioned Gradient Ascent has its major drawback in the convergence time since as the algorithm approaches to the maximum of the function, it usually overshoots the top and ends up on the other side of the hill.

All meta-heuristic algorithms use a certain tradeoff of randomization and local search basing their strategy on a methodology of intensification and diversification (also commonly referred to as exploitation and exploration). A suitable tradeoff between intensification and diversification must be found during the process defining the appropriate length of the next step in the local search activity hoping to improve the rate of algorithm convergence. Diversification greedily explore the search space intending to provide as diverse high quality solutions as possible whereas intensification involves a more thorough examination in an attempt to discover the optimal solution among the neighbors by exploiting the accumulated search experience. Genetic algorithms are a part of evolutionary computing and hinges on the process of natural selection mimic. In those algorithms, diversification is typically applied by mutation and crossover whereas intensification is implemented by the selection operator. A convenient combination of this techniques usually ensure that global optimality is achievable. In fact, if diversification is too vigorous, zones of interest could eventually be unexplored consequently causing the reduction of the convergence rate of the algorithm. On the contrary, by keeping this operator in slow motion could derive into far-from-optimal solutions due to trapping in local optima.

Meta-heuristics have gained momentum in the last decade and significant amount of research has been published. This trend has given raise to manifold algorithms: from Trajectory optimization, which aims at designing the optimal trajectory that minimizes or maximizes some measure of performance within imposed constraints; to Ant Colony Optimization, based on finding the best path through graphs by interacting with the graph via the fitness function itself. It is still plausible to hybridize two meta-heuristic algorithms aiming at having one at the top level optimizing the parameters of the dependent one in the search for the optimal mutation rate, crossover type or any other controlling parameter. These methods were originally coined as meta-genetic algorithms, or more generally meta-optimization, techniques in the oddly-named family of hyper-heuristics. The reader is referred to [38, 39, 40, 41] for recent, comprehensive surveys on meta-heuristic optimization.

2.3.1 Harmony Search

Harmony Search (hereafter HS) is a meta-heuristic algorithm which operates under the paradigm of the musical improvisation process. First proposed by Geem et al. [42], it has been widely studied and applied to a vast number of problems since then, from structural design [43, 44, 45] and image processing [46, 47] to energy dispatching [48, 49, 50], Telecommunications [51, 52, 53], indoor localization [54], biosciences [55], economics [56] and robotics [57, 58], among many others. This profitable literature, along with thorough application-oriented reviews [59, 60, 61, 62], evince the excelling performance of this nature-inspired algorithm in the field of combinatorial optimization.

HS was conceived aiming at mimicking the expertise of musicians who intent to produce a certain pleasant harmony according to aesthetic standards measured by the adequacy of the

pitch, the timbre and amplitude to make pleasant noises please our brains. Plenty of popular algorithms have been in fact developed by obtaining the inspiration from the idiosyncrasy of natural processes or strangest of places. The analogy between the theory behind HS and optimization is hence obvious: optimization concerns finding the best solution by deciding on the best choice from a variety of candidates which could be compared to the music creation, particularly when jazz musicians play together attempting to select the concrete musical notes to join the rest of participants by providing the best overall harmony. In other words, one musician plays certain tune while the rest of the counterparts mean to memorize it in order to decipher what notes best fit together with the song so as to simultaneously create the optimal composition. Each instrument is capable of playing a defined range of notes and the objective lies in discovering the right set for an outstanding performance.

Similar to notes in a tune for a certain instrument, an optimization problem has its correspondingly feasible values delimited by the objective function or by the collection of individually settled values. Rather than seeking for the most harmonious melody, the current purpose is to solve a specific problem by iteratively *composing* better solutions by landing (resorting to diversification or intensification tactics) in different points from the search space. Improvisation is not as an effortless task as it could be deemed initially and requires expertise (knowledge and ability evolved during past experiences) in addition to some constraints or good practices which are strongly recommended to reach that aesthetic standards such as the reasonable choice of adding over F major (D the 6th, G the 9th, and E the major 7th) or the use of chord tones particularly thirds and sevenths in order to achieve stability and a proper acoustic resonance. In fact, ancient Greeks were already aware that two tones whose frequencies were related by a simple ratio like 2:1 (an octave) or 3:2 (a perfect fifth) yielded the most agreeable or consonant musical intervals.

HS is a population-based algorithm then involving maintenance of a memory which plays the role of a songbook or archive of harmonies. Candidate solutions Ψ or Harmony Memory (HM) allows for an iterative improvement which is expected to converge to the optimal solution by the application of a set of optimization operators. Thence, new and enhanced alternatives are added to the HM as a result of an evolution carried out by global-search and local-search mechanisms. This improvement in the aesthetic quality is assessed by a fitness metric which decides on the optimality of the produced harmony and will be adjusted by the competence of the explorative and exploitative behavior of the HS heuristics. The appropriate balance of both operators will be redefined in every simulation under the intuition that local search must be intensified when global search has spotted a region close to the desired optimal solution and minor or slighter perturbations are consequently demanded.

Figure 2.4 illustrates the flow diagram of the HS algorithm, which can be summarized in four steps: (i) initialization of the HM; (ii) improvisation of a new harmony; (iii) update the HM with the new generated harmony outperforms the worst stored in HM; and (iv) returning to step (ii) until a termination criteria is fulfilled. The main probabilistic parameters which manage the improvisation procedure by being sequentially applied to each note are the following:

- The Harmony Memory Considering Rate (HMCR) standing for the probability $\varphi \in [0, 1]$ that an existing value for a note from a memorized harmony is imitated in the new vector. Otherwise (i.e. with a probability $1 - \varphi$), a randomly selected value within their alphabet $\mathcal{X}$ is introduced. Some contributions leave the random consideration as a third, individual operator (Random Selection Rate, RSR) represented by its own probabilistic variable ε.

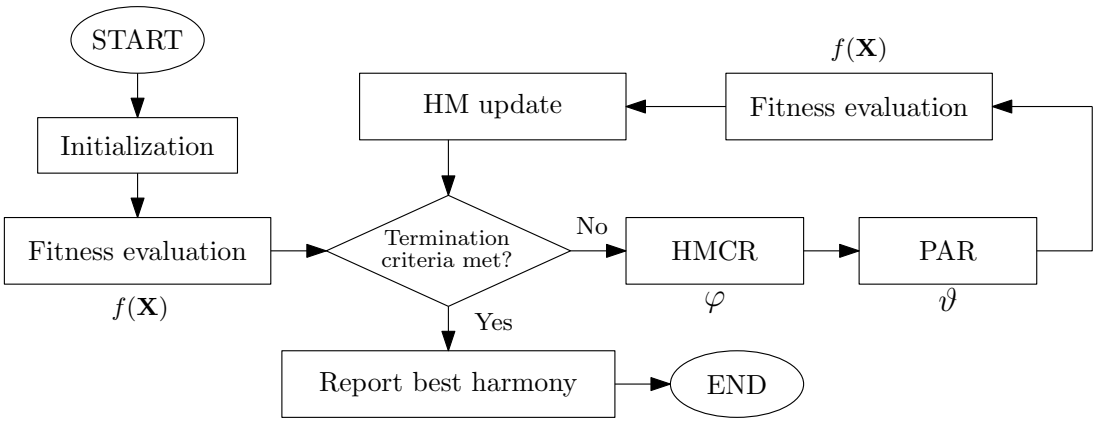

Figure 2.4: Block diagram of a generic HS algorithm.

Following the notation introduced in this chapter, if the HM is denoted as $\{\mathbf{X}(\psi)\}_{\psi=0}^{\Psi-1}$, with $\mathbf{X}(\psi) \triangleq \{X(\psi,0), X(\psi,1), \ldots, X(\psi,K-1)\}$ being the ψ-th harmony of the HM, and K the number of notes, then

$$\varphi \triangleq Pr\{X(\psi,k) \rightsquigarrow \omega_{\text{HMCR}}\}, \tag{2.11}$$

where ω_{HMCR} is a discrete random variable uniformly distributed in $\{X(\psi,0), \ldots, X(\psi,k-1), X(\psi,k+1), \ldots, X(\psi,K-1)\}$. It should be mentioned that this operator is exploited note by note in the HM, i.e. the above operation is repeated for every $k \in \{0,\ldots,K-1\}$ and $\psi \in \{0,\ldots,\Psi-1\}$.

- The Pitch Adjusting Rate (PAR) poses the probability $\vartheta \in [0,1]$ that the new value $\widehat{X}(\psi,k)$ for a given note value $X(\psi,k)$ is computed via a controlled random perturbation, i.e.

$$\vartheta \triangleq Pr\left\{X(\psi,k) \rightsquigarrow \widehat{X}(\psi,k)\right\}, \tag{2.12}$$

where

$$\widehat{X}(\psi,k) = \begin{cases} \omega_{\text{PAR}} & \text{if } |\mathcal{X}| < \infty, \\ x_{old} + \eta \cdot z & \text{otherwise,} \end{cases} \tag{2.13}$$

with $\eta \in \mathbb{R}^+$ representing the pitch bandwidth, and z being drawn from an uniform distribution with support $[-1,1]$. One can deduce that PAR and bandwidth are intimately bound up with the degree of diversification of the algorithm which determines the convergence rate of the overall solver. Nevertheless, a high pitch adjusting rate with a high value of ω_x will cause leaps between the candidates in the search space which could become useful when trapped in a local optimum but may be detrimental when reaching areas with potentially near-optimal solutions. As for the case of discrete alphabets, ω_{PAR} is a random binary variable with equal probability of choosing a neighboring value on $X(\psi,k)$ in $\mathcal{X}$. In order to boost the performance of this operator, the alphabet $\mathcal{X}$ is usually sorted under some vicinity ordering criterium, in general based on the characteristics of the metric $f(\cdot)$ defining the problem at hand.

In HS, diversification is controlled by PAR and 1-HMCR parameters in charge of refining the values obtained previously whereas intensification is represented by HMCR responsible for stimulating the local search by reusing already memorized harmonies in an elitist fashion.

Thus, the intensification is mainly driven by the HMCR operator and has to be properly tuned so as to both prevent from a φ too low which would bring about a slow convergence and dissuade the algorithm from remain confined in a local optimum by applying far too slight changes. On the contrary, the complementary randomization procedure executed with probability $1-\varphi$ in the HMCR operator will force the algorithm to investigate and evaluate distinct candidate solutions of the search space.

HS have been proven to be faster than classical evolutionary optimization in relation with the convergence rate since the latter is constrained to monogamy in the cross-over procedure, giving birth to a new restricted offspring which occasionally does not satisfactorily develop the exploration operation [63]. A number of variants have sprung such as Improved Harmony Search [64], Global-best Harmony Search [65, 66], HS with dynamic subpopulations [67], Parameter-Setting-Free HS [68] and other hybrid approaches [69, 70, 71, 72, 73]. These improvements of the naïve HS solver aim at 1) introducing a dynamic adjustment of parameters (e.g. an improved PAR operator which is adjusted in an proactive manner in order to better mimic the best harmony of the HM); 2) adopting operators from Particle Swarm Optimization [74] or Genetic Algorithms [75, 76] in the search for reducing the sensitivity to noise in the original algorithm; and/or 3) enhancing the ability of the original approach to deal with a greater dimensionality by refining explorative and exploitative strategies. Overviews and insights on different alternative HS algorithms can be found in [77, 78, 79].

2.4 Natural Language Processing

The breach between computers and humans as for understandability has been long tried to be bridged so as to avoid too much effort by users to translate the orders into recognizable code for machines. In fact, the famous Turing Test was devised to analyze the use of language made by a computer thus considering its understanding as a criterion for intelligence. Computers have been originally designed to deal with difficult and sophisticated mathematical and scientific calculations and ironically they encounter severe adversities to struggle with tasks effortlessly performed by humans such as language acquisition. The "fifth generation" of computers defines present day and those early evolving which are intended to provide a technologically advanced platform for future developments in artificial intelligence. This offspring of computers in nascent stage involves the required infrastructure in terms of massively parallel computing for complex assignments such as Voice recognition or other convoluted problems with necessity for "thinking machines" to surface real human problem-solving heuristics.

The linguistic science means to characterize the manifold linguistic observations leveraging the cognitive science to explain how humans acquire, produce, and understand language. In turn, the cognitive science concerns the scientific understanding of the human language in reference to dynamic comprehension and production of language embracing diverse fields, some of them in a continuous development stage, such as psychology, linguistics, anthropology, from a computer science's view. In The Oxford Handbook of Cognitive Linguistics, Dirk Geeraerts and Hubert Cuyckens state:

> *"Cognitive Linguistics is the study of language in its cognitive function, where cognitive refers to the crucial role of intermediate informational structures in our encounters with the world. Cognitive Linguistics is cognitive in the same way that cognitive psychology*

is: by assuming that our interaction with the world is mediated through informational structures in the mind. It is more specific than cognitive psychology, however, by focusing on natural language as a means for organizing, processing, and conveying that information. Language, then, is seen as a repository of world knowledge, a structured collection of meaningful categories that help us deal with new experiences and store information about old ones."

Specifically and putting above theories into practice once reaching a degree of complexity, Natural Language Processing (NLP) names the discipline of understanding humans via speech recognition and natural language understanding comprising intrinsic tasks derived from information retrieval, information extraction and inference. The goal resides in selecting pertinent facts from the textual resources and drawing valid main conclusions about those known evidences (Figure 2.5). In this point, the debate amongst the practitioners focus on determining the most efficient approach to process the language bringing about a division between the defenders of the formal language theory to model utterances and its implicit structures and those prone to analyze the dialogue by accumulating the statistical characteristics of large corpora. As for the former, NLP systems are driven by sets of rules which establish the correctness of the sentence in terms of grammar adequacy requiring for that purpose 3 main phases of translation: the Lexical Analysis which deals with identifying the specific words for every statement, the Syntax Analysis who parses the text discovering the constituents according to a particular grammar and the Semantic Analysis which finally determines the meaning. On the contrary, corpora-based approaches rely on mining features not formally agreed and typed into a knowledge specification under the assumption of the users tendency of bending the rules to accomplish the communicative needs.

Identifying individual words is usually accomplished by tokenizers in charge of splitting the whole text into lexical features or tokens. Nevertheless, considering ulterior procedures, a morphological process (typically through morphological, orthographic or spelling rules) is then applied to filter out suffixes and prefixes to obtain the word stems or morphemes which are likely to be required for syntactic parsers or whether dictionaries or lexicons (repository for words) are included. Word n-grams or combinations of words of size n permit learning from a word and its neighbors being able to even predict which is the next word to appear given a context by means of probabilistic methods (e.g. Bayesian Inference) once a corpus is statistically assimilated.

Syntactic Parsing, in general, involves the translation of some language input into some structure, seen as constituents or "parts-of-speech" (or POS) such as verbs, nouns and adjectives, consistent with specific grammatical rules. POS tagging is not such a elementary task due to ambiguous terms which may hold distinct functions depending on their usage in the context urging the need for more complex treatments like rule-based or stochastic methods based on hand-written disambiguation rules or a priori word probabilities accordingly. One of the most widely employed tagsets are that mined from the Brown corpus [81] and its counterpart Penn Treebank [80] which feed numerous parsers. In a higher syntactic level, parsers operate on "Context-Free Grammars" or CFGs to unfold the sentences into phrases or groups of words containing at least one noun and arranged as a unit aiming at thoroughly examine the structure of a sentence in order to determine its correctness according to the grammar. The resultant parse tree locates the words at the very end of the branches or leaves arraying the complex constituents through a bottom-up strategy. Grammars are designed as statements where the right-hand side of the production is composed of the actual terms or terminals, in

honor to their position in the tree, and the symbol on the left-hand called non-terminal represents the subsequent tag. Mathematically, a CFG G is a quadruple (V, Σ, R, S) where V is a set of non-terminal symbols, Σ is a set of terminals disjoint from V ($V \cap \Sigma = \emptyset$), R stands for the rules or relations ($R : V \rightarrow (V \cup \Sigma)^*$) and S as the start symbol. By means of deriving production rules, CFG describes how sentences are recursively built up from minimal grammar constituents up to largest grammar block structures. Are they denominated as constituency grammars in contrast to dependency ones which include both grammatical and semantic functions to arrange items by capturing the dependency structures deemed as binary asymmetrical relations between words linked semantically.

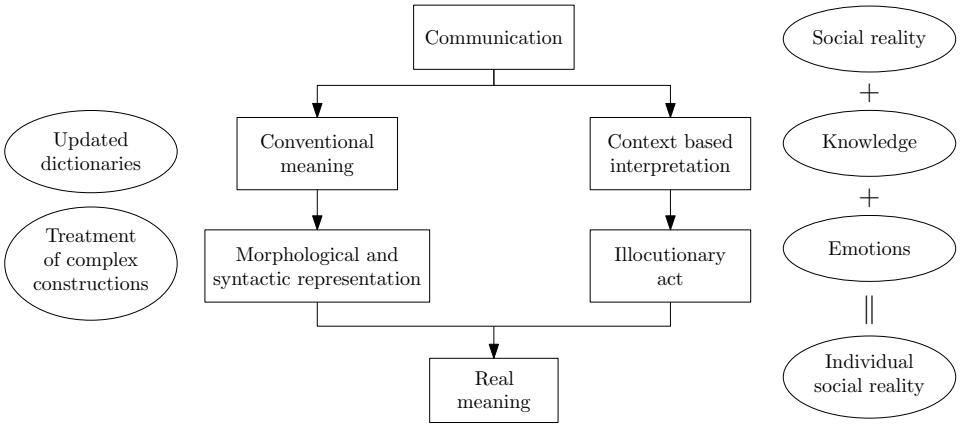

Figure 2.5: Communication act with tools and resources needed for a proper interpretation.

Semantics instead deals with meanings, with the way syntactic structures are interpreted by associating connotative signs with denotative senses. Notwithstanding, interpretation is also a matter of subjectivity due to the individual speaker perspective or the implied significance driven mostly by personal experiences, even though any utterance meaning is anyway domain-dependent. Under this background, pragmatics is defined as a subfield of linguistics and semiotics (the science of signs) in charge of studying the contribution of the context to the meaning, the effect of the domain in the interpretation to overcome apparent ambiguity attributable to the time, manner, place of the utterance. The semantic of a sentence is commonly represented by means of "First-Order Predicate Calculus" which provides, thanks to symbolic logic, a universal and standardized formalization with the purpose of modeling the world. In certain disciplines such as Sentiment Analysis, semantics and pragmatics play a crucial role begin conclusive in determining the polarity. Thus, a word, although it may contain a prepolarity defined in a polarity lexicon, should be considered from the view of the context in which it is framed being likely to be reassigned an inverse sentiment when the domain is taken into account. Under this assumption, researchers opt for creating semantic models for specific knowledge domains to handle their peculiarities and singularities by means of sophisticated rules and dictionaries.

"We must think things not words, or at least we must constantly translate our words into the facts for which they stand, if we are to keep to the real and the true."

Oliver Wendell Holmes Jr.

Along with this substantial technological progress, communication in social networks has dramatically rocketed then coining the concept of the Web 2.0 which describes the phenomenon undergone in the last decade that embraces platforms and forums to share personal experiences and publish user-generated content. These highly interactive platforms allow for co-created messages, discussions or exchange of ideas supplying the agility and dynamism which traditional communication mechanisms lacked of. In turn, language has been also evolving by adjusting the common formal usage in the absence of 'everydayness' dialect to the highest expression of passion and enthusiasm eventually yielding a no well-formed dialogue obviously far from being compliant with the formal language. For instance, a text preprocessing or normalization, comprising automated tokenization, word matching and replacement techniques, is often required to successfully face noisy input which turns arduous or unfeasible for subsequent NLP methods. Nevertheless, misspelled terms or badly formed expressions are full of semantics (i.e. repetition of punctuation marks or the capital letter usage) and are habitually incorporated as evident signs of authorship and writer tendencies.

NLP has been widely applied to diverse casuistry and is becoming increasingly demanded due to its cross-disciplinary nature. In this Thesis, we do not mean to accomplish an entire NL analysis, but rather to shed light on the authorship attribution discipline by delving into the most discriminative stylistic features for the problem at hand and introducing a novel approach based on aforementioned methods.

Chapter 3

Mining Connections to Discover the User

"I think, at a child's birth, if a mother could ask a fairy godmother to endow it with the most useful gift, that gift should be curiosity."

- Eleanor Roosevelt

This chapter elaborates on the first and least intrusive stage of the impersonation detector specifically focusing on identity thefts, which are especially recurrent both for mere forgers and dangerous fugitives in search for a breach to access sensitive information. Identity theft, as introduced in the first chapter, refers to a masquerade attack which is performed through legitimate access identification in order to accomplish illegitimate activities. 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 [94]. 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 the environment involved in the crime and even the messages sent/left by the impostor in the presence of interaction.

During the last decade the boom of the so-called Web 2.0 and the upsurge of debate platforms and social networks have undoubtedly allowed humans to publicly express their feelings and share their opinions on matters of common interest in a much easier and open manner. In this context, 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. However, the vertiginous data flow, rapid interaction mechanisms and multiple information channels provided by these technologies have also laid a rich substrate for cybercrime, since attackers are provided with a huge, distributed repository of valuable multimedia information that enables a wide spectrum of possibilities for their hidden purposes. Unfortunately, despite the efforts invested by cross-national endeavors such as the European Cybercrime Center or the Commonwealth Cybercrime Initiative, most users are not aware of the implications of exposing personal information through these networks, nor are they informed about the possible security breaches caused by their interaction with the platforms. Indeed, users themselves happen frequently to be both the victim and the root of the security breach by granting criminals an easy way to commit cybercrime. Cyber-attackers, being aware of the advantageous situation, have increase the diversity and effectiveness of their malicious activities [82]. 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 [83]. As evinced by the unauthorized access to the details of approximately 250.000 Twitter users in

early 2013 [84], 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 (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. One of the most reported cyberattacks implying an identity fraud is phishing, which is based on sending apparently legitimate messages to users in order to steal their bank credentials or personal information [86, 88, 89]. Phishing attackers or *phishers* elaborate methods for broadcasting messages from apparently reputable sources to collect users' information and to exploit it for their own profit or as a prior step to perform another sort of cybercrime. A straightforward workaround featured by most of the current Web browsers protects users from phishing attacks by automatically filtering out suspicious messages thanks to stored blacklists. Other approaches analyze certain characteristics extracted from the messages themselves to drive the detection of phishing attacks. Currently the research community centers their attention on this last method, mainly motivated by the capability of phishing sources to mutate and camouflage in a highly dynamic fashion.

In this context noteworthy is to mention the work by [85], where a comprehensive set of phishing indicators are grouped and categorized according to the overall criteria on which they are based. These features have been proven to perform effectively when used to detect phishing attacks. Consequently, most of the related contributions have considered a similar list with very scarce modifications. Nevertheless, the fact that attacks can be hidden in an email allows for more involved detection schemes incorporating textual phishing indicators [102] and information retrieval algorithms such as Latent Dirichlet Allocation, Hidden Markov Models or naive bag-of-words procedures [103]. Other features can be extracted from Internet search engines in an attempt at finding inconsistencies between the fake and the true identity [92]. Legitimate sites usually hold a high position for page ranking and an old age in the domain server [93]. These characteristics, jointly with the aforementioned textual features, have been fed to complex classification model relying on machine learning and statistical inference methods [90, 91, 100]. Within this research line it is interesting to highlight the work by [103], where a diverse portfolio of machine learning techniques are analyzed in terms of their Receiver Operating Characteristic curve, which relates the rates of true and false positives of the detection technique.

Notwithstanding the intense activity around phishing attacks, there are other classes of identity theft whose motivation can be very different and whose traces can be more subtle. The recent work by [96] has focused on compromised and legitimate accounts that have been

stolen by an attacker considering that these accounts will be used to spread the same sort of messages through their contact list. Of relevance under the scope of this research is the exploitation of the so-called user behavioral profiles – i.e. expected behavior of the user inferred from historical information about his/her activities in the social network – to detect noticeable changes in the corresponding account that may correspond to an illegitimated use. Behavioral profiles were built by means of a particular class of SVM's applied to a set of content related features extracted from the retrievable data stream of the network. Unfortunately, this work oversees privacy aspects that lie underneath the proposed use of certain characteristics (i.e. topic of the message, contained URL's and language).

From a more general perspective, identity theft is referred to as masquerade attack when it deals with an unauthorized access to personal computer information by using the user's credentials in order to perform illegitimate activities. Masquerading attacks can be detected by capturing and inferring the legitimate user's profile (e.g. the set of usual commands inside a session, applications initiated and other activities such as mail reading or surfing the net) and by detecting behavioral changes potentially due to a suspicious usage of the user's account. [97] proposed to use a dataset of 15000 keyboard commands on a UNIX platform upon which to build a non-timestamped user profile to be tested against several masquerading detection algorithms. Other contributions [87] propose to utilize bag-of-word techniques to model frequencies and changes over time under the hypothesis that a user profile can be better represented by a histogram than by a sequence of commands. The issue arising from the latter approach lies in a high false positive rate because user's behavior in a computer system (with such a variety of possible actions) might not be uniquely specified by a single model. [86] circumvent this issue by first mapping the commands that a user can perform onto a taxonomy.

Most of the above literature assumes that the impersonation attack is driven by economic interests, but the casuistic is far more extensive. Examples abound: a jealous person steals and uses his/her partner credentials to gossip her social network account to search for proofs of a conjugal infidelity; a teenager resorting to impersonation to destroy a classmate's reputation by posting insults against him/her; or a phisher or a spammer willing to get details of the victim's social circle so as to span further the list of recipients of fraud messages. Such a variety of motivations, the low interaction of the attacker when impersonating the victim and the consequent minimal digital footprint left by the impostor call for the derivation of impersonation detection schemes aimed not at the purpose of the attack, but rather at the identity theft itself. This imposes the exploitation of purpose-agnostic features: specifically, connection statistics of the user along time (duration, frequency and periodicity) are crucial to model the interaction grade of the user with the social network, with more or less meaningfulness depending of the regularity of the user when accessing his/her account. These indicators can be input to an early-warning classifier that will notify the user when a statistically outlying account usage is being registered.

This chapter centers 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 [95] 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 [96, 98], 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.

It is important to note that the detection scheme proposed in this chapter is less intrusive in regards to the user privacy when compared to previous approaches since it only relies on time statistics of the network session of the user. Furthermore, it can be applied to impersonation attacks driven by different purposes, and is fully compatible with more involved detection schemes focused on e.g. the social graph of the user or the characteristics of the data flows generated by the user within the network. In this sense, the present work must be conceived as a first, necessary study to formulate the impersonation attack detection paradigm from a formal perspective. The practical scheme for the identity theft detector is framed within a more involved 2-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 identity theft attack, alarm that could be subsequently fed to the other detection stage. This secondary detector, later introduced in the next chapter, will leverage the content itself by turning to e.g. semantic and natural language processing procedures. The early-warning detection stage proposed 1) transforms the connection time information to a feature space yielding more condensed multidimensional profiles for the user under consideration; 2) trains a binary 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. In this chapter, two algorithmic approaches will be described and subsequently discussed in response to the second point.

3.1 Problem Formulation

The detection of impersonation attacks can be mathematically modeled as follows: we label the user for which impersonation is to be detected as A, with connection times denoted by the feature vector $\mathbf{w}_t^A \triangleq \{w_{t,1}^A, \ldots, w_{t,H}^A\} = \{w_{t,h}^A\}_{h=1}^H$. In this vector indexes t and h reflect the

granularity under which connection statistics are registered by the detection tool. Provided that connection duration is captured every day on an hourly basis, t and h represent day and hour index, respectively. This assumption yields $H = 24$ connection time features if these statistics are captured all day and night long. The detection of impersonation attacks can be regarded as a conventional binary test where two mutually exclusive hypotheses, namely

$$\mathcal{H}_0\text{: user } A \text{ has NOT undergone any attack,} \tag{3.1}$$

$$\mathcal{H}_1\text{: user } A \text{ has undergone an attack,} \tag{3.2}$$

are to be tested by resorting to the connection time information $\mathbf{w}_t^A$ stored during a certain period of length T^*, which will be denoted as $\mathbf{W}_{T^*}^A \triangleq \{\mathbf{w}_t^A\}_{t=1}^{T^*}$. Since hypothesis testing is done based on data captured during a certain period the above two hypotheses can be redefined so as to reflect this time dependency, yielding

$$\mathcal{H}_0^{T^*}\text{: user } A \text{ has NOT been attacked at time } T^*, \tag{3.3}$$

$$\mathcal{H}_1^{T^*}\text{: user } A \text{ has been attacked at time } T^*, \tag{3.4}$$

based on which false alarm and detection probabilities can be defined as $P_{fa}^{T^*} \triangleq Pr(\mathcal{H}_1^{T^*}|\mathcal{H}_0^{T^*})$ and $P_d^{T^*} \triangleq Pr(\mathcal{H}_1^{T^*}|\mathcal{H}_1^{T^*})$, respectively. The problem of verifying whether $\mathcal{H}_0^{T^*}$ or $\mathcal{H}_1^{T^*}$ holds for user A can be conceived as a binary classification problem aimed at determining which class the feature vector at time T^* belongs to: 0 (i.e. `no attack`) and 1 (correspondingly, `attack`). This classification task can be mathematically represented by a mapping function $f : \mathbf{w} \mapsto y$, where $\mathbf{w}$ corresponds to the connection times and y represents the class (`no attack` or `attack`) to which vector $\mathbf{w}$ is assigned.

The above mapping $f(\cdot)$ can be inferred from the history of past connection time features $\mathbf{W}_{T^*}^A$ by assuming that no attack has been performed within such a time frame. Three different approaches can be followed on that purpose, which depend on the available a priori information on the attack patterns to be detected:

A. To opt for one-class classifiers that utilize only $\mathbf{W}_{T^*}^A$ for the training process and fit the parameters of a well-known multi-variable probability distribution to the dataset at hand via statistical methods (e.g. Expectation-Maximization).

B. To use non-parametric one-class schemes rather focused on extracting knowledge about the structural properties of the data under analysis by inferring other underlying characteristics such as variable histograms or K-neighborhoods (i.e. no assumption on the distribution of the data is taken).

C. To construct a naïve binary classifier resorting to expert knowledge on the diverse behavioral patterns followed by impersonation attackers so as to generate synthetic connectivity traces to be fed jointly with $\mathbf{W}_{T^*}^A$ as the training set of the model.

This chapter will propose two techniques: 1) a non-parametric one-class model with synthetic data used for exclusively evaluating its performance; and 2) a naïve two-class classifier being fed with a synthetic approximation of the complementary feature space of the user. The former finds its rationale on the evaluation of the detection performance, which requires traces corresponding to impersonation attacks to be fed to the model, while the latter leverages the a priori knowledge about the time characteristics of the attack to be detected by the classifier.

Once a model for $f(\mathbf{w})$ has been constructed by following any of the above strategies, its performance can be quantified in terms of the error function $E^{T^*}(\mathbf{w}, y)$, which are given by

$$E^{T^*}(\mathbf{w}, y) \triangleq \begin{cases} 1 & \text{if } \begin{cases} \mathbf{w} \in A \text{ and } f(\mathbf{w}) = 1 \text{ (attack)} \bigstar \\ \mathbf{w} \notin A \text{ and } f(\mathbf{w}) = 0 \text{ (no attack)} \end{cases} \\ 0 & \text{if } \begin{cases} \mathbf{w} \in A \text{ and } f(\mathbf{w}) = 0 \text{ (no attack)} \\ \mathbf{w} \notin A \text{ and } f(\mathbf{w}) = 1 \text{ (attack)} \clubsuit \end{cases} \end{cases}$$

from which $P_{fa}^{T^*}$ and $P_{d}^{T^*}$ can be computed by accounting for the classification events marked with $\bigstar$ and $\clubsuit$ in the above expression, respectively. The problem can be then formulated as to find a binary classification model (i.e. detector) that simultaneously maximizes $P_{d}^{T^*}$ and minimizes $P_{fa}^{T^*}$ $\forall T^*$, which can be instead reformulated as the joint minimization of $P_{fa}^{T^*}$ and $1 - P_{d}^{T^*} = P_{loss}^{T^*} \triangleq Pr(\mathcal{H}_0^{T^*} | \mathcal{H}_1^{T^*})$, the latter being widely known as *loss probability* or *true negative rate* in the literature regarding binary detection.

3.2 Proposed User Profiling Approaches

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 chapter, 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.

As has been concluded from the previous section, this stage aims at designing a non-intrusive classification scheme that is able to infer potential impersonation attacks in social networks by only requiring connection time information. Such a classification problem can be solved independently by two approaches related to the proposed algorithmic solutions:

A. To find a sufficiently cohesive cluster structure in terms of inter and intra-cluster variance, which allows for a classification phase based on a clustering approach and a subsequent outlier analysis. The classification technique may be devised as an unsupervised algorithm relying on a naive approach such as K-Means or via an unsupervised learning algorithm in charge of building up a more complex model derived from finding patterns from the user connectivity traces.

B. To process the initial connection statistics prior to clustering for subsequently achieving an optimized cluster structure in terms of attack detection performance.

On the other hand, 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. 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.

Without loss of generality and for the sake of understandability 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 3.1.

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, users 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.

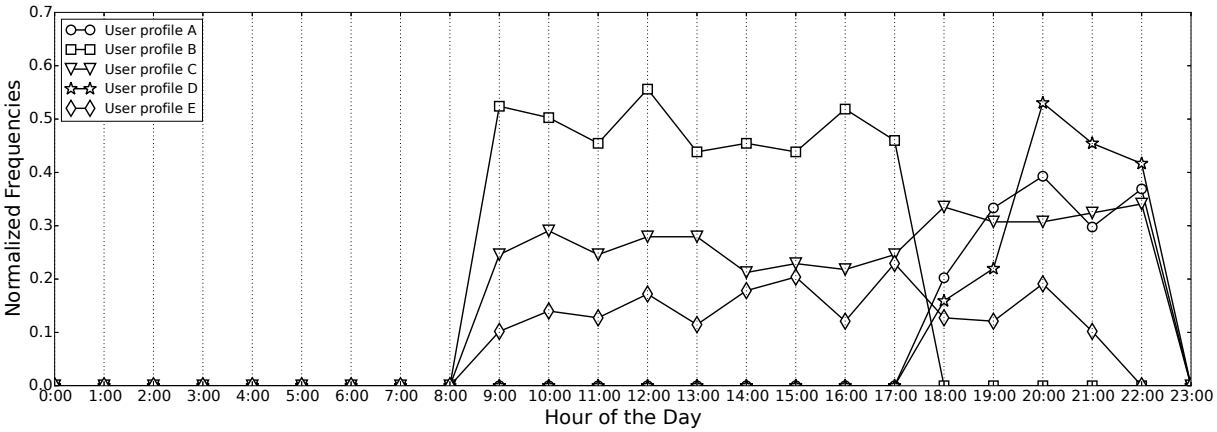

Figure 3.1: Normalized frequencies of the user profiles considered in the experiments. For instance, 50 % of the total number of days a user following profile A would connect to the social network at 9:00.

We believe that by using statistical distributions to generate realistic traces this work will stimulate further research aimed at capturing and building real connection time datasets. Nevertheless, we have attempted at validating these realistic assumptions on the regularity of social network users by analyzing the time-stamped wall post dataset published in the context of the WOSN 2009 conference [101]. This repository contains Facebook wall posts of 46952 users, based on which aggregated posting frequencies of all users were computed and normalized on an hourly basis. The purpose of this aggregation is to find behavioral patterns that justify our validation strategy under the assumption that posting implies a connection of the user generating the message. The reciprocal of this assumption does not necessarily hold in practice, i.e. a user may not generate any interaction traces beyond exploration (gossiping). Nevertheless, this prior exercise provide a valuable insight on the real usage frequencies of the users within the database. In this context, the results shown in Figure 3.2 proved that the overall social network usage agrees with the main modeling assumption taken in this chapter: most users tend to connect and interact with the social platform at the same afternoon and evening hours, observation that buttresses the construction of behavioral models hinging on this fact. If such an assumption did not hold, the usage histogram depicted in Figure 3.2 would rather obey a uniform function, from where no overall connection pattern could be inferred. By performing the same aggregation and normalization procedure in a set of synthetic connection traces generated by the above A, B, C and D models, the resulting frequencies are compliant with those from the dataset in [101], yet sharper due to the behavioral differences established among such synthetic models.

In what follows the analysis focuses on working days due to the more expected connection

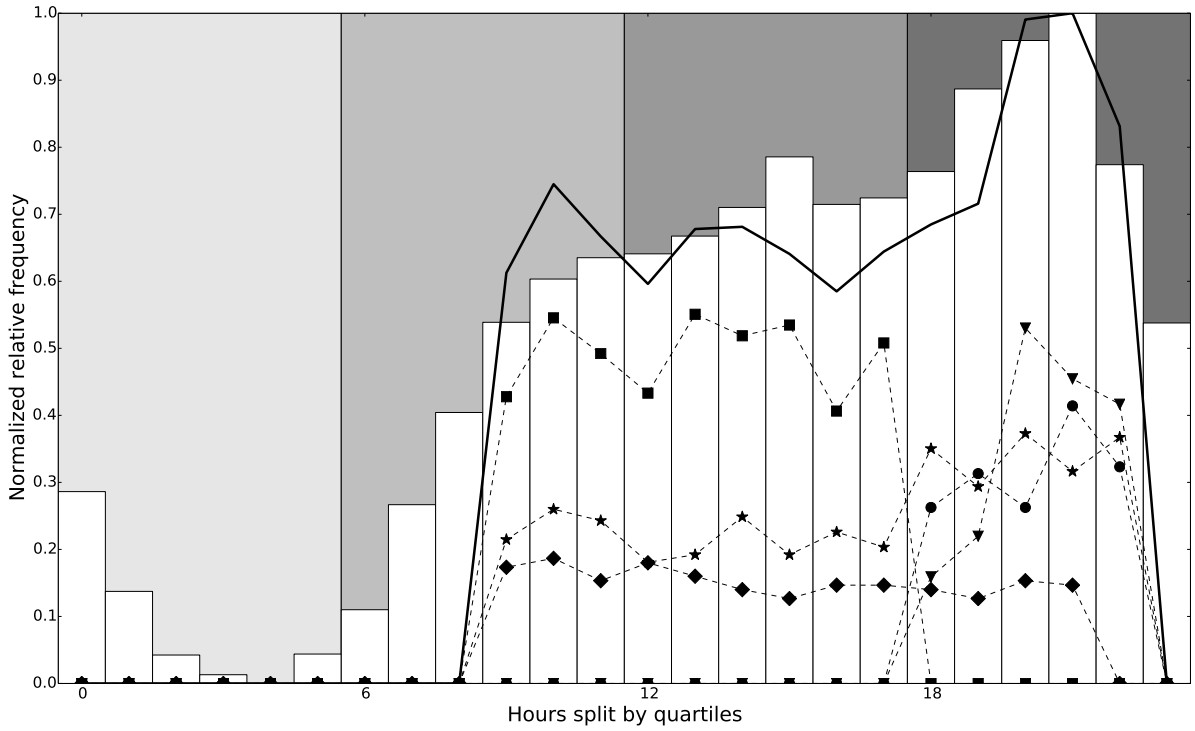

Figure 3.2: Real (white bars) and emulated (black lines with different markers per model) connection time frequencies ordered by quartiles. The black bold line corresponds to the overall connection model computed by aggregating patterns of all models. Models A, B, C, D and E as explained in the text are marked with ▼, ■, ★, • and ♦ markers, respectively.

regularity of the users with respect to the weekend, when the user has more leisure time and their behavioral pattern could be different and irregular. Whichever the algorithmic approach, connection time information 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 overfitted 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.

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 classifiers, especially 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 $\mathbf{w}^A_{T^*}$ 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 3.1 the new feature space is defined by $\mathbf{w}^A_t \triangleq \{w^A_{t,n}\}^N_{n=1}$ 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.

The following two sections will hereafter explain each selected algorithm and present the performance of each proposed detector in relation with the two mentioned approaches earlier discussed.

3.2.1 SVM as a Supervised Learning Algorithm for the Impersonation Detector

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 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 [100]. This section 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 3.3 and discussed in Section 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.

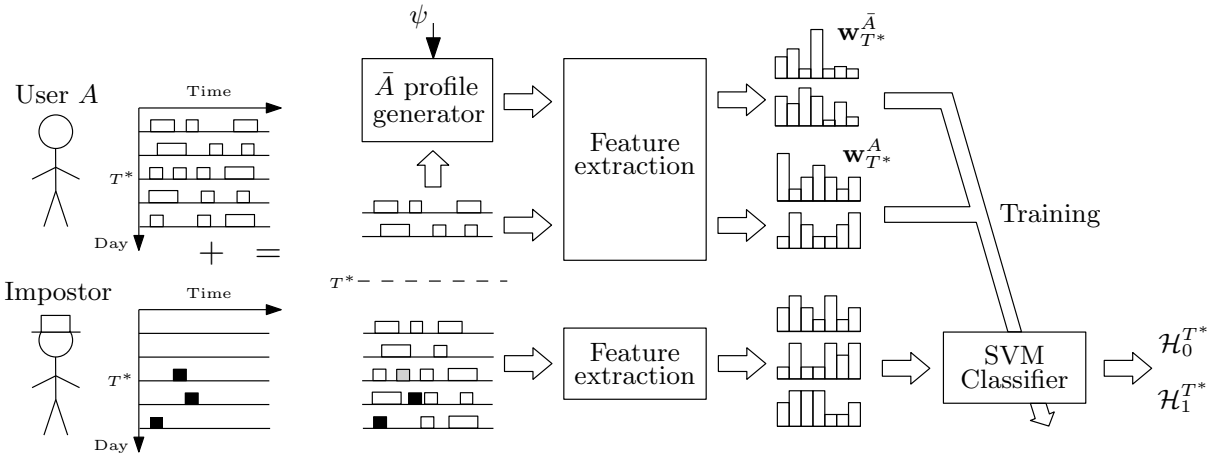

Figure 3.3: Proposed connection time based detector of impersonation attacks.

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 3.4.a, when $\psi = 0$ distant traces (light solid lines) from the user ones (light dashed lines) are produced, hence representing better detectable identity theft attacks. However, as ψ increases (Figure 3.4.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.

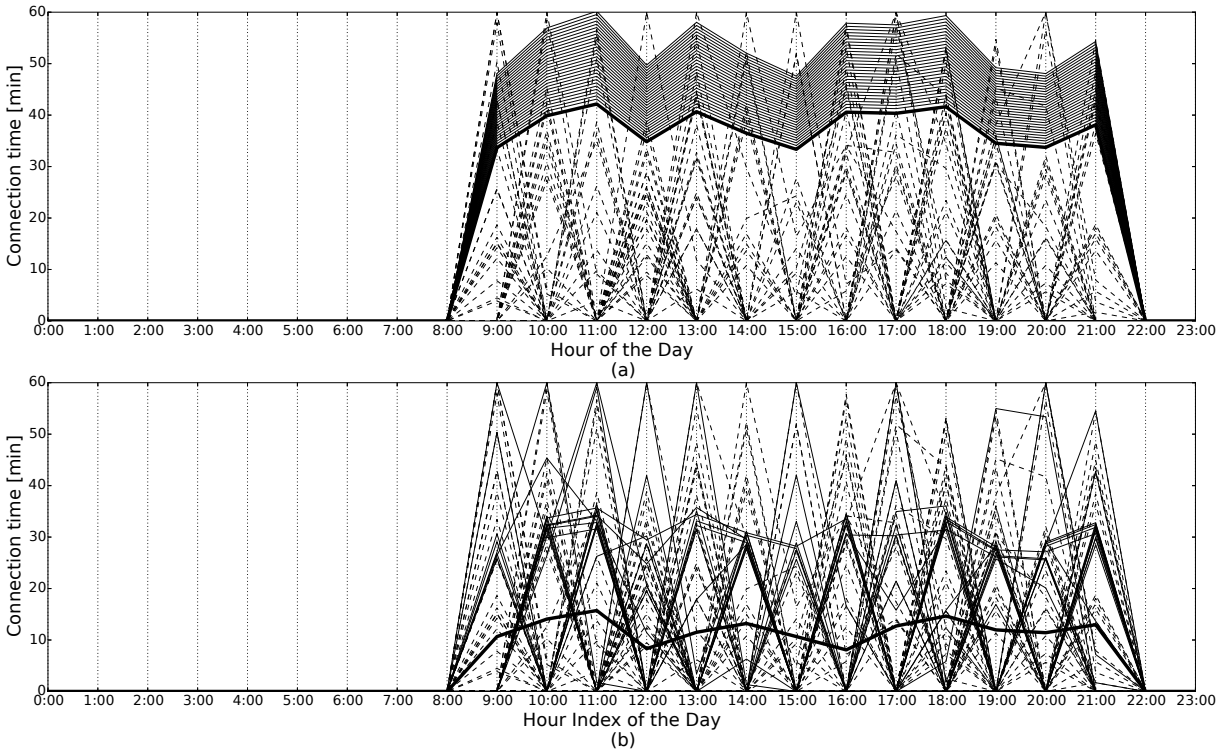

Figure 3.4: Exemplifying realization of user connection time traces (light dashed lines) overlapped with a complementary space (light solid lines) generated with a) $\psi = 0.0$; b) $\psi = 0.5$. The bold solid line represents the upper statistical limit of the trace region spanned by the traces of the user (given by the sum of their mean and standard deviation).

The complementary space in Figure 3.4.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.

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 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 $1 - \psi$).

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. Therefore, 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.

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 3.4.a to 3.4.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 3.4.a, 3.4.c and 3.4.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.

When implementing a practical detector based on the setup proposed in this section, 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,

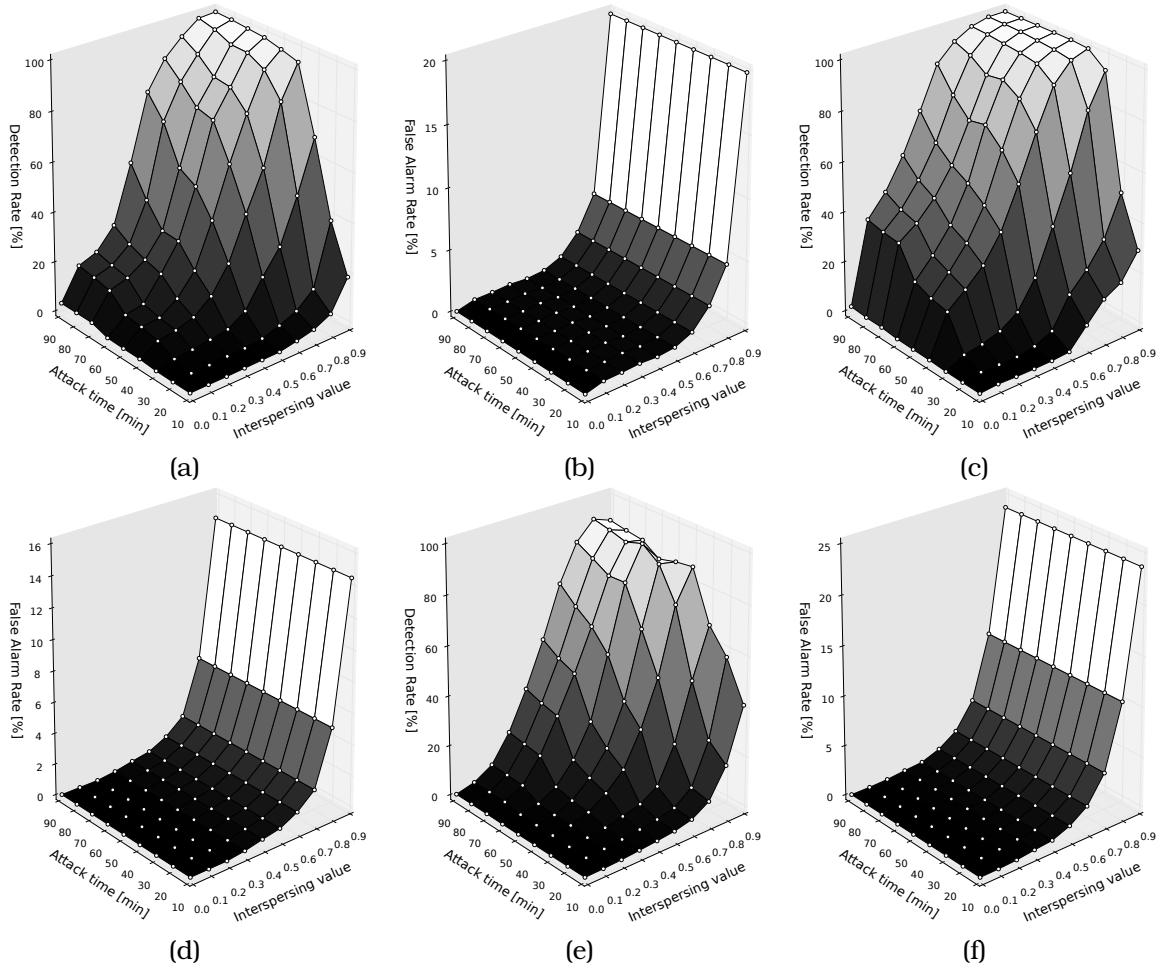

Figure 3.5: Performance results in terms of average detection ($P_d^{T^*}$) and false alarm ($P_{fa}^{T^*}$) rates for profile A (subplots 3.4.a and 3.4.b), profile B (subplots 3.4.c and 3.4.d) and profile E (subplots 3.4.e and 3.4.f). Results are depicted as a function of the interspersing parameter ψ and attack time T_{attack}.

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.

3.2.2 HS as a Meta-heuristic Learning Algorithm for the Impersonation Detector

According to the earlier introduced approach B as the possible algorithmic solutions for the impersonation detection, the current solver outlines two interrelated preprocessing stages: first, connection traces are mapped to the alternative, reduced feature space; secondly, such features are weighted prior to the clustering stage. The reason behind this two-fold approach hinges on the fact that an ideal representation for an outlier-based detection would imply training samples yielding a very concentrated feature space and as far as possible from other potentially similar behavioral patterns that might correspond to impersonation attackers. A detection approach should trade between creating many concentrated clusters (which could lead to an overfitted detection model with an increased rate of false alarms) and few yet broad clusters

(correspondingly yielding an underfitted model with a low rate of detected attacks). The use of an alternative feature space along with its processing through an optimized set of weighting coefficients is postulated as an efficient scheme to build an optimal cluster space not only in terms of the aforementioned structural variance trade-off, but also for maximizing the detection performance of the overall scheme.

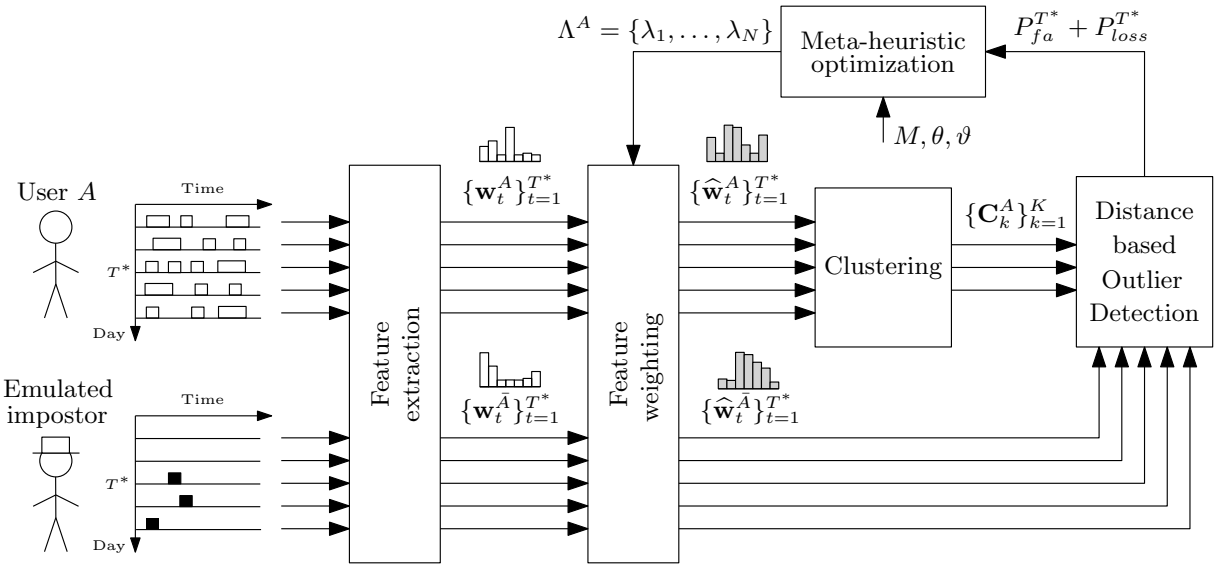

Figure 3.6: Block diagram of the proposed connection time based detection for impersonation attacks.

In line with the above explanation, the proposed system depicted in Figure 3.6 builds upon two interrelated processing subsystems, each in charge for a different functionality. On one hand, unsupervised learning (clustering) allows characterizing behavioral patterns based on a set of distinctive features derived from the connection traces aimed at uniquely profiling users' habits. The well-known K-Means algorithm [26] will be adopted to partition the weighted features in K clusters according to a euclidean distance similarity measure. The challenge when resorting to this specific clustering algorithm is to select the value of K that optimizes a certain measure of the structural fitness of the corresponding cluster space. As such, the so-called Elbow method determines the optimal number of clusters by measuring either the compactness or the separation of the clustering arrangement – in terms of the ratio of either within-cluster or between-cluster variance correspondingly to the total variance of the dataset – when new clusters are progressively considered. On the other hand, a bio-inspired meta-heuristic algorithm optimally modulates the weighting coefficients imposed on the aforementioned clustering features towards an overall optimized system in terms of detection performance. In discriminant analysis, formulae such as the Fisher's linear approach are often utilized to determine the weight based on the ability of a variable to distinguish between two distinct groups. However, in order to dismiss any assumption on the linear separability between the user's and an eventual attacker's connection traces, in our case the estimation of the optimal weights is delegated to a bio-inspired meta-heuristic solver, which iteratively refines the values of such weighting coefficients based on the successive application of several operators to a population of solution candidates. The subsequent clustering of any given weighted set of features allows for simple attack detection strategies based on distance-based outlier analysis over the group of inferred connection patterns.

Likewise, the literature has so far elaborated on different versions of this popular clustering algorithm incorporating methods to infer the number of clusters K from the input data, which vary from greedy constructions based on cluster-structural hypothesis tests [16, 18] to combinations of evolutionary meta-heuristics and special solution encoding strategies [104]. This work takes the assumption that synthetic attack traces will be produced and utilized based on a priori knowledge on the casuistry of impersonation attacks. This supervised information can be of great help when deciding the value of the K parameter. Consequently, enumerating different values of K and opting for the value maximizing the detection performance metric will be selected as an straightforward method to incorporate and benefit from the availability of supervised data in contrast to conventional counterparts based on structural properties of the produced cluster sets.

The above clustering scheme operates on a set of connection time features $\widehat{\mathbf{w}}_t^A$ preprocessed through a set of multiplicative coefficients $\mathbf{\Lambda}^A = \{\lambda_n^A\}_{n=1}^N$, which are used as optimization variables for a subsequent optimization algorithm aimed at refining their value towards maximal impersonation attack detection performance. This solver will tune the values of these weighting coefficients so as to minimize the fitness metric $\xi(\mathbf{\Lambda}^A)$ given by

$$\xi(\mathbf{\Lambda}^A) = P_{fa}^{T^*} + P_{loss}^{T^*}, \tag{3.5}$$

i.e. the sum of the loss and false alarm probability of the overall system when computed over connection time traces over a history time range T^*.

To efficiently minimize the above metric through optimally setting the values of $\mathbf{\Lambda}^A$ the system relies on a continuous version of the Harmony Search (HS) algorithm which, since its invention by [42], has been proven to be competitive and in most cases, outperforming with respect to other evolutionary heuristics for optimization paradigms arising in very diverse fields. As explained in Chapter 2, this superior performance finds its roots in their operators, which blends together evolutionary elements such as polygamy and metric-driven differential mutation in a probabilistic fashion.

Following the notation posed in [42] and employed in the previous chapter, the search procedure of the HS meta-heuristic algorithm is controlled by three different improvisation operators iteratively applied to a pool of candidate solutions. Such operators emulate the musical improvisation process of a jazz band when all its members jointly seek a melodious harmony from an aesthetic criterion. As such, improvisation based on previously played musical notes is referred to memory consideration, whereas low-range pitch tuning is denoted as pitch adjustment. Thereby, notes for a certain instrument are improvised based on the experience of the musician and small pitch changes based on its current value. When translated to the optimization realm, improvised harmonies correspond to candidate solutions, whose notes – optimization variables – are refined via crossover and mutation operators that imitate musical memory consideration and pitch adjustment.

This being said, notes in the HS solver particularized to the problem tackled in this chapter represent the values of the weighting coefficients applied to the connection time features prior to subsequent clustering and evaluation. For the sake of notational coherence we will refer to the M-sized pool of iteratively improvised harmonies as

$$\{\mathbf{\Lambda}_m^A\}_{m=1}^M \triangleq \{\{\lambda_{m,n}^A\}_{n=1}^N\}_{m=1}^M \tag{3.6}$$

which are progressively refined and evaluated in terms of the fitness function $\xi(\mathbf{\Lambda}_m^A)$ from Expression (3.5). The refining operators are controlled by two different probabilistic parameters:

$\eth \in [0, 1]$, which drives the note-wise rate at which the so-called Harmony Memory Considering Rate (HMCR) operator is applied; and $\vartheta \in [0, 1]$, which acts correspondingly for the Pitch Adjusting Rate (PAR) procedure. Such processes are defined as follows:

- HMCR: the algorithmic parameter $\eth$ establishes the probability that the new value for a given weighting parameter $\lambda^A_{m,n}$ is uniformly drawn from the values of the same coefficient in the other $M-1$ candidate solutions, e.g.

$$\eth \triangleq Pr(\lambda^A_{m,n} = Z_{m,n}) \tag{3.7}$$

  where $Z_{m,n}$ is a discrete random variable uniformly distributed in the set

$$\{\lambda_{1,n}, \ldots, \lambda_{m-1,n}, \lambda_{m+1,n}, \ldots, \lambda_{M,n}\}$$

  with $m \in \{1, \ldots, M\}$.

- PAR: in this case ϑ establishes the probability that the new value for a given weighting coefficient $\lambda^A_{m,n}$ is picked from its vicinity. To this end, a maximum variational bandwidth parameter $BW \in \mathbb{R}^+$ is defined so that the new value for weight $\lambda^A_{m,n}$ will be given by

$$\lambda^A_{m,n}\ (\text{new}) \quad = \begin{cases} \lambda^A_{m,n}\ (\text{old}) + BW\text{rand}(-1, 1) & \text{probability } \vartheta, \\ \lambda^A_{m,n}\ (\text{old}) & \text{probability } 1\text{-}\vartheta, \end{cases} \tag{3.8}$$

  i.e. by a subtle differential mutation based on the previous value for the note at hand. In the above expression $\text{rand}(a, b)$ denotes a realization of a continuous random variable uniformly distributed in the interval $[a, b]$.

At every iteration of the HS algorithm the above two processes are independently applied to every note $n \in \{1, \ldots, N\}$ and harmony $m \in \{1, \ldots, M\}$, yielding a new set of M potentially refined candidate solutions. These newly produced harmonies are evaluated, concatenated to those from the previous iteration and sorted in increasing in terms of their fitness metric. Finally, the M *best* (first in the ordered set) harmonies are kept for the next iteration, whereas the rest of candidate solutions are discarded due to their comparatively worst fitness.

The above weight optimization wrapper is integrated in the overall algorithmic procedure of the proposed detector summarized in Algorithm 1. When evaluating the fitness of any given harmony or candidate solution, it should be made clear that every such harmony undergoes the K-Means clustering stage, followed by an analysis of outliers over test data incorporating synthetic attack connection traces. In other words, as shown in Figure 3.6 the values of $\mathbf{\Lambda}^A_m$ are applied to the set of features $\widehat{\mathbf{w}}^A_t$ as

$$\lambda^A_{m,n} \cdot w^A_{t,n} \triangleq \widehat{w}^A_{t,m,n}, \tag{3.9}$$

which are input to the clustering algorithm, which produces a set of K clusters representing behavioral patterns of the above weighted connection time features. A simple rule set based on statistical outliers is henceforth applied to the aforementioned test data, which incorporates ground-truth labels thanks to a priori knowledge on the characteristics of impersonation attacks. This is an important aspect that will be later elaborated through the discussion of the results.

Algorithm 1 Proposed impersonation attack detector

Require: Number of features N; Valid values of the HS parameters (M, ∂, ϑ, BW); number of iterations $\mathcal{I}$; transformed connection time traces $\{\widehat{\mathbf{w}}_t^A\}_{t=1}^{T^*}$ for the user; a fraction of the user traces for measuring false alarms; synthetically generated attacks $\{\widehat{\mathbf{w}}_t^{\bar{A}}\}_{t=1}^{T^*}$ for computing true positives; maximum number $K_\circ$ of clusters to be considered.

Ensure: A set of coefficients $\{\lambda_n^{A,opt}\}_{n=1}^N$ and an optimal number of clusters K^{opt} that minimize the aggregate rate of false alarms and missed attacks (true negatives) over the input datasets.

1: Initialize the memory of candidate weight vectors $\{\{\lambda_{m,n}^A\}_{n=1}^N\}_{m=1}^M$ with values drawn uniformly at random from the set $\mathbb{R}[-\lambda_\circ, \lambda_\circ]$, with $\lambda_\circ = 1$ without loss of generality.
2: **for** iteration in $\{1, \ldots, \mathcal{I}\}$ **do**
3: Apply HMCR controlled by ∂ to $\lambda_{n,m}^A$ for $n \in \{1, \ldots, N\}$ and for $m \in \{1, \ldots, M\}$
4: Apply PAR controlled by ϑ and BW in the same fashion as in the HMCR procedure.
5: Obtain weighted features $\{\widehat{w}_{t,m,n}^A\}_{n=1}^N$ for every candidate solution $\{\lambda_{m,n}^A\}_{n=1}^N$ in the harmony memory as stated in Expression (3.9).
6: Evaluate $\xi(\mathbf{\Lambda}_{m,k}^A) = P_{fa,m,k}^{T^*} + P_{loss,m,k}^{T^*}$ by performing an outlier analysis on the cluster space generated by K-Means for each $k \in \{1, \ldots, K_\circ\}$ (considered number of clusters).
7: For each $m \in \{1, \ldots, M\}$ (candidate solution within the memory of the HS solver), keep the solution corresponding to the fitness value within the set $\{\xi(\mathbf{\Lambda}_{m,1}^A), \ldots, \xi(\mathbf{\Lambda}_{m,K_o}^A)\}$ that best balances between the number of clusters and the marginal gain in terms of detection performance. Refer to this solution as $\xi(\mathbf{\Lambda}_m^A)$.
8: Concatenate the updated harmony memory $\{\{\lambda_{m,n}^A\}_{n=1}^N\}_{m=1}^M$ with the memory from the previous iteration (if any), and sort the concatenated memory in increasing order of their associated fitness $\xi(\mathbf{\Lambda}_m^A)$.
9: Filter out the *worst* (last) ordered M harmonies, and keep the remaining ones for the next iteration.
10: **end for**
11: The sought optimal $\{\lambda_n^{A,opt}\}_{n=1}^N$ and K^{opt} are given by $\mathbf{\Lambda}_0^A \triangleq \{\lambda_{0,n}^A\}_{n=1}^N$ and the number of clusters k associated to their assigned metric $\xi(\mathbf{\Lambda}_0^A)$ (Step 7).

In this first experiment a synthetic set of two-dimensional data have been produced so as to analyze the functionality of the algorithm in a more affordable and simplified visual representation. Specifically, 48 individuals represent one of the classes (marked with ◯, □, ◇ and ⋄, which reflect usage patterns of the actual user), whereas 12 individuals correspond to the other class (impersonation attacker, marked with ☆). As depicted in Figure 3.7, no information is apparently rendered by the first feature (horizontal axis) in terms of discrimination between both classes. Nevertheless, the feature represented by the vertical axis for the attacker class is forced to be slightly (0.05 %) higher than the value of the vertical coordinate of the closer individual belonging to the other class. In other words, the vertical axis contains pertinent yet difficult-to-exploit knowledge for the differentiation task.

This first set of simulations addresses a comparison between the naive K-Means scheme that uses the Elbow method to estimate the number of clusters and the K-Means approach combined with HS meta-heuristics herein proposed. The Elbow method evaluates the accumulated intracluster variance enhancement over increasing values of K to yield an optimum value of $K = 4$ clusters, which is set as an input parameter to the K-means algorithm to produce the cluster arrangement shown in Figure 3.7. A distance-based outlier criterion can be imposed on

this set of clusters representing the legitimate user so as to complete the detection process, e.g. to declare that an individual belongs to the attacker class if its distance to the closest centroid (depicted in black for each cluster or pattern) is higher than the maximum distance from the latter to the members of its represented cluster. In this case, the naive system fails to correctly identify all attacker individuals, resulting in 12 wrongly classified samples. The reason being that the similarity metric to construct the cluster space and the variance criterion to select the number of clusters are inherently structural and not coupled to the overall performance of the detection scheme where it is embedded.

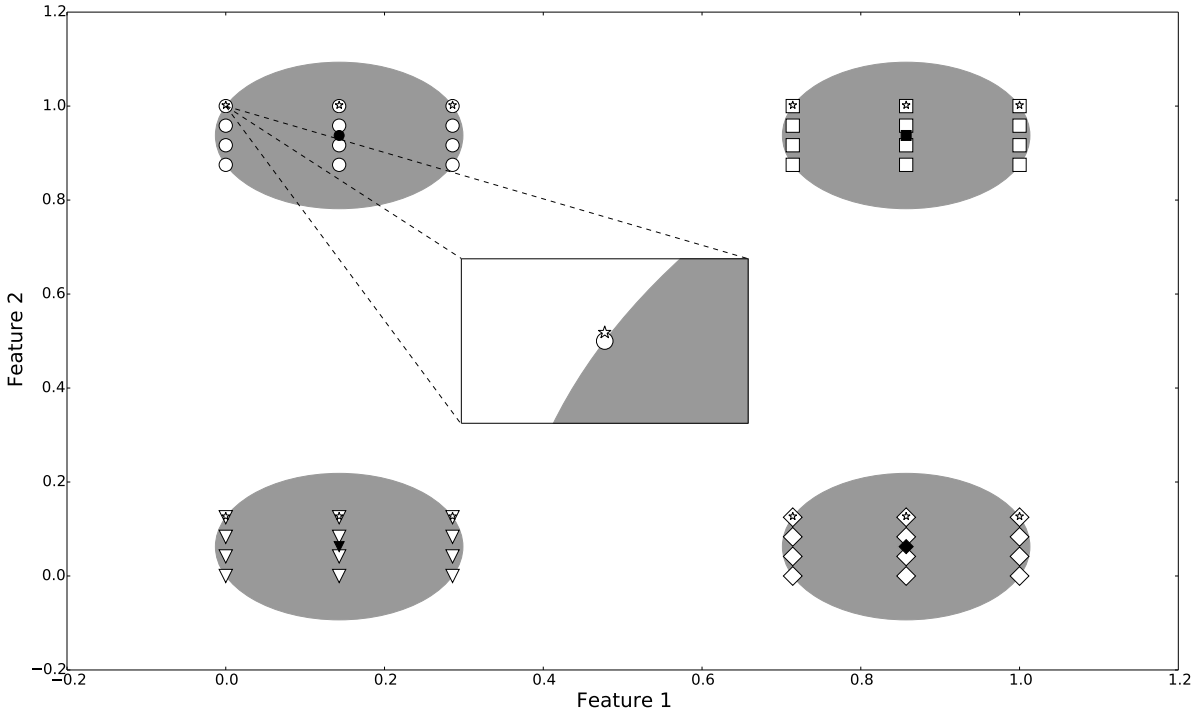

Figure 3.7: Cluster arrangement for the first experimental scenario provided by the combination of the K-Means algorithm and the Elbow method. The area shaded in gray corresponds to the space where connection time patterns are declared as legitimate.

On the other hand, the HS solver optimally determines that the weights applied to all individuals should *stretch* the vertical axis so as to generate a higher distance from the attacker individuals to the closer centroid. Accordingly, the produced weights $\{\hat{\eta}_1, \hat{\eta}_2\} = \{0.016, 4.801\}$ after 18 iterations of the meta-heuristic optimization loop results in the $K = 4$ cluster arrangement in Figure 3.8, where the same outlier-based criterion as the one used above identifies correctly all attacker individuals (namely, $P_{loss}^{T^*} = P_{fa}^{T^*} = 0$).

We now proceed with a second simulation setup where the detection performance of the proposed scheme is evaluated over connection time traces generated by the 5 realistic behavioral models presented at the beginning of this Section. To this end an impersonation attack will be emulated by assuming that the impostor connects once at a time drawn uniformly at random during every day under evaluation. As before, the attacker access the victim's account for a certain time T_{attack}: intuitively, the higher T_{attack} is, the better the resulting connection time trace of the user should be declared as potentially subject to an impersonation attack. The experiments discussed herein aim at validating this intuitive claim, as well as at assessing

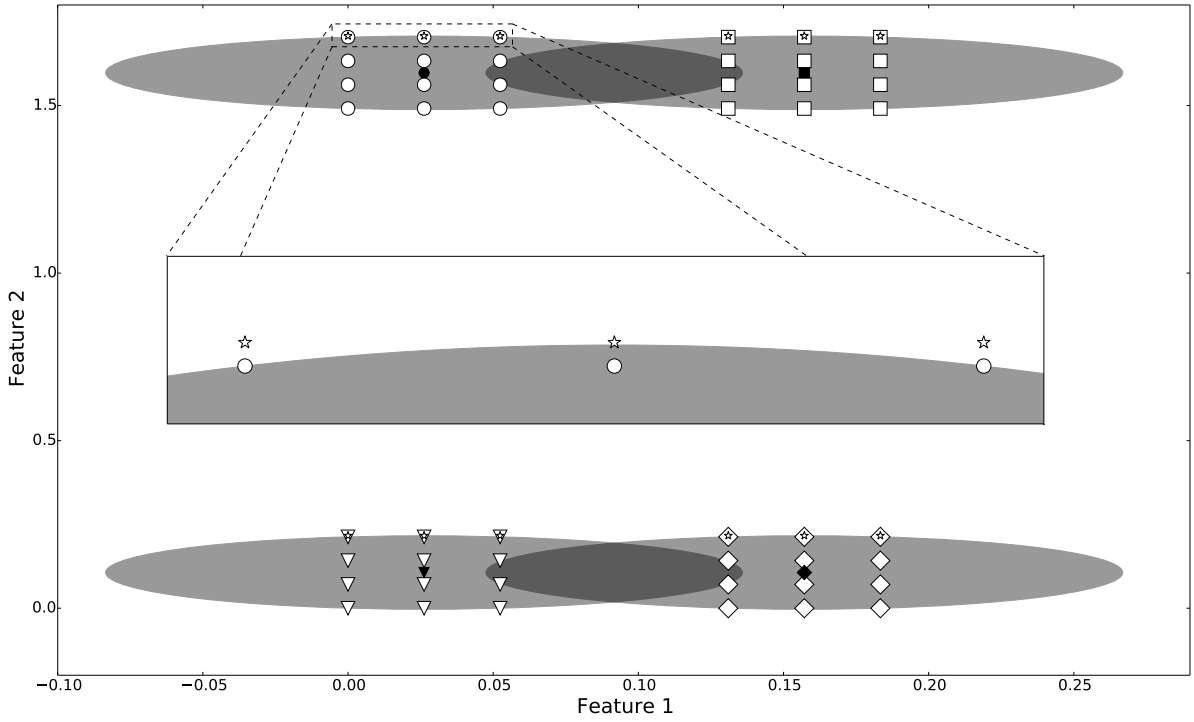

Figure 3.8: Cluster arrangement for the first experimental scenario provided by the combination of the K-Means algorithm and the HS solver. It is important to notice that axis have been scaled for a better understanding of the plot.

the performance improvement of the proposed unsupervised detector due to the performance-driven meta-heuristic optimization of the cluster space and the alternative feature space to which original traces are projected.

To this end, several detectors are considered:

- Detector 1: this first scheme operates on the original, untransformed set of connection time features (i.e. number of connections per hour and their duration). In particular, the system comprises naive K-Means clustering supported by the conventional Elbow method for the selection of the number of clusters, as well as a distance-based outlier detection based on the maximum sample-to-centroid distance within each produced cluster.

- Detector 2: this approach is identical to Detector 1, but operating on the set of transformed features proposed in the previous section. This second scheme permits to assess the contribution of such alternative features to the ultimate performance of the impersonation classifier.

- Detector 3: this third system implements the overall meta-heuristically optimized detection chain in Figure 3.6 and summarized in Algorithm 1.

Before proceeding with the presentation and discussion of the obtained results, it should be emphasized that since we are dealing with unsupervised detection, synthetic attacks are used exclusively for the evaluation of $P_d^{T^*}$, i.e. the rate of true positives. Furthermore, it is

important to note that the distance-based criterion for isolating outliers is set the same for the three detectors under comparison, namely, to declare a potential attack if the distance of the sample being analyzed to the centroid of the cluster to which it belongs is higher than 10 % of the maximum intra-distance within the cluster. Finally, the parameters of the meta-heuristic wrapper have been set to $M = 50$, $K_\circ = 10$, $\eth = 0.5$, $\vartheta = 0.1$, $BW = 1$, and $\jmath = 50$ iterations, whereas detection statistics have been computed over 100 samples corresponding to the user (for evaluating $P_{fa}^{T^*}$) and 100 samples undergoing a randomly placed attack of duration T_{attack} (respectively, $P_d^{T^*}$). Due to the stochastic nature of the HS search operators, statistical results have been computed over 30 Monte Carlo simulations.

We start the discussion by Figure 3.9, where detection performance statistics for users following Model A – regular, relatively long connections in the evening – are depicted as a function of T_{attack} for the three detectors under comparison. The comparison also includes $T_{attack} = 0$, which corresponds to the extreme case where no attack is included in the validation set. This permits to check the capability of each scheme to isolate changes in the behavioral patterns of the user that correspond to the statistical randomness of the model itself rather than to an impersonation attack.

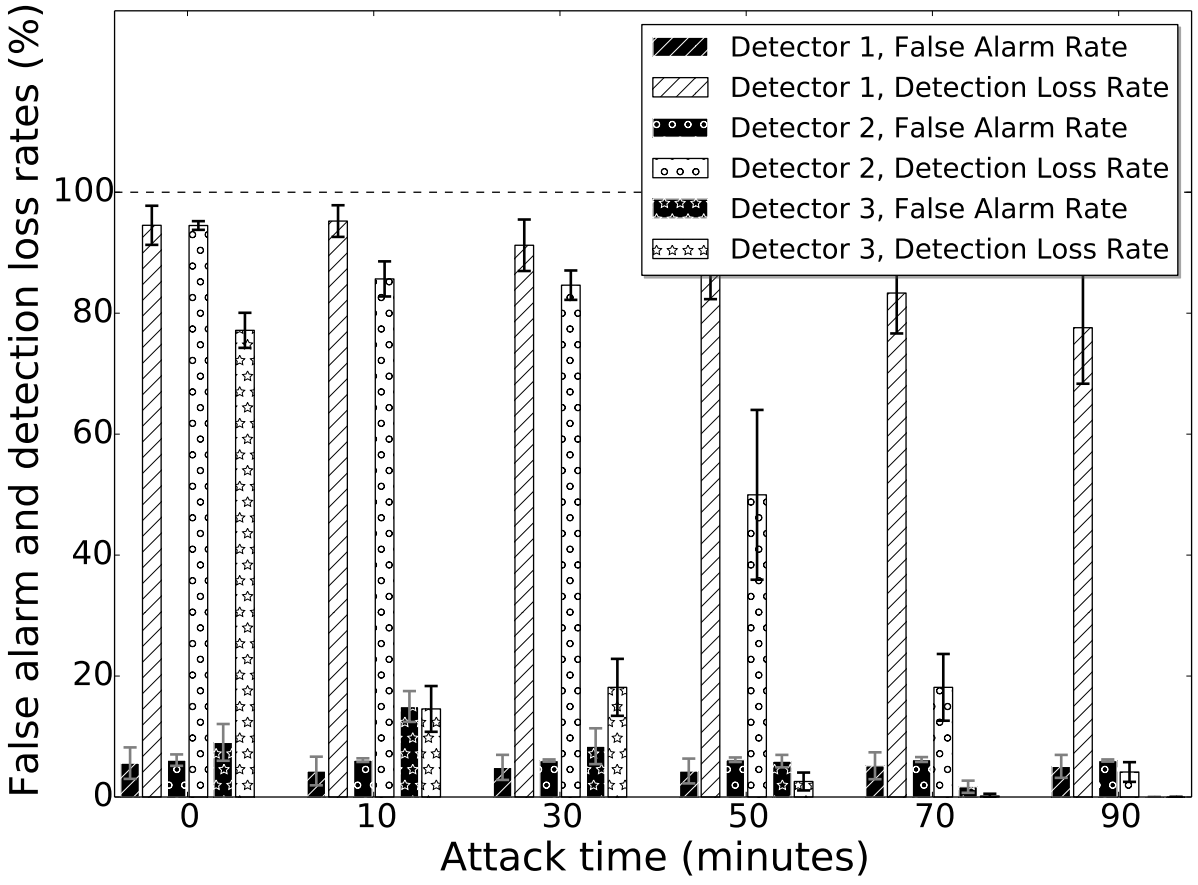

Figure 3.9: False alarm ($P_{fa}^{T^*}$) and detection loss rates ($P_{loss}^{T^*} = 1 - P_d^{T^*}$) for model A as a function of the duration of the attack T_{attack} (in minutes). Bar amplitudes and whiskers represent the mean and standard deviation of the detection metrics computed over the Monte Carlo simulations.

This being said, it can be observed in the plot that the false alarm rate $P_{fa}^{T^*}$ of all schemes

under comparison are significantly lower than their detection loss rate $P_{loss}^{T^*}$, which finds its rationale in the condensed cluster structure within the model. On the contrary, the detection rate $P_d^{T^*}$ increases for all the detectors as the attack time ζ increases, as expected due to the accordingly higher changes on the behavioral model of the user. Note, however, that the detection performance increases sharply for detectors 2 and 3. In the case of detector 2, the inclusion of the transformed feature set produces a cluster space where impersonation attacks are more easily detected as outliers. In regards to detector 3, the performance improves further to yield a probability of detection above 80 % for 10-minute impersonation attacks. A similar analysis on other considered connectivity models comes to the same conclusions. As exemplified in Figure 3.10 for model B, the unsupervised scheme incorporating HS meta-heuristics outperform significantly both the naive K-Means clustering approach (detector 1), even if utilizing the transformed set of features (detector 2).

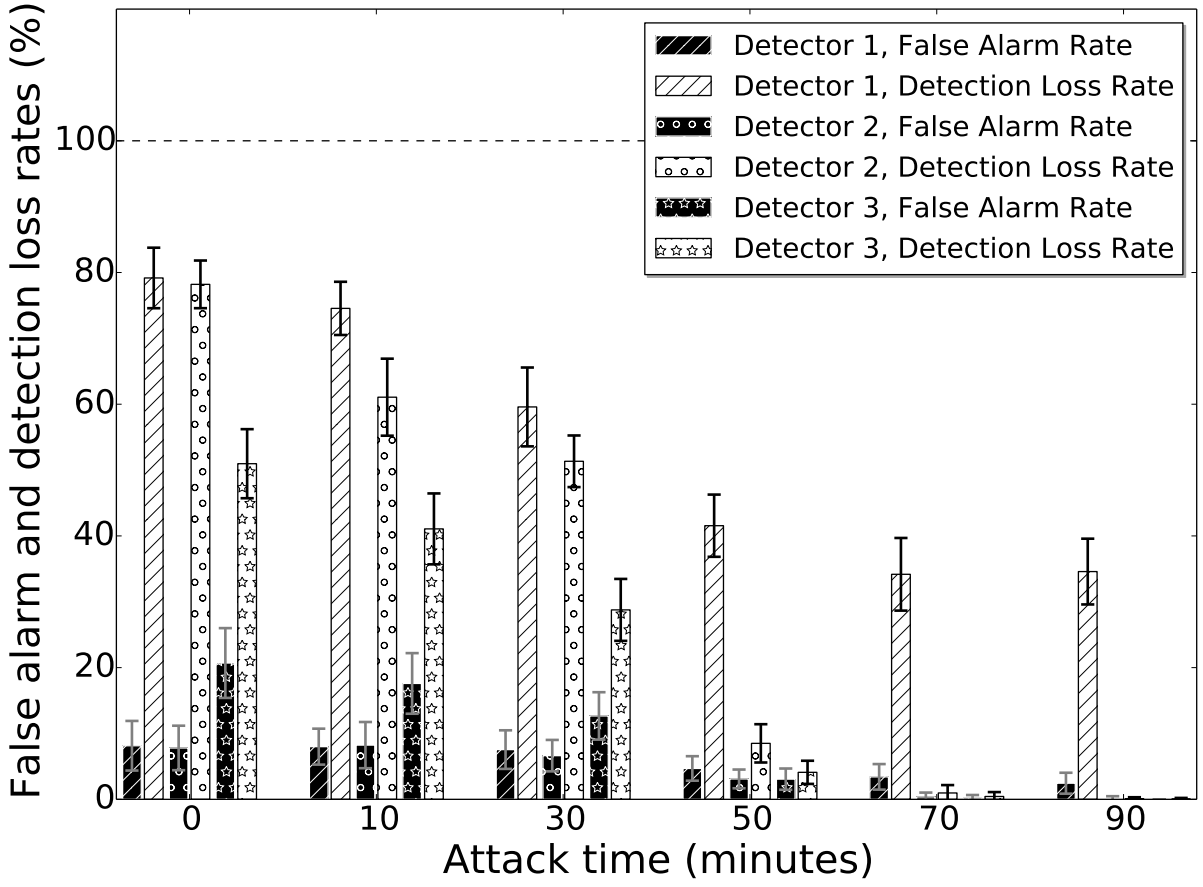

Figure 3.10: False alarm and detection rates for model B as a function of T_{attack} (in minutes).

Interestingly, when addressing model E (i.e. relatively long connections all day long distributed on a uniform random basis), Figure 3.11 elucidates that in this case the transformed features of detector 2 do not render any substantial gain with respect to the original feature space (detector 1). The reason lies on the fact that the uniform randomness of the user connections along the daytime and their subsequent mapping onto the transformed feature set fail to provide K-Means with valuable information about the heterogeneity of the model. Note in the same plot, however, that the meta-heuristically empowered approach proposed in this work

(detector 3) excels at overcoming this issue by optimally equalizing the input features.

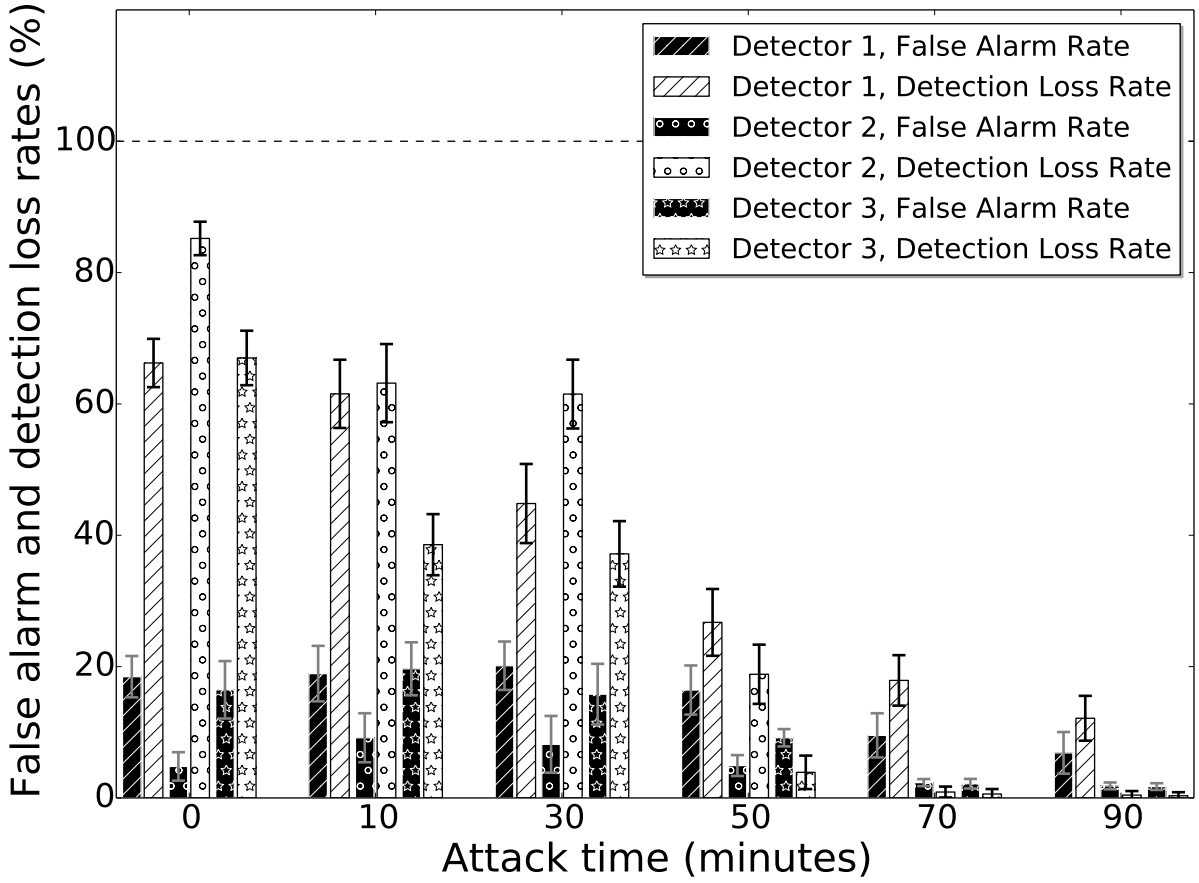

Figure 3.11: False alarm and detection rates for model E as a function of T_{attack} (in minutes).

3.3 Comparison between the Proposed Schemes

In order to comparatively assess the performance of the impersonation detection schemes proposed in this Chapter, Table 3.1 collects the previously discussed detection metrics P_d and P_{fa} for each of such detectors, different attack times $T_{attack} \in \{10, 30, 50, 70, 90\}$ (in minutes). All figures are presented in %. The parameters characterizing the simulations run to obtain these results do not vary with respect to the ones utilized in prior subsections.

In order to establish a comparison baseline, the discussion starts by establishing a minimum detection performance threshold P_d^{th} that permits associating different models. If we aim at stringent detection requirements (e.g. $P_d^{th} = 0.9$), one can notice that Detector 3 (namely, the hybrid HS-based model) attains such a performance level for $T_{attack} \geq 50$ minutes over all users under analysis (i.e. A, B and E). On the contrary, for the SVM approach such a detection threshold is achieved for $T_{attack} \geq 70$ minutes (user A, interspersing value $\psi = 0.7$) and $T_{attack} \geq 50$ minutes (user B, interspersing values $\psi = 0.7$ and $\psi = 0.9$). Even if the performance gain is clear when focusing on the average attack time that can be detected with

one family of approaches or the other (e.g. for user E the SVM scheme is not able to reach $P_d > 0.9$ no matter what the duration of the attack is), a closer insight reveals that the same detection performance level in the SVM case entails a noticeably higher rate of false alarms due to a strongly overfitting of the model to the training feature space characterizing the user at hand. This effect due to overfitting is observed through all the considered models, not only between the SVM and the HS-based model, but also between different interspersing values in the former model.

Table 3.1: Comparison between the proposed impersonation detectors based on connectivity time information in terms of average probability of detection (P_d) and average probability of false alarm (P_{fa}), both measured in %.

			SVM					HS + K-Means		
	T_{attack}	%	0.1	0.3	0.5	0.7	0.9	Det. 1	Det. 2	Det. 3
User Model A	10	P_d	0.63	1.33	3.60	12.31	54.23	5.21	16.60	85.50
		P_{fa}	0.00	1.00	5.80	16.50	47.60	4.33	5.11	12.23
	30	P_d	3.26	10.03	25.26	50.96	61.60	10.81	17.62	82.13
		P_{fa}	0.00	1.00	5.80	16.50	47.60	4.21	5.23	8.23
	50	P_d	9.20	27.86	60.76	84.63	77.63	13.31	52.32	94.80
		P_{fa}	0.00	1.00	5.80	16.50	47.60	4.82	5.00	5.11
	70	P_d	17.86	43.36	69.06	90.61	85.81	17.31	83.89	98.81
		P_{fa}	0.00	1.00	5.80	16.50	47.60	4.80	5.22	2.80
	90	P_d	21.61	50.96	79.03	96.43	87.13	22.63	95.76	99.61
		P_{fa}	0.00	1.00	5.80	16.50	47.60	4.91	5.41	0.53
User Model B	10	P_d	0.56	0.53	3.83	19.03	41.63	25.62	39.93	59.81
		P_{fa}	0.46	0.83	2.66	10.31	29.62	7.25	7.36	16.83
	30	P_d	6.33	13.20	37.81	76.46	87.24	41.81	50.58	72.63
		P_{fa}	0.46	0.83	2.66	10.31	29.62	7.05	6.71	11.30
	50	P_d	20.76	43.61	81.40	97.03	97.16	59.73	90.67	95.39
		P_{fa}	0.46	0.83	2.66	10.31	29.62	4.66	3.96	4.11
	70	P_d	26.66	45.96	82.73	97.77	97.41	67.84	98.16	99.19
		P_{fa}	0.46	0.83	2.66	10.31	29.62	4.11	0.04	0.02
	90	P_d	32.16	60.13	93.76	99.02	98.61	67.45	99.68	99.81
		P_{fa}	0.46	0.83	2.66	10.31	29.62	2.83	0.02	0.01
User Model E	10	P_d	0.10	0.96	5.93	24.78	41.04	38.71	36.17	63.33
		P_{fa}	0.00	0.63	2.13	12.16	22.21	18.54	8.41	18.82
	30	P_d	0.26	2.96	13.46	40.76	65.92	55.16	39.51	65.81
		P_{fa}	0.00	0.63	2.13	12.16	22.21	19.14	8.13	15.20
	50	P_d	0.66	8.46	36.93	73.96	84.82	74.33	81.40	94.11
		P_{fa}	0.00	0.63	2.13	12.16	22.21	15.81	5.46	9.67
	70	P_d	1.51	20.76	56.36	82.73	87.58	82.73	97.77	98.88
		P_{fa}	0.00	0.63	2.13	12.16	22.21	9.82	3.71	3.87
	90	P_d	2.00	25.61	62.76	86.31	89.81	89.32	98.54	99.56
		P_{fa}	0.00	0.63	2.13	12.16	22.21	7.43	3.12	3.31

Likewise, it is interesting to note that Detector 3 significantly outperforms any other scheme in terms of P_d. Although some other configurations may yield better false alarm figures (specially for short attacks, e.g. Detectors 1 and 2 for User A versus Detector 3), such alternative schemes

are extremely conservative. Interestingly, when the impersonation attack is short (small values of T_{attack}) the false alarm rate of Detector 3 is generally higher – or in the same order – than that of Detectors 1 and 2 for any given user. This is again due to the short distance in the feature space between licit and non-licit connection time patterns; in this situation, Detector 3 prioritizes the detection of subtle attacks over false alarms. As the attack time T_{attack} increases, so does the detection probability of these alternative schemes as a result of more discriminable attack patterns; however, the false alarm rates featured by Detector 3 decrease sharply, hence dominating the benchmark in terms of both P_d and P_{fa}, with scores close to error-free detection for the highest values of T_{attack} considered in the benchmark. This conclusion, along with a better interpretability of the HS-based schemes with respect to their SVM counterparts, makes Detector 3 of the hybrid HS-based the best performing approach of this comparison study.

3.4 Conclusions

This chapter 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 chapter 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, 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 two distinct classifiers, SVM with a novel complementary space formulation depicting the negative class for a binary classification fashion, and a bio-inspired meta-heuristic algorithm (Harmony Search) which provides the required parameters for an optimized K-Means clustering stage in terms of attack detection performance.

Nevertheless, even though a meaningful comparison can be made in terms of performance, it is important to firstly note the difference between the number of samples entailed in the both classifiers training phases (tuned in response to a minimum level of detection rate) being the second approach more appropriate when the detector has not recorded enough connection time traces to capture the habits of the user with respect to the usage of his/her social network account. The K-Means based approach is postulated to fit best in the early operational stages of the algorithm by virtue of its capacity to accommodate fluctuations in the behavioral patterns of the user under analysis. 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 schemes proposed in this chapter must be conceived as early warning approaches 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 (introduced in the next chapter).

CHAPTER 4

MINING CONTENT TO DISCOVER THE AUTHOR

"There are two sorts of curiosity - the momentary and the permanent. The momentary is concerned with the odd appearance on the surface of things. The permanent is attracted by the amazing and consecutive life that flows on beneath the surface of things.´´

- Robert Wilson Lynd

Behavior not only comprises the physical and observable actions carried out by the individuals, but also their emotions and therefore the aggregate of responses to internal and external stimulus. Nevertheless, a behavior may be related to a certain context, to a specific emotion and to a particular stimulus but it could give rise to the wrong perception and incorrect assessment that such attitude or activity is commonly adopted by the person. Under this basis, this chapter attempts to shed some light in the authorship attribution research and, as a rebound effect, in impersonation detection approaches by mining textual features from a novel perspective centered on discerning the essence, the unvarying behavioral pattern closest to the unconscious cognitive processes by relying primarily on Pragmatic theories concerned with the context of the discourse.

Authorship attribution refers to the discipline aimed at uniquely distinguishing the author or writer of a certain text by processing and analyzing features extracted from the content under consideration. Although early studies in this field date back to more than 100 years ago, this area has undergone a sharp activity increase in the last decade as the result of several groundbreaking advances in machine learning and Natural Language Processing (NLP) schemes, further ignited by new data management and processing trends under the so-called Big Data paradigm. Similarly, related subareas such as author profiling and authorship deception have furnished the literature with a plethora of contributions dealing with the application of supervised models to these problems. Disregarding the ultimate aim of such contributions, most of the works reported to date resort to similar text representation strategies, writer-specific features and classification models, mostly at the pace dictated by the progress in information retrieval and computational intelligence.

Traditionally the scope of authorship attribution has been mostly focused on long pieces of text delivered over unidirectional and/or non-interactive communication means (e.g. from books to letters and research articles), which intuitively implies a stylistic content homogeneity. However, the advent of more dynamic messaging applications such as the Short Message Service (SMS), chats, micro-blogging and social networks has given rise to recent experiments over these interactive, bidirectional channels [105, 106, 107]. All contributions tackling this

particular communication scenario state that short-length textual contents pose a technically involved challenge due to the shortage and low diversity of vocabulary that determines the predictive richness and uniqueness of the extracted features and ultimately, the accuracy of the subsequent classification model [108, 109].

Vocabulary richness can be indeed measured at different levels. To begin with, stylometry refers to the application of linguistic and tonal style analysis striving to unambiguously identify the writer. According to [110], stylistic features include a vast number of lexical, syntactic, structural, content-specific, and idiosyncratic style markers. Despite its inherent computational complexity and reduced scalability, off-the-shelf text parsing can exploit lexical, token-based and syntactic features, yet they result in a high variance and inaccuracy when dealing with undersized text samples. Vocabulary richness can be also measured by counting either unique terms or those word forms appearing exclusively once or twice in the entire text under analysis (i.e. the so-called *Hapax* and *Dis Legomena*), which can be often approximated by a Zipf distribution [111]. Other measures of vocabulary richness have been extensively used as word-base methods to characterize the variance and diversity of a given glossary[1]. Measures such as Yule's, Simpson's, Honoré's, Sichel's or Brunet's are devised to diverge from prior biased procedures (e.g. the archetypical type/token ratio measure) in relation to the text size [113]. Nonetheless, when dealing with short texts in non-formal contexts a convenient option to characterize the lexical register is to adopt the character n-grams scheme by representing the elocution in n-sized splits of letters, being this at the same time a robust and noise-tolerant artifice to contend with spelling errors.

Furthermore, the morphology of the written record may be justified by the evaluation of the structural aspect of language. Syntactic features comprise sentence and word length distribution, the use of pronouns, conjunctions or other parts-of speech of interest, which reveal the developed and desired structure to heighten aspects such as the connections amongst the constituents, subject and clause shifts and broadly speaking the mixture of diction and grammatical complexity. Parts-of-Speech (PoS) tagging groups together those words with similar grammatical function, being usually combined in bigrams or trigrams to register the discourse assemblage [108]. Other novel approaches have opted for gathering and arranging them into rewrite rules manifesting, in terms of frequency, the hierarchy and composition preferences [114]. On the other hand, connective particles, as conversational connections, reflect the nature of the segmentation from a structural or grammatical view, as well as the writer temper and inclinations [115].

In line with the vocabulary richness and uniqueness required for authorship identification, it is important to note that certain grammatical categories empower the meaning and the semantics conveyed inside the prose specifying the attitude or mood of the author, whereas others are typically structural and bear little significance or connotation. Topical elaboration is well represented by adverbs and adverbial expressions, whereas slightly appreciable semantic information can stem from: 1) the tense and aspect of the employed verbs; 2) the relations between a PoS node and its children in the parsing tree [116]; and 3) the use of lexical resources such as synonyms or hyperonyms, which are used whenever ambiguity or abstraction is required [117]. However, not only strict and order-driven grammar must be captured by language parsers, but also phrase and predicate-argument structures can be functionally inferred by deep parsers to avoid wrong PoS miss-classifications when facing the recurrent ambiguity of attachment in

[1]The authors refer to [112] for a more thorough description.

long-distance relationships [118].

As mentioned before, most literature resorting to the above measures of linguistic richness for authorship attribution has focused on literary records rather than interactive communications in which the speech is generated in a dynamic basis with a two-way flow of information. In this alternative communication scenario the majority of the previously surveyed measures are deemed short and insufficient in terms of prediction accuracy. The rationale behind this statement lies in the intuition that in such a dyadic communication, the noise factor distorts the message by introducing elements such as inattention, disinterest or cultural differences, which finally produce as many diverse dialogues as receivers are involved. These multiple channels subject to different contexts and influences call for an independent analysis of their heterogeneity in an attempt at preventing authorship attribution models from counting on exceptional, occasional or receiver-influenced linguistic features (as pointed out in the very first paragraph of this chaptr).

When approaching authorship identification from a machine learning perspective, the ideal scenario is that where instances (i.e. texts) belonging to the same category (correspondingly, author) are confined within compact clusters in the space spanned by the utilized features. In such an idealized setup clusters should be restricted to the feature *essence* of the writer so as to avoid building strongly adjusted predictive models capable of capturing multiple yet infrequent patterns of the exchanged messages. This may eventually lead to overfitting and consequently, to a poor predictive performance in terms of the generalization properties of the classifier. This manuscript gravitates precisely on how to isolate in practice the feature essence of the writer from the context and the influence of the receiver on the message so as to exploit it in authorship attribution. This hypothesis has a particular application focus on impersonation and identity theft attacks in social networks; previous contributions have commonly highlighted the increasing incidence and severity of this class of subtle cybercrimes, particularly within the teenage community [96, 95]. The requirements posed on an impersonation detector in terms of scalability and sender diversity may find a suited technological response in the concept of feature essence tackled in this chapter.

An experimental setup has been designed to explore our hypothesized concept of feature essence and its impact on an authorship attribution scenario. At this point it is relevant to recall that our hypothesis postulates that content features coming from distinct yet distinguishable senders allows for a new feature selection criterion based on isolating the linguistic pattern of the sender that is invariant with respect to the receivers. This essential feature set is expected to impact positively on the subsequent authorship attribution task in terms of 1) the scalability and computational complexity of the utilized classifier, due to a vastly reduced feature space; and 2) the generalization properties of the model as the number of users grows, with more relevant, predictive characteristics being fed to its learning procedure.

We begin by briefly delving into the dataset utilized within the experiments. The selection was based on the main premise that interactive communication channels are more likely to develop contextual and receiver influences onto the exchanged messages. Intuitively short messages lay the foundation of the dyadic and interactive discourse, being thus a suitable corpus for analyzing senders' linguistics when communicating to multiple receivers. This motivates the myriad of datasets chosen in previous contributions, mostly encompassing extracts from newspapers [115, 119] or books [108]. Other data sources have been lately explored as a result of the growing proliferation of Social Media, which fosters the creation and exchange of

user-generated content via new highly interactive channels such as blogs [116], public on-line forums or message boards [109, 118] IRC chatting systems [105, 106, 109] or micro-blogging environments [120], among many others.

Unfortunately, to the best of the authors' knowledge no dataset containing messages retrieved from social networking platforms is publicly available in the Internet. However, it is important to note that nowadays user habits in terms of social media have evolved towards an ubiquitous usage in mobile phones. This is especially frequent within teenagers, which gets even more usual by the latest proliferation of interactive communication means (e.g. chat) embedded in the application itself. As a result linguistics in Social Media have progressively converged to those of traditional schemes. Thereby, for our experiments we have opted for the NUS SMS Dataset [122], which is a collection of 65296 English SMS messages compiled by researchers from the School of Computing of the National University of Singapore between 2011 and 2014. Among all 65 senders within this dataset, those with at least 4 receivers with more than 100 messages have been selected for the experimental phase, accounting for a total of 13036 messages. First a minimum of 100 messages was imposed between every sender-receiver pair so as to ensure enough data to characterize the linguistic usage in the communication process, yielding a total of 27 eligible senders. Out of them only 6 users met the requirements of at least 4 different receivers. This filtering permits analyzing the concept of essence posed in this manuscript without any eventual side effect due to a low number of messages and/or receiver diversity. It should be also emphasized that to the knowledge of the authors, no other contribution has been previously made with this specific dataset apart from spam filtering (see e.g. [123, 124, 125]).

4.0.1 Feature Selection

Once the dataset has been selected, we delve into the concept of linguistic essence. In formal contexts, singularities are often derived from the frequency of word, n-gram or syntactic elements being considered as specific author's stylistic choices. The variance or the information contained in such distinctive elements determine the precision in the identification of the sender. Nevertheless, in more dynamic environments as the one considered in this chapter other characteristics must be addressed such as the usage of emoticons and/or punctuation marks, which usually evince the emphasis or the intensity of the text. Based on this rationale, a selected set of features has been assembled so as to compile the linguistic peculiarities of every sender within a diversity of contextual communication scenarios. The overall set of linguistic predictors comprises the following items:

1. Word-based features: word length distribution of the message and the character trigrams.
2. Grammatical features: adverbs, adjective and first-person frequency.
3. Syntactic features: PoS Bigrams, sentence complexity (measured as the number of composite – coordinated or subordinate conjunctions – clauses) and function words distribution.
4. Social media and instant messaging based features: punctuation, distribution of emoticons and slang abbreviations (own compiled dictionaries with 181 and 1137 regular expressions, correspondingly).

Before any further grammatical or syntactic processing, trivial procedures have been applied to normalize the messages within the dataset: removal of capital letters, repeated characters and slang abbreviations, the latter after annotation[2]. The word length embraces the so-called concept of readability as a text-inherent factor quantifying the lexical involution. The syntactic features have been extracted by means of a model trained on Twitter driven by the Stanford PoS Tagger [126], which allows for a sophisticated treatment of the inconsistency and ungrammaticality of the messages within this particular dataset. The analysis of adverbs and adjective usage represents the topic elaboration and the grade of quality description implemented by the sender. In turn, sentence complexity refers to the tendency to construct subordinated or dependent and coordinated phrases, which implicitly quantifies the intricacy of the syntax structure of the message at hand.

Once these features have been computed and collected for each message, it should be noted that their cardinality may increase exponentially with the number of distinct senders. For instance, the number of different trigrams compiled over the dataset depends on the diversity and similarity of the messages exchanged among different sender-receiver pairs, and is closely linked to the concept of essence postulated in this chapter. Indeed, an empirical analysis of the total number of features, PoS bigrams and trigrams reveals that they all grow as the number of senders increases. However, as shown in Figure 4.1 differences are minimal when considering the last sender within the selected database. Interestingly this manifests the fact that in a dynamic corpus with short messages as the one utilized in this Thesis, linguistics are more likely to become homogeneous and less diverse. Nevertheless, from the plot it should be inferred that when determining whether any given message corresponds to a given sender, any criterion focused on reducing the overall number of features should be of interest in order to avoid subsequently overfitted classification models and decrease the computational complexity of their training process.

When dealing with classification tasks, information gain, odds ratios or tests such as the Kullback-Leibler divergence [127] or Chi-Squared [128] are commonly applied in the search for the most discriminatory predictors. Nevertheless, these typical feature selection algorithms are exploited as an early and independent stage and regardless the context or the problem at hand. In this chapter we propose a rather different feature selection algorithm well-suited for multi-class authorship attribution models composed by independent OvO (One Versus One) classifiers.

For the sake of understandability, in what follows *essential* and *influential* features will stand for the selected feature subsets by the proposed technique, which springs from theoretical concepts of linguistics. When aggregating features from dyadic dialogues in an attempt to discern the sender of a message a numerous of sporadic, context-dependent linguistic elements can be captured, which are likely to generate over-sized collections of features and potentially overfitted classification models. This expansion phenomenon will become sharper when dealing with hundreds of senders and thousands of messages as in social networks; the detection of impersonation attacks in this scenario requires a fine-grained characterization of the linguistic feature essence of each user in order to avoid very intricate decision regions for the classifier that could eventually lead to a high rate of false positives. We assume that essential patterns can be more productive in the long run when considering several aspects:

[2]This annotated slang corpus can be made available on demand.

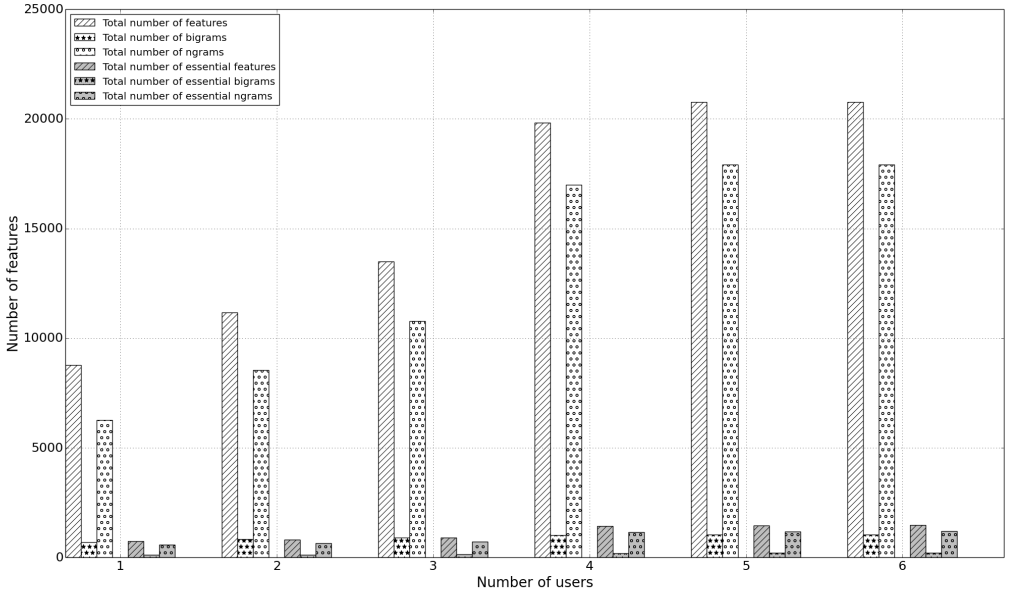

Figure 4.1: Feature growth rate for both total and essential features versus the number of senders to be identified.

- The instability was a criterion for feature selection introduced in [121] under which those terms that can be changed by other alternative terms (synonyms) conform the stylistic choices of the author, as opposed to those ones of forced usage such as some prepositions or monosemic lexicon with no feasible variants. In an informal communication context as the one held through social networks nearly the opposite approach holds: authors do not elaborate on the opinions nor use complicated language to polish the message content. Frequent discourse choices are broadly representative of the sender, and may become advantageous to discern them from each other.

- As the number of users grows, so does the cardinality of the feature set and potentially, the amount of behavioral patterns that must be discerned by the classifier. This gives rise to a higher complexity of the model due to the need for partitioning regions in the feature space that lead to well-generalized decisions in regards to the authorship of the messages under test. Furthermore, it should be assumed that the short average length of the messages and their usage context could unavoidably homogenize their features and consequently, imply a loss of predictability in regards to their authorship that cannot be overridden (not even by a social psychology specialist). In addition, we cannot expect linear patterns related to topics for a specific individual over the time; conversations through the considered communication channels use to be more topically diverging from each other, often implemented over more diverse vocabulary than in books, articles and more static media.

Based on the above two observations, a more flexible feature selection algorithm has been devised. It should be clear that without loss of generality, the proposed scheme can be applied to both one-class and multi-class authorship attribution models. While the former corresponds

to the detection of impersonation attacks in social networks, the latter is deemed appropriate for the characteristics of the selected dataset. When dealing with one-class classifiers the performance assessment usually becomes more involved than the multi-class set, in part due to the lack of ground of truth to which to benchmark the obtained predictive outcomes. This being said, the derived feature selection method is hereafter contextualized and put to practice over a multi-label classification scenario where the sender for the tested SMS's must be discriminated. Nevertheless, discussions will be held on the extrapolation of the conclusions extracted therefrom to the one-class case.

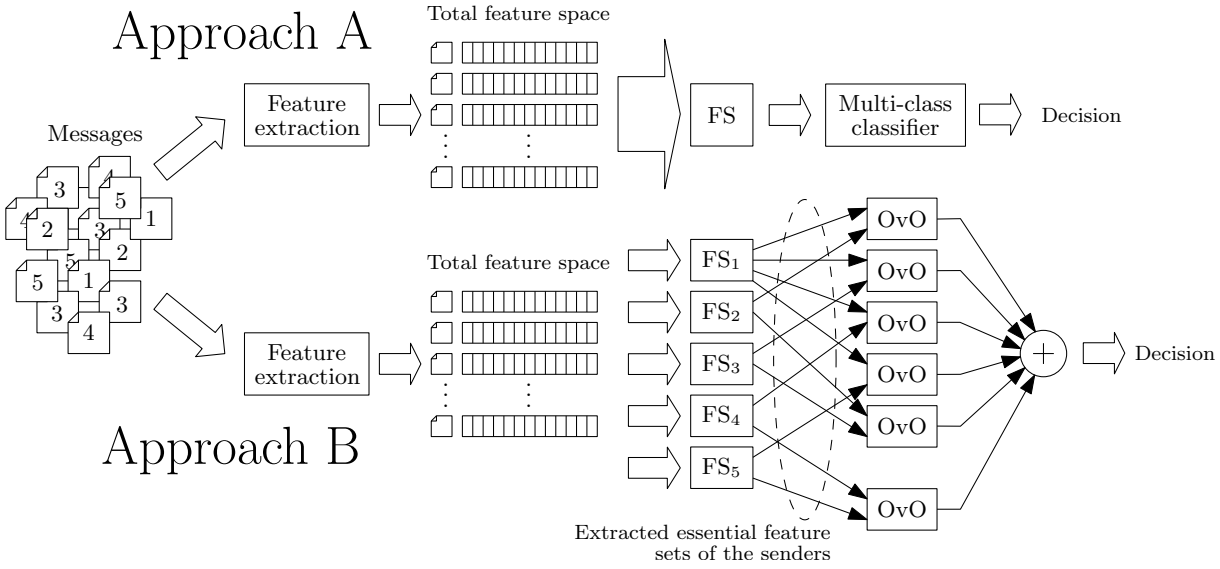

Figure 4.2: Diagram showing the two considered sender identification approaches.

The feature selection scheme proposed in this Thesis aims at isolating the feature essence of each user, and exploiting the union of the essential feature sets of the pair of senders to be discriminated by each OvO classifier. In reference to Figure 4.2, two approaches can be arranged in this envisaged application scenario:

- Approach A: this correspond to a naive concatenation of a feature selection (FS) stage and a multiclass classifier. This preprocessing stage is in charge of discriminating the most predictive features from the overall set of extracted characteristics. As aforementioned in the introduction, this can be performed in very diverse ways. In this context, it is not the purpose of this work to comparatively assess the performance of such a plethora of techniques in the application scenario at hand, but rather set a baseline classification approach to which to compare the proposed detector of each sender's essential feature space. Having said this, the FS stage in Approach A will restrict to a simple feature variance thresholding, i.e. those features whose standard deviation across sample is zero will be discarded. The approach is completed by a multi-class classifier, which may be implemented by resorting to any supervised learning model.

- Approach B: now the feature selection algorithm is split in several stages FS_i (one per sender), each in charge of collecting all samples in their original size sent by user i, and selecting exclusively those features (essence) that are used recurrently along the entire set of messages sent by sender i to any receiver. Then a multiclass classifier is built

by deploying $S(S-1)/2$ OvO classifiers (with S denoting the overall number of distinct senders), each fed with the union of the essential feature set of the users being classified. The final decision results from voting the outputs of the OvO classifiers. The intuition behind this approach resides in the hypothesis that the essence of a sender consists of those less linguistic singularities that hold in every communication between him/her and any third party. Consequently, these essential features are invariant and are not affected by the influence of any receiver or the context, so they should remain present in future messages sent by the same sender. Therefore, the proposed procedure first splits the message set of each sender in disjoint sets depending on the receiver to whom they are sent, and next computes a occurrence frequency histogram of each feature over each of such subsets. The essential set for such a sender-receiver pair results from discarding those features whose frequency of occurrence falls below a given threshold. These selected features belonging to a certain sender-receiver conversation are then intersected with the rest of the filtered sender-receiver feature subsets on the basis that those commonly shared features delimit the interlocutor essence and are context-insensitive.

In mathematical terms and in reference to Figure 4.2, let the k-th message from sender i to receiver $j \in \mathcal{J}_i$ be represented by $\mathbf{m}_{i,j}^k \doteq \{m_{i,j}^{k,0}, \ldots, m_{i,j}^{k,N-1}\}$, with $N = |\mathcal{N}|$ denoting the number of originally extracted features and $\mathcal{J}_i$ the set of receivers of sender i. The purpose of the feature selection algorithm is to compute a feature subset $\mathcal{N}_i \subseteq \mathcal{N}$ such that given two different senders i and i', the union set $\mathcal{N}_{i \cup i'} \doteq \mathcal{N}_i \cup \mathcal{N}_{i'}$ can be used at the OvO classifier yielding a better predictive performance and/or lower computational complexity by virtue of its reduced size. To this end, a N-sized vector $\mathbf{f}_{i,j} \doteq \{f_{i,j}^n\}_{n=0}^{N-1}$ containing the frequency of occurrence of each feature between each sender-receiver pair (i,j) is computed as

$$f_{i,j}^n \doteq \frac{\sum_{k=1}^{K_{i,j}} \mathbb{I}(m_{i,j}^{k,n} > 0)}{K_{i,j}}, \tag{4.1}$$

where $K_{i,j}$ represents the number of messages from sender i to receiver j, and $\mathbb{I}$ is an indicator function taking value 1 if its argument is true and 0 otherwise. Once this vector has been computed, a minimum frequency threshold $\Psi_{i,j}$ determines the number of retained features for the sender-receiver pair as

$$\mathcal{N}_{i,j} = \left\{ n \in \mathcal{N} \ : \ f_{i,j}^n \geq \Psi_{i,j} \right\}, \tag{4.2}$$

from which the set of essential features for sender i is given by

$$\mathcal{N}_i = \bigcap_{j \in \mathcal{J}_i} \mathcal{N}_{i,j}. \tag{4.3}$$

From the above formulae and Figure 4.3 it should be obvious that $\Psi_{i,j}$ plays a crucial role in determining the minimum occurrence support that a feature should meet to be essential for the user at hand. This threshold should be adapted to the particular occurrence profile of the features over the different communication channels of the sender. In other words, it should capture potential inflection points within an ordered occurrence histogram beyond which the remaining feature subset becomes almost uniform hence statistically irrelevant for subsequent classification tasks.

A stand-alone, self-adjusting method to detect this point $n_{i,j}^*$ starts by sorting $\mathbf{f}_{i,j}$ by index n in decreasing order, yielding an index mapping $\hat{n} : \mathcal{N} \to \mathcal{N}$. By defining points $\mathbf{p}_0 = [0, f_{i,j}^{\hat{n}(0)}]$,

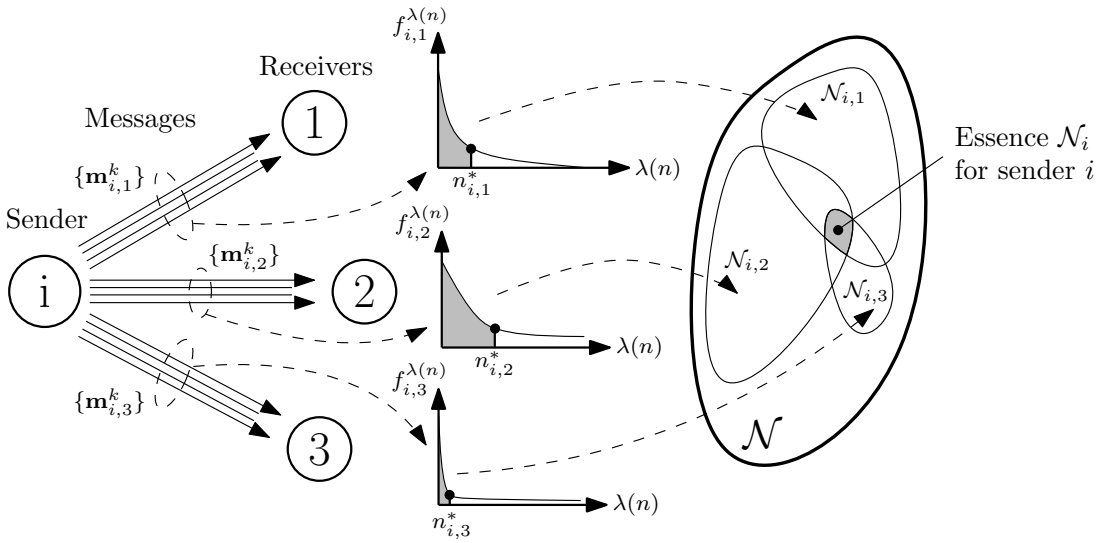

Figure 4.3: Schematic diagram showing the essence extraction procedure for sender i.

$\mathbf{p}_{N-1} = [N-1, f_{i,j}^{\lambda(N-1)}]$ and $\mathbf{p}_n = [n, f_{i,j}^{\lambda(n)}]$, the inflection point $n_{i,j}^*$ is given by

$$n_{i,j}^* = \lambda^{-1}\left(\arg\min_{n\in\mathcal{N}} \arccos \frac{(\mathbf{p}_0 - \mathbf{p}_n)\times(\mathbf{p}_{N-1} - \mathbf{p}_n)}{|\mathbf{p}_0 - \mathbf{p}_n|\cdot|\mathbf{p}_{N-1} - \mathbf{p}_n|}\right) \tag{4.4}$$

where $\times$ denotes vectorial dot product and $|\cdot|$ stands for L_2-norm. The above expression seeks the point from which the angle between the straight lines connecting points $\mathbf{p}_0$, $\mathbf{p}_{N-1}$ and $\mathbf{p}_n$ is minimum. Once this point has been computed, the set of essential features in the messages between sender i and receiver j is given by

$$\mathcal{N}_{i,j} = \left\{ n \in \mathcal{N} \ : \ f_{i,j}^n \geq f_{i,j}^{n_{i,j}^*} \right\}, \tag{4.5}$$

from which the essential feature set for sender i is inferred by resorting to Expression (4.3). This computation of $n_{i,j}^*$ allows for a higher flexibility and adaptability of the threshold $\Psi_{i,j}$; in fact, it should be clear that $\Psi_{i,j} = f_{i,j}^{n_{i,j}^*}$.

4.1 Results and Discussion

Several experiments have been performed to test the average accuracy of our proposed scheme when applied to the selected dataset. To this end, two different machine learning models will be utilized for implementing the multi-class (approach A) and OvO (approach B) classifiers:

- A Random Forest (RF) classifier, which is built by an ensemble of simple decision trees each trained with a bootstrap sample drawn from the overall training set. In addition, the split in such compounding trees is not decided among all features, but instead among a random feature subset. This randomness involves a slightly increased bias with respect to a non-randomized single decision tree. However the variance decreases more significantly to usually compensate for the increase in bias, ultimately yielding a model with enhanced

predictive generalization capabilities. Random Forests are a response to those classifiers which tend to generalize without ruling out outliers or noisy patterns eventually creating models with high variance and then sensible to minor fluctuations in the training set. This model has been selected to check our prior – yet later described – intuitions behind the comparative performance between the selected models and approaches.

- A Support Vector Machine (SVM), which constructs a hyperplane – or set of hyperplanes – in a feature space of increased (if required) dimensionality so as to map different category instances into maximally separated decision regions. By maximizing the margin between the closest points belonging to different categories the generalization error is minimized. SVM permits working on features that when combined or rendered onto a larger space, become significant and decisive in terms of predictive significance. In these models the trade-off between generalization and overfitting is mainly controlled by the penalty parameter C: the larger C is, the less the final training error will be, but a higher risk is assumed to jeopardize the generalization properties of the classifier. In addition, a large C usually entails a higher training time of the classifier.

The experiments discussed in what follows aims at validating the predictive performance of these two models when utilized as baseline classifiers in approaches A and B. Based on this intended scope and for the sake of fairness, the classification model in approach A will implement a OvO ensemble when adopting SVM as the baseline classifier. This will allow comparing both approaches under the same model configuration, hence minimizing any eventual influence due to differing ensembles. In regards to approach B and denoting as $D^k_{i,j}$ the authorship decision made for message k in the OvO classifier deciding between senders $i \in \{1, \ldots, S\}$ and $j \in \{i+1, S\}$, two voting schemes will be further considered:

- Hard voting: the final decision about the authorship of each message results from the statistical mode computed over $\{\{D_{i,j}\}_{i=1}^{S}\}_{j=i+1}^{S}$, i.e. the most frequent value over the outputs of all the OvO classifiers.

- Soft voting: the final decision is furnished by fusing the likelihoods produced for each sender by every OvO classifier. In the case of RF the predicted probabilities of an input sample for a certain class (i.e. sender) is estimated by averaging the predicted class probabilities of the trees in the forest, where the class probability of a single tree is given by the fraction of samples of the same class in a leaf. As for the SVM classifier, Platt scaling [129] is utilized for transforming the hard output of the model into a distribution of probabilities over classes. If $p(i|m_k)$ denotes the overall likelihood about the authorship of sender i estimated for message k, by assuming conditional independence between the OvO classifiers and uniformity among the senders when authoring the message it can be proven that

$$p(i|m_k) \propto \prod_{s=1}^{S(S-1)/2} p_s(i|m_k), \tag{4.6}$$

  i.e. as the product of the output probabilities of those OvO classifiers where the authorship of sender i is compared to every other sender. The soft voter will opt for the sender i^* with the highest total likelihood among all possible senders, i.e.

$$i^* = \arg\max_{i \in \{1,\ldots,S\}} p(i|m_k). \tag{4.7}$$

We begin the discussion by analyzing Table 4.1, where results in terms of classification accuracy are shown for different models (RF and SVM), classification approaches (A with variance thresholding and B with sender essence discrimination with hard or soft voting as depicted in Figure 4.2). The performance of each scheme under comparison has been averaged over 10 stratified folds to assess the statistical stability of the score. Furthermore, parameters controlling the utilized models (e.g. C and γ for the SVM) have been all optimized via grid search and a local 5-fold cross-validation.

Table 4.1: Precision score for different supervised learning techniques and authorship classification approaches. Scores are given as *mean/standard deviation* computed over 10 stratified folds.

Approach	SVM	RF
Approach A (no feature selection)	0.718 / 0.001	0.607 / 0.014
Approach B (hard, $\Psi_{i,j} = 0.8$)	0.684 / 0.003	0.624 / 0.004
Approach B (hard, $\Psi_{i,j} = 0.6$)	0.649 / 0.001	0.599 / 0.002
Approach B (hard, $\Psi_{i,j} = 0.3$)	0.587 / 0.006	0.557 / 0.003
Approach B (hard, self-adjusted $\Psi_{i,j}$)	0.665 / 0.004	0.610 / 0.002
Approach B (soft, $\Psi_{i,j} = 0.8$)	0.710 / 0.003	0.629 / 0.002
Approach B (soft, self-adjusted $\Psi_{i,j}$)	0.703 / 0.004	0.618 / 0.002

As anticipated by the above table, SVM without any feature selection scheme slightly outperforms our proposed scheme. However, the obtained results deserve further analysis: the optimized SVM parameters for Approach A (namely, $C = 10^4$ and $\gamma = 10^{-4}$) give rise to a strongly overfitted model which excels at classifying its input messages due to the unique usage of more than 6000 features in a dataset comprising 13036 messages. This explanation is supported by the results obtained with the RF classifier with pairwise essential features: it outperforms its corresponding Approach A, which operates over the entire set of 20779 features which is reduced to 19095 with non-zero variance over users (i.e. with potential ability to discriminate among users). As earlier discussed, RF splits the training set into smaller groups aiming at avoiding high generalizing error scores due to a complex underlying model which has memorized the training data rather than learned the underlying patterns. The tailored selection of relevant features based on our proposed linguistic essence makes the avoidance of overfitting easier to achieve.

Table 4.2: Absolute and relative number of features used for each approach (A & B) and threshold selection method.

Author	Approach A	Approach B			
		$\Psi_{i,j} = 0.8$	$\Psi_{i,j} = 0.6$	$\Psi_{i,j} = 0.3$	Self-adjusted $\Psi_{i,j}$
0	19095 (100%)	418 (2.19%)	156 (0.81%)	30 (0.15%)	215 (1.12%)
1		223 (1.17%)	92 (0.48%)	20 (0.10%)	157 (0.82%)
2		318 (1.66%)	147 (0.77%)	30 (0.16%)	184 (0.96%)
3		749 (3.92%)	305 (1.59%)	63 (0.33%)	288 (1.51%)
4		241 (1.26%)	107 (0.56%)	29 (0.15%)	135 (0.71%)
5		275 (1.44%)	129 (0.67%)	34 (0.18%)	161 (0.84%)

The performance of the proposed approach can be further analyzed from the perspective of the computational complexity of the model training, which relates directly to the number of

input features. Table 4.2 shows the absolute number of essential features obtained for each threshold selection method in Approach B, and the percentage they represent with respect to the overall number of predictors utilized by Approach A after variance thresholding. It can be noticed that while attaining comparable precision scores to Approach A, Approach B vastly reduces the number of utilized features (in particular, at most 1.51% for self-adjusted thresholding). This observation buttresses the intuition that the proposed essential feature detector not only reduces the computational complexity of the training process for SVM models, but also helps the inherent feature selection method of Random Forests in the discrimination of good predictors, to the point of achieving better prediction results. This conclusion gets further reinforced by assessing the computation time taken by each of the above schemes: when implemented in Python on a Pentium Core i7 Pro with 16 Gigabytes of RAM, the execution of each fold in Approach A takes on average 310.4 times longer than the longest variant of Approach B (self-adjusted $\Psi_{i,j}$).

Further interesting observations can be drawn if the average precision scores shown in Table 4.1 are broken down into the individual metrics attained by each of the compounding OvO classifier. For the sake of simplicity we will focus on Approach B with RF as core learning model and self-adjusted essence feature selection. Table 4.3 depicts the confusion matrix of the overall approach. Therein it can be noticed that while the precision for users 0 to 3 are very satisfactory bearing in mind the short length and limited content of the processed messages (with user 3 amounting up to a precision of 87%), confusion appears between users 4 and 5. This bad classification result is supported in part by Figure 4.4, which depicts the accumulated number of features (discriminated by type of feature) when the number of users increases. This plot evinces that when considering the last user jointly with the rest of possible authors an upper bound in the number of total features is achieved. In other words, this unveils a *linguistic* limit of the SMS messages contained in the dataset under consideration: when dealing with short-length SMS messages it is very likely that a high fraction of them share the same n-gram set. This implies that for certain users it becomes necessary to resort to information of other nature so as to uniquely identify their authorship, such as the connection usage approach described in Chapter 3.

Table 4.3: Normalized confusion matrix corresponding to Approach B, soft voting, self-adjusted $\Psi_{i,j}$.

		Predicted label					
		0	1	2	3	4	5
True label	0	0.68	0.08	0.02	0.06	0.15	0.01
	1	0.02	0.76	0.04	0.09	0.08	0.01
	2	0.04	0.11	0.65	0.08	0.11	0.01
	3	0.00	0.04	0.03	0.88	0.05	0.00
	4	0.03	0.04	0.06	0.05	0.33	0.48
	5	0.01	0.02	0.05	0.02	0.87	0.03

Our approach has been tested considering every message as an independent message providing heretofore certain confidence on our assumptions about the underlying essential behavioral patterns due to the obtained accuracy. However, this stringent policy is itself a detriment to our hypothesis, since most of such dynamic dialogues are short to be mined towards extracting fruitful properties [116] as preceding work has demonstrated in the past working with block sizes ranging from approximately 2000 words [114, 115] to a minimum of 200 words [108].

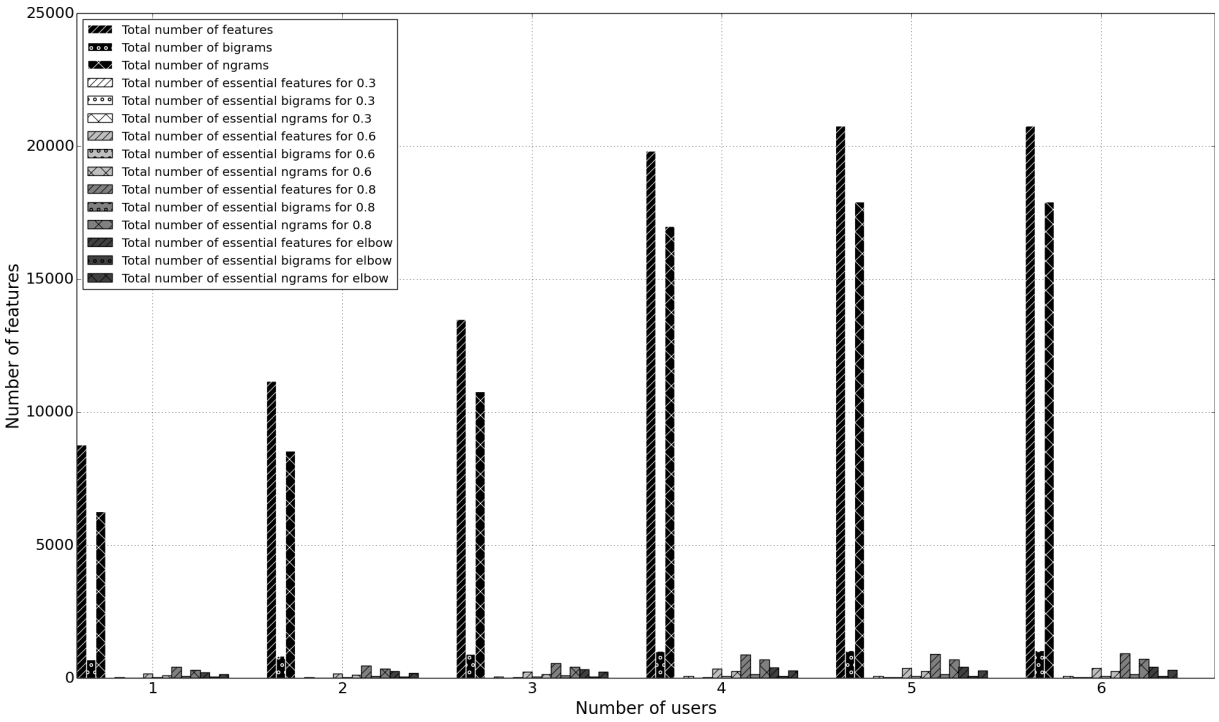

Figure 4.4: Progression of the accumulated number of features per type (3-gram, PoS bigram, essential) for different essence selection threshold schemes and number of considered users. It can be seen that the overall number of features does not increase when considering the last user (user 5), fact that unveils that a single message may not be sufficient for uniquely discriminating among different possible authors, especially when dealing with datasets containing messages of reduced length.

A practical yet even more realistic workaround to the above eventuality can be straightforwardly implemented by voting the results of several consecutive model outcomes. In this manner the impersonation detection system would not focus on accurately classifying a single, hopefully discriminable message, but rather a series of texts in a much more similar fashion to chat sessions and other interactive message exchanging tools alike. In order to illustrate the performance of this scheme we have included results from this proposed voting approach as a function of the number of consecutive classification outcomes voted by majority.

Table 4.4: Precision score (mean /std) for majority voting of successive message and its comparison to the figures of merit of Approaches A and B.

Approach	SVM	RF
Approach A (no feature selection)	0.718 / 0.001	0.607 / 0.014
Approach B (hard, self-adjusted $\Psi_{i,j}$)	0.665 / 0.004	0.610 / 0.002
Approach B (soft, self-adjusted $\Psi_{i,j}$)	0.703 / 0.004	0.618 / 0.002
Approach B (soft, self-adjusted $\Psi_{i,j}$, 3-message voting)	0.753 / 0.003	0.667 / 0.005
Approach B (soft, self-adjusted $\Psi_{i,j}$, 5-message voting)	0.801 / 0.001	0.715 / 0.003
Approach B (soft, self-adjusted $\Psi_{i,j}$, 7-message voting)	0.832 / 0.006	0.747 / 0.001
Approach B (soft, self-adjusted $\Psi_{i,j}$, 9-message voting)	0.845 / 0.005	0.788 / 0.002

As shown in the above table, voting over the decisions taken for successive messages enhances significantly the overall precision score. Moreover, this strategy paves the way towards concatenating different messages before the essence extraction algorithm so as to better discriminate linguistically the set of possible authors, as hypothesized throughout the future research lines derived from this Thesis and described in the following chapter.

4.2 Conclusions

Authorship attribution is conceived by the research community as the problem of identifying the origin of a text among different authors by solely analyzing its content. This paradigm and the interesting technical approaches to efficiently solve it have embodied a very active research area so far, with a sharp multidisciplinary flavor due to the convergence of techniques and methods from Computational Intelligence, Machine Learning and Natural Language Processing. This paradigm has been mostly addressed from a literacy perspective, aiming at identifying the stylometric features and writeprints which unequivocally typify the writer patterns and allow their unique identification.

On the other hand, the upsurge of social networking platforms and interactive messaging have undoubtedly made the anonymous expression of feelings, the sharing of experiences and social relationships much more easier than in other traditional communication media. Unfortunately, the popularity of such communities and the virtual identification of their users sets a rich substrate for cybercrimes against unsuspecting victims. In this context this chapter has hypothesized and analyzed the identification of the sender of a message as an useful approach to detect impersonation attacks in interactive communication scenarios. In particular conventional yet innovative characteristics of messages have been extracted via NLP techniques and selected by means of a newly devised feature selection algorithm based on the dissociation between essential traits of the sender and receiver influences. The proposed selection method has been shown to be promising with real SMS data in terms of identification accuracy, performance further enhanced by means of a more elaborated voting scheme using 1) soft estimates of the one-versus-one classifiers underlying beneath the overall authorship detection scheme; and 2) voting along estimates of different messages corresponding to a single communication session, as could be applied to e.g. chat sessions and message series.

These results pave the way towards future research lines focused on applying the concept of language typicality in the discourse analysis field, summarized in the closing chapter of this Thesis.

CHAPTER 5

CONCLUDING REMARKS AND FUTURE RESEARCH LINES

"A good scientist is a person in whom the childhood quality of perennial curiosity lingers on. Once he gets an answer, he has other questions."

- Frederick Seitz

This Thesis has focused on discovering new schemes to tackle the identity theft problem in Social Networks, gathering and learning from past research experiences. We follow the assumption that simplicity and generic solutions, as opposed to ad-hoc schemes, could provide a better insight and results in the long term, due to the changing, dynamic behavior of criminals in their attacks, who are indeed constantly looking for new security breaches. On this basis, we have proposed to resort to profiling so as to properly confine the user behavior into precise and demarcated personality traits in an attempt to find clear inconsistencies if an attack is perpetrated. From the outcomes of the research activity developed within this Thesis, several relevant conclusions can be drawn as summarized in what follows:

- Chapter 3 has elaborated on a novel approach for detecting impersonation attacks in social networks based exclusively on connection time statistics. To overcome an eventual lack of content-related traces (e.g. gossiping or spying) 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. Specifically, this work has focused on creating a behavioral user profile in terms of connection time information, which is deemed an essential feature in social networks as it represents the degree of dependency, availability and regularity of the user at hand. To this end, two classification schemes have been formulated: a supervised classification model with an innovative method to address the lack of labeled instances for the category representing the suspicious activity (synthetically generated feature set representing any behavior not observed in the user connection record and emulating a potential impersonation attack); and an unsupervised classification model empowered by means of a bio-inspired meta-heuristic algorithm (Harmony Search) which provides the required parameters for an optimized K-Means clustering stage in terms of attack detection performance.

- Chapter 4 has analyzed the context as one of the most crucial variables in the discourse surrounding any verbal or spoken communication and helping to determine its interpretation. This chapter has gravitated on the identification of the sender of a message as an

useful approach to detect impersonation attacks in interactive communication scenarios. Specifically, conventional yet innovative characteristics of messages have been extracted via NLP techniques and selected by means of a newly devised feature selection algorithm based on the dissociation between essential traits of the sender and receiver influences. The proposed selection method has rendered promising results with real SMS data in terms of identification accuracy, and paves the way towards future research lines focused on applying the concept of language typicality in the discourse analysis field.

All in all, the main technical conclusion of the Thesis lies on the practical assessment of hybrid machine learning models and novel feature selection schemes as detectors of subtle impersonation attacks in social networks. However, the impact of this Thesis goes beyond the technical scope around which it has gravitated: the findings in Chapter 3, which rely solely on the processing of connection time logs, can be easily extrapolated to other scenarios such as e.g. money withdrawal at cash points; as for Chapter 4, the essential feature set not only allows discriminating among different authors of a certain message, but also – as introduced above – permits to quantify the influence of the context in the communication. By clustering such a contextual information different profiles of the message sender could be discovered which, when collected over different senders and supervised with external label information, could unveil linguistic patterns characterizing different behavioral roles. This spans a myriad of open research lines, some of which will be outlined within the last subsection of this chapter.

5.1 Research Outcomes

The research results summarized in this dissertation have been presented at several international conferences and submitted to renowned scientific journals with measurable impact factor by the Journal Citation Reports. Along with previous results published by the author of this Thesis in the field of Natural Language Processing, in total 4 manuscripts have been published in international journals and 6 contributions have been presented in specialized workshops and technical conferences:

- Esther Villar-Rodríguez, Javier Del Ser, Sergio Gil-Lopez, Miren Nekane Bilbao and Sancho Salcedo-Sanz, "A Meta-heuristic Learning Approach for the Non-Intrusive Detection of Impersonation Attacks in Social Networks", International Journal of Bio-inspired Computation, accepted, May 2015. JCR: 3.969 (Q1).
- Esther Villar-Rodríguez, Javier Del Ser, Miren Nekane Bilbao and Sancho Salcedo-Sanz, "A Novel Machine Learning Approach to the Detection of Identity Theft in Social Networks based on Emulated Attack Instances and Support Vector Machines", Concurrency & Computation, accepted, July 2015. JCR: 0.784 (Q3).
- Ana I. Torre-Bastida, Esther Villar-Rodríguez, Sergio Gil-Lopez and Javier Del Ser, "Design and Implementation of an Extended Corporate CRM Database System with Big Data Analytical Functionalities", Journal of Universal Computer Science, Vol. 21, N. 6, pp. 757–776, 2015. JCR: 0.401 (Q3).
- Ana I. Torre-Bastida, Marta González-Rodríguez and Esther Villar-Rodríguez, "Linked Open Data (LOD) and its Implementation in Libraries: Initiatives and Technologies", El profesional de la información, Vol. 24, N. 2, pp. 113–120, 2015. JCR: 0.356 (Q4).

- Esther Villar-Rodríguez, Javier Del Ser and Sancho Salcedo-Sanz, "On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks", International Symposium on Intelligent Distributed Computing, published in Intelligent Distributed Computing VIII, Springer Studies in Computational Intelligence, Vol. 570, pp. 259–268, ISBN 978-3-319-10421-8, 2015.

- Ana I. Torre-Bastida, Esther Villar-Rodríguez, Javier Del Ser and Sergio Gil-Lopez, "Semantic Information Fusion of Linked Open Data and Social Big Data for the Creation of an Extended Corporate CRM Database", International Symposium on Intelligent Distributed Computing, published in Intelligent Distributed Computing VIII, Springer Studies in Computational Intelligence, Vol. 570, pp. 211–221, ISBN 978-3-319-10421-8, 2015.

- Ana I. Torre-Bastida, Esther Villar-Rodríguez, Javier Del Ser, David Camacho and Marta Gonzalez-Rodriguez, "On Interlinking Linked Data Sources by using Ontology Matching Techniques and the Map-Reduce Framework", International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), published in Springer Lecture Notes in Computer Science, Vol. 8669, pp. 53–60, ISBN 978-3-319-10840-7, 2014.

- Esther Villar-Rodrǵuez, Ana García Serrano and Marta González-Rodríguez, "Uso de un Enfoque Lingüístico para el Análisis de Sentimientos", TASS Workshop, XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). pp. 200–205, ISBN: 978-84-695-8349-4, 2013.

- Esther Villar-Rodríguez, Ana García-Serrano and Marta González-Rodríguez, "Análisis Lingüístico de Expresiones Negativas en Tweets en Español", poster contribution published in the proceedings of the XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), pp. 4–11, ISBN: 978-84-695-8349-4, 2013.

- Esther Villar-Rodríguez, Ana I. Torre Bastida, Ana García-Serrano and Marta González-Rodríguez, "Using Linked Open Data Sources for Entity Disambiguation", Participation in the competition *Online Reputation Management - RepLab*, contribution included in the Working Notes for the CLEF (Conference and Labs for the Evaluation Forum) conference, 8 pags. ISBN: 978-88-904810-5-5, 2013.

5.2 Future Directions

At the same time that this work has been evolving, future research lines have been outlined as a response of several factors: the technical discussions carried out, the exciting practical possibilities which the conducted work in both stages had triggered and the continuous movements in SN security measures adopted by the administrators as a consequence of the increasing user demand and concern. Such research lines can be summarized as follows:

- As for Chapter 3, specially in reference to the supervised approach, this work has introduced a novel procedure to depict the missing category aiming at leveraging the good performance of the SVM algorithm for multi-class classifications. In identity theft systems, an optimal precision is a hard constraint and strictly depends on the capacity of defining proper boundaries in intricate spaces particularly by virtue of high dimensions. By assuming spherical shapes from traditional one-class classifiers we could be falling into the trap

of overlooking potential and exhaustively emulated attacks. However, these techniques will be thoroughly investigated and discussed in the future. A close look will be taken at the latest advances in the self-configuration of the Harmony Search meta-heuristic solver; instead of predefining a set of fixed values for this algorithm in the context of the hybrid clustering approach proposed in the chapter, parameter-setting-free schemes adjust in execution time the probabilistic parameters driving the meta-heuristic search procedure depending on the fitness evolution of the individuals compounding the harmony memory. Other more autonomous evolutionary schemes will be also under investigation, e.g. the Estimation of Distribution Algorithm [130, 131] and Coral Reefs Optimization [132].

- The feature essence isolation approach proposed in Chapter 4 also provides information about the circumstances or conditions in which the discussion is framed. Thus, being able to discern user essential and commonly used from extrinsic or less primary features out of a conversation can shed light on the scenario description. Let us imagine an individual maintaining several dyadic communications. After discarding all the inherent and unconscious behavioral patterns, one can discover the different roles the user is deliberately developing and therefore pave the way for secondary tasks such as context clustering or audience categorization. These unsupervised methods may then assist in profiling activities by analyzing the shared discourse inside or within the boundaries of the clusters. Unfortunately, we have faced a hub-shaped set of unidirectional messages containing no receiver in the role of sender or the other way around. This is, no receiver has hold any conversation with any other user inside the whole dataset. This disjointed collection of senders and receivers has inevitably forced us to deviate in the analysis and application of the same approach in the receiver´s side which had enabled to discern more actors in the communicative act by desegregating the receiver´s essence from the heretofore called context. After discarding all the inherent and unconscious behavioral patterns one can discover the different roles the user is deliberately developing, with very avant-garde application such as the detection of pedophiles [133].

- In regards to the overall scope of the Thesis, a third stage would be considered so as to complete the whole original notion of a safe platform. Lately, SNs have become the major phishing portals due to the attractiveness of the implemented social graph. The main purpose is to spread phishing links or even malware trusting in the effectiveness of using real hijacked accounts rather than ad-hoc bot ones postulating that more clicks on a fake bank site will be carried out if the link comes from his/her social network friend instead of from an unknown person. In terms of interactivity, such illegal campaign would be translated into bulk messages sent en masse over the whole social graph and could be seemingly detected by analyzing such recurrent pattern in contrast to the characterized behavior related to how and with whom the legitimate user interacts (via e.g. dynamic link grouping or centrality metrics). In this chapter, second-stage features are easily displayed over a graph where vertices, also called nodes, represent the essence of a user, and edges or arcs stand for contextual peculiarities. In future research lines, we will delve into the possibility of adding interactivity features to capture the user's habitual community and its normal evolution over the time so as to recognize dramatic changes or strange enhancements in his/her social skills. Likewise, repeated messages could be efficiently withheld if the platform identified copies. Another matter of interest would be to reinforce our contribution to the analysis of the discourse in the field of linguistics.

Bibliography

[1] Webdesigner Depot,http://www.webdesignerdepot.com/2009/10/the-history-and-evolution-of-social-media/ published on October 7, 2009.

[2] Digital Trends, http://www.digitaltrends.com/features/the-history-of-social-networking/, published on August 5, 2014.

[3] The Guardian, http://www.theguardian.com/world/2009/aug/07/georgian-blogger-accuses-russia, published on August 7, 2009.

[4] Daily Mail, http://www.dailymail.co.uk/news/article-2280625/Burger-King-Twitter-feed-hacked-McDonalds-fan.html, published on February 18, 2013.

[5] The Washington Post, https://www.washingtonpost.com/news/checkpoint/wp/2015/01/12/centcom-twitter-account-apparently-hacked-by-islamic-state-sympathizers/, published on January 12, 2015.

[6] Wired, http://www.wired.com/2012/05/flame/, published on May 28, 2012.

[7] The Washington Post, http://www.washingtonpost.com/wp-dyn/content/article/2009/07/08/AR2009070800066.html, published on July 9, 2009.

[8] The Guardian, http://www.theguardian.com/global/2008/nov/07/obama-white-house-usa, published on November 7, 2008.

[9] Wired, http://www.wired.com/2009/08/tjx-hacker-charged-with-heartland/, published on July 17, 2009.

[10] Dayton Daily News,http://www.daytondailynews.com/news/news/local/student-campaign-raises-awareness-of-internet-dang/nkqBH/, published on April 8, 2015.

[11] Cybertip.ca, https://www.cybertip.ca/app/en/media_release_ccaice_action_plan_highlights, retrieved on August 8, 2015.

[12] Social Times, http://www.adweek.com/socialtimes/facebook-photos-verify/243870, published on July 26, 2010.

[13] J. Hong, "The State of Phishing Attacks", Communications of the ACM, Vol. 55, N. 1, pp. 74–81, 2012.

[14] The Futon Critic, http://www.thefutoncritic.com/news/2015/06/23/usa-network-nation-under-a-hack-survey-results-30313/20150623usa01/, published on June 23, 2015.

[15] Hubspot, http://blog.hubspot.com/marketing/social-media-roi-stats, published on June 6, 2014.

[16] D. Pelleg, A. W. Moore, "X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters", International Conference on Machine Learning, pp. 727-734, 2000.

[17] H. Bischof, A. Leonardis, A. Selb, "MDL Principle for Robust Vector Quantisation", Pattern Analysis & Applications, Vol. 2, N. 1, pp. 59–72, 1999.

[18] G. Hamerly, C. Elkan, "Learning the K in K-Means", Advances in Neural Information Processing Systems, Vol. 16, pp. 281, 2004.

[19] T. Calinski, J. Harabasz, A Dendrite Method for Cluster Analysis", Communications in Statistics-Theory and Methods, Vol. 3, N. 1, pp. 1-27, 1974.

[20] D. L. Davies, D. W. Bouldin, "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-1, N. 2, pp. 224–227, 1979.

[21] J. C. Dunn, "Well-Separated Clusters and Optimal Fuzzy Partitions", Journal of cybernetics, Vol. 4, N. 1, pp. 95–104, 1974.

[22] Q. Zhao, V. Hautamaki, P. Fränti, "Knee-point Detection in BIC for Detecting the Number of Clusters", Advanced Concepts for Intelligent Vision Systems, pp. 664–673, 2008.

[23] G. W. Milligan, M. C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set", Psychometrika, Vol. 50, N. 2, pp. 159–179, 1985.

[24] E. Dimitriadou, S. Dolnicar, A. Weingessel, "An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets", Psychometrika, Vol. 67, N. 1, pp. 137–159, 2002.

[25] P. J. Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", Journal of Computational and Applied Mathematics, Vol. 20, pp. 53–65, 1987.

[26] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, N. 14, pp. 281–297, 1967.

[27] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, N. 8, pp. 832–844, 1998.

[28] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, "Classification and Regression Trees", Wadsworth. Belmont, CA, 1984.

[29] L. Breiman, "Random Forests", Machine Learning, Vol. 45, N. 1, pp. 5–32, 2001.

[30] F. Rosenblatt, “The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain”, Psychological Review, Vol. 65, N. 6, pp. 386, 1958.

[31] R. E. Kass, A. E. Raftery, “Bayes Factors”, Journal of the American Statistical Association, Vol. 90, N. 430, pp. 773–795, 1995.

[32] H. Akaike, “Information Theory and an Extension of the Maximum Likelihood Principle”, Selected Papers of Hirotugu Akaike, pp. 199–213, 1998.

[33] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. Van Der Linde, “Bayesian Measures of Model Complexity and Fit”, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 64, N. 4, pp. 583–639, 2002.

[34] J. R. Quinlan, “Induction of decision trees”, Machine Learning, Vol. 1, N. 1, pp. 81–106, 1986.

[35] T. Mitchell, “Machine Learning”, McGraw-Hill, 1997.

[36] A. Aizerman, E. M. Braverman, L. I. Rozoner, “Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning”, Automation and Remote Control, Vol. 25, pp. 821-837, 1964.

[37] F. Glover, “Future Paths for Integer Programming and Links to Artificial Intelligence”, Computers & Operations Research, Vol. 13, N. 5, p. 533-549, 1986.

[38] C. Blum, J. Puchinger, G. R. Raidl, A. Roli, “Hybrid Metaheuristics in Combinatorial Optimization: A Survey”, Applied Soft Computing, Vol. 11, N. 6, pp. 4135–4151, 2011.

[39] J. A. Parejo , A. Ruiz-Cortes, S. Lozano, P. Fernandez, “Metaheuristic Optimization Frameworks: a Survey and Benchmarking”, Soft Computing, Vol. 16, N. 3, pp. 527–561, 2012.

[40] I. Boussaïd, J. Lepagnot, P. Siarry, “A Survey on Optimization Metaheuristics”, Information Sciences, Vol. 237, pp. 82–117, 2013.

[41] X. S. Yang, “Recent Advances in Swarm Intelligence and Evolutionary Computation”, Studies in Computational Intelligence, Vol. 585, 2015.

[42] Z. W. Geem, J. H. Kim, G. V. Loganathan, “A New Heuristic Optimization Algorithm: Harmony Search”, Simulation, Vol. 76, N. 2, pp. 60-68, 2001.

[43] K. S. Lee, Z. W. Geem, “A New Structural Optimization Method based on the Harmony Search Algorithm”, Computers & Structures, Vol. 82, N. 9-10, pp. 781-798, 2004.

[44] M. P. Saka, “Optimum Geometry Design of Geodesic Domes Using Harmony Search Algorithm”, Advances in Structural Engineering, Vol. 10, N. 6, pp. 595-606, 2007.

[45] Z. W. Geem, “Harmony Search Algorithms for Structural Design Optimization”, Studies in Computational Intelligence, Vol. 239, 2009.

[46] J. Fourie, S. Mills, R. Green, “Harmony Filter: a Robust Visual Tracking System using the Improved Harmony Search Algorithm”, Image and Vision Computing, Vol. 28, N. 12, pp. 1702-1716, 2010.

[47] J. Li, H. Duan, “Novel Biological Visual Attention Mechanism via Gaussian Harmony Search”, Optik, Vol. 125, N. 10, pp. 2313-2319, 2014.

[48] V. R. Pandi, B. K. Panigrahi, S. Das, Z. Cui, "Dynamic Economic Load Dispatch with Wind Energy using Modified Harmony Search", International Journal of Bio-Inspired Computation, Vol. 2, N. 3-4, pp. 282–289, 2010.

[49] R. Arul, G. Ravi, S. Velusami, "Chaotic Self-Adaptive Differential Harmony Search Algorithm based Dynamic Economic Dispatch", International Journal of Electrical Power & Energy Systems, vol. 50, no. 1, pp. 85-96, 2013.

[50] B. Jeddi, V. Vahidinasab, "A Modified Harmony Search Method for Environmental/Economic Load Dispatch of Real-World Power Systems", Energy Conversion and Management, Vol. 78, pp. 661-675, 2014.

[51] R. Zhang, L. Hanzo, "Iterative Multiuser Detection and Channel Decoding for DS-CDMA using Harmony Search", IEEE Signal Processing Letters, vol. 16, N. 10, pp. 917–920, 2009.

[52] J. Del Ser, M. Matinmikko, S. Gil-Lopez, M. Mustonen, "Centralized and Distributed Spectrum Channel Assignment in Cognitive Wireless Networks: A Harmony Search Approach", Applied Soft Computing, Vol. 12, N. 2, pp. 921-930, 2012.

[53] I. Landa-Torres, S. Gil-Lopez, J. Del Ser, S. Salcedo-Sanz, D. Manjarres, J. A. Portilla-Figueras, "Efficient Citywide Planning of Open WiFi Access Networks using Novel Grouping Harmony Search Heuristics", Engineering Applications of Artificial Intelligence, Vol. 26, N. 3, pp. 1124-1130, 2013.

[54] D. Manjarres, J. Del Ser, S. Gil-Lopez, M. Vecchio, I. Landa-Torres, S. Salcedo-Sanz, R. Lopez-Valcarce, "On the Design of a Novel Two-objective Harmony Search Approach for Distance- and Connectivity-based Localization in Wireless Sensor Networks", Engineering Applications of Artificial Intelligence, Vol. 26, N. 2, pp. 669-676, 2013.

[55] M. H. Scalabrin, R. S. Parpinelli, C. M. Benítez, H. S. Lopes, "Population-based Harmony Search using GPU applied to Protein Structure Prediction", International Journal of Computational Science and Engineering, Vol. 9, N. 1-2, pp. 106–118, 2014.

[56] I. Landa-Torres, E. G. Ortiz-Garcia, S. Salcedo-Sanz, M. J. Segovia-Vargas, S. Gil-Lopez, M. Miranda, J. M. Leiva-Murillo, J. Del Ser, "Evaluating the Internationalization Success of Companies Through a Hybrid Grouping Harmony Search—Extreme Learning Machine Approach", IEEE Journal of Selected Topics in Signal Processing, Vol. 6, N. 4, pp. 388-398, 2012.

[57] H. Xu, Z. Zhang, K. Alipour, K. Xue, X. Z. Gao, "Prototypes Selection by Multi-objective Optimal Design: Application to a Reconfigurable Robot in Sandy Terrain", Industrial Robot, Vol. 38, N. 6, pp. 599-613, 2011.

[58] H. Xu, X. Z. Gao, G.-L. Peng, K. Xue, Y. Ma, "Prototype Optimization of Reconfigurable Mobile Robots based on a Modified Harmony Search Method", Transactions of the Institute of Measurement and Control, Vol. 34, N. 2-3, pp. 334-360, 2012.

[59] G. Ingram, T. Zhang, "Overview of Applications and Developments in the Harmony Search Algorithm", Music-Inspired Harmony Search Algorithm, Studies in Computational Intelligence, Vol. 191, pp. 15-37, 2009.

[60] Z. W. Geem, "Music-Inspired Harmony Search Algorithm: Theory and Applications", Studies in Computational Intelligence, 2009.

[61] D. Manjarres, I. Landa-Torres, S. Gil-Lopez, J. Del Ser, M. N. Bilbao, S. Salcedo-Sanz, Z. W. Geem, "A Survey on Applications of the Harmony Search Algorithm", Engineering Applications of Artificial Intelligence, Vol. 26, N. 8, pp. 1818–1831, 2013.

[62] X. Z. Gao, V. Govindasamy, H. Xu, X. Wang, K. Zenger, "Harmony Search Method: Theory and Applications", Computational Intelligence and Neuroscience, Article ID 258491, 2015.

[63] S. Das, A. Mukhopadhyay, A. Roy, A. Abraham, B. K. Panigrahi, "Exploratory Power of the Harmony Search Algorithm: Analysis and Improvements for Global Numerical Optimization", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 41, N. 1, pp. 89-106, 2011.

[64] M. Mahdavi, M. Fesanghary, E. Damangir, "An Improved Harmony Search Algorithm for Solving Optimization Problems", Applied Mathematics and Computation, Vol. 188, N. 2, pp. 1567-1579, 2007.

[65] M. G. H. Omran, M. Mahdavi, "Global-Best Harmony Search", Applied Mathematics and Computation, Vol. 198, N. 2, pp. 643-656, 2008.

[66] M. El-Abd, "An Improved Global-Best Harmony Search Algorithm", Applied Mathematics and Computation, Vol. 222, pp. 94-106, 2013.

[67] Q.-K. Pan, P. N. Suganthan, J. J. Liang, M. F. Tasgetiren, "A Local-Best Harmony Search Algorithm with Dynamic Subpopulations", Engineering Optimization, Vol. 42, N. 2, pp. 101-117, 2010.

[68] Z. W. Geem, K.-B. Sim, "Parameter-setting-free Harmony Search Algorithm", Applied Mathematics and Computation, Vol. 217, N. 8, pp. 3881-3889, 2010.

[69] Y. M. Cheng, L. Li, T. Lansivaara, S. C. Chi, Y. J. Sun, "An Improved Harmony Search Minimization Algorithm using Different Slip Surface Generation Methods for Slope Stability Analysis", Engineering Optimization, Vol. 40, N. 2, pp. 95-115, 2008.

[70] O. Hasanebi, F. Erdal, M. P. Saka, "Adaptive Harmony Search Method for Structural Optimization", ASCE Journal of Structural Engineering, Vol. 136, N. 4, pp. 419-431, 2010.

[71] C.-M. Wang, Y.-F. Huang, "Self-Adaptive Harmony Search Algorithm for Optimization", Expert Systems with Applications, Vol. 37, N. 4, pp. 2826-2837, 2010.

[72] R. Enayatifar, M. Yousefi, A. H. Abdullah, A. N. Darus, "LAHS: A Novel Harmony Search Algorithm based on Learning Automata", Communications in Nonlinear Science and Numerical Simulation, Vol. 18, N. 12, pp. 3481-3497,2013.

[73] M. Castelli, S. Silva, L. Manzoni, L. Vanneschi, "Geometric Selective Harmony Search", Information Sciences, Vol. 279, pp. 468-482, 2014.

[74] Z. W. Geem, "Particle-Swarm Harmony Search for Water Network Design", Engineering Optimization, Vol. 41, N. 4, pp. 297-311, 2009.

[75] W.-W. Shi , W. Han, W.-C. Si, "A Hybrid Genetic Algorithm Based on Harmony Search and its Improving", Lecture Notes in Electrical Engineering, Vol. 204, pp. 101-109, 2012.

[76] M. A. Al-Betar, A. T. Khader, F. Nadi, "An Analysis of Selection Methods in Memory Consideration for Harmony Search", Applied Mathematics and Computation, Vol. 219, N. 22, pp. 10753-10767, 2013.

[77] O. M. Alia , R. Mandava, "The Variants of the Harmony Search Algorithm: an Overview", Artificial Intelligence Review, Vol. 36, N. 1, pp. 49-68, 2011.

[78] J. Fourie, R. Green, Z. W. Geem, "Generalised Adaptive Harmony Search: A Comparative Analysis of Modern Harmony Search", Journal of Applied Mathematics, Vol. 2013 Article ID 380985, 2013.

[79] A. Yadav, N. Yadav, J. H. Kim, "A Study of Harmony Search Algorithms: Exploration and Convergence Ability", Advances in Intelligent Systems and Computing, Vol. 382, pp. 53-62, 2015.

[80] P. M. Marcus, M. A. Marcinkiewicz, B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics, Vol. 19, N. 2, pp. 313-330, 1993.

[81] W. N. Francis, H. Kucera, "The Brown Corpus: A Standard Corpus of Present-Day edited American English", Providence, RI: Department of Linguistics, Brown University [producer and distributor], 1979.

[82] S. M. Abdulhamid, S. Ahmad, V. O. Waziri, F. N. Jibril, "Privacy and National Security Issues in Social Networks: The Challenges", International Journal of the Computer, the Internet and Management, Vol. 19, N. 3, pp. 14–20, 2011.

[83] Social Media Stats in 2014, `http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html`, retrieved on January 2015.

[84] "Keeping our users secure", entry in the blog `https://blog.twitter.com/2013/keeping-our-users-secure`, retrieved on January 2015.

[85] A. Martin, N. Anutthamaa, M. Sathyavathy, M. M. S. Francois, D. V. P. Venkatesan, "A Framework for Predicting Phishing Websites using Neural Networks", International Journal of Computer Science Issues, Vol. 8, pp. 330–336, 2011.

[86] M. B. Salem, S. J. Stolfo, "Modeling User Search Behavior for Masquerade Detection", Recent Advances in Intrusion Detection, pp. 181–200, 2011.

[87] M. B. Salem, S. J. Stolfo, "Detecting Masqueraders: A Comparison of One-class Bag-of-words User Behavior Modeling Techniques", Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, Vol. 1, N. 1, pp. 3-13, 2010.

[88] I. Fette, N. Sadeh, A. Tomasic, "Learning to Detect Phishing Emails", 16th International Conference on World Wide Web, pp. 649–656, 2007.

[89] R. Dhamija, J. D. Tygar, "The Battle against Phishing: Dynamic Security Skins", Symposium on Usable Privacy and Security, pp. 77-88, 2005.

[90] D. Miyamoto, H. Hazeyama, Y. Kadobayashi, "A Proposal of the AdaBoost-based Detection of Phishing Sites", Joint Workshop on Information Security, 2007.

[91] Y. Zhang, J. I. Hong, L. F. Cranor, "Cantina: a Content-based Approach to Detecting Phishing Web Sites", 16th International Conference on World Wide Web, pp. 639–648, 2007.

[92] G. Xiang, J. I. Hong, "A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval", 18th International Conference on World Wide Web, pp. 571–580, 2009.

[93] G. Xiang, J. Hong, C. P. Rose, L. Cranor, "Cantina+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites", ACM Transactions on Information and System Security (TISSEC), Vol. 14, N. 2, 21, 2011.

[94] T. Lane, C. E. Brodley, "Sequence Matching and Learning in Anomaly Detection for Computer Security", AAAI Workshop: AI Approaches to Fraud Detection and Risk Management, pp. 43–49, 1997.

[95] E. Villar-Rodríguez, J. Del Ser, S. Salcedo-Sanz, "On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks", Springer Studies in Computational Intelligence, Vol. 570, pp. 259–268, 2015.

[96] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, "COMPA: Detecting Compromised Accounts on Social Networks", Symposium on Network and Distributed System Security (NDSS), 2013.

[97] M. Schonlau, W. DuMouchel, W. H. Ju, A. F. Karr, M. Theus, Y. Vardi, "Computer Intrusion: Detecting Masquerades", Statistical Science, pp. 58-74, 2001.

[98] H. Gao, Y. Chen, K. Lee, D. Palsetia, A. N. Choudhary, "Towards Online Spam Filtering in Social Networks", Symposium on Network and Distributed System Security (NDSS), 2012.

[99] C. Cortes, V. Vapnik, "Support-Vector Networks", Machine learning, Vol. 20, N. 3, pp. 273-297, 1995.

[100] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, X. Deng, "Detection of Phishing Webpages based on Visual Similarity", 14th International Conference on World Wide Web, pp. 1060–1061, 2005.

[101] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, "On the Evolution of User Interaction in Facebook", ACM Workshop on Online Social Networks, pp. 37–42, 2009.

[102] A. Bergholz, J. H. Chang, G. Paass, F. Reichartz, S. Strobel, "Improved Phishing Detection using Model-Based Features", Proceedings of the Conference on Email and Anti-Spam (CEAS), 2008.

[103] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection", Anti-phishing Working Groups, 2nd Annual eCrime Researchers Summit, pp. 60–69, 2007.

[104] L. E. Agustin, S. Salcedo-Sanz, S. Jiménez-Fernández, L. Carro-Calvo, J. Del Ser, J. A. Portilla-Figueras, "A New Grouping Genetic Algorithm for Clustering Problems", Expert Systems with Applications, Vol. 39, N. 10, pp. 9695–9703, 2012.

[105] L. van der Knaap, F. Grootjen, "Author Identification in Chat Logs using Formal Concept Analysis", 19th Belgian-Dutch Conference on Artificial Intelligence, pp. 181–188, 2007.

[106] G. Inches, M. Harvey, F. Crestani, "Finding Participants in a Chat: Authorship Attribution for Conversational Documents", International Conference on Social Computing (SocialCom), pp. 272–279, 2013.

[107] M. L. Brocardo, I. Traore, S. Saad, I. Woungang, "Authorship Verification for Short Messages using Stylometry", IEEE International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–6, 2013.

[108] G. Hirst, O. G. Feiguina, "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts", Literary and Linguistic Computing, Vol. 22, N. 4, pp. 405–417, 2007.

[109] N. Graham, G. Hirst, B. Marthi, "Segmenting Documents by Stylistic Character", Natural Language Engineering, Vol. 11, N. 4, pp. 397–415, 2005.

[110] A. Abbasi, H. Chen, "Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace", ACM Transactions on Information Systems, Vol. 26, N. 2, Article 7, 2008.

[111] D. Jurafsky, J. H. Martin, "Speech & Language Processing", Pearson Education India, 2009.

[112] F. J. Tweedie, R. H. Baayen, "How Variable may a Constant be? Measures of Lexical Richness in Perspective", Computers and the Humanities, Vol. 32, N. 5, pp. 323-352, 1998.

[113] E. Stamatatos, N. Fakotakis, G. Kokkinakis, "Automatic Text Categorization in Terms of Genre and Author", Computational Linguistics, Vol. 26, N. 4, pp. 471-495, 2000.

[114] H. Baayen, H. Van Halteren, F. Tweedie, "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution", Literary and Linguistic Computing, Vol. 11, N. 3, pp. 121–132, 1996.

[115] P. M. McCarthy, G. A. Lewis, D. F. Dufty, D. S. McNamara, "Analyzing Writing Styles with Coh-Metrix", FLAIRS Conference, pp. 764-769, 2006.

[116] L. Pearl, M. Steyvers, "Detecting Authorship Deception: a Supervised Machine Learning Approach using Author Writeprints", Literary and Linguistic Computing, Vol. 27, N. 2, pp. 183–196, 2012.

[117] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for information Science and Technology, Vol. 60, N. 3, pp. 538-556, 2009.

[118] M. Fissette, F. A. Grootjen, "Author Identification in Short Texts", B.Sc.Thesis, Raboud Universiteit Nijmegen, 2013.

[119] J. Karlgren, G. Eriksson, "Authors, Genre, and Linguistic Convention", Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, 2007.

[120] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, B. Maia, "'twazn me!!!;(' Automatic Authorship Analysis of Micro-blogging Messages", In Natural Language Processing and Information Systems, pp. 161–168, 2011.

[121] M. Koppel, N. Akiva, I. Dagan, “Feature Instability as a Criterion for Selecting Potential Style Markers”, Journal of the American Society for Information Science and Technology, Vol. 57, N. 11, pp. 1519-1525, 2006.

[122] T. Chen, M. Y. Kan, ”Creating a Live, Public Short Message Service Corpus: the NUS SMS Corpus”, Language Resources and Evaluation, Vol. 47, N. 2, pp. 299-335, 2013.

[123] T. A. Almeida, J. M. G. Hidalgo, A. Yamakami, “Contributions to the Study of SMS Spam Filtering: New Collection and Results”, Proceedings of the 11th ACM symposium on Document Engineering, pp. 259–262, 2011.

[124] J. M. G. Hidalgo, T. A. Almeida, A. Yamakami, “On the Validity of a New SMS Spam Collection”, IEEE International Conference on Machine Learning and Applications (ICMLA), Vol. 2, pp. 240-245, 2012.

[125] T. Almeida, J. M. G. Hidalgo, T. P. Silva, “Towards SMS Spam Filtering: Results under a new Dataset”, International Journal of Information Security Science, Vol. 2, N. 1, pp. 1–18, 2013.

[126] K. Toutanova, D. Klein, C. D. Manning, Y. Singer, “Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network”, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1, pp. 173–180, Association for Computational Linguistics, 2003.

[127] I. S. Dhillon, S. Mallela, R. Kumar, “A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification”, Journal of Machine Learning Research, Vol. 3, pp. 1265–1287, 2003.

[128] G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification”, Journal of Machine Learning Research, Vol. 3, pp. 1289–1305, 2003.

[129] J. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”, Advances in Large Margin Classifiers, Vol. 10 (3), pp. 61–74, 2000.

[130] M. Hauschild, M. Pelikan, “An Introduction and Survey of Estimation of Distribution Algorithms”, Swarm and Evolutionary Computation, Vol. 1, N. 3, pp. 111–128, 2011.

[131] P. Larrañaga, J.A. Lozano (Eds.), “Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation”, Kluwer, Boston, MA, 2002.

[132] S. Salcedo-Sanz, J. Del Ser, I. Landa-Torres, S. Gil-Lopez, and J. A. Portilla-Figueras, “The Coral Reefs Optimization Algorithm: A Novel Metaheuristic for Efficiently Solving Optimization Problems”, The Scientific World Journal, Vol. 2014, Article ID 739768, 15 pages, 2014.

[133] Mirror News: a real impersonation attack in Facebook, `http://www.mirror.co.uk/news/uk-news/facebook-paedophile-posed-teen-attempt-5563126`, published on April 22nd, 2015.