



**Intelligent Information System Supporting
Observation, Searching and Detection for
Security of Citizens in Urban Environment**



European Seventh Framework Programme
FP7-218086-Collaborative Project

D4.3. Report on current state-of-the-art of
machine learning methods for behavioural
profiling

The INDECT Consortium

AGH — University of Science and Technology, AGH, Poland

Gdansk University of Technology, GUT, Poland

InnoTec DATA GmbH & Co. KG, INNOTECH, Germany

IP Grenoble (Ensimag), INP, France

MSWiA — General Headquarters of Police (Polish Police), GHP, Poland

Moviquity, MOVIQUITY, Spain

Products and Systems of Information Technology, PSI, Germany

Police Service of Northern Ireland, PSNI, United Kingdom

Poznan University of Technology, PUT, Poland

Universidad Carlos III de Madrid, UC3M, Spain

Technical University of Sofia, TU-SOFIA, Bulgaria

University of Wuppertal, BUW, Germany

University of York, UoY, Great Britain

Technical University of Ostrava, VSB, Czech Republic

Technical University of Kosice, TUKE, Slovakia

X-Art Pro Division G.m.b.H., X-art, Austria

Fachhochschule Technikum Wien, FHTW, Austria

©Copyright 2010, the Members of the INDECT Consortium

Document Information

Contract Number	218086
Deliverable Name	Report on current state-of-the-art of machine learning methods for behavioural profiling
Deliverable number	D4.3.
Editor(s)	Suresh Manandhar, University of York, suresh@cs.york.ac.uk
Author(s)	Ioannis Klapaftis, University of York, giannis@cs.york.ac.uk
Reviewer(s)	Alan Frisch, University of York, frisch@cs.york.ac.uk
Dissemination level	Public
Contractual date of delivery	30 June 2010
Delivery date	23 July 2010
Status	Final version
Keywords	Behavioural Profiling



This project is funded under 7th Framework Program

Contents

Document Information	1
1 Executive Summary	5
2 Introduction	6
2.1 Objectives	6
2.2 List of participants & roles	7
2.3 The task of behavioural profiling	7
2.3.1 Application of behavioural profiling	8
2.4 Behavioural profiling & INDECT	8
2.5 Overview & structure of the report	9
3 Geographical profiling methods	10
3.1 Introduction	10
3.2 Crime generators, attractors, hotspots	10
3.3 The concept of re-victimisation	12
3.4 Statistical methods for geographical profiling	12
3.5 Clustering similar crimes	24
3.6 Summary	27
4 Offenders characteristics profiling methods	28
4.1 Introduction	28
4.2 Language modelling	29
4.3 Authorship identification & characteristic induction	33
4.4 Detecting deceptive identities	37
4.5 Summary	39
5 Intrusion detection profiling methods	40
5.1 Introduction	40
5.2 Machine learning methods to intrusion detection	41
5.3 Summary	47
6 Conclusions	48

List of Figures

1	Description of a simulation of the model presented in Sort et al. [1]	20
2	Plot of function $Y = e^{-x*0.1}$	21
3	Plot of function $Y = e^{-x*0.1}$ with 2 distance units for <i>plateau</i> and 2 distance units for <i>steps</i>	22
4	Linear fitting	23
5	Self-Organising Map	26
6	An example of a ROC curve	32
7	Functional word taxonomies	35

List of Tables

1	Offence types & application of behavioural profiling. GP: Geographical Profiling, TP: Text-based Profiling, IP:Intrusion Detection Profiling . . .	9
2	Time intervals between offences in days.	14
3	Distances between offences in metres.	14
4	Random permutation of time intervals between offences.	15
5	Knox standardised residuals - Contingency table of observed values. . . .	15
6	Knox standardised residuals - Contingency table of expected values. . . .	16
7	Knox standardised residuals	16
8	Parameters of the model presented in Short et al. [1]	19
9	Bache et al. [2] crime types & datasets	29
10	Soundex numbers & represented letters. Letters A, E, I, O, U, H, W, and Y are disregarded.	38
11	Document Updates	55

(This page is left blank intentionally)

1. Executive Summary

Security is becoming a weak point of energy and communications infrastructures, commercial stores, conference centers, airports and sites with high person traffic in general. Practically any crowded place is vulnerable, and the risks should be controlled and minimised as much as possible. Access control and rapid response to potential dangers are properties that every security system for such environments should have. The INDECT project is aiming to develop new tools and techniques that will help the potential end users in improving their methods for crime detection and prevention thereby offering more security to the citizens of the European Union.

Behavioural profiling aims at analysing different types of solved crimes in order to identify behavioural patterns of known offenders. The induced patterns can then be exploited by police forces in order to: (1) enhance their crime prevention abilities, and (2) increase their crime detection rate. In the context of the INDECT project, Work Package 4 (WP4) is responsible for the Extraction of Information for Crime Prevention by Combining Web Derived Knowledge and Unstructured Data. This document is the fourth deliverable of WP4 and provides a detailed review of the current-state-of-the-art in the field of behavioural profiling. The study of current methods allows us to identify their main strengths and weaknesses that will guide us in the steps of the project.

2. Introduction

The general aim of WP4 is the development of key technologies that facilitate the building of an intelligence gathering system by combining and extending the current-state-of-the-art methods in Natural Language Processing (NLP). One of the specific goals of WP4 is to propose NLP and machine learning methods that learn the behavioural profiles of known offenders or offending groups. A key requirement for the development of such methods is the analysis of the current-state-of-the-art methods for *behavioural profiling*.

2.1. Objectives

In this report, we provide a critical survey of the field of behavioural or offender profiling. Based on our review, offender profiling methods are divided into the following categories:

- **Geographical profiling methods**
This category of methods focuses on analysing structured data regarding the locations and the time of a series of connected crimes in order to identify patterns in space and time that predict when and where new crimes are likely to occur.
- *Offenders characteristics* profiling methods
This category of methods focuses on analysing the free text description of a large number of solved crimes, in order to build behavioural models that can identify different offender characteristics, such as age, occupation, gender and others. These models can assist police officers to narrow down the list of suspects for an unsolved crime and prioritise police resources.
- **Intrusion detection profiling methods**
This category of methods focuses on analysing the behaviour of legitimate users of a computer system, in order to build their legitimate profile and detect any intrusion effort as a deviation of the legitimate profile. Equivalently, similar methods attempt to model patterns of known attacks, in order to match those against illegal future actions.

The objective of this deliverable is to identify the advantages of current offender profiling methods, as well as their theoretical and practical limitations that we aim to address in

the next stages of project by developing novel behavioural profiling methods.

2.2. List of participants & roles

This report has been produced by the University of York (UOY), and will be utilised by INNOTEC for the purpose of dissemination (D9.9)

2.3. The task of behavioural profiling

The electronic lexical database of WordNet [3] defines *criminal offence* as an act that is punishable by law. Recorded crime statistics have been collected since 1857, hence the amount of collected data is the most important resource for discovering trends in crime and analysing crime patterns [4].

Despite that, the manual analysis of collected data is difficult, time-consuming and most importantly error-prone, given the high volume of crime statistics being collected daily. Behavioural profiling aims to add computational support on the task of crime data analysis in order to overcome the strain put on human resources.

This computational support focuses on discovering and learning patterns of known offenders based on the solved crimes that have been committed by them, in order to predict the location and the time crimes are likely to occur, and to identify their perpetrators [5]. Note that behavioural profiling does not refer to racial profiling, which is both illegal and ineffective [5]. In contrast, it refers to capturing and modelling certain characteristics of offenders that are expressed during an illegal activity.

In the next section of this report we will focus on three distinct, yet similar, areas of behavioural profiling, in which different Machine Learning (ML) and Natural Language Processing (NLP) methods were applied with a varied level of automation. These methods are summarised below:

- Machine learning (neural networks, statistical models, clustering) approaches for formulating models of criminal behaviour, classifying new activities or clustering similar crimes.
- Language modelling & distributional similarity approaches for capturing behavioural

characteristics (e.g. age, gender, etc.).

2.3.1. Application of behavioural profiling

The application of behavioural profiling in criminal offences is not a straightforward task, since different types of crimes require utilising different types of features. For example, the common crime offence of burglary is largely different from a traffic violation or from a terrorist attack, since these offences have different motives.

Table 1 provides a list of the main types of offences, in which behavioural profiling has found an application. The specific methods applied to these types of offences are described in the next sections of the report. In Table 1, the first column refers to the type of crime, the second to its textual description and the third to the type of behavioural profiling application.

The application of behavioural profiling to any type of crime offence has two main targets. The first one is to increase the crime prevention rate. As it has already been mentioned, behavioural profiling aims at learning patterns of criminal behaviour based on past data of solved crimes. The learned patterns can suggest or predict the location, time and other attributes of future criminal offences, hence enhance the preventive abilities of police forces.

Secondly, the learned patterns can also be applied to unsolved cases of the past, in order to reveal information that had not been captured, in effect assisting in solving these cases. Moreover, behavioural patterns and models can be applied to present cases, in order to narrow down the lists of suspects, so that police officers perform their tasks more effectively.

2.4. Behavioural profiling & INDECT

Within the context of INDECT, behavioural profiling is a key factor, because it allows the detection of characteristics of known criminals or criminal groups, offering the opportunity of creating methods for the automatic detection of threats and recognition of abnormal behaviour or violence.

Additionally, the characteristics extracted by a behavioural analysis system are important

Criminal offence type	Description of offence	Application
Burglary	Entering a building unlawfully with intent to commit a felony or to steal valuable property.	GP/TP
Robbery	Larceny by threat of violence.	GP/TP
Homicide	The killing of a human being by another human being.	GP
Terrorism	The calculated use of violence (or the threat of violence) against civilians in order to attain goals that are political or religious or ideological in nature.	GP
Serial killing	Murdering three or more people[6, 7] over a period of more than thirty days, with a cooling off period between each murder, and whose motivation for killing is largely based on psychological gratification.[7]	GP
Theft	The act of taking something from someone unlawfully.	TP
Shoplifting	The act of stealing goods that are on display in a store.	TP
Assault	A threatened or attempted physical attack by someone who appears to be able to cause bodily harm if not stopped.	TP
Vandalism	Willful wanton and malicious destruction of the property of others.	TP
Criminal identity deception	Intentional falsification of identity.	TP
Computer/network/email deception	Malicious activity such as denial of service attacks, port scans or even attempts to crack into computers by monitoring network traffic.	IP

Table 1: Offence types & application of behavioural profiling. GP: Geographical Profiling, TP: Text-based Profiling, IP:Intrusion Detection Profiling

features that can be exploited by methods aiming at the detection of websites, blogs, or forums, that promote illegal activities. The identification of these sources of illegal activity is one of the main target of WP4.

2.5. Overview & structure of the report

We begin our review by first presenting methods focusing on geographical profiling. Next, we describe methods exploiting free-text in order to build behavioural models, and then we describe the application of offender profiling in intrusion detection. For each of these categories we describe in detail a number of methods and discuss their strengths and weaknesses in the summary of each section.

3. Geographical profiling methods

3.1. Introduction

Geographical profiling is an investigative methodology that focuses on the analysis of the location and time of a series of crimes, in order to: (1) identify when and where a crime is likely to occur, and (2) estimate the threat risk associated with a specific *area*, where the concept of *area* is typically empirically defined, i.e it can be a small neighbourhood around a victimised house, a village, a city and so on.

Current methods focusing on geographical profiling make use of concepts and evidence that were empirically-derived during the nineties. The most significant of these concepts are the following:

- Crime generators
- Crime attractors
- Crime hotspots
- Re-Victimisation

In the next two sections, we describe each of these concepts in detail, in order to form the context, in which automatic geographical profiling methods have been and are being developed.

3.2. Crime generators, attractors, hotspots

Crime generators are places that attract a large number of people for reasons unrelated to any level of criminal motivation they might have [8]. Such places include shopping centres, entertainment centres, sport stadiums, large metro or train station and others. These places generate crime by creating appropriate concentrations of people, where a subset of them are potential offenders with levels of criminal motivation sufficient to commit a crime [9]. Note that these potential offenders do not necessarily visit these locations to commit a crime. In contrast, they might perform a criminal act, when criminal opportunities are available.

Crime attractors are places, areas, and other locations, in which strongly motivated, conscious and most importantly intending offenders are attracted by the well-known criminal-opportunities created in these places [8]. Examples of such locations are bar districts, drug dealing areas, insecure parking areas or shopping malls and others.

For instance, a shoplifter would consciously look for insecure shopping areas to perform his/her criminal act, while a drug dealer would also look for less policed areas to buy or sell illegal substances. These places can also generate other types of crime that are the by-products of the primary-crime of an offender.

Apart from crime generators and attractors most urban environments also include crime-neutral areas. These areas neither attract criminal offenders nor generate crimes by creating criminal opportunities [10, 8]. Criminal offences take place sporadically and occasionally.

Having defined these three crime-related concepts, it is important to note that there is no clear or hard distinction between them. In contrast, some areas might be crime generators for a specific type of crime and crime attractors for another one.

Crime hotspots are areas of concentrated crime suffering from a very high rate of crime incidents as opposed to crime neutral areas [11]. The definition of a crime hotspot encompasses the concepts of crime generators and attractors, provided that a hotspot is geographically defined to be a small specific place.

However, hotspots can also be defined at a higher level, i.e. *street hotspots*, *neighbourhood hotspots* or *large-area hotspots*. Different types of crime require the focus on different type of hotspots [11].

For instance, for small thefts and shoplifting incidents it is more reasonable to focus on street and place locations, in which such incidents occur, rather than larger areas hotspots. In contrast, when considering the incidents of a serial killer, then it would be more beneficial to concentrate on larger areas than the exact street address and number, in which the offences took place.

3.3. The concept of re-victimisation

Repeat victimisation is one of the most important concepts used in behavioural profiling. It refers to the phenomenon, in which multiple offences are committed against the same person or location. Research on repeat victimisation has identified that prior victimisation is a very good predictor of future risk [12, 13]. This risk is greater in places with high crime rates or in areas being characterised as crime hotspots [14].

There are two conceptual approaches that attempt to clarify the phenomenon of repeat victimisation [15]. The first one, referred to as event-dependent, states that the risk of victimisation increases as a direct result of an initial criminal incident. For example, if there was a burglary in a house at time T , then then the risk of a second burglary in the same house increases within a period of time starting from T . Johnson & Bowers [15] note that the victimisation *boosts* the chance of further victimisation.

The second category of approaches, referred to as risk heterogeneity approaches, states that future victimisations are independent of past victimisations. In contrast, the vulnerabilities that became apparent after the first victimisation are the ones that possibly contribute to re-victimisation.

Finally, Johnson & Bowers [15] note that the balance between these approaches is not clear, although there is significant evidence in favour of event-dependent approaches. Specifically, most repeat incidents are within a few day range, while interviews with offenders have revealed that they return on the same property because: (1) they already know their way around, and (2) they want to get what was left out [16].

3.4. Statistical methods for geographical profiling

Johnson & Bowers [15] exploit and further extend the concept of re-victimisation by investigating the risk associated with non-victimised houses, once a nearby and similar house was burgled. If the risk is substantially different than in the past, then preventive actions could be taken in order to cover possible vulnerabilities.

To achieve their aim, they examine the extent to which burglaries cluster in geographical space and time, as well as, how the type of burgled property and its surrounding area influence the development of such clusters.

Specifically, their primary starting point is an analogy with epidemiology, i.e. the study of factors affecting the health and illness of populations¹. They hypothesize that house burglaries could exhibit the same communicability as diseases, where a disease is communicable if people catch it within a short period of time after the exposure to the disease agent. Hence, communicability is inferred from closeness in space and time of manifestations of the disease [15].

A similar phenomenon could also be exhibited in the case of house burglaries. The main presumption is that once an offender commits a burglary in a house, then he/she obtains knowledge regarding the surrounding area, the neighbouring houses features, such as vigilance, possible vulnerabilities and others. In other words, the offender's knowledge base of that area is updated, in effect making it worth repeating the offence at the same or a nearby similar property.

In their setting, multiple offences at the same property are referred as *repeats*, while offences at nearby properties (after an initial burglary took place) are referred *near-repeats*. Time is formulated as an interval between two offences either in terms of days or weeks.

The distance between two properties is measured using the Euclidean distance. Specifically, let x_1, y_1 be the coordinates of a property A and x_2, y_2 the coordinates of a property B . The distance, $D(A, B)$, between them is defined in Equation 1.

$$D(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

The data of their experiment were taken from the county of Merseyside and a set of pre-processing steps were taken in order to clean the records and distinguish between single and repeat offences. At the end of this stage, each offence record consisted of a set of 7 fields:

- Reference number
- Address of offence
- Address coordinates (x, y)
- Date of offence

¹<http://en.wikipedia.org/wiki/Epidemiology>

- Phonetic code of victim’s name (to ensure anonymity)
- Time of offence
- Type of victimised property (e.g detached house)

Analysis using the Mantel z-statistic

The Mantel z-statistic [17] evaluates the correlation between two distance (or similarity or dissimilarity) matrices. In their setting, the Mantel test computes a correlation between two distance matrices, where one matrix represents distances between two offences (Table 3), while the other represents time intervals between two offences (Table 2). The null hypothesis is that the observed relationship between the two matrices could have been obtained by any random arrangement in space (or time) of the observations.

	Offence 1	Offence 2	Offence 3
Offence 1	0	10	25
Offence 2	10	0	30
Offence 3	25	30	0

Table 2: Time intervals between offences in days.

	Offence 1	Offence 2	Offence 3
Offence 1	0	100	300
Offence 2	100	0	600
Offence 3	300	600	0

Table 3: Distances between offences in metres.

This statistic measures the clustering of burglaries in space and time, and then compares the observed distribution with an expected distribution on the basis of chance. Given a total number of n burglaries, let d_{ij} be the distance between burglaries i and j (Table 3) and t_{ij} be the time passed between these two burglaries (Table 2). The z -statistic for the observed data can be calculated using Equation 2:

$$z = \sum_{i=1}^n \sum_{j=1}^n d_{ij} \times t_{ij}, i \neq j \quad (2)$$

In the next step, the test generates an expected distribution of events based on chance, i.e by randomly choosing pairings between all times and distances separating events, and

calculates the mean z -statistic of this random distribution. A random permutation of the time (Table 2) would involve substituting one row with another row or one column with another column. Table 4 shows such a permutation, where we have substituted the first row of Table 2 with its third row.

	Offence 1	Offence 2	Offence 3
Offence 1	25	30	0
Offence 2	10	0	30
Offence 3	0	10	25

Table 4: Random permutation of time intervals between offences.

A z -test is then carried out in order to compare the observed and expected values for significance. A positive z -score, i.e higher than 1.96 (95% confidence level) would indicate strong evidence of a relationship between time and distance of offences, hence strong relationship between space-time clustering of burglary.

Analysis using Knox standardised residuals

The previous statistical analysis produces one score for the entire set of data. The Knox standardised residuals method [18] uses a contingency table showing the number of offences that occur within a certain distance of each other within a certain time interval. Table 5 shows such a contingency table. In the table, we observe for example that 25 offences took place within an one week interval, while the distances between offences were greater than 100 and less than 500 meters.

Time/Distance	0 – 100 (metres)	100 – 500 (meters)	500 – 1000 (meters)
1 (week)	30	25	20
2 (weeks)	20	15	14
3 (weeks)	5	3	1

Table 5: Knox standardised residuals - Contingency table of observed values.

The use of Knox standardised residuals allows to calculate these residuals for each cell of the contingency table, in effect being able to observe if each cell has a significantly different value from that expected on the basis of chance. In other words, the standardised residual is a measure of the degree to which an observed chi-square cell frequency differs from the value that would be expected on the basis of the null hypothesis.

Table 6 shows the contingency table of expected values under the model of independency.

Time/Distance	0 – 100 (metres)	100 – 500 (meters)	500 – 1000 (meters)
1 (week)	24.43	20.07	15.55
2 (weeks)	15.49	11.01	10.40
3 (weeks)	3.07	1.29	-0.53

Table 6: Knox standardised residuals - Contingency table of expected values.

The standardised residual of each cell i, j can be calculated using Equation 3, where $O(i, j)$ refers to the i, j cell of observed values table and $E(i, j)$ refers to the expected values table.

Time/Distance	0 – 100 (metres)	100 – 500 (meters)	500 – 1000 (meters)
1 (week)	-23.94	19.95	7.98
2 (weeks)	-7.98	-27.93	41.23
3 (weeks)	87.78	6.65	-18.37

Table 7: Knox standardised residuals

$$z = \frac{O(i, j) - E(i, j)}{\sqrt{E(i, j)}} \quad (3)$$

Table 7 shows the standardised residuals of our example. In the Table high positive standardised residuals indicate more space-time clustering than would be expected on the basis of chance, while negative values indicate fewer than expected offences.

The two statistical analysis methods were applied on 1692 domestic burglaries committed with an one year interval (April 1999 - April 2000). The result of the Mantel z -score was 5.49, which is significantly greater than the value of 1.96 for a 95% confidence level. This result revealed more space-time clustering in the collected data than one would expect under the model of independence.

For the Knox method, the categories used for time were months and the distances tenths of kilometers. They report that 2734 out of the 2861172 offence combinations occurred in less than a month and 0.1 kilometers apart, while the expected value on the basis of chance was 2090 (Knox standardised residual was equal to 6.5). This result validated the first experiment, i.e. that burglaries cluster in space and time.

Furthermore, the residual method showed that most burglaries take place within 1 month and 400 metres of a burgled property, and then the rate of burglaries falls sharply. The central conclusion of their experiments is that an initial burglary event is a good predictor

of next burglaries within a period of 1 to 2 months and within a distance of 300 to 400 metres.

Another approach to the problem of geographical profiling is presented in [1]. Short et al. [1] present a model to study the emergence dynamics and equilibrium states of crime hotspots. Specifically, they present a quantitative mathematical model that attempts to identify the formation of hotspots by considering a set of sociological factors such as the type of urban environment, the crime rate in the environment and the rate of near-repeats.

Their model is based on a two-dimensional lattice, i.e. a rectangular grid, in which houses are imagined to exist there with a constant lattice spacing l . Each house in the lattice can be identified by its coordinates, $s = (i, j)$, while each house is also characterised by a degree of criminal attractiveness $A_s(t)$ at time t . The quantity $A_s(t)$ is a measure that quantifies the burglar's perception of the attractiveness of the house at the point s .

This quantity is modeled to be equivalent to the statistical rate of burglary at point s , when a burglar is actually present. Formula 4 formally defines $A_s(t)$, where A_s^0 is a static component of attractiveness and $B_s(t)$ is the dynamic component which is associated with repeat or near-repeat victimisation.

$$A_s(t) \equiv A_s^0 + B_s(t) \quad (4)$$

The second component of the lattice is the set of criminal agents that commit the crimes against the houses of the lattice. These criminal agents may perform one out of two actions during a time interval:

- commit a burglary at the house they are currently located

The event of a burglary is a stochastic event characterised by the probability of occurrence of a criminal agent at house s between times t and $t + \delta t$. Equation 5 defines that probability in accordance with a Poisson process, in which the expected number of burglary events during the time interval with duration δt is equal to $A_s(t)\delta t$ [1].

$$p_s(t) = 1 - e^{-A_s(t)\delta t} \quad (5)$$

Once a house is burglarised, the criminal agent responsible for this is removed from the lattice, assuming that he/she needs to return to his base and to avoid committing crimes for a period of time. In order to simulate the return of these burglars back to the lattice, criminal agents are generated at constant rate Γ .

- move to another neighbouring house (without burglarising that house).

When the criminal agent does not commit an offence, then it moves to a neighbouring house in the grid. This movement is a random walk process biased towards the attractiveness of neighbouring houses.

Short et al. mention three reasons for their random walk choice. Firstly, criminals tend to commit crimes in areas surrounding locations that they frequently visit, e.g. home, work etc [19]. Secondly, other studies [20] have shown that the frequency of crimes of offenders decreases monotonically with the distance of the offender from its primary residence increasing.

Thirdly, in the case of burglaries the tendency to stay close to the primary residence outweighs the profits of more distant and desirable targets [21, 22] Hence, the probability of movement from a house at point s to another house at point m at time t is defined in Equation 6, where $N(s)$ is the set of houses neighbouring house s .

$$q_{s \rightarrow n}(t) = \frac{A_n(t)}{\sum_{s' \in N(s)} A_{s'}(t)} \quad (6)$$

Returning on the attractiveness $A_s(t)$ of a house at point s and time t , recall that it was modeled according the dynamic component $B_s(t)$ that reflects the phenomena of repeat and near-repeat victimisations. Formally, the dynamic component $B_s(t + \delta t)$ is defined in Equation 7. We observe that the dynamic attractiveness depends on three factors.

The first one is the parameter θ multiplied by $E_s(t)$, i.e. the number of burglary events in the house at point s and time t . The higher $E_s(t)$, the higher the value of this factor, hence the value of dynamic attractiveness.

The second is a quantity $(1 - \eta)B_s(t)(1 - \omega\delta t)$, which is the previous value of the dynamic component scaled by the length of $(1 - \omega\delta t)$. Note that ω is the dynamic attractiveness decay rate parameter, which models the decrease in the attractiveness as time increases with no burglary events.

The third factor is the quantity $\sum_{s' \in N(s)} B'_s(t)(1 - \omega\delta t)$, i.e. the previous values of the dynamic components of the neighbouring houses scaled again by the $(1 - \omega\delta t)\delta t$. The third factor allows to model the concept of near-repeat victimisation, since the dynamic attractiveness of each neighbouring house is spread according to a factor $\frac{\eta}{Z}$, where η is a parameter between 0 and 1 and Z is the number of neighbouring houses.

$$B_s(t + \delta t) = [(1 - \eta)B_s(t) + \frac{\eta}{Z} \sum_{s' \in N(s)} B'_s(t)](1 - \omega\delta t) + \theta E_s(t) \quad (7)$$

Table 8 shows all the parameters of this method that are used during simulation.

Parameter	Description
l	lattice spacing
δt	Time step
ω	Dynamic attractiveness decay rate
η	Neighbourhood effects (range from 0 to 1)
θ	Increase in attractiveness due to a burglary event
A_s^0	Initial attractiveness of house at point s
Γ	Rate of generating criminal agents

Table 8: Parameters of the model presented in Short et al. [1]

Short et al. [1] note that a spatially homogeneous equilibrium is achieved when all houses in the grid have the same attractiveness \bar{A} and the same number of criminals \bar{n} . To achieve this equilibrium then the average dynamic attractiveness should be equal to $\bar{B} = \frac{\theta\Gamma}{\omega}$ and the average number of criminals should be equal to $\bar{n} = \frac{\Gamma\delta t}{1 - e^{-\bar{A}\delta t}}$.

In their simulations, they initially put their model in a homogeneous equilibrium state and then they varied the parameters of their model (Table 8) to investigate whether their model will remain in such state. Figure 1 illustrates the pipeline of a set of discrete simulations. The results of their simulations have exhibited three types of behaviour with respect to the attractiveness variable $A_s(t)$:

1. No hotspots, i.e. all houses have the same attractiveness and any local fluctuations disappear very quickly [1].
2. Dynamic hotspots, i.e. their model forms spots of increased attractiveness that remain in such a state for a varying length of time depending on the actual parameter values [1].

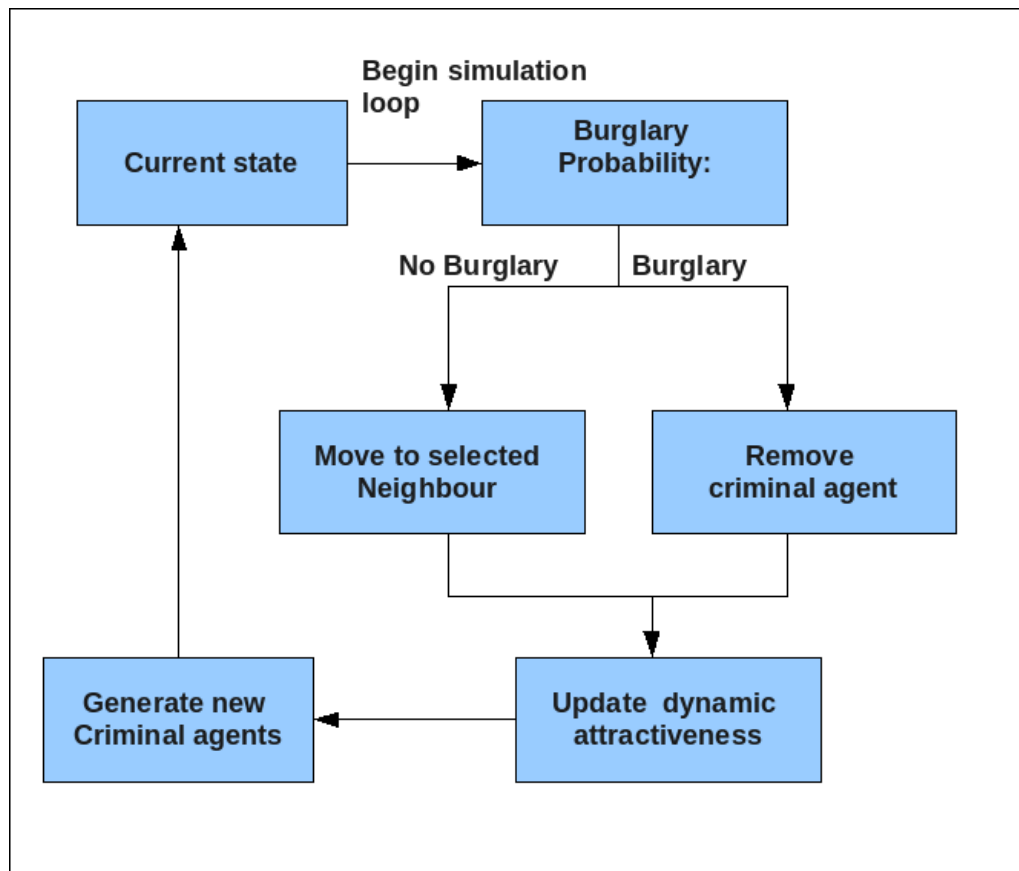


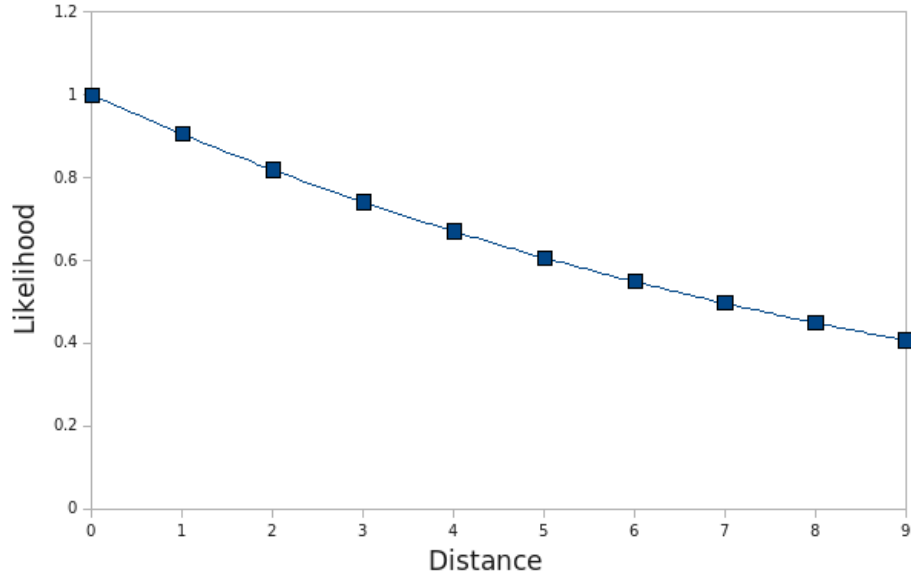
Figure 1: Description of a simulation of the model presented in Sort et al. [1]

- Stationary hotspots, i.e. their model forms spots of high attractiveness surrounded by places of low attractiveness. The size of these hotspots depends on the actual parameter values.

A quantitative evaluation of this model is left as future work.

Canter et al. [23] present another method for geographical profiling, in which they attempt to identify the location of the base residence of a serial killer. Their method is based on the assumption that serial offenders tend to live within an area circumscribed by their offences [19]. Their main assumption is also supported by earlier work [24], which has shown that 87% of the 45 serial rapists from South England had their permanent residence within a circle whose diameter was defined by their furthest offences.

In their work, the location of residence of an offender is sought within a rectangular map

Figure 2: Plot of function $Y = e^{-x*0.1}$

that is divided into 13300 square regions. Based on the work of Rhodes & Conly, they explored a family of negatively exponential decay function types defined in Equation 8, where Y is the likelihood that a particular location contains the residence of an offender, x is the distance of that location from an offence site and β is the exponential coefficient. In their setting they explored 19 different parameter values for β , in effect coming up with 19 different exponential functions.

$$Y = e^{-\beta x} \quad (8)$$

Figure 2 shows one of the selected functions, in which we have set β equal to 0.1. As can be observed, as the distance from the offence site increases the likelihood of the particular location decreases.

Canter et al. [23], have also experimented with the concepts of *plateau* and *steps*, where the first one refers to the setting in which a decay function returns always 1, while the second refers to the case, in which a decay function returns 0. These values are inserted in front of an exponential function.

In their setting, they have experimented with a maximum of 4 distance units to define

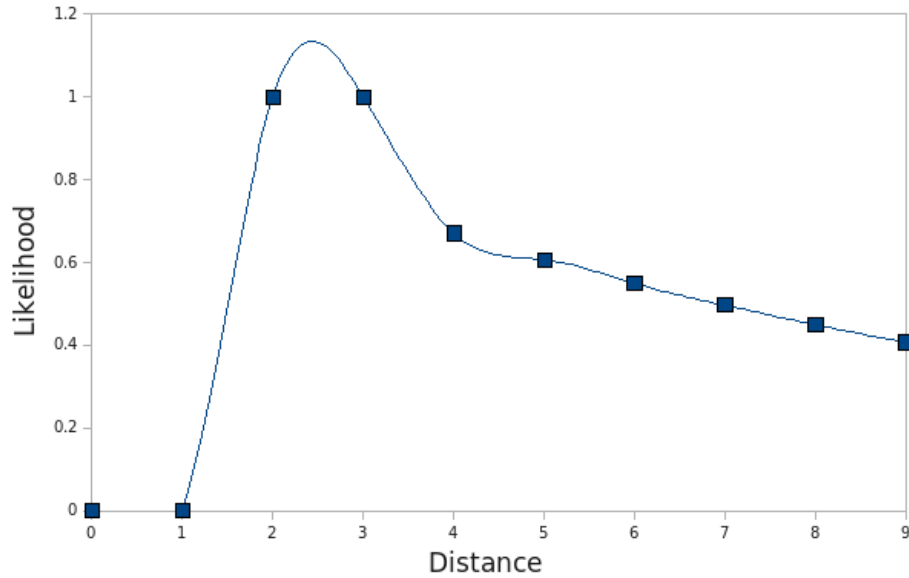


Figure 3: Plot of function $Y = e^{-x*0.1}$ with 2 distance units for *plateau* and 2 distance units for *steps*

both *plateau* and *steps*, e.g. 3 distance units for steps and 1 distance unit for the plateau or 2 distance units for steps and 2 distance units for plateau and so on. In total, there were 15 different combinations for each of the initial 19 decay functions, in effect yielding a total of 285 decay functions. Figure 3 shows the decay function of Figure 2, in which we applied two distance units for *steps* and two distance units for *plateau*. As can be observed, the likelihood is 0 for distance points 0 and 1, and then the likelihood is 1 for distance points 2 and 3.

Each of the selected 285 functions is applied over a distance (x), which is moderated by a normalisation parameter linked to an offenders offence distribution [23]. Two normalisation parameters were tried. The first one was the mean interpoint distance between all offences (MID). The second, QRRange, is calculated by fitting a linear function to the crime scenes points, where each crime scene is defined by two coordinates. Figure 4 shows such a case, in which we have generated a set of 10 offence points and then we fitted a line ($f(x) = 4x + 0.8$) to these points using the method of least squares. Finally, QRRange is calculated by taking the average of the perpendicular distance of all offence points to the fitted line. In total, each of the 285 functions was tested with both normalisation parameters, in effect yielding a total of 570 functions.

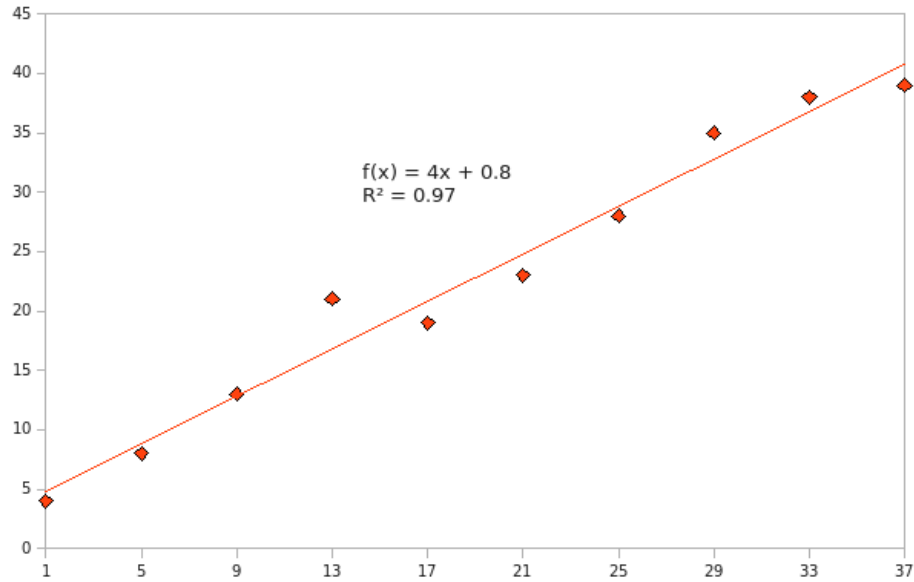


Figure 4: Linear fitting

To test one offence series, i.e. a set of offences by the same offender, they use one of the 570 functions. Initially, the normalisation parameter is calculated either with QRange or MID, and then distance x is set equal to the normalisation parameter multiplied by 2. The function is then applied in a radial fashion, which means that the diameter of the circle is four times the normalisation parameter. It is straightforward that each location in the perimeter of the circle is assigned the same likelihood. This process is repeated for each offence site and finally, the likelihood for each location in the grid is the mean of each of the likelihood values assigned to that location by those functions applied to that location. For example, if the application of the function in Figure 3 at offence site s provides a likelihood for location l equal to 0.6 and a likelihood of 0.3 for the same location l when applied at offences site s' , then the average likelihood value would be equal to 0.45.

To test the effectiveness of their method, the list of locations is ranked according to their average likelihood values, and each location is searched for the offender's residence. When the offender's home base is reached, a cost value is returned. This value is equal to the proportion of all possible locations searched before the offender's home base was found. For example, if the sorted locations list consisted of 100 locations, and the offender's residence was found in the fiftieth location, then the cost value would be equal to 50%,

i.e half of all locations should be searched in order to identify the desirable location.

Their evaluation was performed on 79 U.S. serial killers. Each serial killer had more than one series of crime, and the associated body disposal sites were used as offence sites. The best results for MID and QRange are achieved using the decay functions with $\beta = 1$ and $\beta = 2$ respectively, with associated search costs 0.19 and 0.11. Furthermore, for functions it was clear that the use of QRange always resulted in lower search costs than MID.

Additionally, their evaluation showed that the use of *steps* and *plateaus* was not as effective as expected. Specifically, the most economic function was the one with no *steps* or *plateaus* using the QRange (search cost of 11%).

3.5. Clustering similar crimes

Adderley [25] presents a different approach compared to the ones described in the previous sections. In this work, the target is to exploit the spatial, temporal and other features of crime offences in order to cluster similar ones. The solved crimes within a cluster can then be used to suggest potential suspects for unsolved crimes that belong to the same cluster. The type of crime offences they deal with is burglaries in licensed premises, such as bars, public houses, pubs and others.

Three types of features are used in this method. The first type are temporal features. In consultation with police analysts 11 time bands (within a day) were defined, where each one is represented by a dichotomous variable [25]. These bands are the following:

- Over night
- Early hours
- Take to school
- Lunch
- Get from school
- Evening
- Morning (other)
- Afternoon (other)

- Evening (other)
- Short Wend
- Long Wend

The second type of features is the spatial feature of the average distance that an offender travels from his/her home in order to commit a crime offence. The third type of features includes data extracted by the police reports that describe the crime. Adderley [26] defined a set of 33 dichotomous variables, such as which part of the building afforded entry, whether there was an alarm in the building, how was the alarm activated and others. Two more variables were added in the set of 33, in order to include specific attributes for the crime they were studying. The literature does not provide a full list of these 35 variables.

A total of 1121 individual crimes (feature vectors) were then given as an input to Self-Organising Map (SOM) [27], which clustered these crimes. A self-organising map consists of components called nodes, where each node is a weight vector that has the same dimension as the input vectors and a position in the map space. The number of nodes in their setting was 100, i.e. he chose a 10 by 10 arrangement [26]. Figure 5 shows an example of a SOM with 3 by 3 feature map. As can be observed, not all connections between nodes are shown for clarity.

In the first step, the map's nodes' weight vectors are initialised randomly. Next, in each iteration an input vector is chosen, and its Euclidean distance from each node weight vector is calculated. The node weight vector producing the smallest distance is selected, and its neighbouring nodes weight vectors are updated to be closer the selected input vector. This process is repeated until convergence. At the end of the process, the input weight vectors (crimes) will be assigned to one of the 100 clusters (10 by 10 map) of the map.

In their evaluation, they selected six crimes that were committed by an offender X , and then they selected the clusters in which these 6 crimes were placed. All 433 crimes that appeared in those clusters were analysed and 404 were filtered out by setting a threshold on the distance between X 's residence and the offence site. This threshold was set to 6 miles, since none of the 6 crimes that were attributed to X occurred more than 6 miles from X 's home residence.

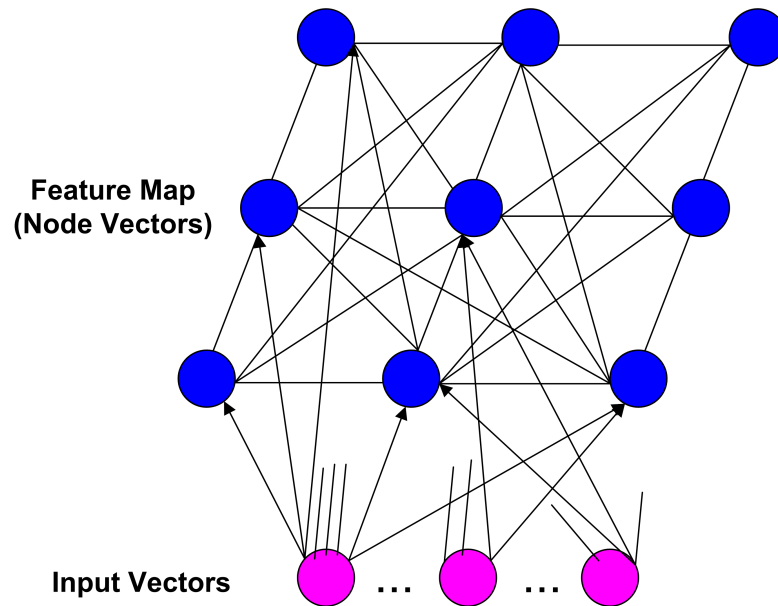


Figure 5: Self-Organising Map

The remaining 29 crimes were again filtered by removing those who were already attributed to another offender. Finally, the remaining 24 crime offences were analysed by police officers in order to determine if they could be attributed to offender *X*. In their opinion all of the 24 crimes could have been committed by the particular offender [25].

3.6. Summary

This section presented a thorough review of a number of methods applied to geographical profiling. Specifically, we described methods that focus on modelling the concepts of repeat and near-repeat victimisation in order to identify significant relationships between the location and time of a series of offences.

Additionally, other methods exploited these features in order to identify the location containing the home residence of an offender, while others targeted at identifying the emergence dynamics of crime hotspots. Furthermore, temporal and spatial features were also used in a framework of clustering similar crimes, in order to suggest possible offenders for unsolved cases.

It is evident from the description of this section that the concepts of repeat and near-repeat victimisation, as well as the travel distance an offender is willing to travel, are the most important features exploited by geographical profiling methods within a statistical framework. Despite that, it is unclear whether these features are portable across different crime types, since they have only been applied on a specific one.

Additionally, depending on the type of crime, new features could also be exploited. For instance, in terrorism it would probably be beneficial to have a feature measuring the experience of a terrorist group, e.g. by counting the number of attacks, in order to assist in predicting future attacks and their severity.

Regarding the evaluation of presented methods, we have observed that most evaluations are performed using a sample of data taken by the police. Given the large range of crimes and crime types, it is unclear whether the produced results would be portable to different larger datasets, different locations of crimes and so on.

Additionally, there is a need for a common evaluation framework, although this highly depends on the subproblem of geographical profiling considered. For example, a method targeting at identifying hotspots would possibly need different evaluation metrics than a method identifying clusters of similar crimes. Despite that, common evaluation settings could be established among methods targeting the same subproblem.

4. Offenders characteristics profiling methods

4.1. Introduction

Modus operandi (Method of operation) is a Latin phrase that is typically used in criminal investigation to describe a crime committed by an offender as well as the methods employed for committing such a crime. The description of a crime is typically written in a police record that might contain a number of structured and unstructured data including the following:

- Free text describing the method employed for committing a crime
- Feature code for the type of a particular offence
- Feature code for the presense or absence of a specific aspect of behaviour
- Feature code for the gender of the offender
- Feature code for the age of the offender
- Feature code for the ethnic appearance of the offender

We can collect the values for the above offender characteristics by solving the corresponding criminal case. Methods exploiting the police reports, and especially the wealth of information provided by free text, aim at identifying specific personal characteristics of a particular offender in order to narrow down the list of suspects for a particular offence, in effect helping the police officer to perform his/her duties more effectively.

In the same vein, other sources of free text, such as notices, emails, and other documents produced, for example, by terrorist organisations can also be exploited to extract personal or group characteristics.

Of particular interest to WP4 is the development of methods that exploit free text available on the web in order to model criminal behaviour and identify cases in which specific behavioural patterns lead to actions that deviate significantly from legal boundaries.

In this section, we present and critically review the state-of-the-art methods exploiting mainly free text, possibly in combination with structured data, in order to identify behavioural characteristics significant for crime prevention and detection.

4.2. Language modelling

Bache et al. [28, 2] presented a language modelling method for inferring the characteristics of offenders from an existing police archive. Based on the information included in police reports of solved cases their target was to link behavioural features with characteristics of offenders. They have focused on the following offender characteristics:

1. Gender

The gender of a criminal agent can either be male or female.

2. Age

The age of a criminal age is defined to be either below or above the median of the ages of offenders committing a series of crimes.

3. Ethnic appearance

The ethnic appearance of an offender can either be white European or Afro-Caribbean.

4. Occupation

The occupation of an offender can only take two possible values, i.e. employed or unemployed.

Additionally, their focus is on a set of eight crime types. Table 9 shows the types of these crimes, along with the number of crimes for each type and the size of the vocabulary of the police reports. For more information, readers are referred to [2].

Crime type	Number of crimes	Vocabulary size
Theft from Vehicles	317	418
Other theft	380	808
Shoplifting	2381	1057
Assault	2230	1576
Criminal damage	934	1183
Damage to vehicles	253	471
Burglary	1292	1226
Robbery	352	632

Table 9: Bache et al. [2] crime types & datasets

The main assumption in their method is that an offender committing a crime generates a document that describes his/her actions, even though the actual document is produced

by a trained police officer. The vocabulary in this document can then be exploited to link that offender with one or more of the personal characteristics mentioned above.

For each of the four personal characteristics a different language model can be learned from a set of reports of solved cases (e.g. a language model for an offender to be male). This language model can then be applied to new unsolved cases and predict the personal characteristic of the offender (e.g. that the offender is male or female). In the following, let us assume a language model for females (f).

Given a document d , the target is to calculate the probability $p(f|d)$, i.e. the probability of the offender to be female given the document d describing the offence. This probability can be easily calculated using Bayes theorem (Equation 9).

$$p(f|d) = \frac{p(f)p(d|f)}{p(d)} \quad (9)$$

The probability $p(d)$ is the probability that document d has been generated and acts as a normalising constant. Hence, it can be eliminated, when we are trying to predict whether the offender is female or male, since it will not change the result of the maximisation. In other words, $p(f|d) \propto p(f)p(d|f)$.

The probability $p(f)$ is the prior probability of the offender to be female. This probability is used to reflect our prior knowledge about the gender of offenders. For example, if police forces already know from their recorded shoplifting crimes that 60% of those are committed by females, then $p(f)$ can be set to 0.6. If such knowledge is not present, then it is usual to assume a uniform prior, i.e. $p(f) = p(m) = 0.5$.

Bache et al. calculate the prior probabilities by the training set they already have. The assumption here is that the training set is representative enough to calculate a reliable estimate without overfitting to the training set.

Regarding the probability $p(d|f)$, i.e. the probability of generating a document d given that the offender is female, they use two different models. The first one, is the common multinomial model used in language modelling, while the second is the binomial model. The multinomial model can be used to calculate $p(d|f)$ in Equation 10, where $p(t|f)$ is the probability that the female language model will generate term t and $fr(t, d)$ is the frequency of term t in d .

$$p(d|f) = \prod_{t \in d} p(t|f)^{fr(t,d)} \quad (10)$$

The probability $p(t|f)$ can be easily calculated using the training data (Equation 11), where F is the set of training documents describing crimes committed by females.

$$p(t|f) = \frac{\sum_{d \in F} fr(t,d)}{\sum_{d \in F} |D|} \quad (11)$$

For the binomial model similarly, $p(d|f)$ and $p(t|f)$ can be calculated using Equations 12 and 13, where S_t is the set of documents containing term t .

$$p(d|f) = \prod_{t \in d} p(t|f) \times \prod_{t \notin d} (1 - p(t|f)) \quad (12)$$

$$p(t|f) = \frac{|S_t|}{|F|} \quad (13)$$

A significant problem that language models have to deal with is data sparsity. This refers to the case where the probability of a particular term in a given model is zero due to its absence from the training corpora. However, the absence of a term from the training data does not mean that the term will also be absent from a new unseen document describing a new criminal case. In general, when dealing with text corpora data sparsity is always an issue [29].

A typical method to deal with data sparsity is smoothing. Applying a smoothing technique involves the subtraction of mass from highly frequent events and the addition of the subtracted mass to zero or low frequency events. Bache et al. [2] exploit the smoothing method of Jelinek-Mercer (interpolation method) [30].

In their setting, they use the set of both solved and unsolved crimes (*universal*) to define the probability of a term t given a language model as the linear interpolation of the maximum likelihood estimate and the marginal probability of t in the *universal* set. This is shown in Equation 14, where λ is the parameter of smoothing that was set equal to 0.5.

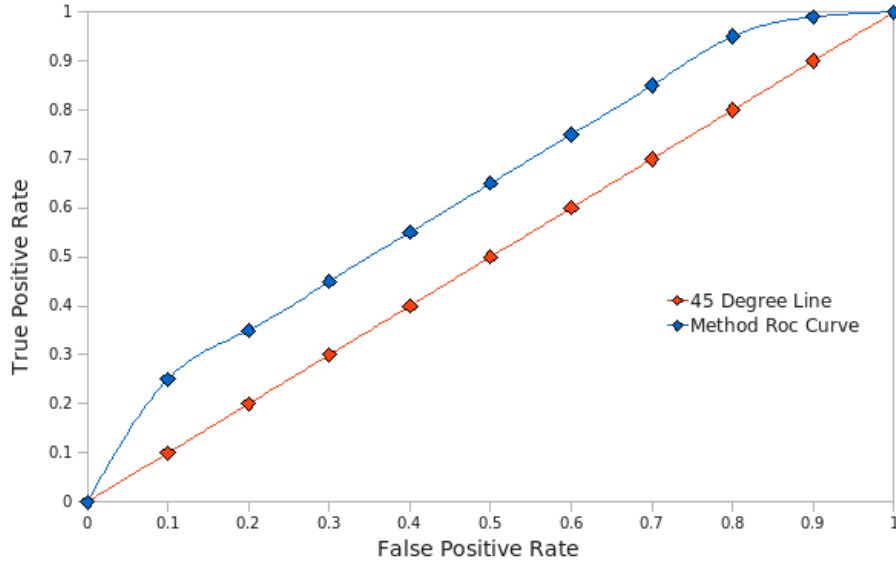


Figure 6: An example of a ROC curve

$$p'(t|f) = (1 - \lambda) \times p(t|f) + \lambda \times p_{universal}(t) \quad (14)$$

For training and evaluation, they use an adaptation of Leave-One-Out (LOO) method. In LOO, a single observation from the data acts as test data, and the remaining observations acts as the training data. This is repeated such that each observation in the sample is used once as the validation data. In their setting, all crimes of an offender are removed in turn and the rest are used for training the models.

For evaluating each of their models, they use the Receiver Operating Characteristic (ROC) curve. ROC curve, is a plot of the true positives versus the false positives of a binary classifier as its discrimination threshold is varied. In their setting, given the language model (f), the discrimination threshold would be the value of $p(f|d)$. Values above $p(f|d)$ would predict females. True positives are cases in which the system correctly predicts females, while false positives are cases, in which the system predicts females but the actual solution is males.

Figure 6 shows an example of such a curve. The straight line indicates a random binary classifier, i.e. a classifier with no predictive power. In contrast, curves above the straight

line indicate predictive power. To calculate a summary statistic, they computed the AUC (Area Under Curve) measure of bowedness and judged whether their models deviate significantly from the straight line (AUC=0.5) by applying the Wilcoxon Ranked Sign test [2].

Their results showed a significance relationship (95%) with ethnic appearance, occupation using the multinomial model and a significance relationship (99%) using the binomial model. For both models there was a significant relationship with age (99%) and gender (95%). Furthermore, the comparison between the multinomial and binomial model showed that although there are no striking differences, the latter performs slightly better.

In the second evaluation setting they experimented with, they attempted to do actual classification, e.g. by predicting that an offender is female or employed if the corresponding probability was above 0.5. Then they compared their predictions with the gold standard answers using the Chi-squared test.

Their results showed that age is a significant characteristic helping to predict all crimes they considered apart from burglary. In the same vein, ethnic appearance seems to be strongly associated with theft, shoplifting, criminal damages, vehicle damage, burglary and robbery, while employment status is associated with assault, vehicle damages and robbery. Finally, the gender seems to be more associated with shoplifting, theft and burglary.

4.3. Authorship identification & characteristic induction

Authorship identification is the task of associating a document with its original author. In the context of computer forensics, authorship identification of emails is an important area, since emails can be used not only for legitimate activities, but also for illegal ones. For instance, emails may be misused for distributing threatening, offensive or even terrorism-promoting material.

De Vel et al. [31] present a method for authorship identification of emails based on a Support Vector Machine (SVM) classifier [32]. SVMs are supervised learning methods that have been used in different NLP tasks including text classification [33], question classification [34], word sense disambiguation [35], and more recently to relation extraction [36]. SVMs provide a principled approach to classification and provide good generalisation

[33]. Deliverable 4.2, *Report on current state-of-the-art methods for relationship mining*, provides a detailed description of classification tasks using SVMs.

De Vel et al. [31] focus on identifying the author of an email irrespectively of the subject or topic of the email. For that reason they select an extensive amount of stylistic and structural features that are neither content nor context dependent. Their features are divided into two categories. The first one includes stylistic features such as the proportion of blank lines in the emails, the average sentence length, the function word distribution and others. The total number of stylistic features was 170.

Structural features are mainly concerned with the particular structure followed when writing email, i.e. whether the email has a greeting or a farewell acknowledgement, if it contains a signature text, the number of attachments, the frequency of html tags and others. The total number of structural features was 21.

Their experiments were performed on a corpus created by the emails of three authors regarding three different topics, i.e. food, travel and movies. The emails were collected by newsgroups and the average number of words per author was around 12000. Since, SVM are binary classifiers, they generated 3 classification models (author 1 versus author 2, author 2 versus author 3, author 1 versus author 3), and each SVM classification was applied 3 times. Performance was then measured using the standard measures of recall, precision and F-Score per author, and the results were then averaged.

They experimented with two evaluation settings. In the first one, all emails were aggregated into one class and 10-fold cross validation was applied for training and validation. Their results for each author varied from an F-Score of 77.6% to an F-Score of 91.6%. A comparison of the stylistic versus the structural features showed that the classification performance depends highly on the former and significantly less on the latter. However, this is expected given that the feature space of stylistic features is significantly richer than the feature space of structural features.

In their second evaluation setting, they experimented with different topic classes. Specifically, the SVM was trained on the movie topic and tested on the food and travel topics. Interestingly, the performance differences between the current and the previous experiment (topics aggregated into one class) were generally small, in effect leading to the conclusion that stylistic and structural attributes can effectively discriminate between authors even when multiple topic categories are involved [31].

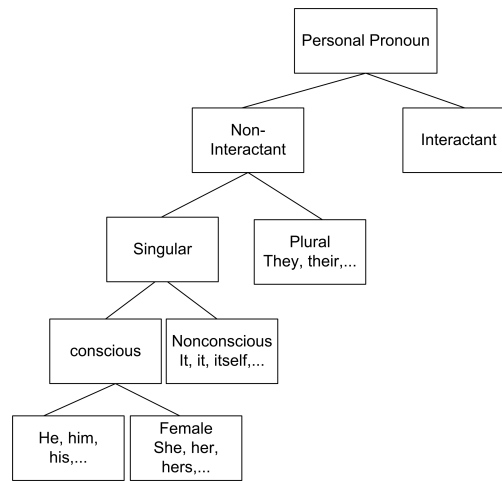


Figure 7: Functional word taxonomies

Recently, another similar approach to authorship profiling was presented in Argamon et al. [37]. Their main difference with the previous method is that their focus is not on identifying the author of a particular document. In contrast, they exploit free text in order to identify an author’s gender, age, native language and neuroticism level.

To perform their task they use a supervised learning method that is trained on vectors having two types of features: (1) stylistic features and (2) content features. For style-based features, they use a novel feature set that subsumes both function words and Part-Of-Speech (POS) tags. The corresponding taxonomies are taken from System Functional Linguistics [38] and provide meaningful distinctions between POS tags and function words [38]. An example of such taxonomy is shown in Figure 7.

The second type of features considered, i.e. content-based features are the top 1000 words that appear sufficiently frequently in the corpus and that discriminate best between the categories according to the application of Kullback-leibler divergence to a hold-out set of training data.

Finally, the supervised classifier that they use is the Multinomial Logistic Regression that is used to estimate the probability of an event by fitting the data to a logistic curve. D4.2, *Report on current state-of-the-art methods for relationship mining* has provided a detailed description and different applications of that classifier.

For evaluation, as it has already been mentioned, they focus on identifying the author’s

gender, age, native language and neuroticism level. The corpus used for gender and age consisted of 19320 blog authors, while the self-reported age and gender of each author was used to create the gold standard. Their classification performance on gender was equal to 76.1% when considering both content and stylistic features. Performance was lower when considering only one of the two feature types, although content-based features perform better than style-based. This result complements the experimental setting of De Vel et al. [31], who did not use any content features.

In the case of age, they created 3 age categories, i.e. (13-17), (23-27) and (33-47). Intermediate ages were not considered as their corresponding blogs were written over a period of several years [37]. The classification results for the age experiments were similar to the gender experiments results. Specifically, the classification accuracy was 77.7% when considering both style and content features, 75.5% for content features only and 66.9% for stylistic features.

Similarly, for identifying native language, they used the International Corpus of Learner English² (ICLE) that contains over 3 million words of writing by learners of English from 21 different mother tongue backgrounds. In their experiment, they selected the writings of 258 authors from each of the following countries: Russia, Czech Republic, Bulgaria, France and Spain. Their objective, was to identify which of the six languages is the native tongue of the author of a text written in English.

The performance in this experiment was better when using only content-based features (82.33%). Stylistic features were not as effective as in the previous experiments achieving an accuracy of 65.1%. Despite that, Argamon et al. [37] note that authors of different languages were assigned topics from an ICLE standard, and since the author instructions may have varied in each country, the differences in content usage might be an artifact of the experimental setup.

In their last experiment, Argamon et al. [37] attempted to identify the level of neuroticism of undergraduate students at the University of Texas. Specifically, their subjects were given a 20-minute period to write an essay describing their current thoughts and feelings. Each student completed a questionnaire, whose aggregate scores define a scale of neuroticism. To formulate a classification task, participants with scores in the upper third were considered as neurotic and those in the lower third as not. Interestingly, their method achieved an accuracy of 65.7% in detecting neuroticism, 15.7% above the

²<http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm>

random baseline. Additionally, they mention that the most discriminating features for neurotics is that they tend to focus on themselves (e.g. using words such as *myself*, *I*), they frequently use subject pronouns and prepositional phrases involving some action (e.g. *in order to*).

4.4. Detecting deceptive identities

Identity deception is an intentional effort to falsify the true identity of a person in order to prevent criminal investigation [39]. According to Wang et al. [39], the Federal Bureau of Investigation (FBI) discovered that conventional investigation methods run into a difficulty when criminal agents use fraudulent identities. Identity deception manifests itself when criminal offenders lie about their name, date of birth, address, social security number and so on [39].

To deal with identify deception Wang et al. [39, 40] presented a method for automatically identifying patterns of identity deception. Their method is based on two basic components, i.e. a spelling string comparator and a phonetic string comparator. The first one is the edit distance or Levenshtein distance [41] between two strings. This distance is a metric that measures how dissimilar those strings are by counting the minimum number of edits that are needed in order to transform one of the strings into the other. An edit operation is considered to be the insertion, substitution and deletion of a character.

The phonetic string comparator used is Soundex ³, an algorithm for indexing names by their sound. The goal of this method is to encode homophones in the same representation, so that they can be matched despite differences in spelling. According to the rules of Soundex, each name is transformed to a Soundex code, where a Soundex code consists of a letter and three digits. The letter is the first letter of the name, while the numbers are assigned to the rest of the letters according to Table 10. For example, the surnames *Mcqueen* and *Mqueen* would be assigned the same code, i.e. M205.

In their setting, they focused on four identity attributes, i.e. name, address, date of birth and social security number. For names they used separately both edit distance and the Soundex algorithm. In addition, edit distance was used for date of birth, address and social security number. Given two records containing the above attributes, the disagreement values for each field were normalised and then summed up and averaged to

³<http://www.archives.gov/genealogy/census/soundex.html>

Number	Represented letters
1	B, F, P,
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Table 10: Soundex numbers & represented letters. Letters A, E, I, O, U, H, W, and Y are disregarded.

represent an overall disagreement value.

Evaluation was performed on a set of 120 deceptive criminal records involving 44 criminals, where 80 of these records were used for training and the remaining 40 for testing. During training, a disagreement matrix was created, in which each cell (i, j) contained the disagreement value between record i and record j . The highest accuracy (97.4%) in the training set was achieved by setting the disagreement value to 0.48. This threshold was also used in the testing dataset, where records having a disagreement values less than 0.48 were considered to belong to the same criminal agent. They achieved an accuracy of 94% showing that their method generalised well on unseen data.

4.5. Summary

This section has provided a thorough description of approaches to inducing offender characteristics using unrestricted text. Specifically, we have presented a number of methods that exploit the free text, in order to identify the age, gender, ethnic appearance, occupation and neuroticism level of possible offenders as well as the author or a given document e.g. email. The methods presented are typically based on supervised classifiers such as support vector machines and logistic regression.

Additionally, one of the presented methods applied techniques from language modelling associating each of the desirable characteristics with a different language model, while another one focused on a weakly-supervised method to identify deception using string and phonetic distance.

Within the context of INDECT, it is crucial to be able to continuously capture and update behavioural models of different offenders based on their committed crime types (hooliganism, terrorism and others). As a result, the use of supervised techniques would impose restrictions on the portability of the developed algorithms as well as on the feasibility of their applications, since there might not be enough training data for every crime type or offender.

For that reason, our future work within the domain of behavioural profiling will focus on minimally supervised methods. Additionally, in certain crime types such computer network intrusion, it is possible to create training and evaluation data through simulation of abnormal activity as the next section shows.

5. Intrusion detection profiling methods

5.1. Introduction

Intrusion detection is a subarea of Information Security focusing on the detection of illegitimate actions that target at exploiting the vulnerabilities of information resources in order to gain access to those. Given that many business and government organisations use the Internet to provide public access to information, it becomes necessary to establish a security framework that protects these resources.

Intrusion detection methods attempt to set that security framework in order to protect the confidentiality, integrity and availability of information. These methods can be categorised into *anomaly detection* and *misuse detection* methods.

Anomaly detection methods, also known as outlier detection methods [42], focus on monitoring legitimate users and their actions for a prespecified period of time, in order to build a model of user-oriented normal behaviour [43]. In the next step, they try to identify patterns that deviate significantly from the normal behaviour model they have already learned. In contrast, misuse detection methods focus on learning the models of illegitimate behaviour first, and then try to match the learned models against future attacks.

According to Lee et al. [44], the task of implementing an intrusion detection system is an enormous engineering task, in which system designers exploit their experience in attack scenarios and possible system vulnerabilities.

As a result, the majority of these methods rely on a time-consuming, ad-hoc and error-prone design process. In this section, we present a variety of methods that exploit the wealth of information contained in audit data sources in order to learn anomaly or misuse detection models. These methods reduce or even eliminate the need for manual analysis of threats and vulnerabilities providing a formal framework for applying statistical models to intrusion detection.

5.2. Machine learning methods to intrusion detection

Lee et al. [44] present a framework both for misuse detection by exploiting audit data sources. Their framework consists of the following main components:

- Classifier

The classifier maps a new data item (e.g. sequence of commands) into one of pre-defined categories. In misuse detection, the categories include the *normal* category and the types of attacks. It is straightforward that the classifier is required to use sufficient audit data for each possible category in order learn a reliable model.

- Link analysis

Assuming that audit data consist of a set of records, where each record consists of a set of fields (e.g. host name, IP address, etc.), this component identifies correlations between record fields in the audit data. For example, there might be a high correlation between Eclipse IDE (command) and Java files (argument of command).

- Sequence analysis

For classification Lee et al., [44] use a classification rule learning program, known as RIPPER [45], that generates a set of rules describing a possible attack. An example of such rules is the following:

1. *pass_guess* : *failed_logins* ≥ 3 , if the failed logins are equal or greater than 3, then this is a password guess attack.
2. *DOS* : *IP_hits* ≥ 100 , *Time* $\leq 1sec$, if there are more than 100 visits from the same IP within second, then this is a Denial-Of-Service attack.

The accuracy of the classifier depends on the selected features for the given type of attack. In their setting, a set of patterns are extracted by mining the audit data, and then these patterns are used as guidelines to construct temporal weighted features will be described below. Additionally, their method targets at building adaptable intrusion detection systems. For that reason, they build a set of classifiers, each one for

a different type of attack. The predictions of these classifiers are then combined into a Metal-classification scheme that inductively learns the correlation of predictions from the different basic classifiers [44].

Mining a set of patterns from audit data is the task associated with the component of link analysis. In this component, pattern extraction is cast into the task of mining association rules [46]. Let N be a set of records from the audit data, where each record consists of a set of fields (e.g. host name, time, command, etc.), and X, Y be two sets of fields. An association rule is an expression $A = X \rightarrow Y$, where $X \cap Y = \emptyset$. The support of the rule is $s = \frac{freq(A)}{N}$ and its confidence is $c = \frac{freq(A)}{freq(X)}$, where $freq(A)$ is the number of records that contain both X and Y and $freq(X)$ is the number of records containing X . An example of such rules would be the following:

- $\{\text{Time=am,command=Eclipse}\} \rightarrow \{\text{arg = Java file}\}, [s=0.1,c=0.99]$
If time is morning and the command is Eclipse, then the argument of the command is a Java file. 10% of the audit records contain that rule, while 99% of times that the command Eclipse is executed in the morning, its argument is a Java file.
- $\{\text{Time=am,command=Emacs}\} \rightarrow \{\text{arg = C file}\}, [s=0.05,c=0.8]$
If time is morning and the command is Emacs, then the argument of the command is a C file. 5% of the audit records contain that rule, while 90% of times that the command Emacs is executed in the morning, its argument is a C file.

Since, the audit data do not contain directed relations between fields or sets of fields, the *schema* level information of the domain can be exploited to define the fields that should be positioned on the left part of the association rules. These fields, referred to as essential features, guide the mining process, in effect restricting the produced association rules to only those containing the essential features.

The extracted rules can be combined to form sequential patterns of network events. This is the task of the sequence analysis component. Given a time interval $[t_1, t_2]$, where $t_2 > t_1$ and X, Y, Z are sets of fields, a frequent episode rule is the expression $X, Y \rightarrow Z$ occurring within the specified interval. As in the case of simple association rules, frequent episodes are also assigned two weighting schemes based on support and confidence. The frequent sequential patterns are extracted by first finding the frequent associations using the essential features, and then the time interval is specified to combine these associations into a single rule.

Finally, the frequent sequential patterns are used as guidelines to build a set of features automatically. This process starts by first defining a reference feature value (e.g. *destination_host*), i.e. the one system designers believe should be included in their rules. Given a time interval of w seconds, the following list of features can be added:

- The count of connections to the same *destination_host*
- Let F_1 be another feature (other than F_0) occurring in all item sets of the episode. A feature can be the percentage of connections that share the same F_1 value.
- Let F_2 be another feature (other than F_0 and F_1) occurring in all item sets of the episode. A feature can be the percentage of connections that share the same F_2 value.

Their experiments were performed on the DARPA Intrusion Detection Evaluation Program⁴, where each participating system was required to learn models of intrusion detection using a set of training data and then test the system on a set of testing data. The training dataset consisted of 4 gigabytes of compressed network traffic data, i.e. 5 million connections. The evaluation scheme focused on four categories of attacks, i.e. Denial-Of-Service (DOS), unauthorised access from a remote IP (R2L), unauthorised access to local superuser privileges by a local unprivileged user (U2R) and Probing (Pr). The anomalous behaviour was simulated by having a less privileged user to behave as a more privileged one, e.g. a normal user behaves like a system administrator.

Additionally, they built three classification models, whose predictions were then combined to a meta-level classifier (RIPPER). The first classification model was the *traffic* model that extracts DOS and Pr detection rules. The second classification model was the *host-based traffic* model that extracts Pr rules, while the third model was the *content* model that extracts both *R2L* and *U2R* rules.

The test data contained 38 attack types with 14 types occurring in the test set only. Their performance was evaluated on a ROC curve (e.g. Figure 6), where the x-axis denoted the percentage of legitimate connections classified as illegitimate (intrusion), while the y-axis denoted the percentage of intrusions that were detected by their system. The analysis of the curves shows that their system performs better than random for all models, although they mention that their performance in R2L attacks is poor.

⁴<http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html>

For all intrusion types they report an average of 70% detection rate, which however is not satisfactory in a real-world application. Furthermore, their average performance was particularly low on attack types that did not exist in the training set. Specifically, the average detection rate on attack types that were in the training set was around 80.2%, while for those that were not included in the training dataset, performance dropped to 37.7%. This picture reveals the problematic case of having inadequate training data to build classification models and the need for developing weakly supervised methods.

Lane & Brodley [47] present a method for anomaly detection that builds user profiles by exploiting Unix command sequences, and then tests whether a current input sequence is similar to the profile learned using a statistical similarity measure. The underlying assumption is that a user behaves in a similar manner in similar type of situations, in effect leading to patterns of behaviour. Naturally, such an assumption might not valid, in cases where an intruder attempts to emulate a normal user's behaviour.

In their setting, a *sequence* is defined as an ordered, fixed-length set of temporally adjacent actions, where each action is taken to be a Unix command along with their arguments. During extraction of the data, the temporal order of commands was only retained for a single shell, while file arguments were transformed in order to correspond to abstract file types rather than specific file names. For example, the following list of commands would be translated into the token stream $ls -l\ cd\ <1>\ scp\ <2>\ <3>$.

```
> ls -l
> cd research
> scp file1.txt /n/staff/giannis
```

During training each extracted token stream is stored to a database record. This database along with a similarity measure and a set of system parameters define a user's profile. Once training has completed and the profiles of each user have been stored, detection takes place by comparing an incoming input of user sequences to the user profile. To perform this task, all token streams are first segmented into overlapping sequences of tokens of length l , where l is a parameter. Two fixed-length sequences, $X = (x_0, x_1, \dots, x_{l-1})$ and $Y = (y_0, y_1, \dots, y_{l-1})$ can then be compared using Equations 15 and 16.

$$Sim(X, Y) = \sum_{i=0}^{l-1} w(X, Y, i) \quad (15)$$

$$w(x, y, i) = \begin{cases} 0 & ,\text{if } i < 0 \text{ or } x_i \neq y_i \\ 1 + w(x, y, i - 1) & ,\text{if } x_i = y_i \end{cases} \quad (16)$$

Given two sequences $cd <1> ls scp <2>$ and $cd <2> ls rm <1>$, their similarity score would be equal to: $1 + (1 + 1) + (1 + 1 + 1) + 0 + 0 = 5$. This similarity measure has a maximum equal to $\frac{l(l+1)}{2}$ that is achieved when the corresponding sequences are identical. The similarity measure has a strong bias towards high similarity values for identical contiguous tokens, while only one different token greatly influences the final result. Finally, the similarity of one sequence X to a set of sequences S is defined to be the maximum similarity attained at any sequence of S . Lane & Brodley [47] note that appropriate values of l are small integers in the range of 8 to 15. In their setting they experimented with $l = 10$.

Lane & Brodley [47] note that their similarity measure can be affected by normal deviations of the users during their routine as well as by random events. For that reason, they apply a smoothing method, in which the similarity of input sequence i to the learned profile is the average of the similarities of the previous $i - w$ sequences to the learned profile. This is defined in Equation 17, where L is the set of records (sequences) of the database storing the user profile. Finally, the actual classification is performed by setting a threshold, where $m_w(i, L)$ values above the threshold are considered as normal and below as intrusions.

$$m_w(i, L) = \frac{1}{w} \sum_{j=i-w}^i Sim(Seq_j, L) \quad (17)$$

For training and evaluation, the Unix commands data was collected from four users within a period of four months, an average of 16544 tokens per user. The sequence length l was set to 10, the window length w was set to 80, while the classification threshold was set to 15. Furthermore, two thirds of the data were used for training and one third for testing. In addition, they experimented with different training set sizes, where the number of records (sequences) of the user profile was set to 50, 200, 500, 1000 and 2000. Anomalous situations were simulated by profiling a user k and testing another one o against k . Performance was measured in terms of the ability of the method to detect

both normal and abnormal conditions, i.e. the detection rate was defined to be the percentage of input sequences correctly categorised as normal or abnormal.

Their results demonstrated that their method has a significantly higher true detection rate than false negatives (intrusions classified as normal behaviour). Furthermore, in some cases the false positives (normal behaviour classified as intrusion) rate is higher than the false negatives, which according to Lane & Brodley [47] is desirable, since false alarms make the system less usable.

Regarding the amount of sequences that should be stored in the database, Lane & Brodley [47] note that this tends to be user specific, i.e. performance increases with a variable rate as the number of records increases depending on the profiled user. This essentially means that certain users are characterised by a small set of actions during their routine work, while others tend to perform a wider variety of activities.

5.3. Summary

This section has provided a description of two main approaches to the field of profiling for intrusion detection. The first one focuses on misuse detection and learns illegitimate patterns of attack by mining association rules from training data, combining them to create temporal events and classifying new actions according to a set of rules learned by a rule learning program.

The second method we have seen focuses on anomaly detection and applies an unsupervised algorithm for building models of legitimate behaviour and a statistical similarity measure to classify an unseen sequence of actions as legitimate or not.

The main difference of these two methods is that the first one requires training data describing the patterns of attack, while the second does not. Furthermore, it is much easier to obtain training data of normal behaviour, rather than obtaining training data of illegitimate actions, since these might not even be complete (network attacks might evolve).

Another interesting point in the field of intrusion detection is that there is a standard framework of evaluation (ROC curve), which however is not based on actual hand-tagged data, but on simulation of normal or abnormal behaviour. Our future work focuses on using the same or similar evaluation to other fields of behavioural profiling.

6. Conclusions

This report has provided a detailed review of the current-state-of-the-art in behavioural profiling. Specifically, behavioural profiling approaches were divided into three categories and for each one of them a number of different methods were presented and discussed. Our critical survey has provided the following important factors that will guide us in the next stages of the project.

- Machine learning method & amount of supervision
- Feature selection & portability of features
- Evaluation framework

Supervised methods do not have the ability to overcome the knowledge acquisition bottleneck that manifests itself, when hand-tagged data are not available to train the corresponding classifiers. Given the nature of the problem, a variety of criminal activities may be performed in numerous ways. Hence, the expectation of training data for all these cases might be not realistic.

On the other hand, unsupervised methods might not be applicable in certain contexts of behavioural profiling, i.e. we always need to have a set of offences in order to build learning models upon them. In the next stages of the project we aim to tackle this problem by focusing on the study and development of weakly-supervised methods that are able to exploit a small quantity of training data, in order to make inferences about cases which do not exist in the training data.

Regarding the selection of features, it was evident from our survey that in the case of burglaries or homicides, the concepts of repeat and near-repeat victimisation can be of a good use within a statistical framework. Despite that, it is unclear whether these or other types of features mentioned in the report are applicable to a wide range of crime types. For that reason, it is necessary to analyse each crime type in consultation with police experts, in order to come up both with a set of features that may be used globally (for all crime types) and also with another feature set that is relevant to a particular crime type.

For example in terrorism, having a feature that measures the experience of a terrorist group, e.g. according to the year of its first attack, could be useful for predicting future attacks as well as their severity. In the same vein, in hooliganism, having a feature indicating whether an offender has or had alcohol problems could also be useful for predicting whether he/she can behave in a similar manner.

Finally, our survey has shown that different in-house and small-scale evaluations have been performed on different types of crime. There are two problems associated with this type of evaluation. Firstly, it is unclear whether the obtained results are applicable to another set of unseen data. Secondly, it is unclear whether the particular experiments can be replicated to allow for method comparison.

Given that the majority of presented methods learn a binary or a multinomial classifier, it is possible to setup an Information Retrieval type of evaluation framework, and create standardised datasets, as in the case of DARPA Intrusion Detection Evaluation Program. That way it is possible to have a clear interpretation of results and a transfer of the knowledge acquired for different crime types.

References

- [1] M. B. Short, M. R. D’Orsogna, G. E. Tita, P. J. Brantingham, A. L. Bertozzi, and L. B. Chayes, “A statistical model of criminal behavior,” *Mathematical Models and Methods in Applied Sciences*, vol. 18, pp. 1249–1267, 2008.
- [2] R. Bache, F. Crestani, D. Canter, and D. Youngs, “A language modelling approach to linking criminal styles with offender characteristics,” *Data & Knowledge Engineering*, vol. 69, no. 3, pp. 303–315, 2010.
- [3] C. Fellbaum, *Wordnet: An Electronic Lexical Database*. Cambridge, Massachusetts, USA: MIT Press, 1998.
- [4] V. Grover, R. Adderley, and M. Bramer, “Review of current crime prediction techniques,” in *Applications and Innovations in Intelligent Systems XIV*, . A. T. R. Ellis, T. Allen, Ed., 2007, pp. 233–247.
- [5] J. Mena, *Investigative Data Mining for Security and Criminal Detection*. Academic Pr Inc, April 2003.
- [6] G. Reavill, *Aftermath, Inc.: Cleaning Up After CSI Goes Home*. Gotham Books, 2007.
- [7] R. Holmes and S. Holmes, *Contemporary perspectives on serial murder*. SAGE, 1998.
- [8] P. Brantingham and P. Brantingham, “Criminality of place, crime generators and crime attractors.” *European Journal on Criminal Policy and Research*, vol. 3, no. 3, pp. 5–26, 1995.
- [9] S. Angel, “Discouraging crime through city planning,” *Berkeley (Cal.), Center for Planning and Development Research*, 1968.
- [10] P. Brantingham and P. Brantingham, “Nodes, paths and edges: considerations on environmental criminology,” *Journal of Environmental Psychology*, vol. 13, pp. 3–28, 1993.
- [11] J. Eck, S. Chainey, J. Cameron, M. Leitner, and R. Wilson, *Mapping Crime: Understanding Hot Spots*. United States National Institute of Justice Special Report, 2005.

-
- [12] G. Farrell and K. Pease, *Biting back: Tackling repeat burglary and car crime*. Crime Prevention Unit, Paper 46, London: Home Office, 1993.
- [13] D. Anderson, S. Chenery, and K. Pease, *Biting back: Tackling repeat burglary and car crime*. Crime Detection and Prevention Series, Paper 58, London: Home Office, 1995.
- [14] S. D. Johnson and A. F. G. Hirschfield, “New insights into the spatial and temporal distribution of repeat victimisation,” *British Journal of Criminology*, vol. 37, pp. 224–241, 1993.
- [15] S. D. Johnson and K. J. Bowers, “The burglary as clue to the future: The beginnings of prospective hot-spotting,” *European Journal of Criminology*, vol. 1, no. 2, pp. 237–255, 2004.
- [16] J. Ashotn, I. Brown, B. Senior, and K. Pease, “Repeat victimisation: Offender accounts,” *International Journal of Risk, Security and Crime Prevention*, vol. 3, pp. 269–279, 1998.
- [17] N. Mantel, “The detection of disease clustering and a generalized regression approach,” *Cancer Research*, vol. 27, pp. 209–220, 1967.
- [18] G. Knox, “Epidemiology of childhood leukaemia in Northumberland and Durham,” *British Journal of Preventative and Social Medicine*, vol. 18, pp. 17–24, 1964.
- [19] P. Brantingham and P. Brantingham, *Environmental Criminology*. Waveland Press, 1991.
- [20] G. F. Rengert, A. R. Piquero, and P. R. Jones, “Distance decay reexamined,” *Criminology*, vol. 37, pp. 427–445, 1999.
- [21] W. Bernasco and F. Luykx, “Effects of attractiveness, opportunity and accessibility to burglars on residential burglary rates of urban neighborhoods,” *Criminology*, vol. 41, pp. 981–1001, 2003.
- [22] W. Bernasco and P. Nieuwebeerta, “How do residential burglars select target areas? a new approach to the analysis of criminal location choice,” *British Journal of Criminology*, vol. 45, pp. 296–315, 2005.

-
- [23] D. Canter, T. Coffey, M. Huntely, and C. Missen, “Predicting serial killers’ home base using a decision support system,” *Journal of Quantitative Criminology*, vol. 16, no. 4, pp. 457–477, 2000.
- [24] D. Canter and P. Larkin, “The enviromental range of serial rapists,” *Journal of Enviromental Psychology*, vol. 13, pp. 63–69, 1993.
- [25] R. Adderley, “The use of data mining techniques in operational crime fighting,” in *Second Symposium on Intelligence and Security Informatics*, 2004, pp. 418–425.
- [26] R. Adderley and P. Mugrove, “Modus operandi modelling of group offending: a data mining case study,” *International Journal of Police Science and Management*, vol. 5, no. 4, pp. 265–276, 2003.
- [27] T. Kohonen, “Self-organized formation of topologically correct feature maps,” pp. 509–521, 1988.
- [28] R. Bache and F. Crestani, “Estimating real-valued characteristics of criminals from their recorded crimes,” in *CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 1385–1386.
- [29] G. K. Zipf, *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932.
- [30] F. Jelinek and R. L. Mercer, “Interpolated estimation of Markov source parameters from sparse data,” in *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, May 1980.
- [31] O. de Vel, A. Anderson, M. Corney, and G. Mohay, “Mining e-mail content for author identification forensics,” *SIGMOD Rec.*, vol. 30, no. 4, pp. 55–64, 2001.
- [32] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.
- [33] T. Joachims, “Text categorization with suport vector machines: Learning with many relevant features,” in *ECML ’98: Proceedings of the 10th European Conference on Machine Learning*. London, UK: Springer-Verlag, 1998, pp. 137–142.
- [34] D. Zhang and W. S. Lee, “Question classification using support vector machines,” in *SIGIR ’03: Proceedings of the 26th annual international ACM SIGIR conference*
-

on *Research and development in information retrieval*. New York, NY, USA: ACM, 2003, pp. 26–32.

- [35] M. Joshi, T. Pedersen, R. Maclin, and S. Pakhomov, “Kernel methods for word sense disambiguation and acronym expansion,” in *AAAI’06: Proceedings of the 21st National Conference on Artificial Intelligence*. AAAI Press, 2006, pp. 1879–1880.
- [36] D. Zelenko, C. Aone, and A. Richardella, “Kernel methods for relation extraction,” *J. Mach. Learn. Res.*, vol. 3, pp. 1083–1106, 2003.
- [37] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, “Automatically profiling the author of an anonymous text,” *Commun. ACM*, vol. 52, no. 2, pp. 119–123, 2009.
- [38] M. A. K. Halliday and C. M. I. M. Matthiessen, *An introduction to functional grammar / M.A.K. Halliday*, 3rd ed. Hodder Arnold, London :, 2004.
- [39] G. Wang, H. Chen, and H. Atabakhsh, “Automatically detecting deceptive criminal identities,” *Commun. ACM*, vol. 47, no. 3, pp. 70–76, 2004.
- [40] H. Chen, W. Chung, J. Jie Xu, G. Wang, Y. Qin, and M. Chau, “Crime data mining: A general framework and some examples,” *Computer*, vol. 37, pp. 50–56, 2004.
- [41] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
- [42] H.-P. Kriegel, P. Kröger, and A. Zimek, “Outlier detection techniques (tutorial),” in *13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009)*, 2010.
- [43] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.
- [44] W. Lee, J. S. Salvatore, and K. W. Mok, “A data mining framework for building intrusion detection models,” *Security and Privacy, IEEE Symposium on*, vol. 0, p. 0120, 1999.
- [45] W. W. Cohen, “Fast effective rule induction,” in *In Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 115–123.

- [46] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.
- [47] T. Lane and C. E. Brodley, “An application of machine learning to anomaly detection,” in *In Proceedings of the 20th National Information Systems Security Conference*, 1997, pp. 366–380.

Document Updates

Table 11: Document Updates

Version	Date	Updates and Revision History	Author
20100615	15/06/2010	Introduction	Ioannis Klapaftis
20100618	18/06/2010	Geographical profiling	Ioannis Klapaftis
20100622	22/06/2010	Offenders characteristics profiling	Ioannis Klapaftis
20100625	25/06/2010	Intrusion Detection	Ioannis Klapaftis
20100629	29/06/2010	Conclusion	Ioannis Klapaftis
20100630	30/06/2010	Spelling correction	Ioannis Klapaftis
20100630	23/07/2010	Editorial corrections	Ioannis Klapaftis