Information & Computer Security

Language-independent gender identification through keystroke analysis

Ioannis Tsimperidis Vasilios Katos Nathan Clarke

Received 23 May 2014
Revised 29 July 2014
12 September 2014
8 October 2014
3 November 2014
Accepted 3 November 2014

Language-independent gender identification through keystroke analysis

Ioannis Tsimperidis and Vasilios Katos
Information Security and Incident Response Unit, Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, and
Nathan Clarke
Centre for Security, Communications & Network Research, Plymouth University, UK

Abstract

Purpose – The purpose of this paper is to investigate the feasibility of identifying the gender of an author by measuring the keystroke duration when typing a message.

Design/methodology/approach – Three classifiers were constructed and tested. The authors empirically evaluated the effectiveness of the classifiers by using empirical data. The authors used primary data as well as a publicly available dataset containing keystrokes from a different language to validate the language independence assumption.

Findings – The results of this paper indicate that it is possible to identify the gender of an author by analyzing keystroke durations with a probability of success in the region of 70 per cent.

Research limitations/implications – The proposed approach was validated with a limited number of participants and languages, yet the statistical tests show the significance of the results. However, this approach will be further tested with other languages.

Practical implications – Having the ability to identify the gender of an author of a certain piece of text has value in digital forensics, as the proposed method will be a source of circumstantial evidence for "putting fingers on keyboard" and for arbitrating cases where the true origin of a message needs to be identified.

Social implications – If the proposed method is included as part of a text-composing system (such as e-mail, and instant messaging applications), it could increase trust toward the applications that use it and may also work as a deterrent for crimes involving forgery.

Originality/value – The proposed approach combines and adapts techniques from the domains of biometric authentication and data classification.

Keywords SVM, Bayesian, Gender recognition, Keystroke dynamics, Manhattan classifier

Paper type Research paper

1. Introduction and motivation

When investigating a computer crime, one of the core concepts and main challenges in digital forensics is to "put fingers on keyboard", that is to identify the human who is responsible for the generation or handling of digital data related to the respective criminal or offensive activity in general (Shavers, 2013). The increased complexity and large scope of the underlying problem space make circumstantial digital evidence critical when attempting to identify a user.

Information & Computer Security
Vol. 23 No. 3, 2015
pp. 286-301

2056-4961
DOI 10.1108/ICS-05-2014-0032

Textual communication between users dominates most other forms such as audio, video or image. For every second, there are about 2.4 million e-mails sent (internetlivestats.com). A typical incident relating to textual communication is a masquerade attack, where a malicious user assumes the identity of another legitimate user. To date, e-mail spoofing detection is mainly focused to the examination of the e-mail metadata found on the headers and its correlation with network information, such as IP addresses, geo-location and so forth. Focus is not given to verifying the individual. However, there exists a wealth of research literature on biometric-based authentication of users based on the way a user types (Clarke *et al.*, 2003, Clarke and Furnell, 2007), yet current e-mail applications do not record the biometric characteristics of the author; therefore, biometric-based identification techniques are not applicable in such a context. Hence, to adopt biometric authentication approaches in identification of user characteristics, the authoring applications need to maintain biometric data recording capabilities. Adding such abilities to authoring applications would offer a number of benefits and protection to the users. Consider, for example, a system that can identify the gender of the author. Masquerading is a popular behavior in Internet applications, such as social networks, associated with various offenses (phishing, pedophilia, etc.). A system that would have the ability to warn or inform the unsuspected recipient of a message about the gender of the sender of that message, would act as a deterrent toward offenses leveraging gender misrepresentation, it would increase the protection of the legitimate users and, of course, support investigations in case of a reported incident.

In this paper, we propose the use of keystroke analysis to determine the gender of a user. We investigate the validation of such hypothesis to develop a business case for creating applications that include functionality for allowing gender identification of the creator of the text, while this text is being typed. We envisage that the proposed method will provide a forensic analyst with enough circumstantial evidence to support their investigation and perform e-discovery through informed decision-making.

More specifically, the proposed system is based on a scenario where user authentication has failed (a user account has been hijacked, for example), and there is no biometric database available to identify the attacker by matching the attacker's typing behavior against a set of user biometric profiles. Having limited information and a high level of uncertainty, the identification of the characteristics of the attacker, such as gender in our case, is expected to have an amount of false prediction rate. In addition, privacy requirements prohibit the logging and capturing of the type text. As such, we argue that the monitoring of keystroke durations rather than the actual content would be a privacy tradeoff a user is willing to accept. That is, a legitimate user may be willing to accept some degradation in privacy (which may already be the case for users subscribing to biometric authentication based on keystroke duration) to support an investigation in case of an authentication breach by allowing the creation of circumstantial evidence, which, when combined with other evidence, will eventually allow the identification of the attacker. Such circumstantial evidence, for example, could be fused with other methods, such as those described by Schler *et al.* (2006) and Rangel *et al.* (2013), who focus on the identification of both gender and age of an author by analyzing the content of a message.

The rest of the paper is organized as follows. In Section 2, the related work on gender recognition within written text is examined. The experimental design and methodology

are presented in Section 3, with the accompanying results documented in Section 4. The paper concludes with Section 5, presenting the key findings and future work.

2. Related work

The question of indentifying the gender of a writer intrigued researchers, as texts were only in the handwritten form. Trudgill (1972) highlights the differences in linguistic styles between male and female authors of a handwritten text. Similarly, there are plenty works studying the differences that appear between males and females in speech and informal writing, like the work of Holmes (1988), for instance.

Gender identification is an interesting topic with a number of useful applications, such as better translation between languages, as some languages use different grammatical structures depending on the gender, the targeted advertising for Internet users and, of course, forensic analysis.

The research community has done a significant amount of work on gender and user identification in general, by proposing a number of techniques with varying prediction success rates. In his work, Lai (2006) constructed datasets from books and blogs, and proposed a Naïve Bayes classifier on a selection of words achieving a success rate between 64 and 91 per cent, depending on the particular set of features.

Vel *et al.* (2002) acknowledge the proliferation of e-mail as a modern means of communication, combining old and new characteristics of written speech; the new characteristics refer to the abbreviations – frequently found in text chatting – as well as intentionally misspelled words used by the author of a message to convey emotional information. They collected e-mails from an academic organization that included approximately 15,000 members and trained a system using a large number of features, such as the number of characters, lines, paragraphs, alphabetic characters, upper-case characters, spaces, words ending with "able", words ending with "al", "sorry" words and so forth. They achieved a success rate, depending on the size of e-mail of between 56 and 71 per cent.

More recently, in a similar manner, Cheng *et al.* (2011) used a dataset from e-mails and journalists' articles and utilized three different classifiers to achieve a success rate between 55 and 85 per cent. The features they selected included; the number of lines, paragraphs, sentences, some special characters, like "%", "&", the number of punctuations, articles, pronouns, auxiliary-verbs and so forth as well as the presence of some particular words. Schler *et al.* (2006) used tens of thousands of blogs as dataset, and, based on the appearance frequency of a selection of words, they categorized a user according the gender and age. Rangel *et al.* (2013) present the findings of the Author Profiling Task, as announced at the PAN 2013 workshop for authorship identification. This work contains a consolidated list of the literature as seen by the 21 candidates who tried to classify an author of an English or Spanish text according to gender and age. Argamon *et al.* (2003) used documents from the British National Corpus and checked the frequency of appearance of pronouns, common nouns, proper nouns as well as the frequency of appearance of some particular words, both in fiction and non-fiction texts. Using a Bayesian technique, they achieved a success rate of 80 per cent. Based on this research, a tool named "Gender Genie" was developed (Gender Genie, 2014). The tool allocates a number of points either on a "male" or on a "female" variable, every time a particular word appears. The highest variable of the two indicates the gender of the author. Another tool is the "Gender Guesser" Krawetz (2006), which is based on "Gender Genie" but is more advanced in the sense that it can process informal writing often found

in blogs and casual messages.). However, to perform with an improved prediction rate, the tool requires a minimum of 300 words. Similarly, another tool is "uClassify" Kagstrom *et al.* (2014), which has been trained on 11,000 blogs, half of which were written by male and the other half by female authors.

A common characteristic of all research published to date is that they are language dependent. More specifically, they require that the language is specified – English in particular. Consequently, other languages, such as Spanish, Chinese, Greek, German and so forth, would require additional training with respective datasets. This requirement results in a significant limitation (Doyle and Keselj, 2005). Despite the acknowledgement of the language dependency limitation, the solution proposed by Doyle and Keselj is language dependent, but it is relatively simple to adapt it for other languages. In their work, 500 student essays were used and the appearances of n-grams (a character, a digram, a trigram, etc.) for both male and female authors were obtained offering a success rate between 51 and 81 per cent. Giot and Rosenberger (2012) claim to have introduced the first language-independent gender recognition system. This was achieved using keystroke duration, digram latency and a vector which is the concatenation of the four previous timing values. They achieved a success rate of 90 per cent and used it to improve user authentication. The main limitation of the work by Giot and Rosenberger (2012) is the focus on the protection of passwords. That is, the users were profiled based on their typing of two specific words (Giot *et al.*, 2009).

Summarizing, almost the entire work on gender recognition of an author is based on *what* a user types. The underlying models depend on appearance frequency of particular words, n-grams, punctuations and so forth, making these approaches language dependent. To our knowledge, our proposed work is the first study focusing on *how* users type a complete passage rather than a few words such as a password. The method is language-independent because it is not subject to particular words of a specific language, but to time parameters related to the way a user types.

Our proposed work aims at providing continuous identification throughout a typing session. In addition, provided that our approach involves the collection of a larger volume of user data, we could reduce the number of required attributes, improving the efficiency of the gender identification system.

3. Keystroke-based gender experiment

3.1 Experimental methodology

In this work, we attempt to perform gender recognition of the author of a text using a limited number of attributes and, more particularly, the parameters of keystroke dynamics. This suggests that the proposed recognition system is language-independent because the results are based on the way the user types rather than what they type. Furthermore, to have a system with increased user acceptance and reduced non-compliance risks due to possible privacy regulations, the keystroke features utilized are limited to the keystroke duration. Nevertheless, maintaining a dataset of keystroke duration information will allow the construction of other related attributes such as digram – or n-gram, in general – based latencies, thus supporting further investigation and identification of the attribute of the highest accuracy.

Unfortunately, to date, there seems to be a limited number of public datasets of keystroke dynamic, therefore it was necessary to design a data collection methodology and implement a data collection tool. A key logging application was developed in Visual

Basic. The application upon recording the user with a suitable unique identifier required the user to type a fixed text of 850 characters containing letters, digits and other symbols. Upon completion of the typing exercise, a comma-separated file was created named after the subject's username, with each line containing the character pressed, the keydown and keyup time in milliseconds. While it is possible to capture keystroke characteristics with a greater resolution, prior research within this domain when applied to authentication has typically focused upon milliseconds (Joyce and Gupta, 1990). Each typing session is captured in a single file. That is, if a user, for example, participates in two different sessions (say by running the experiment twice – one on a laptop and one on a desktop), two files will be created. The times recorded are measured as the time elapsed from the execution of the keylogger application.

An initial team of 24 volunteers used the keylogger application both on desktops and laptops. All volunteers typed the same text twice – one on a laptop and one on a desktop – producing two files to cover a wide variety of typing cases. The recording period was undertaken from October 11, 2012 until November 21, 2012 and from July 13, 2014 until July 15, 2014. While the number of participants is not large, careful thought was given to their selection, to ensure an appropriate representation. This involved controlling for gender and left-right handedness. The volunteers' characteristics with respect to the general population representation, are shown in Table I. The number of

		Group 1		
User id	Gender	Age	Handedness	Education level
a1	Male	38	Left-handed	High School
a2	Male	19	Right-handed	High School
a3	Male	39	Right-handed	University
a4	Female	30	Right-handed	University
a5	Male	54	Right-handed	T.E.I.*
a6	Female	26	Right-handed	High School
a7	Male	39	Right-handed	T.E.I.
a8	Male	18	Right-handed	High School
a9	Female	18	Right-handed	High School
a10	Female	20	Left-handed	High School
a11	Male	31	Right-handed	High School
a12	Female	26	Right-handed	T.E.I.
a13	Female	30	Right-handed	T.E.I.
a14	Female	20	Right-handed	High School
a15	Male	37	Right-handed	University
a16	Female	26	Right-handed	University
a17	Male	37	Right-handed	University
a18	Female	39	Right-handed	University
a19	Male	28	Right-handed	University
a20	Female	22	Right-handed	High School
a21	Male	35	Left-handed	University
a22	Female	31	Right-handed	High School
a23	Male	24	Right-handed	T.E.I.
a24	Female	25	Right-handed	High School

Table I.
User profiles

Note: *Technological Educational Institution

participants who were male was equal to the number of female volunteers. The proportion of left-handed volunteers was about 12.5 per cent, closely reflecting the proportion of the whole population. The educational level of the participants corresponds to the ratio of the level of education of a population with a Greek nationality.

3.2 Descriptive statistics

After the recording period, 48 files, 24 corresponding to male authors and 24 to female authors, were available for training the system. From these files, the keystroke durations of every key that was pressed were extracted and all the records from "male" files were merged to produce a single dataset; the same procedure was followed for the "female" files – creating the two classes of interest.

The dataset was sanitized by removing all outliers that corresponded to values exceeding three times the mean value, a standard methodological approach utilized in keystroke analysis studies (Clarke and Furnell, 2007). Subsequently, the statistical features for each character were calculated to find if there are any differences in the way that type the members of the two classes. Although the sample was not too big, some important differences appeared (Table II).

It should be noted that significant differences are not limited to the characters of Table II, but are encountered almost in half of the tested character set. Some characters display significant differences in their mean value, others in the standard deviation and some in both. A general observation of these preliminary findings is that the males hold their fingers a little bit longer on the keys than the females, while the females are not as consistent as males are, due to the slightly higher standard deviations.

The normality and subsequent *t*-test results between the means of the two classes are shown in Table III. As it can be seen from the Shapiro–Wilk normality test, both classes are normal, permitting us to run a paired *t*-test to establish whether the means of the two classes are different. The resulting probability was equal to 0.0001 indicating that the means of the two classes are significantly different and, therefore, the key latency attribute can be used to distinguish the gender of the author.

Table II. Differences in statistical features

Character	Key code	Appearances per user (mean)	"Female" file Mean (ms)	"Female" file SD (ms)	"Male" file Mean (ms)	"Male" file SD (ms)	Percentage difference (%)	
(Space)	32	207	97.62	32.83	92.60	28.02	−5.15	−14.64
Gamma	41	22	77.78	25.57	92.28	21.99	18.63	−13.99
Omicron	79	90	82.02	26.62	92.63	24.81	12.93	−6.79
Pi	80	49	78.37	28.08	96.09	27.94	22.60	−0.53
Upsilon	89	61	80.88	23.68	90.03	26.71	11.31	12.83

Table III. Normality and *t*-test results

	Mean (ms)	SD (ms)	Shapiro-Wilk normality test W [critical 5%]	Shapiro-Wilk normality test *p*-value	paired *t*-test [*p*]
Male	97.2371	5.8802	0.917313 [0.897]	0.115854	
Female	85.4751	5.935	0.941130 [0.897]	0.302940	[0.0001]

3.3 Classifier design and evaluation

In this work, we constructed and evaluated three classifiers, namely, a Naïve Bayes, and two classifiers based on the Manhattan and Euclidean distance, respectively. This is in agreement with the current relevant literature, where a considerable volume of research displayed preference to these classifiers and corresponding methods in a variety of different problem domains – see for example the work by Phyu (2009), Khamar (2013), Kotsiantis (2007). In addition, the rationale for selecting these specific three candidate classifiers was the significant differences in the statistical descriptors as well as the distribution types of the keystroke duration histograms, thus making these classifiers appropriate for the current work More specifically, the Naïve Bayes classifier was a suitable classifier as our attribute follows a normal distribution (Bouckaert, 2005). The significant differences of the standard deviations between the male and female keystrokes warrant the use and evaluation of the Manhattan distance (Li *et al.*, 2005). Finally, the statistically significant differences between the two means are explored and utilized through the Euclidean distance.

Firstly, we examine the performance of a Naïve Bayes classifier. A histogram of keystroke duration appearances of a representative set of keys is shown in Figure 1.

As the keystroke duration attribute follows a normal distribution, the following probability can be attached to every observation:

$$p(x|g) = \frac{1}{\sqrt{2\pi\sigma_{x,g}^2}} e^{-\frac{(x - \mu_{x,g})^2}{2\sigma_{x,g}^2}} \tag{1}$$

where, $p(x|g)$ is the probability of "x" character being typed by user of gender "g", $\mu_{x,g}$ and $\sigma_{x,g}$ is the mean value and the standard deviation of the data, respectively, for the "x" character and for the "g" gender.

Upon calculating the probability for each character, they are binned into the two classes as follows:

$$Male\_Accumulator = p(\text{'A'}|male) + p(\text{'B'}|male) + \ldots$$
$$Female\_Accumulator = p(\text{'A'}|female) + p(\text{'B'}|female) + \ldots$$

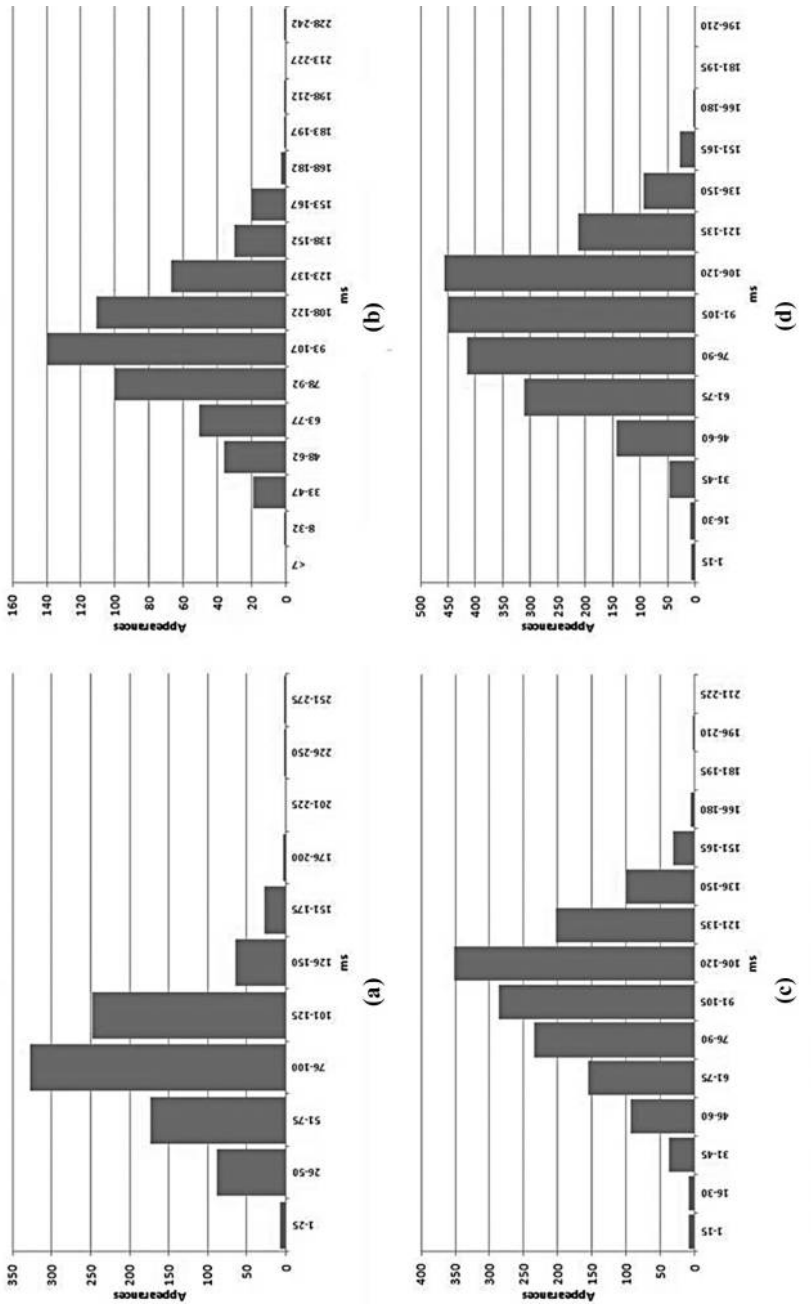
The final probabilities are then obtained from:

$$p(male) = \frac{Male\_Accumulator}{Male\_Accumulator + Female\_Accumulator}$$

$$p(female) = \frac{Female\_Accumulator}{Male\_Accumulator + Female\_Accumulator}$$

It can be seen that $p(male) = 1\text{-}p(female)$. The higher of the two is the final gender guess. The success rate of the proposed system was 61.76 per cent.

As mentioned earlier, the observation that most of the characters exhibit significant differences in their standard deviations between "male" and "female" classes led us to evaluating the performance of the Manhattan distance. The Manhattan distance between the calculated standard deviation from that of "male" and "female" classes indicates the gender of the author of the text. In this experiment, the correct predictions



Notes: (a) Backspace; (b) Enter; (c) key "A"; (d) key "E"

Figure 1.
The distribution of keystroke durations

reached a success rate of 64.71 per cent, but the results showed a bias toward the "male" class.

The third approach was the exploitation of the differences between the means. The mean values of keystroke durations of specific characters were calculated and these values were compared with the respective values of "male" and "female" classes. The shorter distance indicates the author's gender. Once again, the test was performed over all available texts, and, for some characters, the success rate was 67.65 per cent. However, one important drawback of this approach is the dependence upon a single character, thus reducing the reliability of its performance.

To remedy this limitation, consideration was given to the average values for a selection of characters separately rather than an aggregate value for all characters. After this, the Euclidean distances from the values of model are calculated and the shorter of these indicates the author's gender. More analytically, the following comparison was performed:

$$\sqrt{\sum_{x}\left(\frac{tot\_dur\_x}{count\_x} - \mu_{x,f}\right)^2} < \sqrt{\sum_{x}\left(\frac{tot\_dur\_x}{count\_x} - \mu_{x,m}\right)^2} \tag{2}$$

where:

tot_dur_x is the sum of char "x" keystroke durations;

count_x is number of times char "x" was typed; and

$\mu_{x,f}$ and $\mu_{x,m}$ are the mean values of char "x"s keystroke duration from the data that were created by females and males volunteers, respectively.

The success rate achieved was 64.71 per cent, and the system is not dependent on a single character, addressing the aforementioned limitation.

After a thorough study of the above metrics and the corresponding results, a scoring system based on the Manhattan distance was introduced. For each different character, the following comparison is performed:

$$\left|\frac{tot\_dur\_x}{count\_x} - \mu_{x,f}\right| < \left|\frac{tot\_dur\_x}{count\_x} - \mu_{x,m}\right| \tag{3}$$

where, *tot_dur_x, count_x*, $\mu_{x,f}$ and $\mu_{x,m}$ have the same meaning as the terms in expression (2).

The outcome of (3) dictates how the weighted scoring system is to be applied. If (3) holds, the points will be assigned to the probability denoting that the text belongs to a female author. In the opposite case, the points are assigned to the probability of the male author. The number of assigned points depends on the appearance frequency of the character "*x*" in the text and on the difference between the values $\mu_{x,f}$ and $\mu_{x,m}$ of the model.

The classifiers proposed in this work were also compared against SVM and decision tree classifiers. In an SVM classifier, each volunteer file is represented as a data point in an *n*-dimensional space, where *n* denotes the number of attributes. In our case, the attributes are the keystroke duration mean values. The objective of the classifier is to

find a suitable hyperplane for splitting this n-dimensional hyperspace in such a way that male points are separated from female points.

Any hyperplane can be written as the set of points x satisfying:

$$w \cdot x - b = 0 \tag{4}$$

where:

w is a normal vector to hyperplane;

b is a real number; and

$\cdot$ denotes the dot product.

By mapping the two classes to values $+1$ and -1, we can select two hyperplanes that separate the data points without having any other point in between. The hyperplanes will have the following equations:

$$w \cdot x - b = 1 \text{ and } w \cdot x - b = -1 \tag{5}$$

The distance of these two hyperplanes would then be $2/\|w\|$ and, therefore, we seek to minimize $\|w\|$, the norm of w, which can be replaced by $\frac{1}{2}\|w\|^2$. This is a typical example of quadratic programming optimization problem, expressed as:

$$\arg\min_{(w,b)} \frac{1}{2}\|w\|^2 \tag{6}$$

With the constraint:

$$y_i(w \cdot x_i - b) \geq 1 \tag{7}$$

where:

y_i is the value for the class (-1 or 1), for every i = 1, 2, …, n.

By introducing the Lagrange multipliers in the above constraint problem, we seek to obtain:

$$\arg\min_{(w,b)} \max_{\alpha \geq 0} \left\{ \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1] \right\} \tag{8}$$

However, in practice it is not expected to fully separate all male from all female points. In such cases, we seek to identify the hyperplane that achieves the best separation of the two classes. This is achieved by using the soft margin approach that introduces non-negative slack variables, ξ_i, which measure the degree of misclassification of the data x_i. As such, the optimization problem formally expressed by (6) with the constraints of (7) and the Lagrange multipliers (8), is formally written as:

$$\arg\min_{(w,\xi,b)} \max_{\alpha,\beta} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1 + \xi_i] - \sum_{i=1}^{n} \beta_i \xi_i \right\} \tag{9}$$

where:

$\alpha_i, \beta_i \geq 0$.

C a constant set during the training phase.

4. Validation and results

The scoring system considered expression (3) above and included weights for every character, depending on the appearance frequency of the character in the text and its difference between the two classes (male and female). That is, a frequently appearing character with a significant statistical difference would be given the maximum weight (five points in our system), as opposed to a character appearing for a limited number of times and having low differences between the classes (1 point). Group 1 involves the case of the volunteers that were used to both train and test the system. This group included the volunteers described in Table I.

The overall success rate of the proposed classifier for the dataset produced by Group 1 users was equal to 81.25 per cent. More analytically, the success rate for the "female" files was equal to 83.33 per cent, whereas for the male files was equal to 79.17 per cent.

The validation of the constructed classifier was performed with two additional datasets. First, we invited a second group of volunteers. The usernames, gender and age of each member of the second group, as well as the device type that they were recorded on, are shown in Table IV.

This group was asked to type the same control text as the first group, producing a total of 38 files, as two of the volunteers did not provide laptop typing data. The files

Table IV. Second group of volunteers

Group 2			
User id	Gender	Age	Recorded on
b1	Female	27	Desktop
b2	Female	31	Desktop and Laptop
b3	Male	20	Desktop and Laptop
b4	Male	25	Desktop and Laptop
b5	Female	24	Desktop and Laptop
b6	Male	20	Desktop and Laptop
b7	Female	29	Desktop and Laptop
b8	Male	38	Desktop and Laptop
b9	Male	30	Desktop
b10	Female	26	Desktop and Laptop
b11	Female	44	Desktop and Laptop
b12	Male	20	Desktop and Laptop
b13	Male	19	Desktop and Laptop
b14	Male	18	Desktop and Laptop
b15	Male	26	Desktop and Laptop
b16	Male	29	Desktop and Laptop
b17	Male	41	Desktop and Laptop
b18	Female	25	Desktop and Laptop
b19	Female	25	Desktop and Laptop
b20	Female	44	Desktop and Laptop

produced by Group 2 were used to test the system. In this case, the classifier produced a success rate for "female" and "male" files equal to 76.47 per cent, and 71.43 per cent, respectively, with an overall success rate of 73.68 per cent.

The performance of the classifier is similar to that of the first group. That is, the difference in percentage of correct prediction is less than 8 per cent with better results attributed to the "female" class.

A third dataset was used to validate the language independence claim. This is a dataset produced by Bello *et al.* (2010). The dataset contains the keystroke logging of some particular sentences in Spanish and some particular UNIX commands from a team of 40 male and 15 female users, thus producing 55 files in total, which were used to test the system.

After running the proposed classifier against Groups 2 and 3, it was established that the rates for both datasets were slightly higher than 69 per cent, and similar to the first group the classification success rate of the "female" class was equal to 73.33 per cent, which was higher than the "male" class (67.50 per cent), albeit the small number of characters included in the weighted calculation. As such, we conjecture that the more characters used, the better accuracy and reliability of the proposed method.

By observing the classification rates of Groups 1, 2 and 3, it can be noted that the system predicts the gender three out of four users, correctly. This success rate is comparable with those of similar works, see Doyle and Keselj (2005), Argamon *et al.* (2009), Cheng *et al.* (2011). The advantage of the proposed method is that it monitors only the keystroke duration and, hence, the user's privacy is respected.

The classification rates of the Bayesian and Euclidean distance classifiers were 57.69 and 65.38 per cent for Group 2 files, respectively, 58.18 and 61.81 per cent for Group 3 files, respectively, and the aggregate classification rates were 59.13 and 63.48 per cent, respectively. From these results, we conclude that the classifier based on the Manhattan distance yielded the highest classification rate.

The proposed system was further evaluated and compared against an SVM and a decision tree classifier. For doing this, the SVM and C4.5 implementations of the WEKA machine learning software were used. Both systems had the same training set as the classifiers discussed earlier, and the default parameter values were used.

The success rates for the SVM classifier were 72.92 per cent for Group 1, 68.42 per cent for Group 2 and 43.64 per cent for Group 3. From the results, it can be seen that the SVM classifier has slightly worse rates than the Manhattan distance-based classifier for the first two groups, but its performance for the Spanish users is substantially low. As the rate is below 50 per cent, we could construct an oracle with a 100-43.64 = 56.36 success rate by setting it to output the complement answer. However, even in this case, the success rate is low.

With regards to the C4.5 decision tree classifier, the success rates were 93.75 per cent for Group 1, 60.53 per cent for Group 2 and 54.55 per cent for Group 3. Unsurprisingly, the decision tree displayed a very high success rate for the training set, while the results for Groups 2 and 3 were poor. This is due to the linear nature of the decision tree as well as the fact that it only selects those attributes from the training set that are capable of separating the classes of that set, while ignoring all other attributes.

Table V summarizes the results for Manhattan distance based, SVM and C4.5 decision tree classifier for Group 1. Likewise, Table VI summarizes the respective results for Group 2 and Table VII the results of the Spanish group (Group 3).

Table VIII shows the results by aggregating all files from all datasets for Manhattan distance based, SVM and C4.5 decision tree classifiers.

Table V.
Results from the first team of volunteers

	Group 1 (control group)								
	Females			Males			Total		
Classifier	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)
Manhattan distance based	20	4	83.33	19	5	79.17	39	9	81.25
SVM	17	7	70.83	18	6	75.00	35	13	72.92
C4.5 decision tree	23	1	95.83	22	2	91.67	45	3	93.75

Table VI.
Second group results

	Group 2 (same fixed text)								
	Females			Males			Total		
Classifier	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)
Manhattan distance based	13	4	76.47	15	6	71.43	28	10	73.68
SVM	13	4	76.47	13	8	61.90	26	12	68.42
C4.5 decision tree	7	10	41.18	16	5	76.19	23	15	60.53

Table VII.
Third group results

	Group 3 (Spanish users)								
	Females			Males			Total		
Classifier	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)
Manhattan distance based	11	4	73.33	27	13	67.50	38	17	69.09
SVM	8	7	53.33	16	24	40.00	24	31	43.64
C4.5 decision tree	4	11	26.67	26	14	65.00	30	25	54.55

Table VIII.
Overall results

	All of files and Users								
	Females			Males			Total		
Classifier	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)	Successes	Failures	Success rate (%)
Manhattan distance based	44	12	78.57	61	24	71.76	105	36	74.47
SVM	38	18	67.86	47	38	55.29	85	56	60.28
C4.5 decision tree	34	22	60.71	64	21	75.29	98	43	69.50

Overall, SVM and C4.5 decision tree tests yield similar results but their success rates are less than the rate of the Manhattan distance-based classifier. All results are summarized in Table IX.

Finally, to assess the robustness of the proposed method, we split each file from the second group of volunteers into two almost equal segments. We then applied the proposed method independently to each file. We define robustness as the ability of the system to consistently and repeatedly classify a user to the same gender, given any typing subset of that user. It can be seen from the results that the proposed method is consistent for 32 out of 38 files.

The same process was followed for the Spanish dataset. From the total amount of 55 files, those that were smaller than 15 K were excluded to maintain a workable amount of keystrokes, leaving a pool of 48 files. The proposed method was consistent for 38 out of 48 files.

Finally, tests were performed to establish whether statistically significant correlations exist between the success rates and the other attributes such as age and typing medium (laptop or desktop). As the dependent variable (prediction) is binary, we ran a Logit and Probit estimation on the data. None of the tests returned statistically significant results. That is, no valid conclusions can be drawn with regards to age and prediction accuracy from the datasets.

5. Conclusions and future work

Leveraging keystroke dynamics research to construct user profiles in the context of a digital investigation is a promising area of research and a domain with practical importance to electronic discovery. In this paper, we focused on the feasibility of identifying whether a user typed a certain text is male or female, but a complete solution would need to consist of independent tests corresponding to a variety of characteristics or properties. Besides user authentication, keystroke dynamics may be useful to detect the emotional state of the user, or to identify his handedness, or to assess whether the user is typing in their native language or not.

Due to the preliminary yet promising results, the model will be extended to consider other user characteristics or properties to form a concise and concrete solution. The complete approach which is part of our ongoing research involves the identification of correlation of the user properties through latent variables to establish the mutual information between them and the construction of a formal evidence handling framework based on known evidence fusion constructs such as the Dempster-Shafer theory of evidence.

A limitation of the current research was the use of a fixed text to create the reference model, which departs from the realistic behavior of users. An improvement would be to use an agent that logs the user in a real working environment and we conjecture that this would

Classifier	Group 1 Success rate (%)	Group 2 Success rate (%)	Group 3 Success rate (%)	Aggregate Success rate (%)
Manhattan distance based	81.25	73.68	69.09	74.47
C4.5 decision tree	93.75	60.53	54.55	69.50
Euclidean distance based	64.71	65.38	61.81	63.48
SVM	72.92	68.42	43.64	60.28
Bayesian	61.76	57.69	58.18	59.13

Table IX.
Results comparison

increase the success rates. Another parameter increasing the prediction accuracy is a larger user sample. Finally, an improvement that will significantly raise the reliability of the system would be to create feedback mechanism, enhancing in this way the database that generated the equations for the possibilities export and the weights of each character.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A.R. (2003), "Gender, genre and writing style in formal written texts", *Text – Interdisciplinary Journal for the Study of Discourse*, Vol. 23 No. 3, pp. 321-346.

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2009), "Automatically profiling the author of an anonymous text", *Communications of the ACM*, Vol. 52 No. 2, pp. 119-123.

Bello, L., Bertacchini, M., Benitez, C., Pizzoni, C.J. and Cipriano, M. (2010), "Collection and publication of a fixed text keystroke dynamics dataset", *XVI Congreso Argentino de Ciencias de la Computacio*, *Buenos Aires*, pp. 822-831.

Bouckaert, R. (2005), "Naive bayes classifiers that perform well with continuous variables", in Webb, G.I. and Yu, X. (Eds), *AI 2004: Advances in Artificial Intelligence – Lecture Notes in Computer Science, Volume 3339*, Springer, Berlin Heidelberg, pp. 1089-1094.

Cheng, N., Chandramouli, R. and Subbalakshmi, K.P. (2011), "Author gender identification from text", *Digital Investigation*, Vol. 8 No. 1, pp. 78-88.

Clarke, N. and Furnell, S. (2007), "Authenticating mobile phone users using keystroke analysis", *International Journal of Information Security*, Vol. 6 No. 1, pp. 1-14.

Clarke, N., Furnell, S., Lines, B. and Reynolds, P. (2003), "Keystroke dynamics on a mobile handset: a feasibility study", *Information Management & Computer Security*, Vol. 11 No. 4, pp. 161-166.

Doyle, J. and Keselj, V. (2005), "Automatic categorization of author gender via n-gram analysis", *6th Symposium on Natural Language Processing*, *Chiang Rai*.

Gender Genie (2014), available at: www.hackerfactor.com/GenderGuesser.php (accessed 23 May 2014).

Giot, R., El-Abed, M. and Rosenberger, C. (2009), "GREYC keystroke: a benchmark for keystroke dynamics biometrics systems", *3rd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS '09*, IEEE, pp. 1-6.

Giot, R. and Rosenberger, C. (2012), "A new soft biometric approach for keystroke dynamics based on gender recognition", *International Journal of Information Technology and Management*, Vol. 11 Nos 1/2, pp. 35-49.

Holmes, J. (1988), "Paying compliments: a sex-preferential positive politeness strategy", *Journal of Pragmatics*, Vol. 12 No. 3, pp. 445-465.

Joyce, R. and Gupta, G. (1990), "Identity authentication based on keystroke latencies", *Communications of the ACM*, Vol. 33 No. 2, pp. 168-176.

Kagstrom, A., Karlsson, A. and Kagstrom, E. (2014), "uClassify – GenderAnalyzer_v5", available at: www.uclassify.com/browse/uClassify/GenderAnalyzer_v5 (accessed 20 August 2013).

Khamar, K. (2013), "Short text classification using kNN based on distance function", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2 No. 4, pp. 1916-1919.

Kotsiantis, S. (2007), "Supervised machine learning: a review of classification techniques", *Informatica*, Vol. 31 No. 1, pp. 249-268.

Krawetz, N. (2002), "Hacker factor: gender guesser", available at: www.hackerfactor.com/GenderGuesser.php (accessed 18 August 2013).

Lai, C. (2006), "Author gender analysis: I256 applied natural language processing", available at: http://courses.ischool.berkeley.edu/i256/f09/Final%20Projects%20write-ups/Lai Chaoyue_project_final.pdf

Li, W., Wang, K., Stolfo, S. and Herzog, B. (2005), "Fileprints: identifying file types by N-gram analysis", *Sixth Annual IEEE SMC, Information Assurance Workshop, IAW '05, IEEE*, pp. 64-71.

Phyu, T.N. (2009), "Survey of classification techniques in data mining", *International Multi Conference of Engineers and Computer Scientists 2009 Vol. I, IMECS*, Hong Kong, pp. 727-731.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E. and Inches, G. (2013), "Overview of the author profiling task at PAN 2013", *PAN at CLEF 2013*, Valencia.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006), "Effects of age and gender on blogging", *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, Stanford, CA, March, pp. 199-205.

Shavers, B. (2013), *Placing the Suspect Behind the Keyboard: Using Digital Forensics and Investigative Techniques to Identify Cybercrime Suspects*, Elsevier/Syngress, Melbourne.

Trudgill, P. (1972), "Sex, covert prestige and linguistic change in the Urban British English of Norwich", *Language in Society*, Vol. 1 No. 2, pp. 179-195.

Vel, O.D., Corney, M., Anderson, A. and Mohay, G. (2002), "Language and gender author cohort analysis of e-mail for computer forensics", in *Second Digital Forensic Research Workshop*, Syracuse, New York, NY.

About the authors

Ioannis Tsimperidis received Bachelor Degree from Aristotle University of Thessaloniki in 1997, and MSc Degree from Democritus University of Thrace in 2002. Currently he is a PhD candidate at the Department of Electrical and Computer Engineering of Democritus University of Thrace. He has been working as a Secondary Education teacher since 2006. His main areas of research interest are keystroke analysis, user classification and digital forensics.

Vasilios Katos is an Associate Professor of Information and Communications Systems Security at the Department of Electrical and Computer Engineering of Democritus University of Thrace in Greece. Prior to his current post, he was Principal Lecturer at the School of Computing at the University of Portsmouth where he participated in the development of the interdisciplinary Master's course, MSc in Forensic IT. He has worked in the industry as a Security Consultant and Expert Witness in Information Systems Security. His research interests are in information security, privacy, digital forensics and incident response. Vasilios Katos is the corresponding author and can be contacted at: vkatos@bournemouth.ac.uk

Nathan Clarke is a Professor in Cyber Security and Digital Forensics at the Plymouth University. His research interests reside in the area of information security, biometrics, forensics and cloud security. Professor Clarke has over 130 outputs consisting of journal papers, conference papers, books, edited books, book chapters and patents. He is the Chair of the IFIP TC11.12 Working Group on the Human Aspects of Information Security & Assurance. Professor Clarke is a chartered engineer, a fellow of the British Computing Society (BCS) and a senior member of the IEEE. He is the author of Transparent Authentication: Biometrics, RFID and Behavioural Profiling published by Springer and Computer Forensics: A Pocket Guide published by IT Governance.