Statistica Neerlandica (2009) Vol. 63, nr. 4, pp. 432–449
doi:10.1111/j.1467-9574.2009.00434.x

# The reliability of user authentication through keystroke dynamics

Salima Douhou*

*CentER, Tilburg University, The Netherlands*

Jan R. Magnus†

*Department of Econometrics & OR, Tilburg University, The Netherlands*

When typing on a keyboard a user can be authenticated through what he/she types (username, password), but also through how he/she types, that is, through keystroke dynamics. We study whether authentication through keystroke dynamics is sufficiently reliable as a security instrument to be used together with the more standard instruments. Based on a data set of 1254 participants who typed the same username and password, 20 times each, we develop a test statistic and obtain the power of our test. We conclude that keystroke dynamics can be a reliable security instrument for authentication, if used together with other instruments. It seems more suitable for authentication (verification) than for identification. Dwell times (how long a key is held pressed) are more discriminatory and therefore more powerful than flight times (time between consecutive press times).

*Keywords and Phrases:* fraud detection, identity verification, keystroke analysis, biometrics, computer security, empirical power.

## 1 Introduction

People can be authenticated by something they know (password), something they have (credit card), or by something they are (finger prints). When typing on a keyboard a user can be authenticated through *what* he/she types (username, password), but also through *how* he/she types, that is, through keystroke dynamics. The purpose of this paper is to investigate whether authentication through keystroke dynamics is sufficiently reliable as a security instrument to be used together with the more standard instruments.

The study of personal typing behavior (keystroke dynamics) is part of biometrics, where the underlying idea is that certain physical characteristics are (almost) unique and can therefore be used for authentication. Well-known examples are finger prints, voice recognition, and the iris scan.

*s.douhou@uvt.nl
†magnus@uvt.nl

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

The fact that people can be identified through their typing behavior, already known in the early days of the telegraph (Bryan and Harter, 1899), became important during the Second World War. Morse code is made up of dots and dashes, each of which has its described length. But no one replicates those prescribed lengths perfectly. The variation of spacing and the stretching out of the dots and dashes defines a 'rhythm' specific to the operator. This rhythm is called the operator's *fist*. In the Second World War, thousands of British so-called interceptors listened to German military radio broadcasts. These broadcasts were in code, so they could not be understood, but after a short while the interceptors could identify the fists of the German operators, just by listening to the rhythm of the transmission. As the British were also able to locate the radio signals, they could follow the German radio operators around Europe, a very useful piece of war information; see Gladwell (2005). The war experience has proved that a fist emerges naturally and unconsciously, that it reveals itself in even the smallest sample of Morse code, and that it is stable.

A sizeable literature on keystroke dynamics has developed since Gaines *et al.* (1980) reported on an experiment where seven professional typists were each given a paragraph of prose to type, and the times between successive keystrokes were recorded. Since then, various authors have proposed different approaches, more specifically:

*Statistical*: Joyce and Gupta (1990), Bleha, Slivinsky, and Hussien (1990), Song, Venable, and Perrig (1997), Monrose and Rubin (1997, 2000), Bergadano, Gunetti, and Picardi (2002), Guven and Sogukpinar (2003), Kacholia and Pandit (2003);

*Data mining*: Brown and Rogers (1993), Obaidat and Sadoun (1997), Cho *et al.* (2000), Gutiérrez *et al.* (2002), Yu and Cho (2004).

The basic idea of the *statistical* approach is to compare a reference set of typing characteristics of a certain user with a test set of typing characteristics of the same user or a test set of a hacker. The distance between these two sets (reference and test) should be below a certain threshold or else the user is recognized as a hacker. *Data mining* is a collection of techniques from the field of Artificial Intelligence and Machine Learning, and includes also neural networks. A data mining process typically first builds a prediction model from historical data, and then uses this model to predict the outcome of a new trial (or to classify a new observation). In contrast to statistics, data mining makes no assumption about the data. The key difference between the statistical and data mining methods is therefore the information that is used. For example, in a data mining approach, not only the similarities between the patterns of the same user are considered but also the differences of this pattern with all the other patterns observed in building the model. Thus, Lee and Cho (2007) develop a retraining framework by employing not only the user's but also the hackers' characteristics. Our approach will be statistical.

Leggett *et al.* (1991) and Hoquet, Ramel, and Cardot (2005) propose *dynamic* authentication, where the system continuously monitors a user's typing pattern. If the pattern does not match the profile of the logged-on user the computer shuts down or asks the user or hacker to type a password. With this method one continuously updates and monitors a logged-on user's profile.

An excellent review on statistics and fraud is given by Bolton and Hand (2002). More specialized reviews on keystroke dynamics can be found, *inter alia*, in Lipton and Wong (1985) and Peacock, Ke, and Wilkerson (2004). Distinguishing between real users and hackers can also be viewed as a one-class classification problem where one tries to distinguish one class of objects (real users) from all other possible objects (hackers) by learning from a training set containing only the objects of that class; see Duin and Tax (2004), Loog and Duin (2004), Zeng *et al.* (2006), and Kwak and Oh (2009) for discussion and examples of one-class classification problems.

One problem with the empirical applications is the lack of data. Gaines *et al.* (1980) have 7 participants, and the studies by Umphress and Williams (1985), Obaidat and Sadoun (1997), Gutiérrez *et al.* (2002), Monrose, Reiter, and Wetzel (2002), Hoquet, Ramel, and Cardot (2005), and Kang *et al.* (2008) employ between 15 and 25 participants. Monrose and Rubin (1997) and Clarke and Furnell (2007) – in a study on mobile devices – employ around 30 participants. Somewhat larger are the studies by Schonlau *et al.* (2001), Bergadano, Gunetti, and Picardi (2002), and Bartlow and Cukic (2006) – studying shift-key patterns – who employ around 50 participants.

In contrast, our data set consists of 1254 participants who typed *the same* username and password, 20 times each. Of course, mistakes were made and not all participants completed the full session of 20 logins. Nevertheless, the data set is large enough to be informative. The fact that each participant has the same username and password is important, because this allows us to consider each as a possible hacker to the other.

In section 2 we describe the data. In section 3 we develop a test statistic and obtain *theoretical* critical values for this test statistic. In section 4 we obtain *empirical* critical values, which lead to better sizes and are therefore preferable in our study. In section 5 we study the power of the tests, and section 6 concludes.

## 2 The data

The data were collected in May and June 2007 by three students of the Systems and Network Engineering Group of the Faculty of Science at the University of Amsterdam; see van Abswoude, Tavenier, and van der Schee (2007). The students created a website (no longer in existence), which they advertised through the website of the weekly magazine of the University of Amsterdam (http://www.folia.nl), the principal Dutch website read by those with an interest in security systems (http://www.security.nl), and other channels.

When a potential participant hits the website, a 'session' is started. In total, 3476 sessions were started in this way. The first step for the participant is to click the relevant link and download a *flash applet* (developed by the students) to his/her own computer. The purpose of the flash applet is to record the necessary timings during the session, based on the clock of the participant's computer. The main activity thus takes place on the participant's computer and not on the website's server, and therefore technical problems such as network latency or overloading of the server are avoided. Understandably, many potential participants did not download the flash applet or logged off immediately afterwards, without recording any timings. This happened in 64% of the sessions. This leaves us with 1254 sessions where timings have been recorded.

The participants were given a username (*patrick*) and a password (*water83*), the same for all participants. They were then asked to type their username and password 20 times. For each of the 20 login attempts, the press ($P$) and release ($R$) clock times of each of the 14 characters were recorded. This gives $(P_i, R_i)$ for $i = 1, \dots, 14$. From these data, we can calculate dwell times ($D$) and flight times ($F$) as

$$D_i := R_i - P_i, \quad F_i := P_i - P_{i-1}.$$

Hence, the dwell time records the time that each key is held pressed, and the flight time records the time between two consecutive press times. Clearly, $F_1$ has no meaning. We also disregard $F_8$, because we attach no significance to the time elapsed between the last letter of the username and the first letter of the password. This gives us 14 dwell times and 12 flight times per login attempt.

It might seem more natural to define flight time as $F_i^* = P_i - R_{i-1}$, so that the login duration is broken up in 'independent' non-overlapping pieces. This is not, however, a good idea, because $F^*$ can be (and often is) negative. Although the flash applet records both press and release times, characters registered by the computer are controlled only by the moment the key is pressed, not by the moment the key is released, and one may (and often will) press the next key when the previous key is not yet released.

If all participants would complete their session (20 logins) and would make no typing errors, then we would have $26 \times 20 \times 1254 = 652,080$ data points. In fact, some participants quitted voluntarily (they closed their browser) or involuntarily (their computer crashed), so that they did not complete all 20 logins. In addition, participants made typing errors. If a typing error is made in a username (or password), then all dwell and flight times for that attempted username (password) are deleted. Errors cannot be corrected using *backspace*, as this would confuse the interpretation of the dwell and flight times. If an error is made in the username but not in the password (or vice versa), then the correctly typed password (or username) data are *not* deleted.

Some information about early exits and error rates is provided in Figure 1. Of the 1254 participants who started, 104 made a mistake in both username and password in the first login; 1150 participants 'half-successfully' completed the first login (by

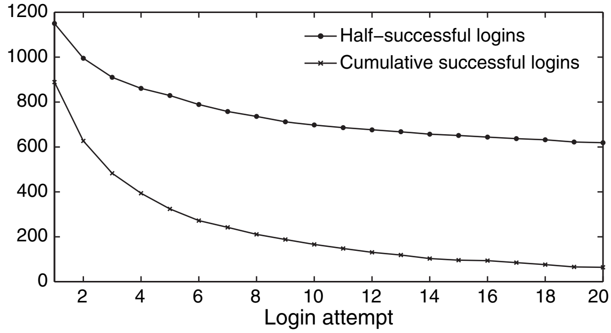

Fig. 1. Number of 'half-successful' and cumulative 'successful' login attempts.

being error-free in either username or password or both). This is the first point on the upper graph. Then, 995 participants half-successfully completed the second login: the second point. Finally, 619 participants half-successfully completed the twentieth login: the last point. Hence, at least 619 participants completed the whole session – at least, because some made an error in both username and password in the final login. The curve is decreasing because some participants drop out during the session. Of the 1150 participants who were half-successful (error-free in either username or password or both) in their first login, 889 were 'successful' (error-free in both username and password). Of those, 627 were successful in the first two logins, and only 64 were successful in all 20 logins. Hence, in contrast to the upper graph, the lower graph in Figure 1 provides cumulative information.

The participants are taken from a small group, consisting primarily of Dutch students, university employees, and those interested in security systems. We do not claim that this is a representative sample. It is possible that the typing behavior of the people in our sample differs from that of individuals with less computer experience. If there is a difference, then the people in our sample are expected to be more homogeneous than the average population, making it more difficult to detect differences in their typing patterns. Hence if we find that we can detect differences in typing patterns in our sample, then it should be easier in a less homogeneous group.

The username and password were chosen to reflect common practice. Both username and password have seven characters. They are simple and easy to remember. The addition of two digits (83) in the password is also quite common (typically, year of birth). We note that there are no repeated characters within the username or the password, which has practical importance in our experiment because it means that difficult issues of identification are avoided. All letters are lowercase symbols.

All participants were given the same username and password. As we shall see, this is of great practical use in our analysis, because we can consider each participant as a possible 'hacker' to everybody else.

To gain some insight into the dwell and flight times and their variation, we consider all participants with at least six error-free username attempts (898 people) and all participants with at least six error-free password attempts (897 people). For each person we calculate the average dwell and flight times: seven average dwell times and six average flight times per person for username and password separately. These averages define an empirical distribution from which we can calculate quantiles. The 10%, 50% (median), and 90% quantiles are given in Figures 2 and 3. We see from Figure 2 that the median dwell times fluctuate around 90 ms for the username and around 100 ms for the password, and that there is not much difference between the 14 characters. [One millisecond (ms) is one thousandth of a second.] The 10% and 90% quantile lines reveal, however, considerable variation among the participants.

Figure 3 shows that the median flight times fluctuate around 160 ms for the username and around 219 ms for the password. The large difference between average flight times in username and password can be contributed to the time it takes to move from *r* to 8 in the password *water83*, namely 465 ms. Apart from this, there is not much difference between the average flight times. The first four flight times of the password (only letters) fluctuate around 152 ms. Again, there is considerable variation among participants. In fact, there is more variation in flight times than in dwell times, because of individual differences in keyboard control: a person who

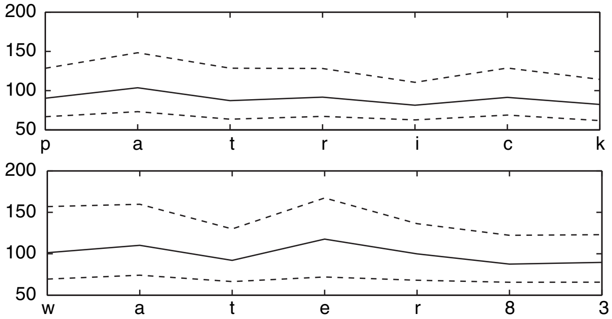

Fig. 2. Median dwell times with 80% bounds for username (upper panel) and password (lower panel).

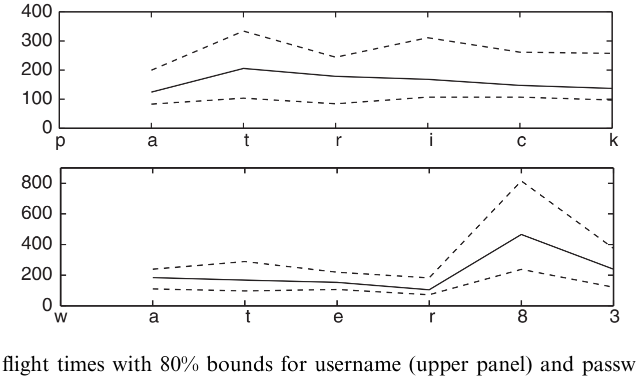

Fig. 3. Median flight times with 80% bounds for username (upper panel) and password (lower panel).

uses only two fingers will have a larger flight time on average than a person who uses ten fingers.

Finally, we comment briefly on the within-person variance. We compare participants from the group where the first login is deleted and exactly 15 of the remaining 19 logins are correct [96 participants, later called group 15(1)] with the group of all participants who have at least 6 error-free attempts. We then calculate for each of the 96 participants and for each character the SD of the dwell times and compare this with the average over 1000 random draws of 15 attempts on the same character from the entire population. The within-person SD is about 47% for the username and 44% for the password compared with the SD in the whole population. We repeat the experiment for a second group where the first five logins are deleted, and all 15 remaining logins are correct [136 participants, later called group 15(5)]. Then, the within-person SD drops to about 42% for the username and 38% for the password compared with the SD in the whole population. The percentages in the second experiment are lower because these participants make fewer errors and are therefore likely to be more consistent typists. A drop in SD of 50–60% may not seem much to develop a powerful test. Nevertheless, we shall see that considerable power can be achieved.

## 3 The test statistic and theoretical critical values

For a given participant we have $n$ observations on each of $m$ characteristics, for example, $n=20$ (number of logins) and $m=26$ (number of characteristics: 14 dwell times and 12 flight times). Let $x_{ij}$ denote the $i$th observation on the $j$th characteristic. If we assume that the $x_i := (x_{i1},\ldots x_{im})'$ are independently and identically distributed as

$$x_i \sim N(\mu, \Sigma), \quad \Sigma := \operatorname{diag}(\sigma_1^2,\ldots,\sigma_m^2), \tag{1}$$

so that the characteristics are independent of each other, then the maximum likelihood estimators of $\mu_j$ and $\sigma_j^2$ are given by

$$\hat{\mu}_j = \bar{x}_j := \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad \hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2.$$

Note that the means $\mu_j$ and variances $\sigma_j^2$ are assumed to be individual-specific.

We have made two strong independence assumptions: the characteristics are independent of each other, and the consecutive logins are independent. The first assumption means that if, for example, within one login the first flight time is smaller than expected, this has no impact on the next flight time. This is certainly not entirely true but it seems a reasonable simplification. (We discuss a possible extension to dependence in the ‘Conclusion’.) The second assumption is more difficult to defend and repair, and dependence between consecutive logins could very well be the reason why the critical values are unsatisfactory, as we shall see.

For the moment we adopt these two independence assumptions. Now assume that, in addition to $x_1,\ldots,x_n$, we have one other $m \times 1$ vector $y$, independent of $\{x_i\}$. Under the null hypothesis that $y$ is generated by the same distribution as the $\{x_i\}$, we have

$$\bar{x}_j \sim N\left(\mu_j, \frac{\sigma_j^2}{n}\right), \quad y_j \sim N(\mu_j, \sigma_j^2),$$

so that

$$\frac{n}{n+1}\sum_{j=1}^{m}\frac{(\bar{x}_j - y_j)^2}{\sigma_j^2} \sim \chi^2(m). \tag{2}$$

As our test statistic, we propose

$$T_{m,n} := \frac{n}{m(n+1)}\sum_{j=1}^{m}\frac{(\bar{x}_j - y_j)^2}{\hat{\sigma}_j^2}, \tag{3}$$

whose distribution depends only on $m$ and $n$. As

$$t_j := \sqrt{\frac{n-1}{n+1}} \cdot \frac{\bar{x}_j - y_j}{\hat{\sigma}_j} \sim \text{ Student}(n-1),$$

we can write

$$T_{m,n} = \frac{n}{m(n-1)}\sum_{j=1}^{m} t_j^2. \tag{4}$$

For large $n$, the statistic $T_{m,n}$ can be approximated by a $\chi^2(m)/m$-distribution. For large $m$, it can be approximated by a normal distribution using the exact moments

$$\mathrm{E}(T_{m,n}) = \frac{n}{n-3}, \quad \mathrm{var}(T_{m,n}) = \frac{2}{m} \cdot \frac{n^2(n-2)}{(n-3)^2(n-5)}.$$

However, for values like $n=20$ and $m=26$, the asymptotic behavior is of little use, and we have to resort to simulation.

For given values of $m$ and $n$ and for given significance levels $\alpha$, the distribution of $T_{m,n}$ can be simulated and quantiles $k_\alpha$ satisfying

$$\Pr(T_{m,n} > k_\alpha) = \alpha$$

can be estimated. As shown in Shorack and Wellner (1986, example 1, p. 639), the sample quantiles $\hat{k}_\alpha$ are consistent and asymptotically normal, and

$$\widehat{\mathrm{var}}(\hat{k}_\alpha) \approx \frac{\alpha(1-\alpha)}{r(f_r(\hat{k}_\alpha))^2}, \tag{5}$$

provides a consistent estimate of the variance of $\hat{k}_\alpha$, where $f_r(\hat{k}_\alpha)$ denotes an estimate of the density of $T_{m,n}$ at $k_\alpha$ after $r$ replications. As we want our estimates for $k_\alpha$ to be accurate to two decimal places, we could use equation (5) to determine the number of replications $r$. In practice, it is more efficient to take $N$ independent batches of 100,000 replications each for every combination of $m$ and $n$, and calculate the mean

Table 1. Theoretical critical values of the $T_{m,n}$ test

|  | $\alpha$ | 0.01 |  |  | 0.05 |  |  | 0.10 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $m$ | 12 | 14 | 26 | 12 | 14 | 26 | 12 | 14 | 26 |
| 13 |  | 3.29 | 3.11 | 2.55 | 2.45 | 2.37 | 2.07 | 2.10 | 2.05 | 1.85 |
| 15 |  | 3.07 | 2.92 | 2.41 | 2.33 | 2.25 | 1.97 | 2.01 | 1.95 | 1.77 |
| 17 |  | 2.94 | 2.79 | 2.30 | 2.25 | 2.16 | 1.90 | 1.94 | 1.89 | 1.71 |
| 19 |  | 2.83 | 2.69 | 2.23 | 2.18 | 2.10 | 1.85 | 1.89 | 1.84 | 1.67 |
| $\infty$ |  | 2.18 | 2.08 | 1.76 | 1.75 | 1.69 | 1.50 | 1.55 | 1.50 | 1.37 |

and variance over these $N$ batches. For $N=1000$ we obtain a SD of $\hat{k}_\alpha$ of about 0.0003 for $\alpha=0.05$ and 0.0004 for $\alpha=0.01$, which secures the required accuracy. Thus, we obtain Table 1 with the relevant quantiles (critical values) $k_\alpha$ of the $T_{m,n}$ test statistic for 15 $(m,n)$ combinations and three commonly used values of $\alpha$.

## 4 Empirical critical values

We will see shortly that the critical values of Table 1, dictated by statistical theory under the simplest assumptions, are not accurate enough to make predictions about the power of the test.

Let us distinguish between the $m_1=12$ flight times and $m_2=14$ dwell times in our sample, and consider two test statistics, using equation (3),

$$T_1=\frac{n}{m_1(n+1)}\sum_{j=1}^{m_1}\frac{(\bar{x}_j-y_j)^2}{\hat{\sigma}_j^2},\quad T_2=\frac{n}{m_2(n+1)}\sum_{j=1}^{m_2}\frac{(\bar{x}_j-y_j)^2}{\hat{\sigma}_j^2},$$

for flight times and dwell times separately, together with the combined statistic

$$T=\frac{m_1}{m}T_1+\frac{m_2}{m}T_2.$$

We shall consider four subsets of our data. As the participants are unfamiliar with their username and password, they need some time to practice. In the first three subsets we therefore delete the first of the logins, as follows:

Group 19(1): first login is deleted, all 19 remaining logins are correct (78 participants);

Group 17(1): first login is deleted, exactly 17 of the remaining 19 logins are correct (161 participants);

Group 15(1): first login is deleted, exactly 15 of the remaining 19 logins are correct (96 participants).

These three groups are mutually exclusive. In addition, we consider one further subset where the first five logins have been deleted.

Group 15(5): first five logins are deleted, all 15 remaining logins are correct (136 participants).

Notice that Group 19(1) is a subset of Group 15(5), and that Groups 17(1) and 15(1) intersect with Group 15(5), but are no subsets.

Table 2. Size of the $T_{m,n}$ test based on theoretical critical values

|  |  | $\alpha$ | 0.01 |  |  | 0.05 |  |  | 0.10 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $G$ | $m$ | 12 | 14 | 26 | 12 | 14 | 26 | 12 | 14 | 26 |
| 13 | 15(5) |  | 0.15 | 0.04 | 0.15 | 0.18 | 0.08 | 0.19 | 0.21 | 0.12 | 0.22 |
| 17 | 19(1) |  | 0.15 | 0.04 | 0.15 | 0.18 | 0.09 | 0.19 | 0.20 | 0.13 | 0.22 |

For each of these four groups we perform a small experiment, as follows. Suppose our group is 19(1). For each of the 78 people in this group we select two login attempts, labeled $y_{(1)}$ and $y_{(2)}$, which can be done in $\binom{19}{2}=171$ ways. From the remaining $n=17$ login attempts we calculate $\bar{x}_j$ and $\hat{\sigma}_j^2$ for each $j$. For both $y_{(1)}$ and $y_{(2)}$ separately we then calculate $T_1$, $T_2$, and $T$. If we do this for each of the 78 people in the group, we obtain 156 values for $T_1$, $T_2$, and $T$. Repeating the experiment for each person and each combination provides us with $78\times171\times2=26,676$ values for $T_1$, $T_2$, and $T$. Each test outcome is then confronted with the appropriate theoretical critical value in Table 1 for $n=17$ and $m_1=12$ ($T_1$), $m_2=14$ ($T_2$), and $m=26$ ($T$), respectively, and the proportion of times that the test rejects (the size) is calculated. In Table 2 we report the empirical sizes for two of the four subsets, namely 15(5) and 19(1); the other two subsets behave similarly. We see that the empirical sizes are about 15 ($T_1$) to 4 ($T_2$) times as large as predicted when $\alpha=0.01$, about 3.6 ($T_1$) to 1.7 ($T_2$) times as large when $\alpha=0.05$, and about 2.1 ($T_1$) to 1.3 ($T_2$) times as large when $\alpha=0.10$. The larger is the $\alpha$, the better is the empirical size approximated by the theoretical size. Also, the approximation works better for $T_2$ (dwell times) than for $T_1$ (flight times).

Although the theoretical sizes are possibly acceptable for $\alpha=0.10$, they are not for $\alpha\leq0.05$. Hence we shall obtain better results for the values of interest when we use empirical critical values, instead of the theoretical critical values of Table 1.

The empirical critical values are obtained as follows. Suppose again that the group of interest is 19(1). The calculations are the same as for Table 2 leading to $78\times171\times2=26,676$ values for $T_1$, $T_2$, and $T$. For $\alpha$ equal to 0.01, 0.05, and 0.10 we then estimate the critical value $k_\alpha$ satisfying $\Pr(T^*>k_\alpha)=\alpha$, where $T^*$ takes the values $T_1$, $T_2$, and $T$, respectively. We repeat these calculations for each of the seven groups:

19(1): $n=17$, $n=15$, $n=13$;
17(1): $n=15$, $n=13$;
15(1): $n=13$;
15(5): $n=13$.

For example: Group 17(1) contains all participants where, ignoring the first login, precisely 17 of the remaining 19 logins are correct. From these 17 correct logins we select $n$ (15 or 13) at random. The results are given in Table 3, which confirms that these (empirical) critical values are quite different from the theoretical values in Table 1. We notice that, within each group, we have selected *two* login attempts, $y_{(1)}$ and $y_{(2)}$, and for each separately we have calculated $T_1$, $T_2$, and $T$. Let us denote these statistics as $T_1^{(1)}$, $T_2^{(1)}$, and $T^{(1)}$ for $y_{(1)}$, and $T_1^{(2)}$, $T_2^{(2)}$, and $T^{(2)}$ for $y_{(2)}$. Defining

Table 3. Empirical critical values: one draw

| | | $\alpha$ | 0.01 | | | 0.05 | | | 0.10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $G$ | $m$ | 12 | 14 | 26 | 12 | 14 | 26 | 12 | 14 | 26 |
| 13 | 15(5) | | 86.44 | 5.53 | 57.66 | 14.82 | 2.81 | 8.35 | 5.21 | 2.21 | 3.79 |
| 13 | 19(1) | | 103.35 | 5.99 | 54.41 | 17.58 | 2.93 | 9.88 | 5.81 | 2.27 | 3.97 |
| 13 | 17(1) | | 71.77 | 7.32 | 42.42 | 15.62 | 2.86 | 9.39 | 5.74 | 2.26 | 3.91 |
| 13 | 15(1) | | 90.06 | 6.58 | 54.36 | 18.27 | 2.70 | 10.68 | 6.38 | 2.20 | 4.35 |
| 15 | 19(1) | | 82.41 | 5.20 | 41.26 | 15.47 | 2.69 | 8.54 | 5.24 | 2.14 | 3.56 |
| 15 | 17(1) | | 66.22 | 5.36 | 34.49 | 14.51 | 2.72 | 8.25 | 5.38 | 2.15 | 3.65 |
| 17 | 19(1) | | 85.41 | 4.77 | 42.17 | 15.48 | 2.60 | 8.27 | 4.90 | 2.06 | 3.36 |

Table 4. Empirical critical values: two draws

| | | $\alpha$ | 0.0001 | | | 0.0025 | | | 0.0100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $G$ | $m$ | 12 | 14 | 26 | 12 | 14 | 26 | 12 | 14 | 26 |
| 13 | 15(5) | | 106.20 | — | — | 14.53 | 4.61 | 14.12 | 6.02 | 2.40 | 4.43 |
| 13 | 19(1) | | 137.90 | 131.66 | 92.99 | 25.99 | 3.57 | 18.93 | 7.49 | 2.52 | 4.81 |
| 13 | 17(1) | | 117.22 | — | — | 19.95 | 5.51 | 17.77 | 6.21 | 2.52 | 4.75 |
| 13 | 15(1) | | — | — | — | 25.93 | 4.01 | 20.88 | 8.26 | 2.42 | 5.35 |
| 15 | 19(1) | | 106.79 | 71.73 | 49.71 | 15.39 | 3.14 | 9.36 | 6.42 | 2.31 | 4.09 |
| 15 | 17(1) | | 120.13 | — | — | 17.27 | 3.48 | 10.25 | 5.79 | 2.31 | 3.92 |
| 17 | 19(1) | | 121.46 | 17.81 | 56.44 | 15.59 | 2.90 | 8.29 | 6.07 | 2.24 | 3.75 |

$$T_1^{\min} := \min(T_1^{(1)}, T_1^{(2)}), \quad T_2^{\min} := \min(T_2^{(1)}, T_2^{(2)}), \quad T^{\min} := \min(T^{(1)}, T^{(2)}),$$

we obtain $78 \times 171 = 13{,}338$ values for $T_1^{\min}$, $T_2^{\min}$, and $T^{\min}$. For $\alpha$ equal to 0.0001, 0.0025, and 0.0100 (the squares of the previous $\alpha$-values), we then estimate the critical values $k_\alpha$, and we repeat these calculations for each of the seven groups. This leads to Table 4. As

$$\Pr(T^{(1)} > k_\alpha \text{ and } T^{(2)} > k_\alpha) = \Pr(\min(T^{(1)}, T^{(2)}) > k_\alpha) = \Pr(T^{\min} > k_\alpha),$$

we see that Table 4 contains the required critical values for two consecutive draws. If these two draws were independent (which they might not be), then we would have

$$\Pr(T^{(1)} > k_1 \text{ and } T^{(2)} > k_2) = \Pr(T^{(1)} > k_1)\Pr(T^{(2)} > k_2)$$

for all $k_1$ and $k_2$, and the numbers reported in Tables 3 and 4 would be identical. The results in Tables 3 and 4 suggest that in fact

$$\Pr(T^{(1)} > k_1 \text{ and } T^{(2)} > k_2) \geq \Pr(T^{(1)} > k_1)\Pr(T^{(2)} > k_2)$$

or, what amounts to the same, that

$$\Pr(T^{(2)} > k_2 \mid T^{(1)} > k_1) \geq \Pr(T^{(2)} > k_2),$$

implying that $T^{(1)}$ and $T^{(2)}$ are 'positively quadrant-dependent' (Lehmann, 1966). Although the two draws are clearly not independent, the deviation from independence does not appear to be large.

It is already clear from Table 3 that the critical values for $\alpha = 0.01$ are less stable than those for $\alpha = 0.05$ and $\alpha = 0.10$. This effect is even stronger in Table 4: the critical values for 0.0001 are very unstable. In fact, certain critical values become infinite, owing to the fact that one or more of the $\hat{\sigma}_j^2$ in equation (3) become zero. This can only happen if a participant has $n$ identical logins. If a critical value is infinitely large then we will find an empirical power close to zero, and hence a hacker is always treated as an authorized user. This requires further investigation. We find the following:

Group 19(1): No infinite values were found, but for $n = 13$ and 15 (not for $n = 17$) some may in fact be there because not all possibilities have been examined.

Group 17(1): Two participants recorded 14 (out of 17) identical dwell times on the letter *i* in the username *patrick*, and one recorded 15 identical dwell times on the letter *a* in *patrick*. It is possible that there are more infinite values for $n = 13$ (but not for $n = 15$) because not all possibilities have been examined.

Group 15(1): One participant recorded 13 (out of 15) identical dwell times on the letter *t* in the password *water83*, whereas another participant recorded 13 identical dwell times on the letter *w* in *water83*, and also 13 identical flight times on the passage *t–e* in *water83*.

Group 15(5): One participant recorded 14 (out of 15) identical dwell times on the letter *i* in *patrick*, and also 13 identical dwell times on the letter *r* in *water83*.

This is rather surprising, at least it was surprising to us. Apparently some participants display a very high degree of regularity in typing behavior, which underlines the potential for using keystroke dynamics for user authentication.

## 5 Power of the test

Now that we have computed the empirical critical values for three given sizes $\alpha$, we can consider the power of our test, that is, the probability that a 'hacker' is recognized as a hacker. Suppose one of the other people in the same group 'breaks in'. What is the probability that he/she is found out?

Based on the empirical critical values of Tables 3 and 4, we perform the following experiment. Choose one group, say 19(1) with $n = 17$. Choose two people (ordered) in this group, say $(i,j)$, where person $i$ is the potential victim and person $j$ is the hacker. This can be done in $78 \times 77 = 6006$ ways. Draw randomly $n$ observations from $i$ and two observations from $j$, and calculate the test statistics. Thus, we obtain $6006 \times 2 = 12,012$ values for each of the three test statistics. We then confront these values with the appropriate critical values in Table 3. This will give us the probability that our tests will label a person as a hacker when the login was indeed performed by a hacker, that is, the power of our tests. Table 5 shows that the power for $\alpha = 0.01$ is not good. But for $\alpha = 0.05$ the test based on dwell times ($T_2$) gives a power of about 85%, and for $\alpha = 0.10$ the test based on dwell times ($T_2$) gives a power of about 90% and the overall test ($T$) gives a power of about 87%. This suggests that dwell times are more discriminatory and therefore more powerful than flight times. For the dwell times ($m = 14$), we visualize the trade-off between size and power in Figure 4. All

Table 5. Empirical power of the $T_{m,n}$ test: one draw

| | | $\alpha$ | 0.01 | | | 0.05 | | | 0.10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $G$ | $m$ | 12 | 14 | 26 | 12 | 14 | 26 | 12 | 14 | 26 |
| 13 | 15(5) | | 0.07 | 0.61 | 0.07 | 0.42 | 0.86 | 0.63 | 0.76 | 0.92 | 0.90 |
| 13 | 19(1) | | 0.05 | 0.57 | 0.06 | 0.31 | 0.84 | 0.53 | 0.68 | 0.90 | 0.88 |
| 13 | 17(1) | | 0.06 | 0.46 | 0.07 | 0.30 | 0.84 | 0.48 | 0.64 | 0.90 | 0.86 |
| 13 | 15(1) | | 0.04 | 0.46 | 0.04 | 0.22 | 0.82 | 0.38 | 0.55 | 0.88 | 0.78 |
| 15 | 19(1) | | 0.06 | 0.62 | 0.09 | 0.31 | 0.87 | 0.56 | 0.68 | 0.92 | 0.90 |
| 15 | 17(1) | | 0.06 | 0.58 | 0.08 | 0.30 | 0.84 | 0.52 | 0.64 | 0.90 | 0.86 |
| 17 | 19(1) | | 0.05 | 0.66 | 0.08 | 0.31 | 0.87 | 0.57 | 0.69 | 0.92 | 0.91 |

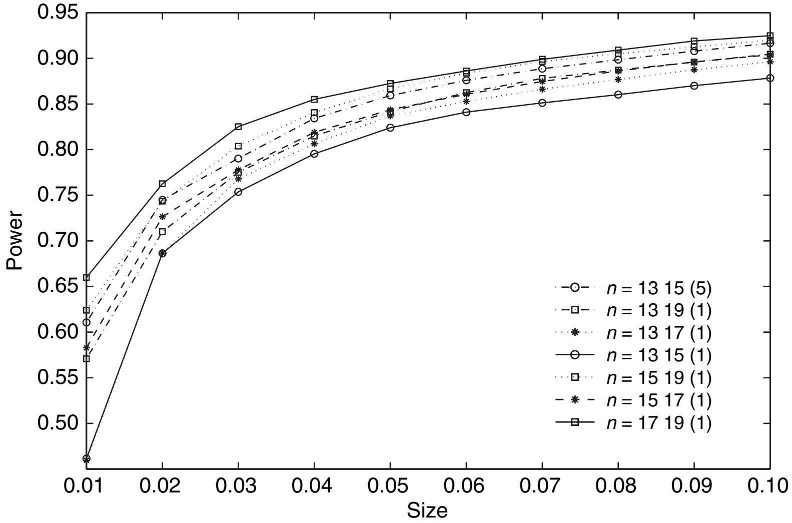

Fig. 4. Empirical power versus size, dwell times: one draw.

data sets behave in the same way, but obviously more information (larger number of successful logins) leads to higher power.

The results in Table 5 assume that a person is labeled as a hacker when the test fails once. In many situations one is allowed a second chance, and a person is only labeled as a hacker when he/she fails twice. The power of test based on two attempts is similarly calculated, now using the critical values in Table 4. The results in Table 6 confirm those in Table 5. For $\alpha = 0.0025$ the test based on dwell times ($T_2$) gives a power of about 67% and for $\alpha = 0.01$ the test based on dwell times ($T_2$) gives a power of about 84%. The overall test ($T$) gives a power of about 76%.

Recall that 19(1) denotes the group where the first login has been deleted and all remaining 19 logins are correct, and that 17(1) and 15(1) denote the groups where again the first login has been deleted and where exactly 17 or 15 of the remaining 19 logins are correct. For $n = 13$ we see that the power increases when we move from 15(1) to 19(1), and for $n = 15$ we see that the power increases when we move from 17(1) to 19(1). This confirms that if a user exhibits more regularity, it will be easier to establish his/her pattern, and it will be more difficult for a potential hacker to break in.

Table 6. Empirical power of the $T_{m,n}$ test: two draws

| | | $\alpha$ | 0.0001 | | | 0.0025 | | | 0.0100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $G$ | $m$ | 12 | 14 | 26 | 12 | 14 | 26 | 12 | 14 | 26 |
| 13 | 15(5) | | 0.03 | — | — | 0.35 | 0.62 | 0.33 | 0.64 | 0.86 | 0.82 |
| 13 | 19(1) | | 0.01 | 0.01 | 0.01 | 0.14 | 0.71 | 0.19 | 0.50 | 0.83 | 0.76 |
| 13 | 17(1) | | 0.01 | — | — | 0.16 | 0.51 | 0.17 | 0.52 | 0.82 | 0.73 |
| 13 | 15(1) | | — | — | — | 0.08 | 0.60 | 0.10 | 0.34 | 0.80 | 0.61 |
| 15 | 19(1) | | 0.02 | 0.02 | 0.04 | 0.22 | 0.76 | 0.44 | 0.52 | 0.86 | 0.81 |
| 15 | 17(1) | | 0.01 | — | — | 0.17 | 0.69 | 0.33 | 0.52 | 0.84 | 0.79 |
| 17 | 19(1) | | 0.01 | 0.14 | 0.03 | 0.22 | 0.78 | 0.48 | 0.53 | 0.87 | 0.83 |

## 6 Conclusions

Based on our experiments, we conclude that keystroke dynamics can be a reliable security instrument for authentication. It appears that dwell times (how long a key is held pressed) are more discriminatory and therefore more powerful than flight times (time between consecutive press times), confirming a similar finding by Obaidat and Sadoun (1997). Our $T_2$-test based on dwell times tells us that:

if we reject a person if the $T_2$-test fails once, then it will reject the true owner 5% of the time and recognize a hacker 85% of the time (Table 5, $\alpha=0.05$, power $=0.85$);

if we reject a person if the $T_2$-test fails twice, then it will reject the true owner 1% of the time and recognize a hacker 84% of the time (Table 6, $\alpha=0.01$, power $=0.84$).

In practice, a biometric test will be used in combination with another test or perhaps several other tests. In such, more realistic, cases we have:

$$\begin{aligned}\Pr(\text{hacker successful}) &= \Pr(\text{our test fails } \textit{and} \text{ current tests fail})\\ &= \Pr(\text{our test fails} \mid \text{current tests fail}) \times \Pr(\text{current tests fail}).\end{aligned}$$

Suppose that the hacker is recognized with the current tests in about 99% of the attempts, so that Pr(current tests fail) $=0.01$. Suppose also that, if the current tests do not recognize the hacker, our test does recognize the hacker in about 85% of the attempts, so that Pr(our test fails | current tests fail) $=0.15$. Then we find that Pr(hacker successful) $=0.0015$, so that the hacker will be unmasked in 99.85% of the attempts. The biometric test thus improved the power from 99.00% to 99.85%, and the cost caused by the hackers will be reduced by 85% if the biometric test is added to the authentication procedure.

It is difficult to compare our results with the literature, because every paper has a different number of participants, a different set-up, and a different statistical method. As a tentative guide we summarize next the Type I errors ($\alpha$) and Type II errors ($\beta$) as reported in the literature:

Umphress and Williams (1985): $\alpha=0.12$, $\beta=0.06$;
Bleha, *et al.* (1990): $\alpha=0.08$, $\beta=0.03$;

Leggett *et al.* (1991): $\alpha = 0.06$, $\beta = 0.05$;
Monrose and Rubin (1997): $0.09 < \alpha < 0.37$;
Bergadano, *et al.* (2002): $0.02 < \alpha < 0.06$, $\beta < 0.01$;
Gutiérrez *et al.* (2002): $\alpha = 0.20$, $\beta = 0.04$;
Kacholia and Pandit (2003): $0.01 < \alpha < 0.08$, $0.01 < \beta < 0.08$;
Gunetti and Picardi (2005): $\alpha = 0.01$, $\beta = 0.05$.

The high power (around 95%) obtained in these studies is a little puzzling given the typically very small number of participants. We have more participants and obtain lower power. Nevertheless, the main conclusion from our analysis is that especially dwell times (how long a key is pressed) can be used to create a powerful test. At a size of 1% the power of our best-performing two-draw test is 84% ($\beta = 0.16$).

Some caution is required in applying our results to different situations. First, our data may not be representative. Mostly students and people interested in security systems take part in our experiment. We do not know whether this affects our analysis, but if it does then the people in our sample are expected to be more homogeneous than the average population, making it more difficult to detect differences in their typing patterns. As we can detect differences in typing patterns in our sample, it should be easier to detect such differences in a less homogeneous group. The reported power can thus be viewed as a lower bound. Second, we only consider the username–password combination, which together contains 14 characters. In an environment where fewer (sometimes only four) characters are required from the user, it is doubtful that the user can be authenticated with sufficient accuracy. Third, the fact that our set-up has no repeated characters may influence our results.

We have developed the test statistic under the assumption that the characteristics are independent. This is probably unrealistic and more power can be obtained by allowing for some dependence, perhaps using Markov models (Jiang, Shieh, and Liu, 2007). Suppose, as before, that for a given participant we have $n$ observations on each of the $m$ characteristics, and let $x_{ij}$ denote the $i$th observation on the $j$th characteristic. Assume again that the $x_i := (x_{i1}, \ldots x_{im})'$ are independently and identically distributed, but now as

$$x_i \sim N(\mu, \Sigma), \quad \Sigma := \Sigma(\theta),$$

where $\theta$ is a $k \times 1$ vector of unknown parameters. In equation (1) we assumed that $\theta = (\sigma_1^2, \ldots, \sigma_m^2)'$ and $k = m$, implying that the characteristics are independent of each other. If we drop this assumption, then the maximum likelihood estimator of $\mu$ is again given by $\hat{\mu} = \bar{x} := (1/n)\sum_i x_i$, and the maximum likelihood estimator of $\theta$ is obtained by

$$\min_\theta \left( \log |\Sigma(\theta)| + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})'\Sigma^{-1}(\theta)(x_i - \bar{x}) \right).$$

More explicitly, letting

$$S := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})',$$

the $\hat{\theta}$s are found by solving the $k$ equations

$$\mathrm{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_h}\right) = \mathrm{tr}\left(\Sigma^{-1}S\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_h}\right) \quad (h = 1, \ldots, k),$$

from which we see that the maximum likelihood estimator $\hat{\theta}$ depends on the observations only through $S$. Instead of equation (2), we then have

$$\frac{n}{n+1}(\bar{x} - y)'\Sigma^{-1}(\bar{x} - y) \sim \chi^2(m),$$

and the test statistic becomes

$$T_{m,n} := \frac{n}{m(n+1)}(\bar{x} - y)'\Sigma(\hat{\theta})^{-1}(\bar{x} - y).$$

As we have to estimate more parameters, we would require more data in this case. We recommend this extension only if the number of observations is large, because otherwise the additional noise generated by having to estimate more parameters might outweigh the additional power of the test.

We have taken account of the fact that username and password are unfamiliar to our participants, by deleting either the first or the first five logins. When comparing the power for $n = 13$ and $G = 15(1)$ and $15(5)$, respectively, we see in Table 5 at $\alpha = 0.05$ that the power of the $T_2$ statistic increases from 0.82 when only the first login is deleted to 0.86 when the first five observations are deleted. Similarly, in Table 6 at $\alpha = 0.01$, the power of the $T_2$ statistic increases from 0.80 to 0.86.

We also note that in practice the number of observations on a specific user ($n$ in our analysis) will be larger than what we use in our experiment (maximum 17) and hence will increase the power of our tests. For example, in Table 5 at $\alpha = 0.05$ and $G = 19(1)$, the power of the $T_2$ statistic increases from 0.84 when $n = 13$ to 0.87 when $n = 17$, and, similarly, in Table 6 at $\alpha = 0.01$ and $G = 19(1)$, the power of the $T_2$ statistic increases from 0.83 to 0.87. This confirms that a larger value of $n$ will increase the power of our test.

In practical applications, the user will be familiar with his/her username and password, and also the number of observations will be larger than 17. It seems therefore reasonable to believe that our power estimates are lower bounds, and that the power of our tests will be higher in practice.

Finally, the balance between Type I and Type II errors can be controlled by the company. In a period when many hackers are active, the company may choose to increase $\alpha$, thus increasing the power. Users may be annoyed because they may be denied access to their own accounts, but hackers will find it more difficult to break in.

In conclusion, keystroke dynamics can be a reliable and flexible security instrument for authentication, if used in addition with other instruments. It seems more suitable for authentication (verification) than for identification.

**Acknowledgements**

The authors are grateful to Patrick van Abswoude, Peter Tavenier, and Maurits van der Schee for permission to use their data; to Maurits van der Schee for his patience in explaining technical details; to John Einmahl for helpful discussions; and to the associate editor and two referees for their constructive and helpful comments.

## References

van Abswoude, P., P. Tavenier and M. van der Schee (2007), *Keystroke dynamics and two factor authentication*, Master's thesis, Systems and Network Engineering Group, Informatics Institute, Faculty of Science, University of Amsterdam, The Netherlands.

Bartlow, N. and B. Cukic (2006), Evaluating the reliability of credential hardening through keystroke dynamics, in: *Proceedings of the 17th International Symposium on Software Reliability Engineering*, Washington DC, USA, pp. 117–126.

Bergadano, F., D. Gunetti and C. Picardi (2002), User authentication through keystroke dynamics, *ACM Transactions on Information and System Security* **5**, 367–397.

Bleha, S., C. Slivinsky and B. Hussien (1990), Computer-access security systems using keystroke dynamics, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 1217–1222.

Bolton, R. J. and D. J. Hand (2002), Statistical fraud detection: a review. *Statistical Science* **17**, 235–249.

Brown, M. and S. J. Rogers (1993), User identification via keystroke characteristics of typed names using neural networks, *International Journal of Man–Machine Studies* **39**, 999–1014.

Bryan, W. and N. Harter (1899), Studies in the telegraphic language, *Psychological Review* **6**, 345–375.

Cho, S., C. Han, D. H. Han and H.-I. Kim (2000), Web-based keystroke dynamics identity verification using neural network, *Journal of Organizational Computing and Electronic Commerce* **10**, 295–307.

Clarke, N. L. and S. M. Furnell (2007), Advanced user authentication for mobile devices, *Computers & Security* **26**, 109–119.

Duin, R. P. W. and D. M. J. Tax (2004), Statistical pattern recognition, in: C. H. Chen, L. F. Pau and P. S. P. Wang (eds), *Handbook of pattern recognition and computer vision*, World Scientific, Singapore, pp. 3–24.

Gaines, R. S., W. Lisowski, S. J. Press and N. Shapiro (1980), *Authentication by keystroke timing: some preliminary results*, Unpublished report, Rand Publication Series R-2526-NSF.

Gladwell, M. (2005), *Blink: The power of thinking without thinking*, Little, Brown and Company, New York.

Gunetti, D. and C. Picardi (2005), Keystroke analysis of free text, *ACM Transactions on Information and System Security* **8**, 312–347.

Gutiérrez, F. J., M. M. Lerma-Rascón, L. R. Salgado-Garza and F. J. Cantú (2002), Biometrics and data mining: comparison of data mining-based keystroke dynamics methods for identity verification, in: C. A. Coello Coello, A. de Albornoz, L. E. Sucar and O. C. Battistutti (eds), *Proceedings of the Second Mexican International Conference on Artificial Intelligence (MICAI)*, Merida, Yucatan, Mexico, Lecture Notes in Computer Science, Springer-Verlag, Berlin, pp. 460–469.

Guven, A. and I. Sogukpinar (2003), Understanding users' keystroke patterns for computer access security, *Computers & Security* **22**, 695–706.

Hoquet, S., J.-Y. Ramel and H. Cardot (2005), Fusion of methods for keystroke dynamic authentication, in: *Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, Washington DC, pp. 224–229.

JIANG, C.-H., S. SHIEH and J.-C. LIU (2007), Keystroke statistical learning model for web authentication, in: *Proceedings of the Second ACM Symposium on Information, Computer and Communications Security*, New York, pp. 359–361.

JOYCE, R. and G. GUPTA (1990), Identity authentication based on keystroke latencies, *Communications of the ACM* **33**, 168–176.

KACHOLIA, V. and S. PANDIT (2003), *Biometric authentication using random distributions (BioART)*, Paper presented at the 15th Annual Canadian Security Symposium, Ottawa.

KANG, P., S. PARK, S.-S. HWANG, H.-J. LEE and S. CHO (2008), Improvement of keystroke data quality through artificial rhythms and cues, *Computers & Security* **27**, 3–11.

KWAK, N. and J. OH (2009), Feature extraction for one-class classification problems: enhancements to biased discriminant analysis, *Pattern Recognition* **42**, 17–26.

LEE, H.-J. and S. CHO (2007), Retraining a keystroke dynamics-based authenticator with impostor patterns, *Computers & Security* **26**, 300–310.

LEGGETT, J., G. WILLIAMS, M. USNICK and M. LONGNECKER (1991), Dynamic identity verification via keystroke characteristics, *International Journal of Man–Machine Studies* **35**, 859–870.

LEHMANN, E. L. (1966), Some concepts of dependence, *Annals of Mathematical Statistics* **37**, 1137–1153.

LIPTON, D. L. and H. K. T. WONG (1985), Modern trends in authentication, *ACM SIGSAC Review* **3**, 36–42.

LOOG, M. and R. P. W. DUIN (2004), Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 732–739.

MONROSE, F. and A. D. RUBIN (1997), Authentication via keystroke dynamics, in: *Proceedings of the Fourth ACM Conference on Computer and Communications Security*, Zürich, Switzerland, pp. 48–56.

MONROSE, F. and A. D. RUBIN (2000), Keystroke dynamics as a biometric for authentication, *Future Generation Computing Systems* **16**, 351–359.

MONROSE, F., M. K. REITER and S. WETZEL (2002), Password hardening based on keystroke dynamics, *International Journal of Information Security* **1**, 69–83.

OBAIDAT, M. S. and B. SADOUN (1997), Verification of computer users using keystroke dynamics. *IEEE Transactions on Systems, Man, and Cybernetics* **27**, 261–269.

PEACOCK, A., X. KE and M. WILKERSON (2004), Typing patterns: a key to user identification, *IEEE Security and Privacy* **2**, 40–47.

SCHONLAU, M., W. DUMOUNCHEL, W.-H. JU, A. F. KARR, M. THEUS and Y. VARDI (2001), Computer intrusion: detecting masquerades, *Statistical Science* **16**, 58–74.

SHORACK, G. R. and J. A. WELLNER (1986), *Empirical processes with applications to statistics*, John Wiley, New York.

SONG, D., P. VENABLE and A. PERRIG (1997), User recognition by keystroke latency pattern analysis, retrieved from: http://citeseer.ist.psu.edu/song97user.html

UMPHRESS, D. and G. WILLIAMS (1985), Identity verification through keyboard characteristics, *International Journal of Man–Machine Studies* **23**, 263–273.

YU, E. and S. CHO (2004), Keystroke dynamics identity verification: its problems and practical solutions, *Computers & Security* **23**, 428–440.

ZENG, Z. H., Y. FU, G. I. ROISMAN, Z. WEN, Y. X. HU and T. S. HUANG (2006), One-class classification for spontaneous facial expression analysis, in: *Proceedings of the Seventh International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, pp. 281–286.

Received: August 2008. Revised: July 2009.