# A Replication of Two Free Text Keystroke Dynamics Experiments under Harsher Conditions

Nahuel González, Enrique P. Calot and Jorge S. Ierache[1]

**Abstract:** Replication of experiments lies at the very core of the scientific process. In spite of this, the relevance and relative count of replication studies has fallen sharply in the recent past in several disciplines and a similar trend is apparent in computing science and software engineering. Keystroke dynamics studies have not been exempted from this rule, where replication studies are still no more than a handful and have often lead to astonishingly varying error rates in comparison with the original study, both for static passwords and free texts. The effect of the typing environment and the user emotional state is obviously a concern for free text analysis, as much as his physiological states like stress and tiredness. Thus, the real world performance of a certain method in a specific setting could vary drastically compared with the laboratory setting even though the latter might not be consciously biased. In this paper we examine and compare the performance of two techniques for keystroke dynamics analysis in a free text dataset under evaluation conditions which are harsher than the originally used: the free text extension to the R and A distances method of Bergadano, Gunetti and Picardi and the authors' method based on finite context modeling. Some of their key properties proved to be extrapolatable outside the realm of ideal conditions; the impostor pass rate of A and R distances showed little change, combined distances proved consistently better than pure ones and their best combination remained the same. Others did not, like the relative efficacy of *n*-graphs, the performance of A distances and —expectedly— the false alarm and correct classification rate for both methods.

**Keywords:** Keystroke dynamics, Biometric authentication, Free text keystroke dynamics.

## 1 Introduction

Replication of experiments lies at the very core of the scientific process. It is both useful and necessary, not only to avoid forgery of results —which never lasts long— but also to provide confirmatory evidence. A more subtle purpose is to help establish universality, in the best case, or to find caveats and biases that might have remained unnoticed by the original researchers in the worst. In spite of this, the relevance and relative count of replication studies has fallen sharply in the recent past in several disciplines [HV96]. Even though a specific review for replication studies in security and biometrics research is lacking, a similar trend is apparent in computing science and software engineering [JG12].

Keystroke dynamics studies have not been exempted from this rule. Although this subfield of soft behavioral biometrics has been the subject of renewed interest in recent years, a trend which can be seen in an updated review [TTY13], replication studies of keystroke

[1] Laboratorio de Sistemas de Información Avanzados, Facultad de Ingeniería, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina, {ngonzalez,ecalot,jierache}@lsia.fi.uba.ar

dynamics experiments are still no more than a handful. This is rather unfortunate, because establishing universality of results has proven very difficult due to the inherent complexities of classifying this noisy behavior. Replication and implementation efforts using different training and evaluation datasets has often lead to astonishingly varying error rates in comparison with the original study, both for static passwords [KM09, CRI14] and free texts [Me11, KC15].

Setting aside ill-disposition and bad experimental setups, whose analysis is not the aim of this article, there are many practical reasons which can account for the wide variation of reported results in similar methods. It is well known that typing patterns, as most behavioral biometrics patterns, are modulated by the user emotional state significantly enough to enable the inference of the latter [ELM11]; physiological states like stress and tiredness seems to have an identically strong influence [VZS09]. The effect of the typing environment is obviously a concern for free text analysis, as it can force pauses due to external interruptions which are not linked to the user cadence and also distract him or her, thus forcing a deviation from his natural typing rhythm. Even without considering external effects, keystroke dynamics in free text has intrinsic variations due to the decision process of the following content, and the familiarity —both motor and intellectual— with the words which are being typed.

Each of the previous internal and external influences on the user's typing pattern is ultimately a source of noise. Due to this, the error rate of every method for keystroke dynamics analysis is not a fixed number but can be seen as a random variable with contributions from multiple sources [Ki12] such as the users themselves, the universe of possible impostors, their familiarity with the legitimate users, and the structure of the typing task. Thus, the real world performance of a certain method in a specific setting could vary drastically compared with the laboratory setting even though the latter might not be consciously biased.

In this paper we examine and compare the performance of two techniques for keystroke dynamics analysis in free texts under evaluation conditions which are much harsher than the originally used: the free text extension [GP05] to the R and A distances method of Bergadano, Gunetti and Picardi [BGP02] and the authors' method based on finite context modeling [GC15]. The rest of the article is organized as follows. In section 2 previous replication efforts are reviewed. In section 3 a short description of the methods and dataset is given. In section 4 the results are presented and discussed.

## 2 Previous replications and implementations

An extensive comparison of previously tested keystroke dynamics methods for password authentication by Killourhy and Maxion [KM09] over the same dataset has shown wide differences between the originally reported performance and the replication results. Also,

| | Original dataset | | | | | | |
|---|---|---|---|---|---|---|---|
| a. R Measures | | | | | | | |
| Adopted distance measure | $R_2$ | $R_3$ | $R_4$ | $R_{2,3}$ | $R_{2,4}$ | $R_{3,4}$ | $R_{2,3,4}$ |
| No. of classification errors | 13 | 44 | 61 | 5 | 9 | 29 | 9 |
| % of error | 2.16 | 7.33 | 10.16 | 0.83 | 1.5 | 4.83 | 1.5 |
| 95% CI | 1.22<br>3.57 | 5.45<br>9.63 | 7.94<br>12.78 | 0.32<br>1.82 | 0.74<br>2.72 | 3.33<br>6.77 | 0.74<br>2.72 |
| b. A Measures | | | | | | | |
| Adopted distance measure | $A_2$ | $A_3$ | $A_4$ | $A_{2,3}$ | $A_{2,4}$ | $A_{3,4}$ | $A_{2,3,4}$ |
| No. of classification errors | 44 | 84 | 133 | 41 | 39 | 81 | 41 |
| % of error | 7.33 | 14 | 22.16 | 6.83 | 6.5 | 13.5 | 6.83 |
| 95% CI | 5.45<br>9.63 | 11.40<br>16.95 | 18.98<br>25.62 | 5.02<br>9.06 | 4.73<br>8.68 | 10.94<br>16.41 | 5.02<br>9.06 |
| | Harsher replication dataset | | | | | | |
| a. R Measures | | | | | | | |
| Adopted distance measure | $R_2$ | $R_3$ | $R_4$ | $R_{2,3}$ | $R_{2,4}$ | $R_{3,4}$ | $R_{2,3,4}$ |
| No. of classification errors | 858 | 620 | 825 | 516 | 527 | 455 | 402 |
| % of error | 14.53 | 10.5 | 13.98 | 8.74 | 8.93 | 7.71 | 6.81 |
| 95% CI | 13.65<br>15.45 | 9.74<br>11.30 | 13.11<br>14.88 | 8.04<br>9.48 | 8.22<br>9.68 | 7.05<br>8.41 | 6.19<br>7.47 |
| b. A Measures | | | | | | | |
| Adopted distance measure | $A_2$ | $A_3$ | $A_4$ | $A_{2,3}$ | $A_{2,4}$ | $A_{3,4}$ | $A_{2,3,4}$ |
| No. of classification errors | 3524 | 2973 | 2407 | 2683 | 2225 | 2414 | 2214 |
| % of error | 59.7 | 50.36 | 40.78 | 45.45 | 37.69 | 40.89 | 37.51 |
| 95% CI | 53.34<br>55.88 | 49.09<br>51.64 | 39.53<br>42.03 | 44.18<br>46.72 | 36.46<br>38.93 | 39.65<br>42.15 | 36.28<br>38.75 |

Tab. 1: Experimental Results in User Classification for Different R and A Measures

some highly dissimilar methods both in their structure and complexity (i.e. SVM and z-score) happened to achieve almost similar error rates. In a later article about methodological difficulties in keystroke dynamics studies [KM11] the same authors noted, albeit disconcertingly, that *"a neural network's false-alarm rate changed from 1.0% to 85.9% from one evaluation data set to another"*. Although passwords and free texts use different approaches, the importance of testing with unrelated datasets to establish universality of results is easily seen to extrapolate to the latter case.

Distances R and A were used as the base of a continuous authentication system by Messerman *et. al.* [Me11]. The error rates turned out to be much higher than in the original presentation but a direct comparison is not fair because the authors introduce additional parameters aimed at altering the balance of false alarms and impostor passes, while the decision process is modified to fit an interactive system. What is more, the 3500 keystrokes used for initial enrollment is about one third of the amount originally tested, and each de-

| | Original / Harsher replication | | | |
|---|---|---|---|---|
| a | | | | |
| Adopted distance measure | $R_2 + A_2$ | $R_2 + A_{2,3}$ | $R_2 + A_{2,4}$ | $R_2 + A_{2,3,4}$ |
| No. of classification errors | 6 / 1381 | 4 / 1676 | 4 / 1372 | 14 / 1618 |
| % of error | 1 / 23.39 | 0.66 / 28.39 | 0.66 / 23.24 | 2.33 / 27.41 |
| 95% CI | 0.42 / 22.33<br>2.05 / 24.49 | 0.23 / 27.25<br>1.58 / 29.55 | 0.23 / 22.18<br>1.58 / 24.33 | 1.34 / 26.28<br>3.78 / 28.56 |
| b | | | | |
| Adopted distance measure | $R_{2,3} + A_2$ | $R_{2,3} + A_{2,3}$ | $R_{2,3} + A_{2,4}$ | $R_{2,3} + A_{2,3,4}$ |
| No. of classification errors | 2 / 487 | 4 / 693 | 4 / 539 | 7 / 777 |
| % of error | 0.33 / 8.25 | 0.665 / 11.74 | 0.665 / 9.13 | 1.16 / 13.16 |
| 95% CI | 0.07 / 7.57<br>1.07 / 8.97 | 0.23 / 10.94<br>1.58 / 12.58 | 0.23 / 8.42<br>1.58 / 9.89 | 0.52 / 12.32<br>2.28 / 14.04 |
| c | | | | |
| Adopted distance measure | $R_{2,3,4} + A_2$ | $R_{2,3,4} + A_{2,3}$ | $R_{2,3,4} + A_{2,4}$ | $R_{2,3,4} + A_{2,3,4}$ |
| No. of classification errors | 2 / 339 | 1 / 467 | 1 / 388 | 4 / 529 |
| % of error | 0.33 / 5.74 | 0.16 / 7.91 | 0.16 / 6.57 | 0.66 / 8.96 |
| 95% CI | 0.07 / 5.17<br>1.07 / 6.36 | 0.02 / 7.91<br>0.78 / 8.62 | 0.02 / 5.96<br>0.78 / 7.23 | 0.23 / 8.25<br>1.58 / 9.71 |

Tab. 2: Experimental Results in User Classification for Different R and A Measures Combined

cision phase used less than a fifth.

More recently, Kang and Cho compared the performance of a variety of authentication methods [KC15] for free texts, including R and A distances in their pure and combined form, over different user groups typing on physical, virtual mouse-operated and touch keyboards. A combined R+A distance achieved an equal error rate of 7.87%, being the record holder for physical keyboards yet not reaching the classification performance originally reported. Once again the figures cannot be compared directly, because the latter is an average over all sessions which lengths vary from a hundred to a thousand keystrokes and the reported rates for a fixed session size are averaged over all methods. Both pure and combined distances score considerably better in the physical keyboard than in the virtual and touch ones; the confidence intervals for their error rates do not overlap.

The epistemological basis of replication experiments in computer science and software engineering has been subject of debate [Sh08]. Whether evaluating methods with different datasets is a replication or not, and to which extent test data and implementations can or should vary has been answered differently by different authors with labels like duplication, modification, model comparisons, data re-analysis, virtual, strict, partial and systematic replication, pseudoreplication, etcetera [JG12]. We prefer to take a pragmatic approach here and leave the question aside, stating that we tried to be as faithful as possible to the

| | Original dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Adopted Distance Measure | $R_2$ | $R_{2,3}$ | $R_{2,4}$ | $R_{2,3,4}$ | $A_2$ | $A_{2,3}$ | $A_{2,4}$ | $A_{2,3,4}$ |
| No. of passed impostors | 563 | 324 | 279 | 199 | 590 | 335 | 366 | 331 |
| No. of false alarms | 50 | 32 | 41 | 41 | 92 | 80 | 84 | 79 |
| IPR (%) | 0.125 | 0.072 | 0.062 | 0.044 | 0.131 | 0.074 | 0.081 | 0.074 |
| FAR (%) | 8.333 | 5.333 | 6.833 | 6.833 | 15.333 | 13.333 | 14.0 | 13.166 |
| 95% CI IPR (%) | 0.115 | 0.064 | 0.055 | 0.038 | 0.121 | 0.067 | 0.070 | 0.066 |
| | 0.136 | 0.080 | 0.070 | 0.051 | 0.142 | 0.083 | 0.010 | 0.082 |
| 95% CI FAR (%) | 6.32 | 3.75 | 5.02 | 5.02 | 12.62 | 10.79 | 11.40 | 10.64 |
| | 10.75 | 7.35 | 9.06 | 9.06 | 18.38 | 16.23 | 16.95 | 16.05 |
| | Harsher replication dataset | | | | | | | |
| Adopted Distance Measure | $R_2$ | $R_{2,3}$ | $R_{2,4}$ | $R_{2,3,4}$ | $A_2$ | $A_{2,3}$ | $A_{2,4}$ | $A_{2,3,4}$ |
| No. of passed impostors | 58 | 40 | 41 | 40 | 569 | 377 | 208 | 280 |
| No. of false alarms | 1647 | 1115 | 1136 | 952 | 3920 | 3186 | 2652 | 2696 |
| IPR (%) | 0.054 | 0.037 | 0.038 | 0.037 | 0.532 | 0.353 | 0.195 | 0.262 |
| FAR (%) | 27.90 | 18.89 | 19.24 | 17.70 | 66.41 | 53.97 | 44.93 | 45.67 |
| 95% CI IPR (%) | 0.042 | 0.027 | 0.028 | 0.027 | 0.490 | 0.319 | 0.170 | 0.233 |
| | 0.070 | 0.050 | 0.051 | 0.050 | 0.578 | 0.390 | 0.222 | 0.294 |
| 95% CI FAR (%) | 26.77 | 17.91 | 18.25 | 15.21 | 65.19 | 52.70 | 43.66 | 44.4 |
| | 29.06 | 19.90 | 20.27 | 17.08 | 67.60 | 55.24 | 46.20 | 46.94 |

Tab. 3: Experimental Results in User Authentication for Different Distance Measures R and A

original implementations and reporting methodologies except for the purpose of evaluating them under naturally harsher conditions without introducing additional biases.

## 3 Experiment and results

### 3.1 A short warning on terminology

Please note that in the context of this article, FAR will stand for *False Alarm Rate* and IPR for *Impostor Pass Rate*. The first acronym clashes with the usual one for *False Acceptance Rate* while having a different meaning, and the second one is rarely used —if at all— nowadays. We acknowledge that this terminology is not standard; however, it was the one used by Bergadano, Gunetti and Picardi in their original articles [BGP02] [GP05]. Hopefully, the inconvenience of attaching to non standard terminology here will be offset by the possibility of comparing tables directly between the original and this replication.

| | Original/Harsher replication | | | |
|---|---|---|---|---|
| a | | | | |
| Adopted Distance Measure | $R_2 + A_2$ | $R_2 + A_{2,3}$ | $R_2 + A_{2,4}$ | $R_2 + A_{2,3,4}$ |
| No. of passed impostors | 360 / 104 | 272 / 160 | 272 / 96 | 237 / 159 |
| No. of false alarms | 36 / 1967 | 34 / 2216 | 34 / 1854 | 41 / 2110 |
| IPR (%) | 0.08 / 0.10 | 0.06 / 0.15 | 0.06 / 0.09 | 0.05 / 0.15 |
| FAR (%) | 6.0 / 33.32 | 5.67 / 37.54 | 5.67 / 31.41 | 6.83 / 35.74 |
| 95% CI IPR (%) | 0.07 / 0.08 | 0.05 / 0.13 | 0.05 / 0.07 | 0.05 / 0.13 |
| | 0.09 / 0.12 | 0.07 / 0.17 | 0.07 / 0.11 | 0.06 / 0.18 |
| 95% CI FAR (%) | 4.31 / 32.13 | 4.03 / 36.31 | 4.03 / 30.23 | 5.02 / 34.53 |
| | 8.11 / 34.53 | 7.73 / 38.78 | 7.73 / 32.60 | 9.06 / 36.97 |
| b | | | | |
| Adopted Distance Measure | $R_{2,3} + A_2$ | $R_{2,3} + A_{2,3}$ | $R_{2,3} + A_{2,4}$ | $R_{2,3} + A_{2,3,4}$ |
| No. of passed impostors | 235 / 48 | 205 / 56 | 205 / 46 | 185 / 62 |
| No. of false alarms | 26 / 1039 | 24 / 1264 | 24 / 1057 | 30 / 1330 |
| IPR (%) | 0.052 / 0.045 | 0.046 / 0.052 | 0.046 / 0.043 | 0.041 / 0.058 |
| FAR (%) | 4.33 / 17.60 | 4.0 / 21.41 | 4.0 / 17.91 | 5.0 / 22.53 |
| 95% CI IPR (%) | 0.05 / 0.03 | 0.04 / 0.04 | 0.04 / 0.03 | 0.04 / 0.04 |
| | 0.06 / 0.06 | 0.05 / 0.07 | 0.05 / 0.06 | 0.05 / 0.07 |
| 95% CI FAR (%) | 2.92 / 16.65 | 2.65 / 20.38 | 2.65 / 16.94 | 3.47 / 21.48 |
| | 6.19 / 18.59 | 5.80 / 22.47 | 5.80 / 18.90 | 6.96 / 23.61 |
| c | | | | |
| Adopted Distance Measure | $R_{2,3,4} + A_2$ | $R_{2,3,4} + A_{2,3}$ | $R_{2,3,4} + A_{2,4}$ | $R_{2,3,4} + A_{2,3,4}$ |
| No. of passed impostors | 124 / 48 | 78 / 40 | 78 / 42 | 131 / 45 |
| No. of false alarms | 19 / 824 | 23 / 947 | 23 / 848 | 23 / 1013 |
| IPR (%) | 0.038 / 0.045 | 0.017 / 0.037 | 0.017 / 0.039 | 0.029 / 0.042 |
| FAR (%) | 3.17 / 13.96 | 3.83 / 16.04 | 3.83 / 14.37 | 3.83 / 17.16 |
| 95% CI IPR (%) | 0.02 / 0.03 | 0.01 / 0.03 | 0.01 / 0.03 | 0.02 / 0.03 |
| | 0.03 / 0.06 | 0.02 / 0.05 | 0.02 / 0.05 | 0.03 / 0.06 |
| 95% CI FAR (%) | 1.98 / 13.09 | 2.51 / 15.12 | 2.51 / 13.49 | 2.51 / 16.22 |
| | 4.80 / 14.86 | 5.60 / 17.00 | 5.60 / 15.28 | 5.60 / 18.14 |

Tab. 4: Experimental Results in User Authentication for Different R and A Measures Combined

### 3.2 Dataset

The dataset used for the current replication experiments was acquired in a genuine real world environment. Keystroke capture began in December 17, 2014 and extended until December 24, 2015, spanning more than a year. Users from both sexes range between 28 and 60 years old; their typing skills varied between clumsy and proficient, but none was a single finger typist. Their identity was verified through passwords or conventional biometric means before free texts were captured to avoid mislabeling. Except for the fact that some foreign or technical words might appear sporadically in the text, all content was fully

written in the native language of the subjects. Mixing languages in profiles and evaluation sessions has been shown to degrade classification performance [GPR05b] but our dataset can be considered safe in that sense.

Many of the sessions did not have enough keystrokes to be considered for evaluation due to the length requirement of having at least 700 characters and less than 900 (see next subsection for details). Since the task is unrestricted, some keystrokes might correspond to arrows and other special keys used for edition and navigation; in order not to bias the results negatively, the keystroke count referred above considers only alphanumeric characters. A count of 20 users and 5903 sessions remains after considering the length restriction.

The motives to consider the dataset "harsh" (or at least "realistic") are diverse. First of all, it was acquired in a production environment and not in a laboratory setting. The experimental subjects were not chosen from a universitary population. Users had no fixed office or computer terminal assigned, thus the keyboard used in each session might vary. Some of them worked extended or rotating shifts; the effects of tiredness and stress in the subjects cannot be factored out. Interruptions in their work are common. All things considered, the dataset can be described as being "difficult" or "harsh" in comparison to similar ones captured in laboratory environments, fitting the purpose of the experiment. As will be seen in section *Key observations*, the fact that every combination of every method performs worse than in the original experiments confirms this intuitive appreciation. However, we are not aware of an objective way of measuring such difficulty —a quality vector— in keystroke dynamics datasets, as has been already explored for other biometric modalities [GT07] like face, fingerprint and iris recognition.

### 3.3 Experiment

A replication experiment of Gunetti and Picardi's method was performed as described in [GP05] using the previously characterized dataset. Like the original experiment, session length was restricted to 700–900 keystrokes and 14 sessions were used as the initial template for the user. Using old sessions in the user profiles has been shown by the same authors [GPR05a] to increase error rates due to the inevitable drift in keystroke dynamics as the user typing habits naturally evolve, thus demanding a strategy to deal with them as the data acquisition for the evaluation set lasted for more than a year. In order to remain as faithful as possible to the original implementation we simply ordered the sessions with respect to their recoding timestamp, separated the oldest 14 to create the initial profile and evaluated all the rest in increasing time order updating the profiles to reflect only the most recent sessions. Profiles then become sliding windows which are tested only against sessions near in time, avoiding the aforementioned effect of aging. No additional impostors where added to the 20 legitimate users but each session not used for initial profiles was used to attack every other profile, adding up to $19 \times (5903 - 14 \times 20) = 106837$ impostor login attempts. This calculation follows [GP05].

| Metric | Distance | Average | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| EER | Scaled $\mathscr{L}_1$ | 11.20% | 4.82 | 2.13% | 21.61% |
| | Scaled $\mathscr{L}_2$ | 23.55% | 9.77 | 9.21% | 43.63% |
| | Outliers | 10.63% | 4.92 | 2.12% | 21.37% |
| | Est. best | 10.44% | 4.82 | 2.12% | 21.37% |
| IPR at FAR=13.91% | Scaled $\mathscr{L}_1$ | 10.10% | 10.51 | 0.44% | 35.16% |
| | Scaled $\mathscr{L}_2$ | 41.44% | 23.55 | 3.51% | 85.83% |
| | Outliers | 8.74% | 9.35 | 0.14% | 33.67% |
| | Est. best | 8.13% | 8.99 | 0.14% | 35.16% |

Tab. 5: Replication results for finite context modeling method

To simplify comparison between the original experiment and this replication, tables 1 to 4, and 6, in this article show the same information as in [GP05] with same structure, including original results and the ones obtained with our evaluation dataset: error count and rate in user classification for different R and A measures, pure (table 1) and combined (table 2), passed impostor and false alarm count and rate for a selection of R and A measures, again pure (table 3) and combined (table 4). Confidence intervals that show the statistical significance of the results where added to avoid the common pitfall [KM11] of assuming pointwise error rates.

The replication of our method based on finite context modeling was performed as described in [GC15]. No other provisions apply. Results for equal error rates of the four originally tested classifiers are shown in table 5; a second box which will be used for comparison in the next subsection shows IPR at FAR=13.91%.

### 3.4 Key observations

A few observations which will prove useful in the next subsection can be extracted directly from the results data. Table 7 shows a summary of the results for both datasets and whether they generalize to the harsher one.

- *The dataset used in this study is harsher than the ones used for the original evaluation of the methods*. This should be obvious from the previous description, but it is empirically confirmed by the fact that every variation of both R and A distance and finite context modeling methods performs worse than in the original experiments.
- *Performance of A distances degrades less gracefully than that of R distances with the harsher dataset*. The error rates in classification for A distances are on average 3.2 times higher than those of R distances in the original paper, but the figure rises to 4.6 in the current replication. For the authentication experiment, the respective figures are 2.0 and 2.6. Every A distance scores worse error rates than every R distance in the replication experiment, which is not the case in the original.

- *The relative efficacy of digraphs, trigraphs and* 4*-graphs for both distances varies with the dataset*. In the original experiment, for pure A and R distances, the error rate increases with the size $n$ of the $n$-graph. With the harsher dataset it is no longer the case, with $R_3 < R_4 < R_2$ and $A_4 < A_3 < A_2$, even changing their relative orders. Note that the confidence intervals of $R_2$ and $R_4$ overlap, so their order cannot be established for sure.
- *Combining A and R distances consistently improves classification and authentication performance*. In both the original experiment and the replication lower error rates result from combining A and R measures than those obtained separately. What is more, the optimal combination is consistently $R_{2,3,4}+A_2$ for best FAR and $R_{2,3,4}+A_{2,3}$ for best IPR, while being almost the same for classification, changing $R_{2,3,4}+A_{2,3}$ to $R_{2,3,4}+A_2$ (in the original experiment they differ in only one misclassified session, thus the difference is statistically negligible).
- *Impostor pass rates remain consistently low* in the authentication experiment replication, well under 1% and even 0.1% for most combinations of R and A distances, even though the false alarm rates increase between 2.4 and 3.4 times.
- *The relative efficacy of scaled manhattan distance, scaled euclidean distance and outlier count remain constant both for datasets* while using finite context modeling. Best classifier estimation still beats every simple strategy.

### 3.5 Method comparison

| | Harsher replication dataset | | | |
|---|---|---|---|---|
| Length of Samples | $R_{2,3,4}+A_2$ | | $R_{2,3,4}+A_{2,3}$ | |
| | IPR (%) | FAR (%) | IPR(%) | FAR (%) |
| 1/4 | 0.1797 | 27.19 | 0.2096 | 30.33 |
| 2/4 | 0.0561 | 17.79 | 0.0861 | 19.46 |
| 3/4 | 0.0196 | 17.21 | 0.0252 | 17.48 |
| 4/4 | 0.0004 | 13.96 | 0.00037 | 16.04 |

Tab. 6: Authentication results for different length of the samples

The difference in reporting methodology makes immediate comparison of both methods impossible. For this purpose we include in table 5 a second box of results reporting impostor pass rate at a false alarm rate of 13.91%; this value was chosen to compare the best performing combination of distances, $R_{2,3,4}+A_2$, in the authentication experiment, which happens to reach the former false alarm rate. Our method is far from reaching such spectacularly low impostor pass rates as the one of Bergadano, Gunetti and Picardi as thus would seem inferior at first glance. However, some subtleties should be explored before a final conclusion is reached.

As described by the original authors, the authentication method based on A and R distances is indeed an identification task augmented by a mean distance check with a threshold. The identification stage balances the method towards security by exchanging a higher false alarm rate for a diminished impostor pass rate very efficiently but it inevitably introduces a dependency of the error rates with the amount of users, the quality of their profiles, the individuality of their typing rhythms and the breadth of their variations. Related issues are discussed in the section 6.5 of the original article [GP05]. On the other hand, the authentication process of finite context modeling is based only on the profile of the user being authenticated and thus the results can be extrapolated to any amount of users.

Also, a session length greater than 700 keystrokes favors the methods based on A and R distances, since shorter typing samples increase the error rates rapidly below that point, as can be seen in table XIII of the original article and table 6 in this one. However, a near optimal authentication performance is reached near 150 keystrokes with finite context modeling and a better compromise between error rates can be achieved if user experience (which is very sensitive to false alarms) and not only security is considered as a requirement.

| **Property** | **Original** | **Harsh** | **Generalizes?** |
|---|---|---|---|
| A:R (classification) | 3.2 | 4.6 | ✗ |
| A:R (authentication) | 2.0 | 2.6 | ✗ |
| Relative error rate of n-graphs | $R_2 < R_3 < R_4$ | $R_3 < R_4 < R_2$ | ✗ |
| | $A_2 < A_3 < A_4$ | $A_4 < A_3 < A_2$ | ✗ |
| Combined distances | combined $<$ pure | | ✓ |
| Best class. method | $R_{2,3,4} + A_{2,3}$ | $R_{2,3,4} + A_2$ | $\approx$ ✓ |
| Best class. rate | 0.16% | 5.74% | ✗ |
| Best IPR method | $R_{2,3,4} + A_{2,3}$ | | ✓ |
| Best FAR method | $R_{2,3,4} + A_2$ | | ✓ |
| Best IPR value | $\approx 0.05\%$ | | ✓ |
| Best FAR value | $\approx 3\%$ | $\approx 14\%$ | ✗ |
| FCM error rate | Best $<$ Outlier count $< \mathscr{L}_1 < \mathscr{L}_2$ | | ✓ |

Tab. 7: Summary of properties and their generalizability

## 4 Concluding remarks

Reviews and methodological critiques on keystroke dynamics literature usually state the difficulty of comparing experiments due not only to different reporting methodologies and lack of replications and comparisons but also to the inevitable choice of parameters like session length and the composition of the session dataset, considering not only the users and their typing skills but also the environment of data collection. A parameter which is usually ignored in studies, but is relevant for real world implementations is the “difficulty” of the dataset and whether it matches expected challenges outside the laboratory environment; although it is intuitively clear what both terms mean they can only be measured

indirectly through the error rates of the considered methods. In this paper two methods previously implemented were tested against a harsher set of sessions than the originally used for evaluation and their results were compared in spite of their differences. Some of their key properties proved to be extrapolatable outside the realm of ideal conditions; the impostor pass rate of Bergadano, Gunetti and Picardi method showed very little change —surprisingly regarding the increased difficulty—, combined distances proved consistently better than pure ones and their best combination remained the same. Others did not, like the relative efficacy of $n$-graphs, the performance of A distances and —expectedly— the false alarm and correct classification rate for both methods.

### 4.1 Future research

Our harsher dataset will be used to contrast results on new finite context modeling algorithms as we planned future lines of research on dataset exploitation applied to information security and emotional state detection.

## References

[BGP02] Bergadano, F.; Gunetti, D.; Picardi, C.: User authentication through keystroke dynamics. ACM Transactions on Information and System Security (TISSEC), 5(4):367–397, 2002.

[CRI14] Calot, Enrique P.; Rodríguez, Juan Manuel; Ierache, Jorge Salvador: Improving versatility in keystroke dynamic systems. In (Finochietto, Jorge Raúl; Pesado, Patricia Mabel, eds): Computer Science & Technology Series. XIX Argentine Congress of Computer Science, Selected papers, pp. 289–298. Editorial de la Universidad Nacional de La Plata (EDULP), 2014.

[ELM11] Epp, Clayton; Lippold, Michael; Mandryk, Regan L: Identifying emotional states using keystroke dynamics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 715–724, 2011.

[GC15] Gonzalez, Nahuel; Calot, Enrique P: Finite Context Modeling of Keystroke Dynamics in Free Text. In: Biometrics Special Interest Group (BIOSIG), 2015 International Conference of the. IEEE, pp. 1–5, 2015.

[GP05] Gunetti, Daniele; Picardi, Claudia: Keystroke analysis of free text. ACM Transactions on Information and System Security (TISSEC), 8(3):312–347, 2005.

[GPR05a] Gunetti, Daniele; Picardi, Claudia; Ruffo, Giancarlo: Dealing with different languages and old profiles in keystroke analysis of free text. In: AI* IA 2005: Advances in Artificial Intelligence, pp. 347–358. Springer, 2005.

[GPR05b] Gunetti, Daniele; Picardi, Claudia; Ruffo, Giancarlo: Keystroke analysis of different languages: A case study. In: Advances in Intelligent Data Analysis VI, pp. 133–144. Springer, 2005.

[GT07] Grother, Patrick; Tabassi, Elham: Performance of biometric quality measures. IEEE transactions on pattern analysis and machine intelligence, 29(4):531–543, 2007.

[HV96] Hubbard, Raymond; Vetter, Daniel E: An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. Journal of Business Research, 35(2):153–164, 1996.

[JG12] Juristo, Natalia; Gómez, Omar S: Replication of software engineering experiments. In: Empirical software engineering and verification, pp. 60–88. Springer, 2012.

[KC15] Kang, Pilsung; Cho, Sungzoon: Keystroke dynamics-based user authentication using long and free text strings from various input devices. Information Sciences, 308:72–93, 2015.

[Ki12] Killourhy, Kevin S: A scientific understanding of keystroke dynamics. Technical report, DTIC Document, 2012.

[KM09] Killourhy, Kevin S; Maxion, Roy A: Comparing anomaly-detection algorithms for keystroke dynamics. In: Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on. IEEE, pp. 125–134, 2009.

[KM11] Killourhy, Kevin S; Maxion, Roy A: Should security researchers experiment more and draw more inferences? In: CSET. 2011.

[Me11] Messerman, Arik; Mustafic, Tarik; Camtepe, Seyit Ahmet; Albayrak, Sahin: Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. In: Biometrics (IJCB), 2011 International Joint Conference on. IEEE, pp. 1–8, 2011.

[Sh08] Shull, Forrest J; Carver, Jeffrey C; Vegas, Sira; Juristo, Natalia: The role of replications in empirical software engineering. Empirical Software Engineering, 13(2):211–218, 2008.

[TTY13] Teh, Pin Shen; Teoh, Andrew Beng Jin; Yue, Shigang: A survey of keystroke dynamics biometrics. The Scientific World Journal, 2013, 2013.

[VZS09] Vizer, Lisa M.; Zhou, Lina; Sears, Andrew: Automated stress detection using keystroke and linguistic features: An exploratory study. International Journal of Human-Computer Studies, 67(10):870–886, 2009.