Keyboard Behavior Based Authentication for Security

Patrick Juola
Juola & Associates
pjuola@juolaassociates.com

John I. Noecker Jr.
Juola & Associates
jnoecker@juolaassociates.com

Ariel Stolerman
PSAL, Drexel University
ams573@cs.drexel.edu

Michael V. Ryan
Juola & Associates
mryan@juolaassociates.com

Patrick Brennan
Juola & Associates
pbrennan@juolaassociates.com

Rachel Greenstadt
PSAL, Drexel University
greenie@cs.drexel.edu

Abstract

We developed a large corpus of keyboard behavior based on temporary workers employed in a simulated office environment. Analysis of this corpus using stylometric techniques shows good accuracy in distinguishing users.

1 Introduction

How do you know who's at your computer? The problem of the keyboard dilemma (Chaski, 2005) is a major issue in computer security, as once you've logged in, the computer assumes that the user continues to be you. A better system would use "active authentication" to look at how the person is using the computer, and see if their usage pattern matches yours.

2 Background

One of the most important ways of interacting with a computer is through the keyboard (and mouse), in part because there are so many different ways to interact. Keyboard interaction includes not just behavioral data (like typing speed) but also cognitive and linguistic data as well. Language use has been successfully used to infer the authorship of written documents (Juola, 2006; Koppel et al., 2009; Stamatatos, 2009; Jockers and Witten, 2010) although it has not typically been used for authentication before. The theory behind this field is that everyone has their own unique "stylome" (van Halteren et al., 2005), a unique set of idiolectal choices that describe their speaking and writing style. At a group level, this is the kind of choice that causes Brits to walk on "pavements" instead of "sidewalks," and at an individual level can be the kind of choice that causes you to place a fork "to" the left of the plate instead of "on" the left or "at" the left. Quantifying these choices, for example, by making a histogram of function words (Mosteller and Wallace, 1964; Binongo, 2003) or of character n-grams enables investigators to develop a computationally tractable summary of stylistic choices and form judgments based on these summaries.

The application of this technology to authentication is fairly straightforward. Instead of using a training set of documents, one uses a pseudo-document containing the user's long-term behavior, and "verifies" that the recent behavior at the keyboard is consistent with this long-term behavior. A significant inconsistency, of course, would trigger a security response. We describe preliminary findings towards such a system.

3 Materials and Methods

3.1. Subjects and Environment

Any experiment in authentication requires as a base a set of validated data attached to ground truth identities; the simplest way to collect this data was simply to create an environment. We set up a simulated office environment in downtown Pittsburgh and hired temporary workers for a week at a time, presenting them with a set of writing tasks and a standardized computer. On each computer was tracking software including Free Key Logger and GhostSpy (which measured keyboard dynamics, including keypresses and timing; mouse movements and clicks; applications launched and the mapping of text to applications, the use and text of the clipboard, and browsing history). Over the course of 12 weeks, 80 temporary workers were hired to perform a long-term blogging task (research and write blog articles on topics "related to Pittsburgh in some way") over the course of a normal work-week. The last two hours of every day were devoted to "microtasks," small, focused, explicitly-defined tasks such as describing a local landmark or event, which provided data subsets on more directly comparable tasks. We report here on an initial analysis of 63 subjects working over 300 person-days, creating more than 4.2Gb in data. The most commonly used applications were, unsurprisingly, Internet Explorer and Microsoft Word; the most commonly visited sites included www.bing.com, www.google.com, search.yahoo.com, www.facebook.com, dell.msn.com, www.pandora.com, en.wikipedia.org, www.youtube.com, disneyworld.disney.go.com, and www.yahoo.com among nearly 100,000 total sites visited. Looking only at the content generated in Microsoft Word, users generated 1057 documents, containing 47,411 words/2,969,814 characters. We consider this to be among the largest corpora developed for this kind of authentication purpose.

In addition to the raw interaction data, subjects were also asked to take a battery of psychometric tests, including a basic demographic survey (incorporating *inter alia* gender, education level, native language, age, and dominant hand), the Rosenberg Self-Esteem Scale, the Myers-Briggs Personality Inventory (MBTI), the NEO-R, the Multiple Intelligences Developmental Assessment Scales (MIDAS), and the Learning Styles Inventory. These provided higher-level psychometric classifications that we plan to use to further refine the eventual authentication system.

3.2. Materials

The corpus described above was processed to extract various linguistic and behavioral features of interest, features that have proved useful in identification, forensic, or security contexts (Chaski, 2005; Juola, 2006; Koppel et al., 2009; Stamatatos, 2009; Jockers and Witten, 2010; Zheng et al., 2011). Examples of these features include lexical statistics such as word length, characters, character bigrams, percentage of letters/digits/uppercase letters, and numbers, including single digits, 2-digit numbers, and 3-digit numbers. Syntactic statistics include function words and part-of-speech tags as well as n-grams; content features include common words as well as word n-grams. Behavioral features included domain visits, keyboard intervals and dwell times, and mouse dynamics such as direction, distance, curvature angle and distance, and button dwell time.

Depending upon the features, various forms of pre- and post-processing were applied. For example, special [non-printable] characters were replaced by printable placeholders (for example, BACKSPACE was replaced by β), while case unification was applied to character n-grams. To increase the semantic coherent of content features, special characters were applied; for example “heββHeloβlo” becomes “hello.” Some degree of normalization was also applied, for example, calculating frequencies divided by the total window size from which the windows were generated.

3.3. Methods

The data described above was independently analyzed through several different methods. One set of analyses used the JGAAP stylometric analysis (Juola, 2006; Juola et al., 2009) using a centroid-based 1-nearest neighbor classification scheme. Each analysis reserved one day’s worth of keystrokes to be identified and the other four days of keystrokes for that person as positive examples of training data, while using the other 62 subjects as training examples of distractor data. Features used for these analyses included (case-unified) character bi- and trigrams as well as the 100 most common words. In addition to analysis for identity, we report also on analyses for MBTI personality type, gender, and dominant hand.

A second set of analyses were performed using the JStylo framework[1] using smaller amounts of captured text based on a sliding window of 500 or 1000 characters. These character sets were analyzed with a one multiclass SMO Support Vector Machine with a polynomial kernel (derived from WEKA (Hall et al., 2009)).

[1] http://psal.cs.drexel.edu/

4 Results and Discussion

Results are attached as table 1.

Task	Class(es)	Features	#Subj.	Analysis	Data Size	Accuracy
Identification	Subject ID	Writeprints	79	1-vs.-1 SMO SVM 10-fold CV	500 chars	Exact: 55.3%
						In top 3: 70.3%
					1000 chars	Exact: 63.98%
						In top 3: 79.3%
		Clear bigrams	60	1-vs.-all, L_1 dist. Leave-one-out CV	1 day	Exact: 19.6%
						In top 3: 85.3%
MBTI	Extrovert/ Introvert	Char bigrams	60	1-vs.-all, Jaccard dist. Leave-one-out CV	1 day	78.4%
	Sensing/ iNtuition					73.0%
	Thinking/ Feeling					76.5%
	Judging/ Perceiving					80.1%
Gender	M/F	100 MF Words				76.6%
Dominant Hand	Right/Left	Char bigrams (no case unification)				99.2%
	Right/Left/ Ambidex.					96.2%

Table 1: Accuracy of active stylometric analyses

In general, they show that authentication can be performed using keyboard behavior with extremely high accuracy even before sophisticated fusion techniques are applied; in that sense, we have achieved what we consider to be a proof-of-concept of the viability of this approach. Although our system is not yet at the accuracy levels that a commercial product would demand, we expect as a matter of course to improve our accuracy as we incorporate more and more features – and in particular the higher-order cognitive features – into our system.

A key ongoing aspect of this research is the application of decision fusion techniques to improve detection accuracy. In general, we apply a set of weights to the various features in order to develop the most accurate classifier in accordance with the Neyman-Pearson Decision Rule (Thomopoulos et al., 1987). As a proof of concept, our initial analysis focused on mouse and keyboard sensors (Zheng et al., 2011), specifically on mouse movements (direction and distance), button press intervals, and button dwell times. More detailed (ongoing) analyses incorporate mouse curvature angle and distance, keystroke interval and dwell times, website visit frequency, and some of the traditional stylistic features as described above. With this fusion, accuracy can be increased for identity to over 99%.

One major concern at this point is scaling; how many minutes (or seconds) of data are necessary to make an identification? Alternatively, how responsive can we make the system while maintaining acceptable accuracy levels? Clearly, needing a full day's worth of data is too much; even 30 minutes is probably too long in a high-security environment. On the other hand, we understand DARPA's ultimate vision to be a fusion of many different modalities beyond what we ourselves have studied (creating, of course, even more opportunities for data and decision fusion) which will mitigate this issue. A more serious concern, however, is the adversarial user. A sufficiently sophisticated user might, for example, fake the writing style of a real, authorized, user, or use replay-style attacks to duplicate the mouse and keyboard timing from a record of a real session. The analysis of such adversaries is a key aspect of any real-world security system and a key research topic going forward.

5 Conclusions

DARPA's vision[2] is that computer-captured biometrics can be “used to uniquely recognize humans” with high accuracy and minimal intrusiveness. The work presented in this paper confirms this. Using only standard computers and off-the-shelf logging software, we have been able to create a detailed corpus of more than 4.2 GB of computer use data obtained in the course of a normal (simulated) office environment, with detailed data about nearly 80 participants including a variety of psychometric analyses of these participants. Using this data, we are able to authenticate one particular person among this set with as little as ten minutes of data using a wide variety of different data sources, ranging from purely behavioral such as mouse movements to high-order linguistic content analysis such as use of parts-of-speech in typing.

6 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. OCI-1032683 and by DARPA under BAA-12-06. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or DARPA.

References

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17.

Chaski, C. E. (2005). Who's at the keyboard: Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Jockers, M. L. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *LLC*, 25(2):215–23.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Juola, P., Noecker Jr., J., Ryan, M., and Speer, S. (2009). JGAAP 4.0 – A revised authorship attribution tool. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

[2] BAA-12-06

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–56.

Thomopoulos, S., Viswanathan, R., and Bougoulias, D. (1987). Optimal decision fusion in multiple sensor systems. *IEEE Transactions on Aerospace and Electronic Systems*, pages 644–653.

van Halteren, H., Baayen, R. H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Zheng, N., Paloski, A., and Wang, H. (2011). An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS '11, pages 139–150, New York, NY, USA. ACM.

The abstract is reproduced here:

We developed a large corpus of keyboard behavior based on temporary workers employed in a simulated office environment. Analysis of this corpus using stylometric techniques shows good accuracy in distinguishing users.

Suggested keywords include: authentication, stylometry, profiling, applied psychometrics

References are reproduced here:

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17.

Chaski, C. E. (2005). Who's at the keyboard: Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):n/a. Electronic-only journal: http://www.ijde.org, accessed 5.31.2007.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Jockers, M. L. and Witten, D. (2010). A comparative study of machine learning methods for authorship attribution. *LLC*, 25(2):215–23.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Juola, P., Noecker Jr., J., Ryan, M., and Speer, S. (2009). JGAAP 4.0 – A revised authorship attribution tool. In *Proceedings of Digital Humanities 2009*, College Park, MD.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–56.

Thomopoulos, S., Viswanathan, R., and Bougoulias, D. (1987). Optimal decision fusion in multiple sensor systems. *IEEE Transactions on Aerospace and Electronic Systems*, pages 644–653.

van Halteren, H., Baayen, R. H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Zheng, N., Paloski, A., and Wang, H. (2011). An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS '11, pages 139–150, New York, NY, USA. ACM

Author bios follow:

Patrick Juola is the Director of Research, CEO, and a Founder of Juola & Associates, a text analysis firm in Pittsburgh, PA. He is also Associate Professor of Computer Science at Duquesne University, also in Pittsburgh. He holds a double B.S. in E.E. and mathematics from the Johns Hopkins University, an M.S. and Ph.D. in computer Science from the University of Colorado, and worked as a post-doc in experimental psychology at the University of Oxford.   His research interests include security, forensic text analysis, and humanities computing.

John I. Noecker Jr. is a Founder and Staff Scientist at Juola & Associates, a text analysis firm in Pittsburgh, PA. He holds a B.S. in Computer Science and a B.A. in Mathematics from Duquesne University. He is the former technical lead of the JGAAP authorship attribution software package, and now serves as the resident machine learning expert at Juola & Associates. His research interests include authorship attribution, author profiling, and distractorless authorship verification technology.

Ariel Stolerman is a PhD student and research assistant at the Privacy, Security and Automation laboratory in the Department of Computer Science at Drexel University in Philadelphia, PA. He holds a B.S. in Computer Science from Tel-Aviv University in Israel, and an M.S. in Computer Science from Drexel University. His research interests include applications in security and privacy, applied machine learning and text analysis.

Michael Ryan is a Founder and Staff Scientist at Juola & Associates, a text analysis firm in Pittsburgh, PA. He holds a B.S. in Computer Science and a B.A. in Mathematics from Duquesne University. He is the former head of quality control for the JGAAP authorship attribution software package, and now serves as the chief software architect at Juola & Associates. His research interests include authorship verification, high performance computing, and data mining.

Patrick Brennan is the President of Juola & Associates, a text analysis firm in Pittsburgh, PA. He holds a double B.S. in computer science and multimedia from Duquesne University, an M.B.A and an M.S. in management information system from the University of Pittsburgh Katz School of Business.

Rachel Greenstadt is an Assistant Professor of Computer Science at Drexel University. Dr. Greenstadt's research centers on the privacy and security properties of multi-agent systems and the economics of electronic privacy and information security. Her lab -- the Privacy, Security, and Automation Laboratory (PSAL) -- focuses on designing more trustworthy intelligent systems that act autonomously and with integrity, so that they can be trusted with important data and decisions. The lab takes a highly interdisciplinary approach to this research, incorporating ideas from artificial intelligence, psychology, economics, data privacy, and system security. However, a common thread of this work has been studying information flow, trustworthiness, and control. Recently, much of the work has focused on using machine learning to better understand textual communication.

Contact addresses:

Patrick Juola
EVL Laboratory
Duquesne University
600 Forbes Avenue
Pittsburgh, PA 15236
(412)396-2276 (tel)
(412)396-2269 (fax)
pjuola@juolaassociates.com

John I. Noecker Jr
199 Dock Street
Schuylkill Haven, Pennsylvania 17972
Office: (412) 567-8754
Fax: (888) 958-5476
Email: jnoecker@juolaassociates.com

Ariel Stolerman
Dept. of Computer Science
Drexel University

3175 JFK Blvd. Room 144
Philadelphia, PA 19104
Tel (215) 895-2920
Fax (215) 895-0545
ams573@cs.drexel.edu

Michael Ryan
Juola & Associates
276 W Schwab Ave.
Munhall, PA 15120
Office: (412) 475-8548
E-mail: mryan@juolaassociates.com
No Fax

Patrick Brennan
Juola & Associates
276 W Schwab Ave.
Munhall, PA 15120
(412)396-2276 (tel)
(412)396-2269 (fax)
pbrennan@jgaap.com

Rachel Greenstadt
3175 JFK Blvd room 140
Philadelphia, PA 19104
(818)825-0302 (tel)
(215) 895-0545 (fax)
Rachel.a.greenstadt@drexel.edu

There are no figures in the manuscript.

Copyright release form is being sent under separate cover.