



Annual ADFSL Conference on Digital Forensics, Security and Law

2016
Proceedings

May 24th, 1:00 PM

Using Computer Behavior Profiles to Differentiate between Users in a Digital Investigation


Shruti Gupta

Indiana University Purdue University Indianapolis

Marcus Rogers

Purdue University, rogersmk@purdue.edu

Follow this and additional works at: <https://commons.erau.edu/adfsl>

 Part of the [Aviation Safety and Security Commons](#), [Computer Law Commons](#), [Defense and Security Studies Commons](#), [Forensic Science and Technology Commons](#), [Information Security Commons](#), [National Security Law Commons](#), [OS and Networks Commons](#), [Other Computer Sciences Commons](#), and the [Social Control, Law, Crime, and Deviance Commons](#)

Scholarly Commons Citation

Gupta, Shruti and Rogers, Marcus, "Using Computer Behavior Profiles to Differentiate between Users in a Digital Investigation" (2016). *Annual ADFSL Conference on Digital Forensics, Security and Law*. 9. <https://commons.erau.edu/adfsl/2016/tuesday/9>

This Peer Reviewed Paper is brought to you for free and open access by the Conferences at Scholarly Commons. It has been accepted for inclusion in Annual ADFSL Conference on Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

EMBRY-RIDDLE
Aeronautical University™
SCHOLARLY COMMONS

(c)ADFSL



USING COMPUTER BEHAVIOR PROFILES TO DIFFERENTIATE BETWEEN USERS IN A DIGITAL INVESTIGATION

Shruti Gupta, Marcus Rogers
Purdue University

ABSTRACT

Most digital crimes involve finding evidence on the computer and then linking it to a suspect using login information, such as a username and a password. However, login information is often shared or compromised. In such a situation, there needs to be a way to identify the user without relying exclusively on login credentials. This paper introduces the concept that users may show behavioral traits which might provide more information about the user on the computer. This hypothesis was tested by conducting an experiment in which subjects were required to perform common tasks on a computer, over multiple sessions. The choices they made to complete each task was recorded. These were converted to a 'behavior profile,' corresponding to each login session. Cluster Analysis of all the profiles assigned identifiers to each profile such that 98% of profiles were attributed correctly. Also, similarity scores were generated for each session-pair to test whether the similarity analysis attributed profiles to the same user or to two different users. Using similarity scores, the user sessions were correctly attributed 93.2% of the time. Sessions were incorrectly attributed to the same user 3.1% of the time and incorrectly attributed to different users 3.7% of the time. At a confidence level of 95%, the average correct attributions for the population was calculated to be between 92.98% and 93.42%. This shows that users show uniqueness and consistency in the choices they make as they complete everyday tasks on a system, and this can be useful to differentiate between them.

Keywords: computer behavior users, interaction, investigation, forensics, graphical interface, windows, digital

1. INTRODUCTION

In today's electronic age, it would be difficult to find people whose lives do not include any kind of digital devices. In 2011, 76% Americans reported having a computer at home and 72% reported home internet use [File and Ryan, 2014]. Any crime scene that

includes a cell phone, a computer, or even a simple calculator, is dealing with digital evidence. 80-90% of crimes deal with digital evidence of some kind [Rogers et al., 2007]. Digital crimes often use artifacts on the computer as clues to identify the culprit. In most cases, the computer user can be identified using account or login informa-

tion. However, login information is not a very reliable method for user identification since this information is often shared between individuals, resulting in claims of compromise. Hence digital investigations potentially require a means of identifying computer users without the use of login information. To illustrate this, consider the situation in which User A was logged into User B's account when a crime was committed [Peisert et al., 2008]. This was seen in the case of *United States v. Keith Moreland* [Mor, 2011]. The district court's judgment convicting Moreland was reversed in the appellate court because there was no way for the police to determine which one of the family members was using the computer when the illegal images were received. There needs to be a way for law-enforcement to connect the "user-name" to the actual user.

The authors suggest that one can classify users on the same account of a system by their usage patterns. "Usage" in the context of this study refers to how everyday tasks are performed on the computer by the user. The data that was collected of a person's usage was used to create a user 'profile.' These user-profiles were analyzed to identify if they can be useful in distinguishing between two different users on a computer. A behavior profile or a user-profile can be compared to a criminal profile where the person's traits are used to narrow down suspects [Abraham and de Vel, 2002]. This research study identifies a gap in digital investigation procedures and provides the behavioral research that can be exploited in the future to fill this gap.

2. RELATED WORK

The task of modeling user behavior on the computer without relying on login information has the same goals as studies that deal with masquerade attacks. The

way a person interacts with the computer (i.e., a person's computer usage) has been used to differentiate people in earlier studies [Gamboa and Fred, 2004]. Other studies have used computer usage to identify a user, employ methods like system interaction profiling (e.g., command line profiling), keyboard usage analysis, mouse movement analysis and Graphical User Interface (GUI) usage analysis; however, this study attempts to see if there are broad metrics such as the choices users make as they interact with the computer, that can be used to model their behavior on a computer. Without significant research in this area within the psychological domain, the related work focuses on studies within the computer security realm that deal with user behavior to detect masquerade attacks.

2.1 System Interaction Profiling

Lane and Brodley [Lane and Brodley, 1997] proposed a method by which a user's current behavior is matched to a historical profile consisting of the command line history of the user as the main variable. Schonlau et al. [Schonlau et al., 2001] conducted a research study in command line profiling using 50 users, collecting approximately 15,000 commands from each user. The best results that the study reported had a success rate of 69.3%. Maxion and Townsend [Maxion and Townsend, 2002] expanded the research done by Schonlau et al. by using the same dataset. While Schonlau et al. simulated the masquerade attack by injecting the data of the 50 users with random data from outside; this study had every user act as a masquerader for every other user. This study claims a 56% improvement over the results presented by Schonlau et al. [Schonlau et al., 2001].

Shim, Kim, and Gantenbein

[Shim et al., 2008] proposed another study that dealt with creating user profiles based on the command line input. Their dataset consisted of UNIX command histories of nine students. The overall average detection rate of the system was 73.13%. Li and Manikopoulos [Li and Manikopoulos, 2004] used the concept of a system state to create user profiles. They used parameters such as window titles and process table entries to model user behavior. They used four users for training and another four users for testing. They achieved a detection rate of 63% and a false positive rate of 3.7%.

2.2 Keystroke Dynamics

An initial study by the Rand Corporation was the basis of user authentication related to keystroke timings [Gaines et al., 1980]. The authors used digraph times, which is the time used to type certain two-letter combinations, as a metric to distinguish users. Umpress and Williams [Umphress and Williams, 1985] conducted early research in this area, using keystroke latency, and achieved a false positive rate of 12% and a false negative rate of 6%. Joyce and Gupta [Joyce and Gupta, 1990] furthered research in this area. Joyce and Gupta used the same metrics as Umpress and Williams but used login credentials of users for authentication. Joyce and Gupta yielded a false positive rate of around 7% and a false negative rate of less than 1%. In 1996, Obaidat and Sadoun [Obaidat and Sadoun, 1996] used the concept of hold time as an additional parameter. They found that using hold times and latencies together provided best results compared to using the individual parameters.

2.3 Mouse Dynamics

Keystroke dynamics have never really established themselves as an authenti-

cation mechanism [Peacock et al., 2004]. One major factor for this could be the emergence of computers with graphical user interfaces. Pusara and Brodley [Pusara and Brodley, 2004] modeled mouse movement by the two-dimensional screen coordinates each time the mouse moves. They also included other variables like mouse wheel movement, mouse clicks and non-client area movement. Their experiment gave a false positive rate of .43% and a false negative rate of 1.75%. Garg et al. [Garg et al., 2006] conducted research very similar to that of Pusara and Brodley's, except that participants were allowed to work freely while Pusara and Brodley had participants read specific web pages. Garg et al. had a higher false negative error rate of 3.85%. Ahmed and Traore [Ahmed and Traore, 2007] performed very similar studies combining the concept of keyboard dynamics and mouse dynamics. They achieved error rates with a false negative rate of .651% and a false positive rate of 1.312%.

2.4 GUI Usage Analysis

Imsand and Hamilton [Imsand and Hamilton, 2007] proposed another approach for detecting anomalous users. They shifted the focus from what the user does to how the user performs their standard tasks. They modeled the user profile on the interaction of the user with the graphical user interface of the system. The false positive error rate was 26.5% and the false negative error rate was 38%. They expanded this research by customizing attack thresholds for each user. This gave a detection rate of 60% with no false positives. The study, however, showed very high error rates for certain users and very low rates for others, prompting that the usage analysis is very user-dependent for some reason. The positive feature of Imsand and Hamilton's

approach is that it claims to use 72% less data than the previous mouse dynamics methods proposed.

Imsand and Hamilton conducted the same experiment again to explore the cause for the variation of the detection rates among different users. The detection rate fell from 60% in the previous experiment to 52%. The aim of this study was to identify how computer usage affects the detection rate. The only significant conclusion that the author drew from the research study was that participants who spend more than six hours on the computer each day could use this method of masquerade detection more effectively than others. In 2008, Imsand and Hamilton [Imsand and Hamilton, 2008] repeated a similar experiment again. This time they chose a more mixed population by recruiting from a high school and a local engineering and assembly firm. Also, they used a different measurement of similarity known as the Jaccard index. The attack threshold customized for each user gave a false negative rate of 6.27% with a false positive error rate of zero.

3. PROCEDURE

The study involved looking at how different users vary in the manner in which they perform everyday tasks on the computer. As a part of the experiment, different users were required to log on to the same computer environment and perform the same set of tasks. To study the effect of a particular change, it is common practice to keep all other variables constant [Schonlau et al., 2001]. The operating environment on the computers that was in the study included:

- Windows 7 Home Premium, Version 6.1, Service Pack 1
- Mozilla Firefox 7.0.1, Internet Explorer 8.0.7601.17514, Google Chrome

14.0.835.202 m

- Microsoft Word 2010 14.4734.1000

Desktop recording software Camstudio (version 2.6), DRPU PC Data Manager (version 5.4.1.1), and REFOG Free Keylogger 7.2.1.1445 were used for data collection. Camstudio records all desktop activity in the form of videos. The PC Data Manager and REFOG Keylogger records all keystrokes that have been typed on the system. Both tools were used concurrently as there was no single tool that could record desktop activity and the keystrokes entered. The authors chose these tools because they were commercially-used and were available for free. The combination of the tools provided the best picture of the actions that were performed to complete a particular task.

3.1 Participants

The criteria for the selection of participants required that they were above the age of 18 and were comfortable with the use of a computer in a Windows environment. The latter requirement was enforced because there the users' behavior should not be impacted by the learning curve associated with performing new tasks. A sample size of 60 participants was used.

Table 1 shows the demographic data for the participants. Most of the participants were students at Purdue University. This resulted in an age group of 18-25 years mostly.

3.2 Methodology

Each participant was required to perform a set of tasks. They consisted of everyday operation on a computer such as copying and pasting text, renaming folders, etc. Table 2 shows the set of tasks that were given to the participants. As the user performs a task, the desktop recording software and keystroke

Table 1. Frequency Distribution of Sample Demographics

		Percentage (Frequency)
Ethnicity	Asian	48 (29)
	African-American/African	3 (2)
	White	47 (28)
	Pacific Islander	2 (1)
Gender	Male	72 (43)
	Female	28 (17)
Age	18 - 20	3 (2)
	20 - 30	87 (52)
	30 - 40	8 (5)
	40 - 50	0 (0)
	50+	2 (1)
	Total	100(60)

logger record all the actions that the user performed.

The data collection procedure was repeated over six sessions where the first session was treated as a practice session and the data collected in that session was discarded. Imsand [Imsand and Hamilton, 2008] has proposed that an interval of half an hour is enough to ensure the independence of each session. This research study was conducted with a minimum interval of one hour between each session to ensure that sessions are independent.

4. RESULTS

This is a structured observational study that falls into the mode of generalizing, where the observer attempts to model participants' behavior into a general categorical data set [Tjora, 2006]. An observational study can either have predefined categories which would make it structured, or can be free where categories are assigned as data is observed. Using the free approach introduces some degree of subjectivity in the observations [Drury, 1995]. To counter this, the data from the first 30 participants was used to obtain the categories that could be assigned to the variables. The next 30 participants were

used to test the consistency of user behavior using those obtained categories. Previous studies with similar goals have mostly employed a smaller number of participants (approximately 10). However, most researchers have expressed the belief that better results can be obtained with a larger sample size. The aim of data analysis was to see if the set of 150 sessions (five sessions each from 30 users) could be grouped into the correct user sessions. Two-step clustering was used for classifying data, as it is the preferred approach when the number of clusters is known and it can deal with a categorical data set. A log-likelihood distance measure was used that assumes a multinomial distribution for a categorical dataset. Clustering of the data assigned a cluster-ID to each data session. The cluster-ID predicts the group membership for each session. These predicted group memberships are compared to the actual groups (users) that the sessions belong to and the number of accurate predictions is calculated; 147 out of 150 sessions were assigned a cluster-ID corresponding to the user's identity. This means that 98% of the sessions were classified correctly. The measure of how distinctly the groups can be classified is known as Sil-

Table 2. List of tasks given to each participant

1) Create a new folder on the Desktop and name it alpha.
2) Copy the file abc.docx from the Desktop into the new alpha folder that you have just created.
3) Navigate to the My Documents folder. Rename the folder named beta in the My Documents folder to gamma.
4) Create a new Microsoft Word document in the alpha Folder. Name this document study.docx.
5) Open abc.docx. Copy all the content from the file abc.docx to study.docx. Close abc.docx.
6) Perform the following operations on the study.docx document <ol style="list-style-type: none"> a. Delete the first sentence of the document (Not the title). b. Make the title of the document bold. c. Increase the font size of the second paragraph to 14 points. d. Move the first paragraph of the document to the end of the document (so that it becomes the last paragraph of the document). e. Close study.docx.
7) Open a web browser of your choice and go to Purdue College of Technology homepage.

houette Co-efficient. As seen in Figure 1, the Silhouette Co-efficient was obtained to be .6, which indicated that the groups formed had good cohesion and separation.

The good degree of cohesion and separation means that users showed very similar behavior within their own sessions and high variability as compared to other users. In simpler terms, it indicated that users exhibited uniqueness in how they perform the same tasks on the computer. Cluster Analysis also provided an indicator of the contribution of each variable to the group membership. This means that it showed degree to which each of the variables contributed

to the uniqueness. This provides a better picture of computer usage behavior among different users. Figure 2 shows the graph indicating the importance of the different variables.

As seen in Figure 2, different variables had varying degrees of contribution to the group membership i.e. to making clusters, which indicated the user. According to the graph in Figure 2, the five variables that contributed the highest are:

1. Method of locating specified folder.
2. Method of making text bold.
3. Method of increasing font size in Word document.
4. Method of copying text from a Word file to another.
5. Method of opening specified webpage.

Cluster analysis needs a significant amount of data to form meaningful clusters. Studies in the past have tried to detect anomalous users by having every user act as an intruder for every other user [Schonlau et al., 2001]. The authors use a similar approach by computing a similarity score for pair-wise comparisons for all the sessions. The similarity scores were calculated using the Jaccard Index similar to Imsand [Imsand and Hamilton, 2008]. The data from the first 30 users was used to obtain a threshold value, which helped us determine if sessions can be attributed to the same user or not. Data from the next 30 users was used to test if sessions can be attributed to users using the threshold obtained. User sessions from the next 30 users are compared similarly in a pair-wise manner and similarity scores are generated. If the similarity score is above this threshold, the sessions are considered to be from

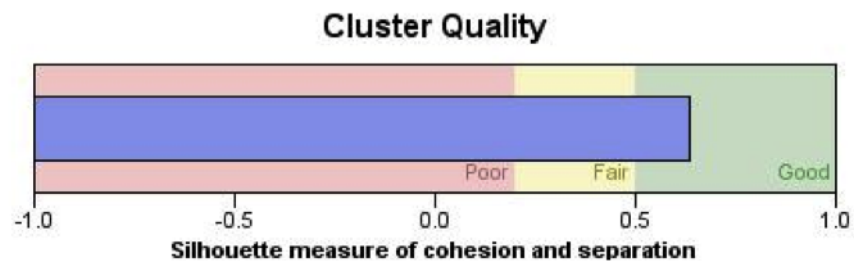


Figure 1. Cluster quality output as obtained from SPSS

the same user. If the similarity score is below this threshold, sessions are attributed to different users. The results have been outlined in Table 3.

Table 3. Table showing the rates at which user sessions were matched

Average rate at which sessions were correctly attributed to either same or different user	93.2%
Average rate at which sessions were incorrectly attributed as belonging to the same user	3.1%
Average rate at which sessions were incorrectly attributed as belonging to different users	3.7%
Confidence interval of correct attributions extrapolated for population at a confidence level of 95% (using Student's t distribution)	(92.98%, 93.42 %).

5. CONCLUSION/DISCUSSION

The results of the experiments show that the computer sessions could be divided into groups, with the groups representing each user. This indicated that users show unique behavior and consistency in their actions. The results also show that the computer sessions could be clustered correctly 98% of the time and when a one-on-one comparison was done, they were attributed correctly 93.2% of the time. These numbers are significant enough to warrant further exploration of computer behavior as a metric to differentiate between users. The other research studies in this area use automated data collection and analysis mechanisms but only employ a small number of participants to test their system. This research study took a behavioral approach to test if users show a high degree of consistency and uniqueness in the choices they make for general use of a personal computer. Even without an automated system, the data collected from 60 participants provides a valuable contribution towards studying user behavior in digital environments.

The results show great promise. The high degree of similarity between sessions belonging to the same user and low similarity levels between sessions belonging to different users shows that computer usage has very good potential to be used to differentiate be-

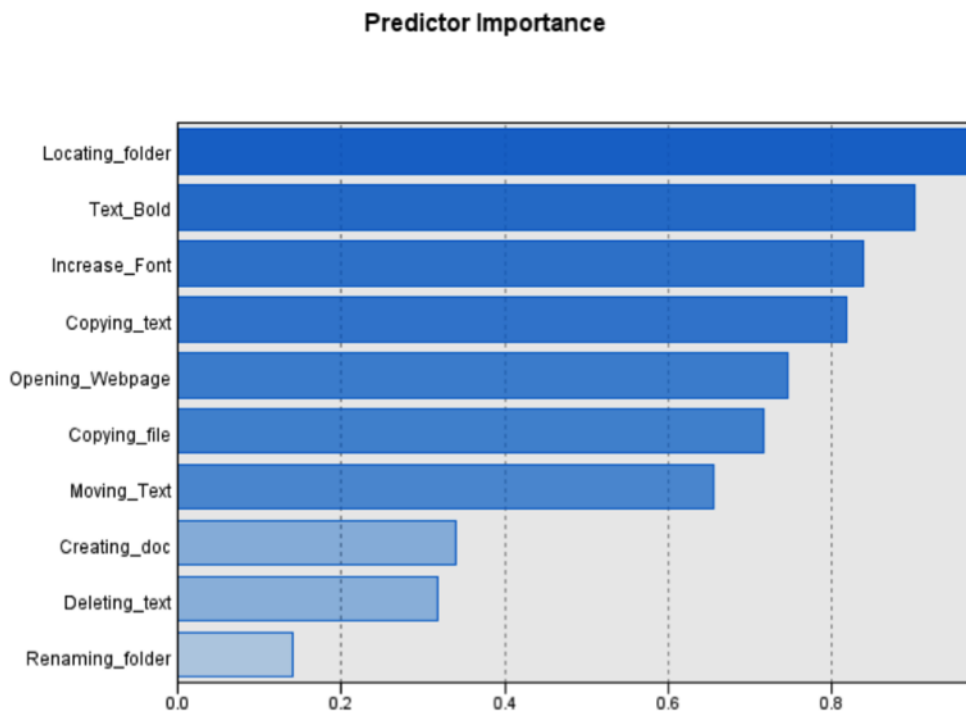


Figure 2. Graph showing contribution of variables to group membership

tween users on the computer. While additional tools need to be developed that can model usage behavior and differentiate between users, logging computer usage can be developed into a soft biometric to give further direction to a criminal investigation. Even though it would have a long way to go to be admissible in court, it can provide investigators with more information about the person using the computer and provide direction for the investigation. The results of the study also identified variables that had a greater impact on the discrimination between users and these variables can be used in future profiles.

6. FUTURE WORK

This was a preliminary study to test whether user sessions could be attributed to correct users by employing a broad metric like the

computer usage behavior of users. An automated logging tool needs to be employed that can effectively log the users' interaction without the users' knowledge. Also, the sample population was constrained by the limitation that no compensation was offered to the participants. Future research in this area should also focus on the relationship between metadata and the user-profiles generated. Metadata like age, gender, occupation etc. might reflect on the computer behavior of individuals and this should be explored so that computer habits of people can be associated with their demographic information. The findings presented in this study can be used to form a more structured framework that can be implemented in digital investigations to identify the correct individual using the computer.

REFERENCES

- [Mor, 2011] (2011). *US v. Moreland*, volume 665. Court of Appeals, 5th Circuit.
- [Abraham and de Vel, 2002] Abraham, T. and de Vel, O. (2002). Investigative profiling with computer forensic log data and association rules. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 11–18. IEEE.
- [Ahmed and Traore, 2007] Ahmed, A. A. E. and Traore, I. (2007). A new biometric technology based on mouse dynamics. *Dependable and Secure Computing, IEEE Transactions on*, 4(3):165–179.
- [Drury, 1995] Drury, C. G. (1995). Methods for direct observation of performance. *Evaluation of human work*, 2:45–68.
- [File and Ryan, 2014] File, T. and Ryan, C. (2014). Computer and internet use in the united states, 2013.
- [Gaines et al., 1980] Gaines, R. S., Lisowski, W., Press, S. J., and Shapiro, N. (1980). Authentication by keystroke timing: Some preliminary results. Technical report, DTIC Document.
- [Gamboa and Fred, 2004] Gamboa, H. and Fred, A. (2004). A behavioral biometric system based on human-computer interaction. In *Defense and Security*, pages 381–392. International Society for Optics and Photonics.
- [Garg et al., 2006] Garg, A., Rahalkar, R., Upadhyaya, S., and Kwiaty, K. (2006). Profiling users in gui based systems for masquerade detection. In *Information Assurance Workshop, 2006 IEEE*, pages 48–54. IEEE.
- [Imsand and Hamilton, 2007] Imsand, E. S. and Hamilton, J. (2007). Gui usage analysis for masquerade detection. In *Information Assurance and Security Workshop, 2007. IAW’07. IEEE SMC*, pages 270–276. IEEE.
- [Imsand and Hamilton, 2008] Imsand, E. S. and Hamilton, J. (2008). Masquerade detection through guiid. In *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, pages 1–5. IEEE.
- [Joyce and Gupta, 1990] Joyce, R. and Gupta, G. (1990). Identity authentication based on keystroke latencies. *Communications of the ACM*, 33(2):168–176.
- [Lane and Brodley, 1997] Lane, T. and Brodley, C. E. (1997). An application of machine learning to anomaly detection. In *Proceedings of the 20th National Information Systems Security Conference*, volume 377, pages 366–380. Baltimore, USA.
- [Li and Manikopoulos, 2004] Li, L. and Manikopoulos, C. N. (2004). Windows nt one-class masquerade detection. In *Information Assurance Workshop, 2004. Proceedings from the Fifth Annual IEEE SMC*, pages 82–87. IEEE.
- [Maxion and Townsend, 2002] Maxion, R. A. and Townsend, T. N. (2002). Masquerade detection using truncated command lines. In *Dependable Systems and Networks, 2002. DSN 2002. Proceedings. International Conference on*, pages 219–228. IEEE.
- [Obaidat and Sadoun, 1996] Obaidat, M. and Sadoun, B. (1996). Keystroke dynamics based authentication. In *Biometrics*, pages 213–229. Springer.

- [Peacock et al., 2004] Peacock, A., Ke, X., and Wilkerson, M. (2004). Typing patterns: A key to user identification. *IEEE Security & Privacy*, 2(5):40–47.
- [Peisert et al., 2008] Peisert, S., Bishop, M., and Marzullo, K. (2008). Computer forensics in forensics. In *Systematic Approaches to Digital Forensic Engineering, 2008. SADFE'08. Third International Workshop on*, pages 102–122. IEEE.
- [Pusara and Brodley, 2004] Pusara, M. and Brodley, C. E. (2004). User re-authentication via mouse movements. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 1–8. ACM.
- [Rogers et al., 2007] Rogers, M., Scarborough, K., Frakes, K., and San Martin, C. (2007). Survey of law enforcement perceptions regarding digital evidence. In *Advances in Digital Forensics III*, pages 41–52. Springer.
- [Schonlau et al., 2001] Schonlau, M., DuMouchel, W., Ju, W.-H., Karr, A. F., Theus, M., and Vardi, Y. (2001). Computer intrusion: Detecting masquerades. *Statistical science*, pages 58–74.
- [Shim et al., 2008] Shim, C. Y., Kim, J. Y., and Gantenbein, R. E. (2008). Practical user identification for masquerade detection. In *World Congress on Engineering and Computer Science 2008, WCECS'08. Advances in Electrical and Electronics Engineering-IAENG Special Edition of the*, pages 47–51. IEEE.
- [Tjora, 2006] Tjora, A. H. (2006). Writing small discoveries: An exploration of fresh observers observations. *Qualitative Research*, 6(4):429–451.
- [Umphress and Williams, 1985] Umphress, D. and Williams, G. (1985). Identity verification through keyboard characteristics. *International journal of man-machine studies*, 23(3):263–273.