Chapter 11

LINKING INDIVIDUALS TO DIGITAL INFORMATION

Shelly Seier, David Greer and Gavin Manes

Abstract As computer crime increases in scope and magnitude, it is imperative to develop techniques that can link individuals to specific computers, computer programs and electronic documents. Unfortunately, scientific techniques that can establish these links are limited at best. This paper demonstrates that computer use characteristics can be employed to establish strong, legitimate links between individuals and digital information. Certain characteristics can be used to identify individuals. Other characteristics may be used to create profiles that assist in eliminating suspects and reducing the scope of investigations.

Keywords: Computer use characteristics, pattern analysis, identifying individuals

1. Introduction

On September 20, 2001 a distributed denial of service attack disabled vital navigation systems at the Port of Houston in Texas. The attack was traced to a computer in Aaron Caffrey's home in England, which contained an attack script with the words "coded by Aaron." Caffrey admitted to being a member of the "Allied Haxor Elite" group, and to hacking his friends' computers to "test their security" [1]. All the evidence pointed to Caffrey: the script contained his name, his machine executed the attack and he had the necessary technical expertise.

However, Caffrey claimed that his computer had been commandeered by another individual via a Trojan virus. Caffrey argued that this individual was responsible because the Trojan was in control of his computer at the time of the attack. Although a forensic examination of Caffrey's computer yielded no evidence of a Trojan, Caffrey was acquitted.

To prove that Caffrey was responsible for the crime, it was necessary to: (i) link Caffrey to the attack script, (ii) link Caffrey to his computer,

and (iii) link the execution of the attack script to Caffrey's computer. Linking the attack script to Caffrey's computer was accomplished by standard digital forensic techniques. However, linking Caffrey to the attack script and to his computer proved to be more difficult. This paper proposes that computer use characteristics can be employed to establish strong, legitimate links between individuals and digital information, which were missing in the Caffrey case.

2. Demonstrating Uniqueness in Individuals

Individuals possess unique characteristics such as fingerprints and DNA, which are often used as objective evidence [9]. When such evidence is not available, characteristics such as handwriting may be used to link an individual to a written document [6]. The uniqueness of an individual's handwriting comes from education, artistic ability, physiological development and preference. The slant, spacing and letter formation embody unique stylistic features that tend to become permanent over time [3, 4, 8]. Linguistics is an important component in written documents as spelling and word choice are distinct stylistic traits [2, 4, 5].

Handwriting samples fall into two categories: requested and non-requested. For requested samples, suspects agree to write a set of predetermined words on paper. Non-requested samples are personal letters and notes created independent of the investigation.

It is important to obtain requested and non-requested writing samples to determine a match. Requested samples may be intentionally altered or may reflect a suspect's nervousness or excitement. Investigators can request lengthy samples to negate attempts at subterfuge; experienced handwriting analysts are usually able to discern intentional alterations of writing style. Typically, free-writing samples are better than requested samples as they manifest the true handwriting style.

Much like handwriting, the way an individual uses a computer is the result of education, artistic ability, physiological development and preference [7]. Because of the similarities between handwriting and computer use, handwriting sampling techniques could be used as a guide for developing computer use sampling techniques.

3. Digital Characteristics

If computer use can be monitored and measured, an investigator should be able to utilize a computer use sample in much the same way as a handwriting sample. A requested computer use sample would involve asking an individual to perform certain tasks on a computer with monitoring tools installed. A non-requested sample could be obtained

by monitoring an individual's computer use surreptitiously. There are several technical methods for monitoring and data collection; they can be categorized as active monitoring and passive monitoring.

Characteristics obtained using active monitoring include keystrokes, mouse use patterns and network use patterns. Other information can be obtained through passive methods such as traditional digital forensic investigations, undo history, application history, passwords and linguistic analysis of previously-typed documents [2, 5].

Active Monitoring

- **Keystrokes:** A keystroke logger can be used to record every keystroke made by an individual. These logs can determine the typing speed and style, including shortcuts, command line operations and program operations. For instance, one individual might use the backspace key to correct typing mistakes immediately, while another individual might leave the error for the spell checker to fix. A keystroke logger can also help determine accuracy by recording the use of the backspace and delete keys.
- **Mouse Distance/Patterns:** Each person has a unique way of using a mouse. Some may fidget while concentrating, others might trace words with the mouse as they read. Some may right-click for options, while others may use menus. Using the scroll bar instead of clicking and dragging to navigate inside a window is another distinctive mouse pattern.
- **Network Use Patterns:** Network monitoring tools can be used to log an individual's network use. Advanced monitoring tools can log every packet that is sent and received. This could help identify frequently visited Internet sites as well as the communications protocols that are used. Also, it could reveal the information content sent across the network.

Passive Monitoring

- **Undo History:** In relation to the frequency of error correction, the undo history may be unique for different individuals. One individual might use the undo backing order to retrieve an earlier paragraph, while another may just retype the entire paragraph. It may also be beneficial to know the set of circumstances under which an individual uses the undo function.
- **Dictionary/Word Choice:** Individuals may be identified based on the complexity of their word choice and use of a dictionary

or thesaurus. The frequencies of certain words and phrases in emails or chats may be a distinguishing characteristic, as well as expressions, shorthand and slang [2].

- **H4x0r T4lk:** Since the majority of computer crime is perpetrated by technologically-savvy individuals, it is important to understand computer-specific languages such as H4x0r T4lk (hacker talk), l33t (Leet) and chat shorthand. Certain groups of users substitute characters for letters, and often a character is substituted for the same letter every time. This consistency allows all the members of the group to understand the language. Knowing the styles used by hacker groups may indicate which groups to investigate in the event a crime has been committed.

- **Application History:** Individuals often favor one program or application over another. Some may use Internet Explorer, while others prefer Mozilla or Netscape. Many different chat clients, email clients and word processors exist, and most users have strong preferences among these programs. The combination of frequently used applications may be an identifying aspect of an individual.

- **Passwords:** Measuring password strength could eliminate certain individuals (e.g., those using weak passwords), thereby narrowing the list of suspects. Since many people use the same password for a variety of programs or functions, tracking password use over several computers might be beneficial in an investigation.

4. Evidence Collection Issues

Many of the same difficulties surrounding the collection of physical evidence hold for digital evidence, including the requirements of search warrants and application of forensically-sound evidence collection techniques. One of the more complex issues involves obtaining evidence from a suspect during an investigation. Active monitoring tools can provide valuable evidence, but these devices are invariably illegal without consent. Unfortunately, the time needed to obtain a warrant negatively affects the amount of monitoring that can be performed. Also, there is little chance that a suspect would agree to monitoring.

Collecting the requested evidence presents another set of difficulties. Typically, obtaining a requested computer use sample would involve asking an individual to perform certain tasks on a computer with the monitoring tools installed, but there is no guarantee the subject will not intentionally or inadvertently adjust his actions during the monitoring. As mentioned earlier, a possible solution is to collect large samples.

It is hard to determine the appropriate comparison of statistics with regard to requested and non-requested samples, as taking samples at different times might create bias. An individual's mood and energy level might affect the samples, as well as external forces such as network traffic loads, which could distort packet speed or chat room statistics. As in any statistical study, large samples and/or multiple samples, provide greater confidence in the results.

5. Experimental Results

To test whether digital characteristics are unique to each individual, an experiment was designed to monitor computer use. A laptop computer was loaded with keystroke monitoring software to capture each keystroke and the elapsed time between keystrokes. Then, a series of tasks was displayed on separate pages so the subject would not be able to prepare for the next task until the current task was completed.

The first task assigned to the subject was:

```
Open a Word document and begin typing a description of the
weather today. Please try to type at least five lines of
text.

If you cannot type five lines about the weather today, then
write about how the weather has been since the beginning of
this week.

Go to the next page.
```

The subject was asked to write about the weather to simulate a free thought process. As mentioned previously, requested handwriting samples do not reflect an individual's style as accurately as non-requested samples. This task allowed for some flexibility in word choice, grammar, sentence structure, punctuation, capitalization and spacing. Since there were no restrictions, the subject would be expected to type in a manner that accurately represents his typical style. From the monitor logs, an investigator would be able observe how the subject opened the word processor, how the subject formed sentences with respect to grammar, spacing, capitalization, punctuation, formatting and word choice, and how quickly the subject typed when he composed free-form sentences.

The second task assigned to the subject was:

```
Go to http://www.nws.noaa.gov/. Type "Tulsa" into the Local
Forecast box on the top left and view the weather today.

Copy and paste the paragraph about today's weather into your
Word document.
```

```
Go back to the Internet browser. Scroll to the bottom of the
page and copy and paste the last paragraph about the weather
a week from today.

Go to the next page.
```

This task was designed to monitor the use of the mouse and shortcut keys. The investigator would be able to observe how the individual opened the Internet browser, switched between windows, scrolled through windows, highlighted text, and copied and pasted text. The investigator could also note how the subject typed the web address into the address bar, and how he typed "Tulsa" in the text box.

The third task assigned to the subject was:

```
Retype the following paragraph into your document below the
text you just pasted.

The preparedness guide explains thunderstorms and related
hazards and suggests life-saving actions can take. With
this information, you can recognize severe weather, develop a
plan and be ready to act when threatening weather approaches.
Contact your local National Weather Service Office for a
variety of weather-related brochures.

Once the paragraph is completely retyped, save the document
to the desktop as <yourname>.doc and close all windows.

You are now finished.
```

The purpose of this task was to demonstrate differences between typing freely-composed ideas and copying predetermined text. Both the typing speed and the attention to detail can be measured. For example, the subject might type "life-saving" with or without the hyphen. If the subject usually types two spaces after each sentence, he would probably not notice that all the sentences only have one space between them. Another detail to note is whether the subject tries to make logical sense of the sentences when retyping them. The first sentence reads "... and suggests life-saving actions can take" instead of "... and suggests life-saving actions you can take."

Another interesting result was the difference in typing speeds for free-form composition and predetermined text typing (Figure 1). Subjects tended to type at different speeds when they composed sentences as opposed to when they copied text. Although the differences in typing speeds varied, no individual had the same speed for both typing activities. It is possible to use this technique to distinguish individuals (and eliminate suspects) based on their typing speeds.

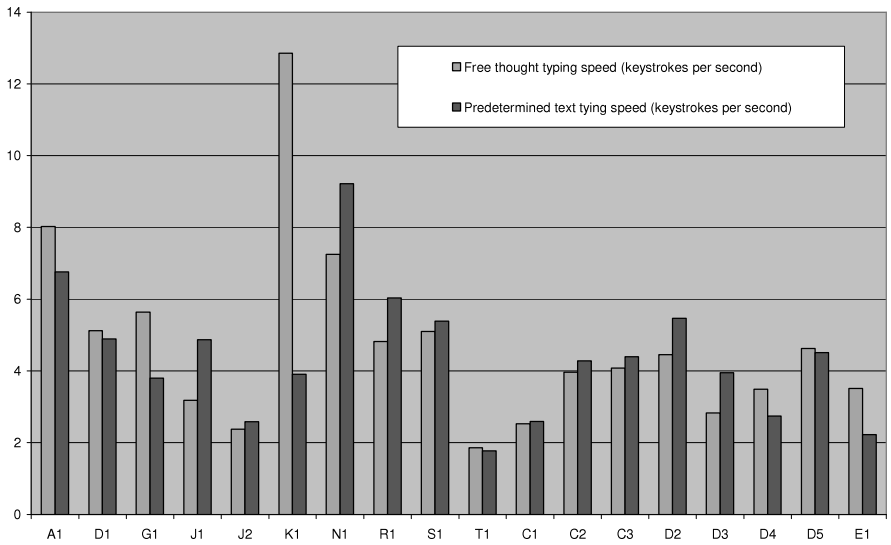

Figure 1. Keystrokes per second.

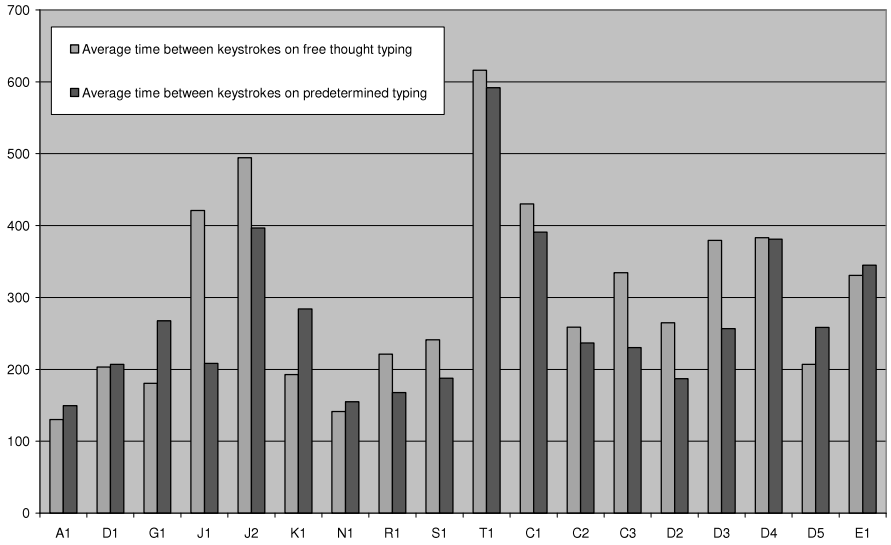

Figure 2. Time between keystrokes.

Figure 2 shows the differences in the time intervals between keystrokes for free-form composition and predetermined text typing. When comparing Figures 1 and 2, each subject differs considerably in terms of typing speed and keystroke intervals; thus, these could be useful identification characteristics.

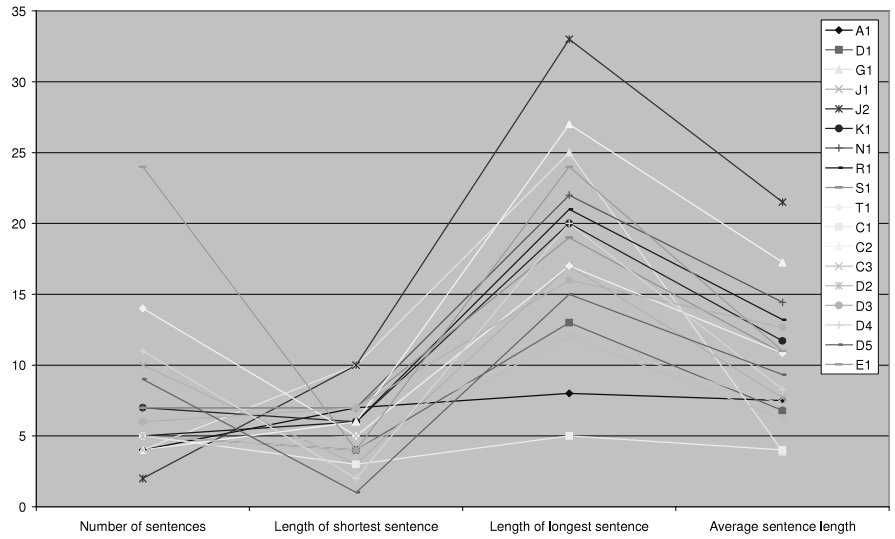

Figure 3. Sentence characteristics.

When given the same typing directions, each individual formatted the text differently. Figure 3 shows the differences in the numbers of sentences and sentence lengths. This particular task only required a small amount of text to be typed. It is expected that the differences would be much more significant for tasks that require a large amount of text to be typed.

Figure 4 indicates that each subject spent a different amount of time and corrected a different number of errors while performing the task. The lightest line represents the number of corrected mistakes compared to the elapsed time; the differences are unique for each subject. The difference between each subject's set of fastest-typed characters is also pronounced, and no two individuals had equivalent sets.

Groups of subjects showed similar results for characteristics such as name formatting when saving documents, use of the shift, control and backspace keys, use of shortcuts such as copy, paste and undo, window navigation (opening and closing), sentence and paragraph formatting, opening web pages, and typing in search criteria. Although these results may not distinguish a particular individual, they can be used to eliminate certain members of a group.

6. Future Work

The graphs presented in the previous section represent a modest subset of the analysis we have performed. Several other profiling points may

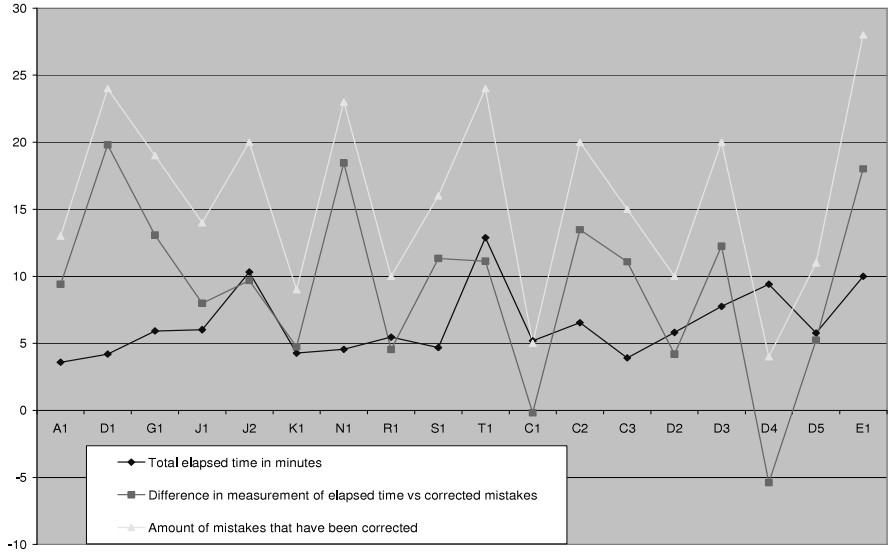

Figure 4. Differences in characteristics.

be developed from the test data. Analyzing the complete set of results from the tests is beyond the scope of this paper; however, additional results are available upon request.

Much like DNA evidence, some computer use characteristics may require the use of population profiling techniques [9]. Even if it is not possible to uniquely identify an individual, computer use characteristics can help identify a group of individuals as the target of an investigation, or exclude certain individuals from consideration.

Other techniques for identifying individuals include chat and network monitoring using keystroke monitors and surveillance tools. Unique characteristics involved in chat analysis include slang, acronyms, chatting style and typing speed, as well as the chat programs used, time of day and length of chat sessions. It might even be possible to identify individuals in chat rooms as they type their communications, and the monitoring could be performed remotely using chat room clients.

To monitor network use, packet analysis could be performed on various machines to track an individual's network use over time. Alternatively, two individuals could be given a specific question and asked to find the answer online. While they search for the answer, information pertaining to their browsing characteristics could be collected for on-line or off-line analysis.

7. Conclusions

The lack of robust scientific techniques for linking individuals to digital information hinders computer crime investigations and subsequent prosecution. Some of the principles underlying handwriting analysis can be applied to computer use characteristics to establish strong, legitimate links between individuals and specific computers, computer programs and electronic documents. The results can be used to identify individuals or to create profiles that may assist in eliminating suspects. Of course, this work is very preliminary; extensive research and statistical analysis are necessary before the techniques can be put to practice.

References

[1] BBC News, Questions cloud cyber crime cases (news.bbc.co.uk/1/hi/technology/3202116.stm), October 17, 2003.

[2] C. Chaski, Who's at the keyboard? Authorship attribution in digital evidence investigations, *International Journal of Digital Evidence*, vol. 4(1), 2005.

[3] J. Olsson, *Forensic Linguistics: An Introduction to Language, Crime and the Law*, Continuum International Publishing Group, London, United Kingdom, 2004.

[4] K. Ramsland, Document analysis (www.crimelibrary.com/forensics/literary), April 23, 2004.

[5] V. Raskin, C. Hempelmann and K. Triezenberg, Semantic forensics: An application of ontological semantics to information assurance, *Proceedings of ACL 2004: Second Workshop on Text Meaning and Interpretation*, pp. 105-112, 2004.

[6] G. Shpantzer and T. Ipsen, Law enforcement challenges in digital forensics, *Proceedings of the Sixth National Colloquium for Information Systems Security Education*, 2002.

[7] E. Spafford and S. Weeber, Software forensics: Can we track code to its authors? *Computers and Security*, vol. 12(6), pp. 585-595, 1993.

[8] S. Srihari, S. Cha, H. Arora and S. Lee, Individuality of handwriting, *Journal of Forensic Sciences*, vol. 44(4), pp. 856-872, 2002.

[9] B. Weir, Population genetics in the forensic DNA debate, *Proceedings of the National Academy of Sciences*, vol. 89, pp. 11654-11659, 1992.