CHAPTER 5

KILLING THE PASSWORD, PART 1: AN EXPLORATORY ANALYSIS OF WALKING SIGNATURES

5.1 Introduction

The very first computers were simple machines built to solve a single, specific problem–they could not do anything else. In 1945, John von Neumann proposed a general-purpose computer architecture that could be repurposed to solve different problems using software. In the 1960s, computers were still extremely large and expensive. The prestigious Massachusetts Institute of Technology's Compatible Time-Sharing System may have been the first computer to make use of passwords, which were used to allow individuals to access the CTTS computer for up to four hours a week [46] [47]. In a January 27, 2012 Wired article, Dr. Allan Scherr admitted to using other students' passwords to gain additional time on CTTS. For over 50 years the password has been used to authenticate users, provide confidentiality, and integrity–with limited effectiveness.

Each of us has numerous passwords that we use every day. But we often fail to follow best practices when it comes to password security. We chose simple passwords [48] because they are easy to remember and easy to use when we want to check our email or log into Facebook. Complex passwords may require multiple attempts to enter correctly and can prevent users from completing their tasks [49]. Perhaps you use a complex password for more risky Internet activities such as online banking, but that complex password is difficult to remember so you have written it down and keep it next to your computer. How long has it been since you last changed your passwords? Do you even know how many passwords you have?

Passwords are not the best way to authenticate people. Even if we follow best practices such as using long, complex passwords, using a different password for each resource, and change our passwords often; there are other problems with passwords. Passwords can be guessed using social

engineering or brute force attacks. Passwords can be intercepted and reused (e.g., man-in-the middle attack), or stolen and cracked. The Honorable Michael Daniel, Special Assistant to the President and Cybersecurity Coordinator at the White House suggested we kill the password dead as a primary security measure in his keynote address at the International Conference on Cyber Engagement on April 27, 2015 [50]. But what can we use to replace the ubiquitous password?

This paper describes the foundation for a biometric method of user authentication that is completely transparent to the user. This method does not require a specialized biometric sensor and strengthens existing security mechanisms instead of replacing them.

In Section 5.2, we propose an alternative method of passive, continuous user authentication. Section 5.3 explains the hypothesis behind our proposed solution. Section 5.4 describes how the initial experiment was conducted and Section 5.5 analyzes the data gathered during that experiment. Related work and future work are briefly covered in Section 5.6 and Section 5.7, respectively. Lastly, our conclusion can be found in Section 5.8. The first author plans to publish additional papers (e.g., Part 2, Part 3, etc.) as future work in this area is accomplished. wjh3

September 30, 2016

5.2 Proposed Solution

Today there are three widely used methods for authentication; something you know (e.g., password), something you have (e.g., RSA token), and something you are (e.g., fingerprint) [45]. Using biometrics (i.e. something you are) requires special sensors. Some biometric methods of authentication have been compromised using simple, inexpensive techniques [51] [52]. Once compromised, biometrics may no longer provide the necessary assurance of a user's identity. You can easily change your password, but you cannot change your fingerprints. Any authentication method using something you have often requires additional cost in materials and management. The solution we are proposing transparently implements best practices for password management, combined with the benefits of biometrics without additional costs. We call this solution "passive,

persistent authentication".

Most of us now carry a smartphone with us everywhere we go. With wireless and mobile networking, a camera, microphone, GPS receiver, inclinometer, and accelerometer, smartphones employ a wide range of sensors combined with relatively powerful computational capabilities. Imagine that your smartphone could learn to recognize your identity based on your patterns of behavior. Also imagine that your cell phone could immediately discern between you and another person.

Alice's smartphone has been in her pocket for the last fifteen minutes, counting her steps and storing metadata (a timestamp) about each step. Her smartphone uses this information to passively and continuously authenticate her. When Alice opens the *Facebook* app, her smartphone automatically authenticates her. When she is finished with *Facebook*, she sets down her smartphone. Alice's little sister, Eve, picks up Alice's smartphone and tries to look at Alice's *Facebook*. However, Alice's smartphone is able to differentiate between Alice and Eve. The smartphone locks itself, then revokes all of Alice's passwords, preventing anyone from accessing those resources, including Alice, until she retakes possession of her smartphone. After picking up her smartphone and taking a few steps, her smartphone recognizes Alice and enables access to all of her resources, including her *Facebook* account.

5.3 Hypothesis

Each of us has a unique "walking signature" that can be recognized by when, where, and how we walk. Additionally, the frequency of our pace, how many steps we take each minute, along with the force of our feet impacting the ground may all combine to create a unique walking signature that our smartphone can continually observe to determine who is using it at any given time.

5.4 Experiment Design

I developed a simple smartphone app to collect the necessary data to evaluate our hypothesis. The app listens for the *StepCounter* sensor change event on a Samsung Galaxy S5 and observes the current timestamp for each step taken. The data is logged to a SQLite database that can be retrieved from the smartphone to derive additional information for analysis.

I plotted a 109-yard long course using a *Garmin eTrex* GPS receiver. I would have preferred 100 yards, but 109 yards made for very clear starting and ending locations. The GPS receiver accuracy was measured at ± 12 feet at the starting location and ± 17 feet at the ending location of the course. I had hoped to use latitude and longitude readings from the smartphone as an additional source of metadata to measure the distance between each step, but the inaccuracies of GPS [53] make it an unsuitable measure of such a small distance. The Galaxy S5 smartphone does have additional location-based sensors that could be used when GPS is unavailable (e.g., GPS, Cell-ID, WiFi) [54] but they added too much complexity to the initial exploratory experiment.

I conducted our initial experiment with four subjects that I expected to have different walking signatures based on variables such as age, gender, height, and weight. Obviously a sample size of $n = 4$ for any study is too small to use to draw significant conclusions. However, Section 5.7 outlines additional experiments to draw statistically significant conclusions using predictive models, functions, and power analysis to reject a null hypothesis. Likewise, in the age of big data, $n = 1$ is conceivably a suitably interesting population for an exploratory study. What type of data and how much of that data is needed to passively and continuously authenticate an individual is something we can and should explore in future experiments.

Before analyzing any data, I suspected each walking signature to be dependent upon certain variables such as terrain, a person's height, weight, exertion level, and choice of foot ware. I was unsure about other variables such as weather conditions like temperature, wind, and precipitation.

Date / Time	Temperature	Humidity	Wind Speed, Mean	Wind Direction	Wind Gust
July 28, 2016 / 9:05 AM (GMT -8)	90^o F	38%	4 mph	WNW	7 mph
July 28, 2016 / 10:18 AM (GMT -8)	93^o F	39%	3 mph	NWN	7 mph

Table 5.1: Weather Conditions During *Walking Signature* Experiment

Identifying dependent and independent variables is a crucial step in developing a model and efficiently computing an individual's walking signature with sufficient accuracy to be used for passive, continuous authentication.

Several academic papers [55] [56] [57] have concluded that the accuracy of smartphone pedometers is not affected by smartphone placement, unless the phone is carried in a back pocket. For this experiment, each subject held the smartphone face up in the left hand, with the elbow bent at a 90-degree angle. The subjects were then instructed to walk the same predefined course at their own natural pace, looking ahead and ignoring the smartphone. When walked west to east, the course had a significant incline. I asked each subject to walk the course multiple times in different directions in an attempt to gather data to identify terrain as a dependent or independent variable.

Weather conditions have not been determined to be independent variables, so information about the weather conditions at the start and end of the experiment can be found in Table 5.1. Observed weather conditions were gathered using an *Acurite* Weather Station ≤ 300 yards from the walking course.

For this paper, we have chosen to not publish specifics about the four test subjects, because initial analysis revealed that their characteristics such as age, height, weight and gender are not dependent variables for calculating an individual's walking signature.

5.5 Analysis

The first step in proving a hypothesis is exploratory data analysis [58]. The only metadata we collected for the first samples was a timestamp associated with each step taken. SQLite timestamps are stored in ISO 8601 format (e.g. 2016-09-30 13:06:45.123). SQLite database files can be copied off

Sample	Mean (μs)	Mean (ms)	Min (ms)	Max (ms)	Subject
1	467792.20779	467.8	358	857	1
2	484118.42105	484.1	339	819	1
3	449639.53488	449.6	260	997	1
4	549500.00000	549.5	358	626	2
5	467690.90909	567.7	517	956	2
6	495481.01265	495.5	10	559	3
7	535202.89855	535.2	70	760	4

Table 5.2: Step Duration Statistics

the smartphone using an USB cable and the "adb pull [/data/data/[package]/databases/[filename].db]" command that is included with Android Studio. The SQLite database file is then exported as comma-separated values using the SQLite Database Browser (version 2.0b1 for OSx). When the CSV file is imported into Microsoft *Excel*, the ISO 8601 date strings can be converted to a number with precision to 10 decimal places to get a real number that represents a fraction of a single day. The resulting number t_i can be subtracted from t_{i+1} to calculate the duration (i.e. elapsed time) between two steps. The difference can be converted back to milliseconds using Equation 1.

$$(t_{i+1} - t_i) \times (24 \text{ hours} \times 60 \text{ minutes} \times 60 \text{ seconds}) \tag{5.1}$$

I also wrote a simple Python program to extract the basic statistics about each sample in Table 5.2. I was expecting a wider variation in average step duration. At first I thought the duration variation was hiding in a lack of precision, but the results were fairly uniform, even at μs (microsecond) resolution.

Next I generated marked scatterplots for each sample. Before my exploratory analysis, I expected a Gaussian (i.e. normal) distribution of step durations for the samples, but was surprised to see the fairly uniform distribution in Figures 1-4 for each sample in the population. I was also surprised to learn that many variables I suspected to be dependent, such as age, height, weight, and

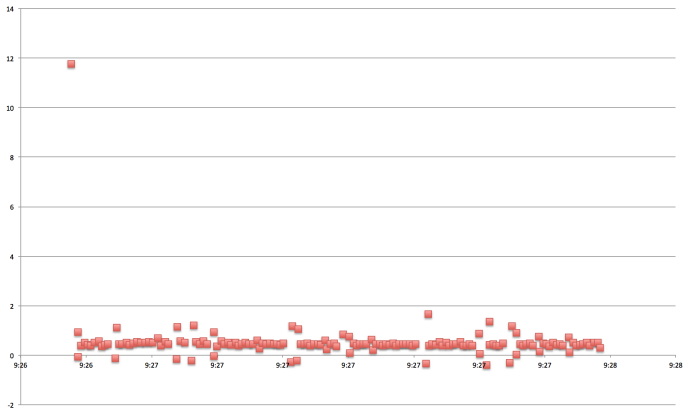

Figure 5.1: Walking Signature Sample #1 (Subject #1)

choice of foot ware, did not appear to influence step duration. Figures 4 and 5 are different samples from the same subject walking the course in opposite directions. Figure 4 contains a significant decline in terrain and Figure 5 contains a significant incline in terrain. Both figures appear to exhibit a very similar pattern, regardless of terrain, which I also expected to be a dependent variable. Figures 1–4, which are all from different subjects each appear to exhibit a different pattern. Figure 4 and Figure 5 appear to exhibit a similar pattern from the same subject. This seems to be further evidence that each individual does indeed have a unique walking signature.

Data from the initial experiment does contain two anomalies, which can be clearly seen the scatterplots. First, there is a significant lag in the first step, which is measured in multiple seconds. Second, $\approx 6\%$ of the calculated step durations are negative, presumably due to a subsequent step having an earlier timestamp than the previous step.

5.5.1 Estimated Free Energy of a Walking Signature

Each walk in our dataset consists of a series of timestamps. Recall that a step duration is a real number in microsecond resolution that measures the duration between steps. A walk is simply a list of step durations. The potential free energy of a walk is very high considering there are so many potential choices of durations between each step. We can reduce the number of the potential

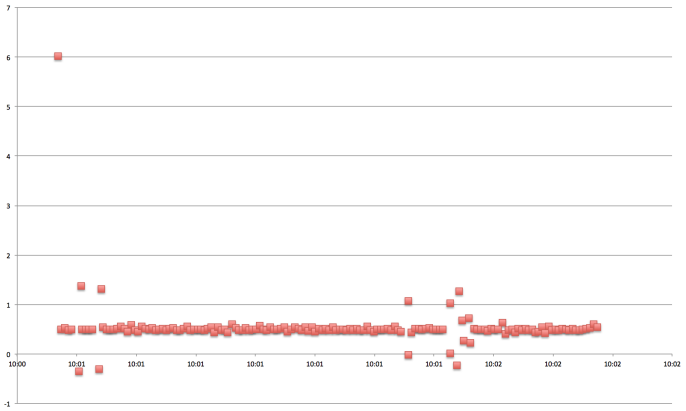

Figure 5.2: Walking Signature Sample #6 (Subject #3)

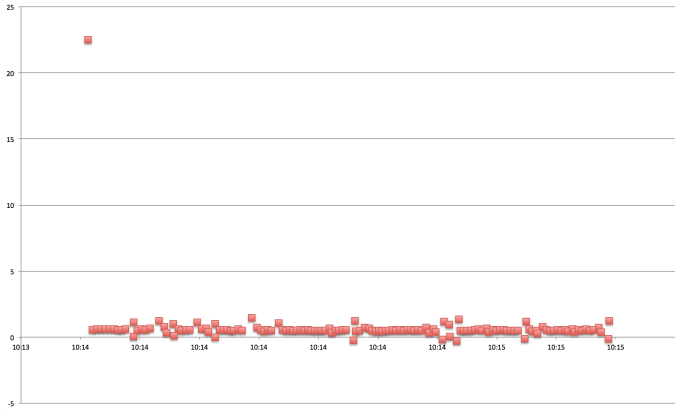

Figure 5.3: Walking Signature Sample #7 (Subject #4)

durations by calculating some statistics about the observed walk, including a mean duration μ, shortest duration, longest duration, and standard deviation σ to use in our interface as an encoding scheme and or weight to be assigned to each step duration.

Obvious patterns are apparent in the plotted step durations shown in the Fig. 5.2 and Fig. 5.3. We created a simple interface to encode each step duration, then calculate the estimated free energy of each subject's walk. Our analysis in Chapter 3 showed that the average weight in Equation 2.15 was not a significant contribution to the estimated free energy of Netflow conversations. So we first calculate only λ_i (i.e. the inverse compression ratio of the compressed and uncompressed encoded

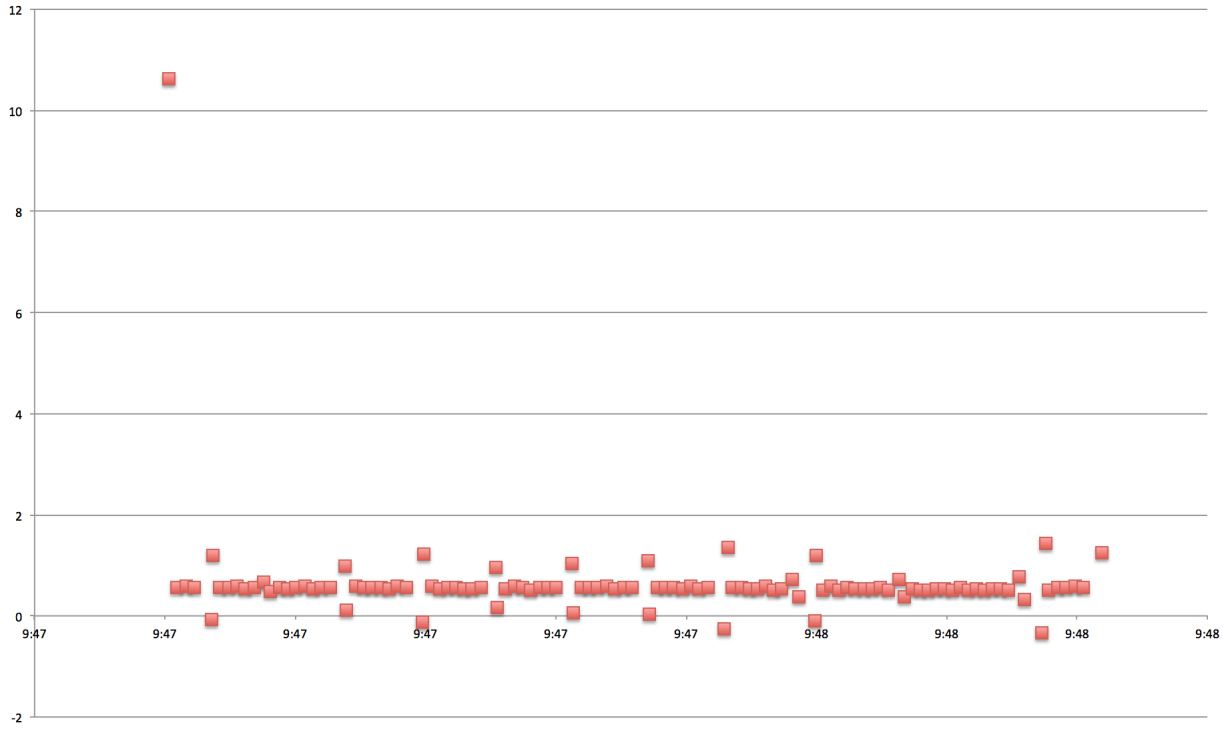

Figure 5.4: Walking Signature Sample #4 (Subject #2)

Subject	λ_i Low	λ_i High	λ_{α_i} Low	λ_{α_i} High
Subject 1	0.219269102990033	0.235621521335807	0.25975284231339596	0.2756427246223165
Subject 2	0.251515151515152	0.267857142857143	0.3200731595793324	0.3707482993197279
Subject 3	0.227848101265823	None	0.2529585798816568	None
Subject 4	0.238095238095238	None	0.27925729115175485	None

Table 5.3: Estimated Free Energy of Each Subject's Walk

durations). When our experimental data contained multiple walks for some subjects, we display a low and high value for each subject to encompass all observed values of λ_i.

For reference the range of estimated free energy values for each subject can be found in Table 5.3.

Put simply, the range of λ_i for Subject 1 is approximately .22 – .24, Subject 2 is approximately .25 – .27, Subject 3 is approximately .23 and Subject 4 is approximately .24. λ_i for Subject 1 and Subject 2 are distinct enough to be used as an authentication signature. But there is a problem with Subject 3 and Subject 4. Their λ_i values are contained within the range of Subject 1. This

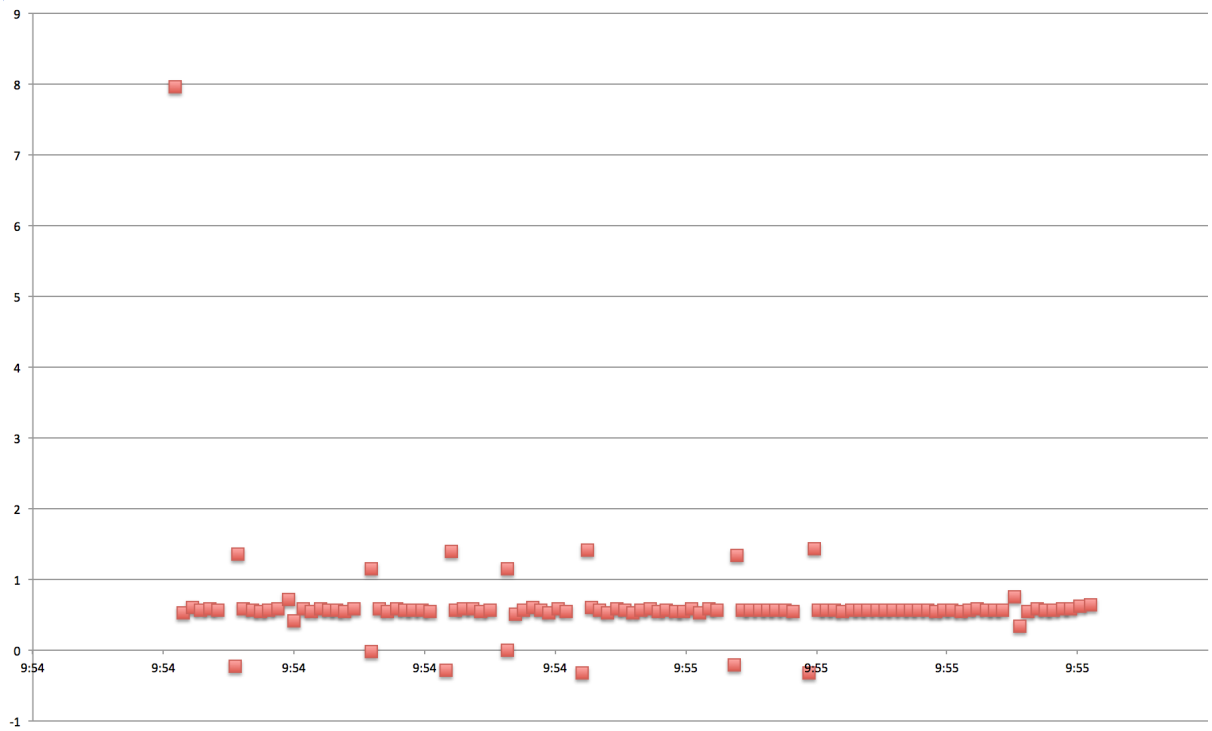

Figure 5.5: Walking Signature Sample #5 (Subject #2)

makes λ_i alone a poor classifier, even for our small population. Next we experimented with adding $W(\alpha_i)$ to λ_i as shown in Equation 2.15. It was our hope that the average weight of a walk would be a significant contribution to the estimated free energy value of a walk when combined with the inverse compression ratio.

We assign $-\lambda_i$ as a low weight for any duration $< \mu - \sigma$ We assign a weight of zero for any duration within σ of μ. We assign λ_i for durations $> \mu + \sigma$. Again we assign a weight to each observed step duration, then calculate a simple average for the entire walk. When $W(\alpha_i)$ is added to λ_i, the result is a better discrimination between Subject 1 and Subject 2 with difference of ≈ 0.07 to 0.10 compared to a difference of just ≈ 0.03 without the inclusion of $W(\alpha_i)$. However, as you can see in Table 5.3, the free energy values of Subject 3 and Subject 4 are still contained within the range of Subject 1 making estimated free energy for walking signatures a poor authentication mechanism on its own.

Clearly additional observations other than just step duration are necessary to further evaluate the effectiveness of our free energy approach to improving the sensitivity of individual walking signatures. Step durations that fell outside the range of σ made up a very small percentage of all the observed steps; just 3 of 77 durations for Subject 1 in Experiment 1 and 8 of 85 for Subject 1 in Experiment 3. 31% of Subject 2's step durations fell outside of σ. Subject 3 had 27Subject 4 had only 7% of their steps with durations outside of σ.

An alternate interface for encoding step durations may consist of a very small Σ to encode each duration as short, average, or long. Experiments of longer walks would be necessary to evaluate this interface. We expect the walking signature of each subject to consist of a combination of μ step duration, along with how often their step durations deviate from μ by more than σ, and if those deviations tend to be shorter or longer.

5.6 Related Work

Methods of biometric authentication have existed for decades, but most require specialized sensors to read fingerprints, iris patterns, or hand geometry. Most biometric readings are static sources, which makes them useful for authentication, but also relatively easy to spoof and difficult, it not impossible to revoke or change. Some research on dynamic biometric signatures has been done, but require specialized, intrusive sensors. For example, in 2014 Bionym received $14M in investment funds to use create a wearable electrocardiogram sensor to unlock devices [59]. My technique of using a dynamic biometric source was conceived independently and the biometric source is different, but it is promising that other work in this area is well funded.

As mentioned above, research regarding the accuracy of smartphone pedometer sensors [55] [56] [57] has already been conducted. However, we could not locate any preexisting research that identified using pedometer data to perform passive, continuous authentication. Therefore we believe our approach is novel.

5.7 Future Work

This paper describes an initial experiment with a very small population and limited number of samples. However, the initial results of the exploratory data analysis give us hope that our hypothesis may be true. Much additional work is necessary to prove our hypothesis. We also expect that additional work will be required to decrease the amount of observational time necessary to identify an individual. It is also likely that additional metadata may be necessary to accurately identify an individual.

Research to prove that this method is extremely difficult, if not impossible to spoof should also be conducted. Research is also necessary to understand the probability of a walking signature collision. A signature collision is a naturally occurring event when two or more individuals have the same walking signature. If the probability of a signature collision is very low, the benefits of this authentication method may well be worth the slight risk of a collision.

5.7.1 Larger Experimental Datasets

Larger experimental datasets are necessary to prove the hypothesis. The initial experiment had a limited population and number of samples per subject. More subjects, preferably with similar physical traits such as age, height, and weight are required to explore dependent variables. Likewise, many more samples from individual subjects are also necessary to determine how consistent an individual's walking signature is in spite of dependent variables. Both datasets should help create a mathematical model that generalizes to all walking signatures.

5.7.2 Additional Metadata

Once a mathematical model of walking signatures has been discovered, we expect that additional metadata will be required to more accurately identify an individual. For example, assume two subjects share the same traits that are known to be dependent variables. These variables will then need to be measured with enough precision to differentiate the individuals. If this cannot be done, different metadata, or additional metadata will need to collected. Height may be sufficient do

differentiate users of different height, but not users of the same height. In that case, data from the accelerometer may be used to differentiate users of the same height but different weight. Users of the same height and weight may be differentiated by the average angle and orientation that they hold their smartphone when in use. This metadata may be available from the inclinometer and magnetometer. Additional behavioral observation techniques such as "geofencing" and network usage may also be helpful in passively authenticating a user.

The concept of "trust decay" is also important. There may be times when a user's behavior changes drastically enough to alter their walking signature, perhaps due to an injury. When this happens, additional authentication methods may be necessary to establish a new, temporary walking signature. Securing this process will be crucial to preventing an attacker from using it to assume the identity of another user. This process may not be necessary if enough behavioral metadata can be collected to verify a user's identity in absence of their walking signature. The primary idea is that the walking signature is computationally inexpensive to compute and almost continually updated by the user's motion throughout the day.

5.7.3 Different Analytical Approaches

Classical statistical approaches to analyzing the initial experiment data appears to be sufficient for revealing distinct patterns that could be used to classify samples by user. When larger datasets are collected statistics may not be sufficiently provide for passive, continuous authentication. In this case, there are several information theory approaches that may be helpful.

One approach involves treating the observes steps as a continuous stream of information. By assigning a unique symbol to each piece of metadata, applying a Burrows-Wheeler transformation [60] then compressing the resulting string using the Lempel-Ziv-Welch algorithm [61], the amount of information in the stream can be compared with another stream. If both streams share a high amount of mutual information, than the smartphone can continue to authenticate the user. If the amount of mutual information changes, the smartphone can determine that it is no longer in the

possession of the authorized user and take steps to protect its original owner. Other information theoretic approaches may also be applied to the problem of passive, continuous user authentication.

5.7.4 Prototype Application

Ideally, this new authentication scheme would be combined with an identity management system and integrate with existing applications through available APIs (i.e. Application Program Interfaces). The application would manage a unique, complex password for each associated resource. It is not necessary for the user to know these passwords if resources are accessed from the smartphone. If the smartphone trusts the current user based on their walking signature, it can relay the complex password to the requested resource.

Of course, complex passwords are still vulnerable to brute force attacks, or theft and cracking. To combat these attacks, the application can change complex passwords frequently. The application could even treat complex passwords as a one-time use scheme, greatly reducing the probability of compromise due to external attacks.

A configurable trust decay setting could be used to re-authenticate a user when the smartphone observes a change in identity or a period of user inactivity elapses.

If the smartphone is unable to authenticate the current user, it can revoke the current set of complex passwords. This would of course require network connectivity. A physical attack on the smartphone may prevent this by placing it in airplane mode, or blocking RF transmissions by placing the smartphone in multiple anti-static bags, a paint can, or simply turning it off. To prevent this, a cloud-based identity management service may be used. The identity management service on the smartphone would communicate with the cloud-based service at a regular interval, configurable by the user based on their level of risk and Internet access. If the cloud service did not receive a heartbeat signal from the smartphone at the expected interval, the cloud service could revoke the user's passwords or at least change them, and continue to change them at an interval that minimized the chance that access to the resource could be compromised.

5.8 Conclusion

Exploratory data analysis using classical statistics appears to reveal unique walking signatures between the subjects in the initial experiment. Further, these walking signatures appear to not be influenced by several variables that were expected to be dependant. These results are sufficient to motivate additional research in this area, including gathering larger data sets, applying additional analytical methods, and developing an identity management system based on the passive, persistent user authentication.

Acknowledgment

William would also like to thank Margaret Pieczkowski for her help debugging his first Android app.