Systems and Models for Secure Fallback Authentication

by

Alaadin Addas

A thesis submitted to the School of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements for the degree of

Master of Science

in

The Faculty of Business and IT Computer Science

University of Ontario Institute of Technology

Oshawa, Ontario, Canada

December 2018

Thesis Examination Information

Submitted by: Alaadin Addas

Master of Science in Computer Science

Thesis Title: Systems and Models for Secure Fallback Authentication

An oral defense of this thesis took place on the 27th of July, 2018 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee: Dr. Shahram Heydari

Research Supervisor: Dr. Julie Thorpe

Examining Committee Member: Dr. Amirali Salehi-Abari

Examining Committee Member: Dr. Stephen Marsh

External Examiner: Dr. Pejman Mirza-Babaei The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments.

The research work in this thesis that was performed in compliance with the regulations of UOIT's Research Ethics Board under REB certificate number:14450.

Alaadin Addas

December 2018

Acknowledgements

My research and this thesis would not be possible without help and support from many individuals including; my supervisor Dr. Julie Thorpe, committee members, family, and friends.

I would like to especially thank my supervisor Dr. Julie Thorpe, for her continued support, guidance, and input throughout the research process and throughout my degree in general. Her commitment to guiding me throughout the research, and her valuable input made this research possible.

I would also like to give a special thanks to my thesis defense committee Dr. Amirali Salehi-Abari, Dr. Stephen Marsh, and Dr. Pejman Mirza-Babaei for their valuable input throughout the defense and the thesis edits process. Their input greatly enhanced the quality of this work.

My sincere gratitude goes to NSERC for funding sources that made my degree funding possible, in addition to the Faculty of Business and IT at UOIT.

Lastly, I would like to thank my parents for their continued support and guidance over the years. Without them I would have never been able to complete my degree.

Abstract

Fallback authentication (FA) techniques such as security questions, Email resets, and SMS resets have significant security flaws that easily undermine the primary method of authentication. Security questions have been shown to be often guessable. Email resets assume a secure channel of communication and pose the threat of the avalanche effect; where one compromised email account can compromise a series of other accounts. SMS resets also assume a secure channel of communication and are vulnerable to attacks on telecommunications protocols. Additionally, all of these FA techniques are vulnerable to the known adversary. The known adversary is any individual with elevated knowledge of a potential victim, or elevated access to a potential victim's devices that uses these privileges with malicious intent, undermining the most commonly used FA techniques.

An authentication system is only as strong as its weakest link; in many cases this is the FA technique used. As a result of that, we explore one new and one altered FA system: GeoPassHints a geographic authentication system paired with a secret note, as well as GeoSQ, an autobiographical authentication scheme that relies on location data to generate questions. We also propose three models to quantify the known adversary in order to establish an improved measurement tool for security research. We test GeoSQ and GeoPassHints for usability, security, and deployability through a user study with paired participants (n=34). We also evaluate the models for the purpose of measuring vulnerabilities to the known adversary by correlating the scores obtained in each model to the successful guesses that our participant pairs made.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Fallback authentication, also known as *secondary authentication*, is the recovery process that occurs when the primary means of authentication such as entering a password has failed. The three most common fallback authentication methods are security questions which are sometimes referred to as personal knowledge questions, Email resets, and SMS resets. The fallback authentication method is very important because it can completely bypass a strong primary authentication, and a weak fallback authentication system undermines the security of our accounts/devices.

The most commonly utilized fallback authentication methods are especially vulnerable to what we call the known adversary. The known adversary is any individual with elevated first hand knowledge of a potential victim and/or elevated access to a potential victim's devices, that uses these privileges with malicious intent. Elevated knowledge of a potential attack victim can make it very easy for an attacker to correctly answer security questions as a fallback authentication method. Elevated access to a potential victim's devices (e.g., mobile device is left lying around) makes Email resets and SMS resets vulnerable to the known adversary. We specify first hand knowledge to make a clear distinction between first hand knowledge and knowledge that can be gained about a person remotely (e.g., from their social media accounts). In this work we aim to design, implement and test alternative fallback authentication methods that are more secure, and we attempt to formally quantify the known adversary for the purposes of improving measurements of its success in security research.

1.1 Motivation

Commonly utilized fallback authentication techniques such as security questions, Email resets, and SMS resets all have proven security and usability flaws [17,31,34,35]. Security questions are low entropy making them easy to guess [17, 31]. Email resets and SMS resets rely on a secure channel of communication and are vulnerable to the known adversary [38, 40, 42, 55]. This motivated our effort to design, implement, and test alternative fallback authentication systems.

An important threat to evaluate for fallback authentication systems is the known adversary. The known adversary is the threat posed to the security of our accounts/devices from social insiders. The known adversary was first identified as the "insider threat" by Muslukhov et al. [42]. We use this term to avoid confusion with the more common use of the term "insider threat", which is most often used to describe trusted individuals within an organization who abuse their power [52]. Security research typically evaluates vulnerability to the known adversary threat utilizing two different tiers (strong and weak) of self-reported social insiders based on self-reported social closeness [13, 34, 35]. We hypothesize this is a flaw, as such a threat is not only posed by a social insider (e.g., a spouse) but could be posed by many acquaintances (e.g., co-worker, acquaintance, boss). Thus, we explore different alternative models to measure the known adversary threat, accounting for the fact that the known adversary is any individual with elevated access to a potential victim's devices and/or elevated first-hand knowledge of the potential victim that uses that privilege with malicious intent.

1.2 Thesis Summary

In this work we design, implement, and test two candidate fallback authentication systems for security, usability and deployability. GeoPassHints is a geographical authentication system, and GeoSQ that uses autobiographical location data. We tested both systems through a user study involving three different sessions. Session 1 had 19 pairs (38 participants). Session 2, which was conducted 7-11 days after Session 1, had 17 pairs (34 participants) who returned. Lastly, Session 3 is scheduled to be 6 months after Session 2 and has not been completed as of this date. However, our results from Session 2 are sufficiently conclusive that we can report on our results at this time. We do not expect any phenomenal improvements in Session 3 due to the memorability results bad usability results obtained for GeoPassHints in Session 2.

We also design and create three models to quantify the known adversary. In order to test whether or not the models for quantifying the known adversary yielded any results, our study included a guessing phase, where each participant actively guessed their pair's GeoPassHints secret notes and GeoSQ autobiographical locations. We attempted a correlation between the scores obtained in the models, and the results of the GeoSQ pair guessing scores. A high correlation meant that the model was successful. Two of the three models we designed and created were successful. All of the models we created had a more positive correlation with the GeoSQ pair guessing scores than simple self-reporting the relationship characterization by participants, which security research has typically relied on [13, 34, 35]. Note that this was not done for GeoPassHints as there were so few successful guesses by the known adversaries.

We concluded that two models were successful (the KAI and the adapted RCI). We also concluded that while GeoPassHints and GeoSQ offered increased security in some categories when compared to commonly utilized fallback authentication techniques, they were unusable due to high error rates, and long login times on average.

1.3 Thesis Statement

This thesis explores one new and one altered fallback authentication system as a possible replacement to the current commonly utilized fallback authentication systems. This thesis also explores models to quantify the known adversary for the purposes of a more formal classification.

The following questions will be addressed in this research:

1. How does the security, usability, and deployability of GeoPassHints compare to the commonly utilized fallback authentication techniques?

2. How does the security, usability, and deployability of GeoSQ compare to the commonly utilized fallback authentication techniques?

3. Do the models created for quantifying the known adversary prove to be promising in showing a correlation between pair guesses and scores obtained from the models?

4. How does utilizing the models we have created to quantify the known adversary compare to self-reported relationship characterization in identifying the known adversary?

1.4 Contributions

Our contributions are as follows: (1) We design, implement, and test a new geographic fallback authentication system called GeoPassHints which is inspired by GeoPassNotes [39], location-based security questions [35], and work on authentication hints [11]. (2) We design, implement and test an altered autobiographical location authentication system we call GeoSQ. (3) We conduct a user study in order to test the usability, security, and deployability of both authentication systems and collect data for evaluating our models to quantify the known adversary. (4) We develop three models to quantify the known adversary and evaluate the correlation between the scores from the models and the GeoSQ pair guessing scores, demonstrating that our models offer an improvement over the typical self-relationship declaration utilized by a previous security research in this area [13, 34, 35].

1.5 Thesis Organization

Following Chapter 1, this thesis is organized as follows. Chapter 2 presents a literature review of the known adversary threat, flaws in current fallback authentication systems, geographic authentication systems, autobiographical authentication systems, measuring relationship closeness, utilizing hints for increased memorability, multiple passwords interference, and lastly a framework for the comparative evaluation of web authentication schemes. Chapter 3 will detail the user study. Chapter 4 will introduce the models for measuring the known adversary that we created, and it will include the comparison as well as the statistical analysis of results. Chapter 5 will evaluate GeoSQ for security, usability, and deployability. Chapter 6 will evaluate GeoPassHints for security, usability, and deployability. Chapter 7 will include the discussion and future work, and lastly, Chapter 8 will conclude the thesis.

Chapter 2

Literature Review

2.1 Overview

Commonly utilized fallback authentication techniques such as security questions, email resets, and SMS resets have many flaws in their security and usability [13,17,31,34,35] that have motivated many attempts to try to find alternatives [13,34,35]. Despite all their security flaws, security questions, email resets, and SMS resets continue to be the dominant fallback authentication techniques used today [17, 31, 35], due to their usability.

A threat that undermines the security of all the common fallback authentication techniques is the threat of the known adversary [42]. The known adversary is any potential attacker with elevated knowledge of a potential victim or elevated access to the devices of a potential victim as a result of any type of relationship (e.g., friend, close friend, spouse, co-worker) that uses that privilege for malicious intent. As such, our research focuses on finding alternative authentication techniques that have the ability to deter the known adversary while at the same time providing increased usability and security across many different criteria.

We begin our literature review with related work that identifies how pertinent of a threat the known adversary is (see Section 2.3).

Next, we explore related work on scales utilized in the field of Social Psychology to measure and quantify relationship closeness (see Section 2.5), which we attempt to do in an effort to formally quantify the known adversary. The next portion of our literature review focuses on current flaws the most commonly utilized fallback authentication techniques (see Section 2.6). Followed by that, we explore related work on alternative authentication techniques that motivated our research, such as autobiographical authentication and geographic authentication (see Section 2.4).

After exploring related work that attempts to measure relationship closeness, we move on to exploring related work on improving the memorability of geographic, autobiographical, and classical password-based authentication techniques (see Section 2.7). We also explore the effect of multiple password interference, which occurs when users utilize an authentication system in a real-world scenario for multiple accounts/devices [11, 22] (see Section 2.8). Multiple password interference occurs when the authentication credentials associated with one account are mistakenly input while attempting to authenticate into a different account (e.g., using your Facebook password while trying to login to your Twitter account).

Next, we explore the Bonneau et al. [18] framework for the comparative evaluation of authentication frameworks. In our work, we utilize this framework as it provides us with guidelines and criteria that formalize the evaluation of the two systems that we designed and implemented (GeoPassHints and GeoSQ). The Bonneau et al. framework provides us with criteria to evaluate our authentication systems for usability, deployability, and security (see Section 2.9). Furthermore, since passwords are the baseline in which the framework for the comparative evaluation of authentication frameworks relies on and they are mentioned often as a reference, Section 2.10 will cover related work on passwords. Lastly we will highlight the focus of our research in relation to previous work in Section **??**.

2.2 Research Focus

Due to the breadth of the related work presented, it is pertinent that we present the focus of this research in relation to the related work.

We will begin this literature review with a focus on the known adversary, quantifying the known adversary, and the flaws that exist within currently utilized fallback authentication systems. This was done to ground our goal of researching alternative fallback authentication methods that are more secure and usable.

We go on to explore the field of alternative authentication techniques. Section 2.6 of the literature review was pertinent to explore what has already been done in the field of alternative authentication techniques. We review this work to ensure that our implementation is not flawed due to some usability or security flaw discovered by other research. Alternative authentication techniques are oftentimes authentication systems that have been researched but not implemented on a large scale by a giant verifier such as Google. In section 2.6 we discuss several different alternative authentication systems relevant to GeoPassHints and GeoSQ.

Furthermore, in order to improve the memorability of GeoPassHints and GeoSQ, we explore what related work has discovered in terms of improving the memorability of autobiographic, geographic, and classic password based authentication schemes. We used that related work in order to attempt to improve the memorability of GeoSQ and GeoPassHints. We attempted to include the lessons learned from related work to the best of our ability in our implementation of GeoSQ and GeoPassHints.

Furthermore, we discuss password interference because any new authentication system that is meant to be utilized on a larger scale needs to account for the user having multiple accounts while verifying in the same or similar manner. The interference effect is important because it could affect the usability (specifically the memorability) in addition to the security of the authentication systems we are testing.

We discussed a comparative framework for evaluating authentication schemes, more specifically the Bonneau et al. framework for comparatively evaluating authentication schemes [18]. This was important to our research because we needed a formal framework to compare the security, usability, and deployability of GeoPassHints and GeoSQ to that of similar authentication schemes or at least to fallback authentication schemes that are currently being widely utilized.

Lastly, we will discuss relevant research in text based passwords because passwords and password variants are the most utilized authentication mechanism in the world. As a result, many security researchers compare their authentication schemes in terms of security and usability to that of the common text based password and its variants. In addition to that, it was important for us to know what constitutes a safe and secure password because GeoPassHints contains a secret note component to authentication which could be considered similar to a text based password. We utilized the related work in that field to set policies which will be explained in later chapters.

Our work builds on previously proposed alternative authentication systems and improves on them by incorporating the important lessons learned from related work. We also utilize the work done in the field of Social Psychology in order to better understand and accurately define the known adversary. We created our models for quantifying the known adversary based on the related work in the field of Social Psychology.

We proposed and tested GeoPassHints; a novel authentication system that was based on GeoPassNotes and GeoPass [39] [48]. GeoPassHints is a significant alteration of GeoPassNotes. GeoPassNotes is an authentication system that utilizes a location input on a map in addition to a secret note in order to authenticate, while GeoPassHints utilizes the location set by the user as a cue to aid in the user's ability to remember the secret note. This is the first study of this kind for GeoPassHints and it is the first comprehensive usability and security analysis for this kind of system. In addition to that, GeoSQ is an autobiographic authentication scheme that was implemented and improved upon by us but was originally implemented and proposed by related work [13] [12] [35]. We improve on GeoSQ by ensuring that we included the lessons learned from related work within the settings of the implementation. We utilize both alternative authentication systems in order to test our models pertaining to the known adversary. The models are a highly important contribution because they are one of the first attempts to classify known adversaries without the need for self declaration. By proposing and implementing GeoPassHints, and implementing and improving GeoSQ as well as running an in person lab study to test the security and usability of both authentication systems, we further contribute to the research conducted by related works by analyzing alternative authentication schemes and conducting a thorough security and usability analysis.

2.3 Known Adversary

Muslukhov et al. [42], to the best of our knowledge, present the first to attempt to study the prevalence of breaches by insiders through the use of an online survey (n=724). The survey results identified that 12% of surveyed users were aware of a breach of their privacy by an individual the user considered an insider, which we refer to as a known adversary (to avoid confusion with the more common use of the term “insider” which is most often used to describe trusted individuals within an organization who abuse their trust [52]). In addition, 9% of survey participants admitted to accessing the device of a socially close individual without consent. Through this survey, Muslukhov et al. demonstrated that the fear of the known adversary threat was indeed a valid one among many users and showed that known adversaries accessing devices and accounts without authorization is indeed a major security flaw. This security flaw indicates that we must enhance our security research in order to account for this threat, since a known adversary can bypass many traditional security mechanisms.

Security research evaluating new authentication systems typically relies on recruiting participant pairs in order to evaluate the threat of the known adversary [12, 13, 34, 35]. Followed by that, they classify the known participating pairs into strong and weak known adversaries.

2.4 Measuring Relationship Closeness

The known adversary is an important threat model in security research discussed in Section 2.5 due to the elevated access to devices and elevated first-hand knowledge of the potential victim. To quantify whether a pair is indeed a known adversary, we must first be able to quantify the relationship between the adversary and the participant in security research. In the quest to quantify relationship closeness, we discuss several different relationship closeness indices/scales used in the social sciences.

Berscheidl et al. [16] proposed one of the most popular relationship closeness measurement tools in the form of a questionnaire (Appendix A). The Relationship Closeness Inventory (henceforth RCI) is used to quantify the closeness of two people to each other, but is not specifically tailored towards security research and quantifying the known adversary in today's world. The RCI in its original format as proposed by Berscheidl et al. does not contain any questions regarding pairs accessing each other's devices, which is more pertinent now that it was in 1989.

The RCI is split into three subscales starting with RCI Frequency which measures the frequency of activities being performed together by the pair. This is particularly important when trying to quantify the known adversary because the more time spent together, the more autobiographical data is less secure (i.e., if a pair spend a lot of time together autobiographical location questions become obsolete because the known adversary could easily know all the answers). The second RCI subscale is RCI Diversity which measures the diverse breadth of activities that pairs undertake together. The third RCI subscale is RCI Strength, which measures the amount of influence an individual has over his/her pair. The sum of all three subscales make up the RCI score with quantifies how close two people are to each other.

We create an adapted RCI model, which includes portions of the RCI in addition to questions to account for modern technology (i.e., access to devices, access to accounts etc... (see Appendix C). In this work, we attempt to establish a correlation between the number of successful guesses and the adapted RCI score using a linear regression analysis, in an attempt to validate the use of the adapted RCI for the purpose of quantifying the known adversary.

Another popular tool from the field of Social Psychology is the Inclusion of the Other in the Self Scale (henceforth IoS Scale) developed by Aron and Smollan [14]. The IoS Scale is a simple pictorial tool that allows the participant to pick the image that most closely resembles the relationship to the pair (see Appendix B). The IoS Scale is a very simple pictorial tool that does not take that long to complete by the participant. It was created with the goal of replacing the RCI, as the amount of time it takes for a participant to complete the RCI questionnaire is often a limiting factor due to time constraints.

Cialdini et al. [24] contributed the We Scale. The We Scale presents a Likert Scale question with seven levels. The question asked is: “Please, select the appropriate number below to indicate to what extent you would use the term “WE” to characterize you and this individual”. The lowest on the seven-point scale is 1 = “not at all” and the highest is 7 = “very much so”. This is another compact way to measure closeness. The twist is that Cialdini suggests what is referred to as a oneness score. The average between the IoS Scale and the We Scale is called the oneness score. In this work we attempt to establish a correlation between the number of successful guesses and the oneness score by means of a linear regression analysis for the purpose of validating the use of Social Psychology frameworks within security research to quantify the known adversary.

Gachter et al. [28] performed a study (n= 772) on Amazon’s Mechanical Turk with the aim of establishing a correlation between the IoS Scale, and several other relationship closeness scales/inventories. The result was positive and a high correlation was found when a principal component analysis was conducted to construct an Index of Relationship Closeness in order to correlate it with the results obtained from the IoS scale. The Index of Relationship Closeness is essentially a table that shows the correlation of many different relationship closeness metrics to the IoS Scale, based on the data from 772 participants. We are most concerned with the high correlation between the IoS Scale and the RCI because the IoS Scale is more compact and does not ask as many questions (making it more efficient in a sure study). In this work, we compare the correlation of the successful guesses and the oneness score obtained with the correlation of the successful guesses and the adapted RCI score (see Appendix B and C). We perform this comparison analysis in order to establish whether a compact tool for quantifying relationship closeness is sufficient instead of using the adapted RCI score which is more time consuming for participants in security research.

2.5 Flaws and Attacks in/on Current Popular Fallback Authentication Techniques

The three most popular fallback authentication mechanisms are security questions, Email resets, and SMS resets. In this section we will list a number of flaws and attacks on these fallback authentication techniques. The weaknesses presented are the motivation for our attempt to design, implement, and test new authentication systems.

2.5.1 Security Questions

The safety of fallback authentication mechanisms was put into question by several different researchers in the field. According to Just et al. [37], security questions often have a very low entropy, even when users are prompted to purposefully input security questions and security answers that would be tough to guess. The study found that despite the prompt, the answers were still considered to be very guessable.

Another study completed by Golla et al. [31] evaluated the security of challenge questions. After completing an analysis on 3.9 million answers to security questions that were leaked from an online dating website, Golla et al. found that the answers were extremely low entropy making them very easy to guess. Furthermore, users would input fake answers in an attempt to make it less guessable. After analyzing that data, Golla et al. found that inputting fake answers actually had a negative impact on the true entropy hence further decreasing the security of security questions.

Bonneau et al. [17] also performed an analysis on the security and usability of security questions.

The findings were in line with Golla et al. [31]. Bonneau et al. found that in 37% of instances, users would attempt to make the answers harder to guess by adding a word to the end or the beginning of an answer. Often times this led to the exact opposite effect, mainly due to the predictability in the methods which users were deploying to make their answers seemingly unpredictable. The study goes on to establish the perceived memorability versus the actual memorability. Through the analysis of millions of password recovery attempts, it was found that 40% of users failed to remember their security questions' answers when needed. This is contradictory to the inspiration behind using security questions, which is their perceived usability and memorability. Bonneau et al. went on to compare SMS resets to security questions and found that an SMS reset code had an 80% chance of success which is significantly higher than the chance of success with security questions (which was 60%).

2.5.2 Email Resets

Email resets have also been proven to be effective in terms of usability for fallback authentication. Garfinkel [29] presented one of the earliest works investigating utilizing email resets instead of personal knowledge questions. However, several security considerations were also discussed. The main security consideration when utilizing email for account password resets was making the email a single point of attack as well as a single point of failure. Should the recovery email account be compromised, many other accounts become compromised and linkable. Accounts are linkable when one account being compromised leads to other accounts being compromised, even if they are registered with different verifiers. This can be as a result of the same credentials being used across multiple different accounts, or because the security of one account is based on the security of another. Garfinkel also noted an important usability concern that occurs when the user forgets the password to the email or has lost access to the email that is designated as the reset email.

Guri et al. [32] describe a man in the middle attack that proves the secure channel that email resets rely on can be compromised. The man in the middle attack involves a malicious application on a smartphone that requests access to resources and applications that it is not supposed to have, and does not need to have. This includes compromising popular emailing applications which could compromise email resets to a great extent. This attack using a malicious application can also be used to compromise SMS resets as well and will be described in greater detail in Section 2.4.3. Malicious applications with the ability to snoop on our communications channels undermine the security provided by email resets and SMS resets.

Email resets are also vulnerable to the known adversary. The primary reason being the fact that email resets are easy to view by a known adversary. By definition, a known adversary is any individual with elevated access to a potential victim's devices and/or elevated first hand knowledge of the potential victim. Email resets can often be seen on the notifications of a locked screen, depending on the settings set by the user on a smartphone.

Furthermore email resets are also vulnerable to the known adversary because man in the middle attacks as described by Guri et al. [32] are easier to execute with increased physical access to a potential victim's devices. The known adversary is also connected to the same networks that the potential victim is connected to. With this advantage many other kinds of snooping attacks on email resets are possible.

2.5.3 SMS Resets

As for SMS resets, they are highly vulnerable to many different types of attacks. Al-Maqlabi and Mitchell [40] noted that most SMS reset codes are only six digits long, which means there are only 10^6 possibilities which makes it possible for an attacker to run a guessing algorithm. As such, Al-Maqlabi and Mitchell note that it is imperative for websites to limit the number of attempts a user is allowed. They also note that SMS resets rely on one factor, which is that the channel in which the SMS is being sent is actually secure. However, the authors refer to successful attacks that compromise the contents of an SMS via the usage of an IMSI catcher or “Stingray” [38]. An IMSI catcher will mimic a cell tower and send signals to the mobile phone to trick it into connecting to it, and trick it into sending information that should only be shared with the cell tower. IMSI catchers do have certain limitations because they have to be in the same area as the target mobile phone.

The use of attacks on telecommunications protocols such as the SS7 protocol [55], which allows telecommunications companies to share information for accurate billing, also compromises the supposedly secure channel that is being relied upon to deliver SMS reset codes for account recovery. The SS7 protocol which allows telecommunications companies to share information for accurate billing, is an old protocol that has not been updated in quite some time; as a result there are some communications protocols that run on SS7 that have no authentication protocol [7] (i.e., a phone with no PIN to secure it). Attackers exploiting the SS7 protocol vulnerabilities can triangulate a cellphones location, view sensitive billing data, and monitor communications (calls and SMS). This attack does not share the limitations posed by IMSI catchers because they can be conducted remotely, without the need to be in the immediate area of the potential victim.

It is important to note that the flaw in the SS7 protocol is promptly being mitigated by major telecommunications companies. However, Welch et al. [55], did manage to successfully demonstrate the exploits available to an attacker as a result of the flaws in the SS7 protocol. In addition to that, many online sources provide detailed instructions on how to take advantage of the exploit [7]. The tools required are pretty simple and easy to obtain. Most types of Linux OS can be utilized to take advatnage of the exploits, in tandem with a free and readily available SS7 software development kit (SDK) [7]. We reiterate that telecommunications networks have already patched some of the exploits at the time of this writing.

Gelerenter et al [30] describe a man in the middle attack on fallback authentication systems that rely on two techniques (i.e., security questions + SMS resets, or security questions + Email resets). The attack is a mixture of a man in the middle attack and phishing. The attack is focused on a malicious website that has the primary purpose of collecting answers to security questions. The second phase of the attack is telling the user to expect an SMS reset code. At that point in time, the attacker is actually on the legitimate website beginning the steps required to send out an SMS reset code (with the power of the answers to the security questions) from the legitimate website which the user actually has an account on. The target then receives that SMS reset code and since the SMS reset code is not labelled (i.e., your reset code is xyz but the message does not say which account this reset code is for), the targeted user will then enter the SMS reset code into the malicious website, providing the attacker with all the information necessary to access the account. Following the same steps this attack was also described as a means to attack Email resets. The authors go on to describe some of the counter measures that verifiers can use to avoid such an attack from being successful, such as inputting the website name the reset code was designated for (e.g., this is your reset code for your Twitter account, so that the user does not input the reset code into the malicious website), and making the code valid for a limited period of time.

Guri et al [32] perform a thorough analysis on password recovery techniques (fallback authentication techniques) for popular internet service providers such as Facebook, Gmail, and Twitter. In their review of the password recovery techniques, they found leakage of private information such as partial Emails and phone numbers that may be useful to attackers. Within their examination of password recovery techniques for popular internet service providers they describe possible attacks on SMS resets and email resets. The attack scenario described in this work is based on a malicious Android application that requests access to resources such as SMS texts and emails. This jeopardizes SMS resets and email resets since many applications can request access to resources they do not need. It is important to note that more recent versions of Android (Android Nougat, and Android Oreo) and iOS have significantly limited the ability to request access to resources and have required explicit consent for each resource being accessed, rather than a one-time acceptance of all resources accessed at the time of installation. Nevertheless, for many devices that still run older versions of Android or iOS, it is highly possible that an application can indeed ask for consent to utilize resources that are not necessarily pertinent to the stated function of the application, and therefore be able to snoop on email reset codes and SMS reset codes.

Much like email resets, SMS resets are also vulnerable to the known adversary for the same reasons. SMS reset codes can oftentimes be viewed form a lock screen, depending on the privacy settings. A known adversary that has elevated access to a potential victim's devices can reasonably initiate a reset and obtain the code from the potential victim's phone (since the known adversary enjoys elevated access).

The preceding work gave us ample evidence that a security risk is posed by security questions, Email resets, and SMS resets which are currently the most commonly utilized fallback authentication techniques [30, 31, 37, 40]. Also, there is a lack of usability when it comes to security questions [17]. As a result, a great deal of research is being done into alternatives to current fallback authentication techniques. One such alternative is the usage of autobiographical authentication; another alternative is the use of geographic authentication.

2.5.4 User Attitudes and Perceptions

Before we delve into the different alternative fallback authentication techniques, it is important to note user attitudes towards current fallback authentication techniques.

Hang et al. [35] performed a user-based survey (n=244) to gauge public opinion on the fallback authentication process that occurs as a result of device lockout, account lockout, and SIM lockout. Device lockout occurs when a user enters a PIN, pattern, or password incorrectly a number of times. Typically, what occurs is a throttling mechanism called exponential back-off where the user has to wait a predetermined time after entering an incorrect credential. After the lockout period, if the credential is entered incorrectly again, then the wait time before the user can re-enter the credentials is increased. In some cases, the exponential back-off is not reversible and will have to be taken to the manufacturer for a reset.

Account lockout also occurs when credentials are entered incorrectly; however, the typical fallback authentication mechanism is a security question or an Email reset. The work by Hang et al. [35] delved into SMS resets which are also commonly used for fallback authentication. The third type of lockout discussed is the SIM card lockout, this is an uncommon type of lockout that occurs only if the user has a PIN set for the SIM card. In the case of a SIM card lockout, the user typically must enter some sort of PUK (PIN Unlock Key).

The work done by Hang et al. [35] is important because it showcases user attitudes towards the usability of current fallback authentication methods. The study concluded that the majority of participants surveyed appear satisfied with current fallback authentication mechanisms. However, the edge cases such as being in a place with no cell reception when initiating a SIM reset are problematic. Participants appeared to be most content with fallback authentication for account lockouts as those tended to be the easiest to recover. Followed by that they ranked recovery from a device lockout as the second easiest with some persistent cases which were particularly annoying (e.g., cases were the user had to go to the manufacturer to unlock the device). Lastly, SIM lockout recovery was the most difficult but not many users in the study had experienced it before. Despite the lack of security current systems offer, the usability of the systems will ultimately decide whether or not a system is utilized. The usability of the systems still has to be matched by relative security. However, as it stands today, the most usable systems are not often them most secure.

2.6 Alternative Authentication Techniques

In the same work where Hang et al. [35] gauged user attitudes towards current fallback authentication techniques, Hang et al. propose an alternate authentication technique which relies on an app based utilization queries. In the case of a device lockout for example, an app would activate that would ask the user regarding apps that are current installed on the smartphone. If the user answered 80% of the questions correctly then the user would be granted access.

Hang et al. [35] incorporated the known adversary into the work by asking recruited participants to bring a strong adversary (someone the participant is close to). The pairs self-declared their relationship to the owner of the phone, and the participants were asked to bring someone along that they would consider close. The participants were asked to bring someone close to them to mimic a strong adversary and were given examples of who a close adversary would be such as a partner or a close friend [35]. This is part of the problem that we attempt to solve by attempting to quantify the relationship between the participant and the adversary that is typically present within security research. The known/strong adversary in this scenario is only being imagined as an individual that is socially close to the participant, and this is being self-reported rather than quantified.

The results of this study were very promising in terms of repelling the threat of the known adversary, because while the threshold for allowing access was 80%, no adversary achieved over 61% correctness. Legitimate users had a 95% success rate with 15 recruited participants, and each recruited participant brought along an adversary. App installation history is a form of autobiographical authentication. The questions take the form of multiple choice questions asking which applications were recently installed, and which applications were recently uninstalled.

Another study that explored alternative authentication techniques, specifically autobiographical authentication was the study completed by Albayram et al. [13]. Albayram et al. focus on the usability, and security of using different types of autobiographical data for authentication. Many different types of autobiographical data are collected and used for the purpose of fallback authentication. Autobiographical data can include anything from mobile device usage data, communication data, or location data.

Call (incoming and outgoing), SMS (incoming and outgoing), location, application, music, activity, and battery data were collected for the purposes of evaluating their uses from a usability and security standpoint. The security standpoint was evaluated through the usage of an adversary model, one that is naïve and the other is a strong adversary. The study included 24 participants recruited in pairs (12 pairs) and the study spanned 30 days.

In the 30-day field study, the user's data was collected using a monitoring application on their mobile devices. Followed by that they were queried about their autobiographical data.

Nine autobiographical questions were asked of each participant (one from each category; call and SMS are considered two categories incoming and outgoing), and the participants were only asked questions about data collected in the last 24 hours. AlBayram et al. noted that the reasoning behind utilizing data collected in the last 24 hours only was the fact that autobiographical data is episodic memory which tends to be more memorable in shorter time spans [25].

The study determined an accuracy score for each type of user in relation to the type of autobiographical data was obtained. Table 2.1 below shows the accuracy scores.

Question Type	Accuracy Score		
	Legitimate	Strong	Naive
Call	0.76	0.13	0.008
SMS	0.46	0.08	0.002
Location	0.69	0.29	0.038
App	0.55	-0.03	-0.549
Music	0.46	-0.71	-1.782
Activity	0.42	0.06	-0.240
Battery	0.53	-0.005	-0.157

Table 2.1: Accuracy in responses for each type of user and each type of autobiographical data collected and queried. Negative scores are as a result of penalty system imposed on incorrect answers. Please note that this is remake of the original table. [13]

The results as shown in Table 2.1 above summarize the security-usability tradeoff. Given a known adversary threat model, the use of autobiographical data for fallback authentication proved highly effective in repelling an attacker. At the same time, the accuracy rates by legitimate users proves that it is not highly usable without a low threshold for authentication. The two types of autobiographical data that yielded the highest accuracy from legitimate users are call and location data. Both those types also yielded the highest accuracy among strong adversaries. This could indicate a relation between the implementation method (e.g., only utilizing location data from the past 24 hours which could be easier to guess) and the method that the accuracy score is calculated, as opposed to true usability and a security flaw.

Negative accuracy score by naive/strong adversaries are as a result of the penalty system put in place in the accuracy score calculation for incorrect answers, in this penalty system it is not simply the number of correct answers out of nine that is taken into consideration. Instead a series of incorrect answers could lead to negative scores because the user is being punished by the system for inputting incorrect answers and not utilizing a simple threshold (e.g., 5 out of 9 correct answers is a successful login).

It is important to note that the adversaries were self-declared and that this the motivation for improving this practice in security research. We would like an empirical study to determine closeness and that is the reason why the models for discovering the known adversary were created.

Hang et al. [34] also conducted a similar study on autobiographical data being used for the purpose of fallback authentication. In this study, seven categories of autobiographical data were examined. SMS out, SMS in, call out, call in, application (usage and installation details), music, and photos. This is slightly different than the autobiographical data collected in Albayram et al's. field studies. The results yielded were significantly different. In Hang's work, communication data such as SMS in, SMS out, calls in, and calls out, as well as Application installation / uninstallation autobiographical data proved to be the most promising. This is in contrast to Albayram et al's results [13], which concluded that, SMS related queries did not yield enough accuracy among legitimate users. The call-related autobiographical data was comparable between the two studies.

Another alternative method for fallback authentication that is being investigated is geographic authentication. Geographic authentications schemes are authentication schemes that utilize a map so that the user can set a location.

MacRae et al . [39] [48] investigated geographic authentication schemes as an alternative to traditional authentication. In MacRae et al. two geographic authentication schemes were designed and investigated for usability and security. The first geographic authentication scheme they designed was GeoPass. GeoPass is a simple geographic authentication scheme were the user is shown a map of the world, the user is asked to zoom to a predefined minimum level (zoom level 16) and is asked to set the location. The location set is then used to authenticate the user later on with a 21 by 21 pixel box error tolerance, therefore if the user makes a small input mistake that will not affect the usability since that answer would be accepted. The second geographic authentication scheme MacRae et al. designed was GeoPassNotes. In this geographic authentication scheme the user is not only asked to set a location but is also asked to associate a note with the location (an annotation to the location without actually labelling the location).

GeoPass and GeoPassNotes were tested using different participants (n=35 and n=30 respectively). The study was run across three sessions. In the first session the participants were asked to set and confirm their login credentials, and after a distraction questionnaire was administered (10 – 20 minutes) they were asked to login again. This session was in person. Session 2 was held online within 2 days. All participants in the GeoPassNotes study returned however only 33 returned for the GeoPass study. Both systems had a 100% memorability rate with very few failed login attempts (they were given up to five attempts).

Session 3 was held in the lab and was held 7 days after Session 2. All GeoPassNotes study participants returned, however only 30 GeoPass study participants returned. Again, nearly perfect memorability across both systems (97% for GeoPass, and 100% for GeoPassNotes) was observed with very few failed login attempts.

The security of GeoPass given the error margin was low when not paired with stringent requirements on the number of guesses therefore making GeoPass more suitable for online authentication with stringent restrictions. GeoPassNotes offered significantly greater resistance to online and offline attacks however, one concern was that the users were setting locations/annotations that were significant life events and thus could potentially be easily guessed by a known adversary.

Hang et al. [34] also developed a geographic authentication technique that was meant to be an alternative to security questions. In this geographical authentication technique users were guided by predefined questions (i.e., Whereto was your first travel by plane?), guided questions (i.e., Please define a location-based question that refers to a travel destination/vacation destination), or open ended questions (i.e., users input the question in a textbox) [34]. After that, the users were asked to select a location on the map. The proposed method of fallback authentication was tested for security as well as usability. The method of testing was through the use of a user study that spanned six months. The study included 32 participants, 28 of the 32 acted as both the close adversary and the main participant, two participants only acted as primary users, and lastly two participants only acted as close adversaries. The participants were asked to bring a pair that was considered close (e.g., partner or close friend). Interestingly, the pairs were asked to state the relationships between each other and in four cases the answers did not match. Thus, underlining the subjectivity of this method for defining the known adversary. The users were first asked to enroll the locations that they would like to use for fallback authentication (Session 1). They were asked to enroll the locations for 3 predefined questions, 3 guided questions, and 3 open ended questions that they set. They were asked to come back four weeks later to evaluate memorability. The results showed that 90% of users were able to successfully remember all the location set after four weeks (Session 2).

Furthermore, to get more thorough results the authors of the paper asked the participants in the study to return after a period of six months had elapsed (Session 3), the results were also promising, and 92% of the participants were able to remember the locations that they had previously enrolled with an error rate of thirty meters around the location that they initially enrolled six months earlier. Our study design to assess usability (specifically the memorability criteria within usability) was motivated by Hang et al's work [34].

Futhermore, the authors of the paper also did a security analysis based on known adversaries attempting to guess the locations that were enrolled by different users. This was done to test if users simply selected locations that were easily guessable by an attacker with knowledge about the user. The results were very promising, and adversaries had limited success in guessing the answers despite the fact that they could use the internet to search information about the victim. Close adversaries that were not allowed to perform an internet search guessed 7.5% (6/80) location questions correctly, close adversaries that were allowed to perform an internet search guessed 10% (8/80), adversaries that did not know study participants were only able to guess 3.75%(3/80) location questions correctly.

In addition to that, another graphical authentication system is PassPoints [56]. PassPoints is a graphical authentication system developed to replace classical text password based authentication schemes.In PassPoints users place points on a single image and utilize those points for authentication in the future. PassPoints was originally only evaluated for usability and it was found that users typically took longer to authenticate using PassPoints and typically had more failed attempts but users did successfully create valid graphical passwords within a short period of time. The study conducted by Wiedenbeck et al. [56] did not include an in depth security analysis and was more focused on the usability aspect.

Thorpe et al. [50] conducted a more in depth security analysis of PassPoints specifically or any graphical password interface that utilizes a single background image generally in order to evaluate their security. Thorpe et al. [50] utilized a human seeded attack which takes in click points from a subset of users and utilizes these clicks on a certain image to determine if hot spots within a background image exist. Hot spots within an image that is being used for graphical authentication are spots in that image that are highly likely to be selected by the user. Thorpe et al. [50] use human seeded attacks in order to prove that in an offline environment graphical passwords are just as vulnerable as text based passwords to targeted guessing attacks. Human seeded attacks yielded a 36% success rate within 2^{31} guesses. In addition to that Thorpe et al. [50] included a security analysis that utilizes purely automated attacks using image processing techniques in order to evaluate the resilience to offline guessing attacks of graphical passwords. The results of this purely automated attack were not as successful as the human seeded attack with 30% of graphical passwords guessed within 2^{35} guesses. This work suggests that graphical passwords are indeed as susceptible to classical text based passwords to offline guessing attacks.

Expanding on the work done on purely automated attacks on graphical passwords Thorpe et al. [20] [43] improve the efficiency and effectiveness of purely automated attacks on graphical passwords to a great extent. In their first subsequent work on purely automated attacks Thorpe et al. [43] utilize click order heuristics and a computational model of visual attention to improve the effectiveness of purely automated attacks on graphical passwords. For two representative images, their method of a purely automated attack on graphical passwords guessed 8-15% of passwords using dictionaries of less than $2^{24.6}$ entries for two representative images being utilized for a graphical authentication scheme (PassPoints style) [43]. Furthermore Thorpe et al. [20] conducted further work on improving automated attacks on click based graphical passwords this time utilizing a graph based algorithm that creates dictionaries based on certain heuristics like click order patterns in order to improve the effectiveness and efficiency of purely automated attacks on graphical click based passwords. As a result the improved purely automated attack on graphical click based passwords yielded 7-16% of passwords using dictionaries of up to 2^{26} in size [20].

Furthermore Al-Badawi et al. [47] conducted further research on graphical passwords and demonstrated how simple changes to a graphical password interface can change the distribution of user chosen passwords and by extension the implied security it provides. Al-Badawi et al. [47] simply changed the way a background image is represented to the user during the creation phase and observed a change in the user distribution of graphical click based passwords through this simple change in the way the background image is represented.

Thorpe et al. [49] also proposed Video Passwords. Video Passwords utilize some feature within video playback to authenticate. More specifically Thorpe et al. [49] proposed four Video Password schemes. The first video password scheme proposed utilizes time stamp information and is called Timeline Video Passwords the user presses a button at various times throughout the video in order to set the password and authentication later. The strength of Timeline Video Passwords is comparable to that of a four number/letter PIN. The second video password scheme proposed is Click Based Video Passwords which asks the user to click on the video much like a graphical password utilizing one image, with the added benefit of a video having many frames and adding the time line information. The third scheme proposed is Tagged Video Passwords. In this scheme the user not only selects a time within a video (similar to the first scheme proposed) but also adds a text tag to it. Both of those the time and the text is used for authentication in the future. Lastly, they proposed a fourth scheme titled Tagged Click Video Passwords, in this scheme the user also clicks on the video at a certain time (of his or her choosing) and then adds text as well, this is similar to the second proposed scheme with the added benefit and security of having a text based input as well. An error tolerance is allowed for each scheme, however, it was found that usability wise it is pertinent to have a slow paced video because that decreases the errors when authenticating.

Video Passwords have the potential to replace text based passwords not only on the usability and security fronts, but they also add a utility to the verifiers in that they offer a potential to include sponsored paid content as the video.

2.7 Improving the Memorability of Autobiographic, Geographic, and Classic Password-Based Authentication Schemes

Many attempts at improving the memorability of authentication schemes in general have been made. Longer passwords have been proven to be less memorable [46], which means users will typically create more memorable (shorter and more guessable) passwords in order to aid in usability (successful and quick logins). This creates a large divide between usability and security. Many attempts at system assigned passwords, which fit the criteria of being hard to guess even with advanced guessing algorithms that can perform millions of guesses a second, have failed the memorability test [46]. Most system assigned passwords often assigned by a password generator have many special symbols and tend to be longer (depending on the policy implemented). However, this means they only tend to be usable when used in tandem with a password manager due to the lack of memorability of such passwords and due to how tough it is to input them without any error. Many attempts were made to make system assigned passwords more memorable most notably by Shay et al [46].

Shay et al. attempt to generate more usable system assigned passwords by utilizing Passphrases that tend to be longer than the average password hence making them more immune to online and offline guessing attacks. Passphrases are whitespace delimited sequences of natural words. Shay et al. investigate the effect of utilizing Passphrases that are made of 3 or 4 word random combinations. The results were disappointing in terms of memorability, but Passphrases offered more security when compared to user assigned passwords. However, they did not perform well when compared to system assigned passwords with the same caliber of entropy in terms of memorability.

Participants in the study found them annoying and many participants wrote down the Passphrase. Therefore, the challenge remains; how do we encourage the creation of authentication credentials that are not easily guessable by way of offline or online attacks and are usable at the same time?

This is where the use of hints and cues to improve memorability is useful. In previous work, GeoPassNotes [39] [48] (see Section 2.4) offered great results in terms of memorability (100% recall rate after one week with very few failed login attempts). GeoPassNotes requires the user to set a location as well as an annotation and both have to be input for authentication.

Albayram and Khan [12] evaluated the use of hints in order to improve the memorability of autobiographical authentication. The authors conducted a study that investigated the use of hints with three types of autobiographical data (SMS logs, call logs, and location logs). The hints can be in multiple formats with the most common one used in the study is “you called this person before at (insert time and date)”. The study involved 24 participants (12 pairs) over a period of thirty days with strong and naïve adversaries to investigate the security implications of providing hints for autobiographical questions. The findings were notable; the hints improved the memorability of all three types of autobiographical data tested for legitimate users. The hints also had a negative impact on the strong adversaries (close adversaries such as a spouse or best friend) and had no impact on the ability of naïve adversaries in guessing the answers to the questions.

Lastly, Passhint is a graphical authentication system designed by Chowdhury et al. [23]. Passhint used a text-based hint in order to help a user remember a graphical password. In this authentication system a user had to remember which image was selected by picking 1 image out of 16 a series of times. The authors found that this increased memorability of graphical authentication systems. Our motivation for attempting to utilize a geographic hint in GeoPassHints comes from the positive results obtained all the work reviewed in this section.

2.8 Multiple Password Interference

Multiple password interference is an effect observed when an authentication system is utilized across multiple different devices/accounts. In real world implementations an authentication system that has been adopted is rarely utilized by only one verifier. Therefore, users tend to have multiple accounts that utilize the same authentication system, a great example of this is passwords; many verifiers utilize passwords for authentication.

Multiple password interference is the byproduct of having multiple verifiers utilize the same authentication system, where the users will mistakenly input the authentication credentials associated with a certain verifier into a different account/device for authentication (i.e., inputting your Facebook password while trying to login to your Twitter account). Multiple password interference does not only pertain to password like authentication schemes but can be observed in any authentication system.

Al-Ameen and Wright [11] performed two multiple password interference studies on GeoPass [39] [48], the first study using the original GeoPass found that users successfully remembered their locations set less in than 70% of all login attempts. This is an astounding finding because the original GeoPass study when not testing for multiple password interference obtained a memorability of 97%. Al-Ameen and Wright divided their user study in three sessions. During the first Session participants (n=18). would select locations for four different accounts. One week later the participants would be asked to come back and attempt to recall the locations they set for each account (referred to as*login* 2). After another week has passed the participants would be asked to come back again and attempt to login again (referred to as *login* 2).

In order to measure multiple password interference Al-Ameen and Wright measured the distance between the location set for a specific account and the location input. This distance was compared to the location's set for the three other accounts. If the location was closer to the location set for a different account, then that failed login attempt was considered a failure due to the effect of multiple passwords interference. As mentioned earlier around 70% of participants managed to login eventually (participants had five attempts per account). However out of 282 login attempts during *login* 1 44.8% of attempts failed due to interference effects. During *login* 2 out of 309 login attempts 38.2% failed due to interference effects [11]. Non-interference related failures were 40.4% and 46.3% for *login* 1 and *login* 2 respectively.

The second multiple password interference study conducted by Al-Ameen and Wright was with a modified GeoPass in an effort to mitigate the effect of multiple password interference. Participants in this second study were prompted to associate a mental story that would be utilized as a cue to remember the location [15, 26]. This study was designed the exact same way that the first study was designed except for the fact that participants (n=38) were not asked to come back to the lab for *login* 1 and *login* 2, instead it was done online. The methodology for counting a failure to login as an interference or simply a failure due to lack of memorability also did not change.

The results offered great improvement in mitigating the failures due to the multiple password interference effect. In *login* 1, out of 237 login attempts, only 4.2% failed due to multiple password interference. During *login* 2, out of 268 login attempts, only 2.6% failed due to multiple password interference. The login attempts required for a successful login were significantly reduced as well because the failures as a result of multiple password interference were significantly decreased. Non-interference related failures were 32.9% and 41.1 for *login* 1 and *login* 2 respectively.

The work done by Al-Ameen and Wright in passwords interference for GeoPass, alongside the work Chowdhury et al. [23] did on hints, motivated our study and system design. We also have four accounts that the participants in our study must set credentials for and attempt to remember. In Al-Ameen and Wright's modified GeoPass, the location set was the credential and the cue set the mental story related to that location. In GeoPassHints, the secret note set is the credential used for authentication which should be related to the location cue.

2.9 Comparative Framework for Evaluating Authentication Schemes

Bonneau et al. [18] proposed a widely adopted framework for comparatively evaluating authentication schemes. The framework is an effective tool for comparatively evaluating the usability, deployability, and security of any authentication scheme. It is important to note that the primary aim of this framework was to evaluate web-based authentication schemes. However, the categories can apply to authentication schemes in different environments.

The framework is divided into three broad categories: usability, deployability, and security. Each of these broad categories have sub categories with certain conditions that have to be met in order for an authentication system to be awarded under that specific subcategory. It is important to note that the Bonneau et al. framework uses web passwords as the baseline for comparison. The framework utilizes a black circle denoting the authentication system offering the benefit, a hollow circle indicating the authentication system almost offers the benefit, and no circle (just a blank space) indicating the authentication system does not offer the benefit. A shade indicates that this authentication system is better than passwords in a specific category, and a red shade indicates that this authentication system is worse than passwords in a specific subcategory.

The broad category Usability has 8 different subcategories that measure different usability features that should be common to all authentication systems. Important metrics to measure are the error rates, login time, and credential set time. A detailed overview and explanation of each subcategory is provided in Appendix F.

The broad category Deployability has 5 different subcategories that measure different aspects of an authentication system that would make it deployable on a larger scale. This includes issues such as browser compatibility and more importantly accessibility. A detailed overview of each subcategory within Deployability is provided in Appendix F.

The broad category Security has 11 different subcategories that measure an authentication system's security features. We added two more categories that we believed are pertinent to our systems. A detailed overview of each subcategory within Security is provided in Appendix F.

This framework for the comparative evaluation of authentication schemes will be utilized in order to evaluate GeoPassHints and GeoSQ.

2.10 Passwords

A large portion of this work compares fallback authentication scheme security to password security as a benchmark. This occurs as a result of passwords being the most commonly utilized means of primary authentication in addition to the Bonneau et al. framework for the comparative evaluation of web authentication schemes utilizing passwords as the primary benchmark and comparing all other authentication schemes to it.

While this may seem irrelevant because passwords are primary authentication while GeoPassHints and GeoSQ are targeted towards fallback authentication, it is still highly important to have a benchmark to measure against.

Bonneau and Preibusch [19] performed a complete analysis on password security in a web based environment and came up with a set of weaknesses and strengths outlined in the Bonneau et al. framework [18] for the comparative evaluation of web authentication schemes (see table 7.1).

According to their work, passwords show great promise in the usability and deployability categories. However, they seldom offer security benefits in terms of being resilient to throttled and unthrottled guessing, and they do not offer protection against internal observation.

In addition to that it was should that passwords are not resilient to physical observation and are highly susceptible to leaks by other verifiers.

Hanamsagar et al. [33] also explore passwords and found that users tend to reuse passwords to a high degree, making passwords highly susceptible to leaks by other verifiers and confirming the work completed by Bonneau and Preibusch [19].

In addition to that the Melicher et al. [41], conducted a study on the usability of and security of text based passwords on a mobile device environment. The findings support the deployment of GeoPassHints on a desktop/laptop environment because it was found that text based passwords are more prone to error on a mobile environment. This effect would be exacerbated if we were using GeoPassHints on a mobile platform due to the fact that Melicher et al. [41] found that creating text based passwords on a mobile environment is more frustrating and takes significantly longer. The secret note that is part of GeoPassHints is exactly the same as the input of a text based password.

Furthermore, Melicher et al. [41] found that password policies seemed to have a different effect on users attempting to set passwords in a mobile environment versus users attempting to set passwords on a desktop/laptop environment. Specifically, users attempted to make the password easier to enter on a mobile environment. As a result, the password policies yielded statistically weaker set passwords on a mobile environment. This further motivates utilizing a autobiographical authentication scheme such as GeoSQ on a mobile environment because it does not require a great deal of detailed input on a mobile device.

In addition to that, Ji et al. [36] noted that cultural considerations or differences in regions can be utilized in order to make training password crackers more efficient hence needing fewer guessing attempts. Ji et al. [36] also note that password leakages tend to make certain password more susceptible to various cracking methods and that password meters need to take that into consideration. Furthermore, building on the work of Castelluccia et al. [21], Ji et al. note that leveraging social media mining or personal information extracted from social media in general is highly effective in reducing the number of guesses that a password cracker requires to successfully crack a subset of passwords. According to Castellucia et al. [21] utilizing personal information that is publicly available and applying that personal information to leaked data sets speeds up password guessing by approximately 30%. Lastly, both Ji et al [36] and Castellucia et al. [21] note that password meters being utilized by verifiers need to utilize this information so that the they can better advise their users on password selection.

Chapter 3

Models for Quantifying the Known Adversary

3.1 Introduction

The known adversary is any potential attacker with elevated first-hand knowledge of a potential victim and/or elevated access to the potential victim's devices. The elevated knowledge or elevated access usually comes as a virtue of having some sort of relationship with the potential victim of an attack (i.e., spouse, co-worker, close friend). The known adversary was first identified by Muslukhov et al. [42] and identified as the insider threat. We coined the term known adversary because the risk of elevated access to a potential victim's devices or elevated knowledge of the potential victim does not only come from social insiders such as close friends or spouses. Therefore, we needed a broader term that encompasses individuals that we would not normally consider to be in our inner social circles (such as superiors at work, classmates, and co-workers).

Muslukhov et al. first identified the threat of the known adversary as an important and pertinent one to the security of our authentication systems through the use of an online survey conducted on Amazon's Mechanical Turk (n=724). 12% of survey participants indicated that they were aware of an instance where their devices/accounts have been breached by a known adversary. 9% of survey participants indicated that they have actively engaged in breaching the account/device of a known adversary.

Security research today accounts for the known adversary by recruiting participants to test authentication systems in pairs. They often label the pair as being a weak or strong pair (or strong and naive) [13, 34, 35, 39]. Security research often relies on the participants to self-declare their relationship and base their classification on the declaration of participating pairs [13,34,35,39]. We believe that this is a flawed model because close/strong adversary pairs as labelled by social parameters/norms are not always the only known adversaries to account for. In addition to that it leaves the security research vulnerable to criticism because in some cases the pairs do not label each other the same way (they see the relationship differently), that occurs when one participant labels the relationship as a close one and the participant's pair labels the relationship as not a very close one. This has occurred before in security research [34]. This is what motivated our quest to develop a model that formally quantifies the known adversary in an empirical manner to be utilized within security research.

This model is highly important to all security research and specifically when geared towards performing a security analysis on authentication systems that rely on autobiographical data, and geographic authentication schemes. Autobiographical data specifically autobiographical location data is naturally more susceptible to be common knowledge to known adversaries who are often in close proximity to the potential victim. Geographical authentication schemes often rely on giving hints or location cues that could make the credentials obvious to a known adversary [34].

The three most common fallback authentication mechanisms are security questions, email resets, and SMS resets. Security questions are often low entropy and easy to guess [17, 31]. This makes security questions highly vulnerable to an attack by a known adversary who has elevated knowledge of a potential victim.

Email resets rely on the on the email account not being compromised and also rely on the device on which the email is setup to be secure. With many email reset implementations, the reset code or reset link can be viewed from the lock screen (depending on the potential victim's settings). This also makes email resets highly vulnerable to a known adversary with elevated access to a potential victim's devices/accounts (e.g., a known adversary who is a co-worker spots a potential victim's phone lying on a desk while the potential victim is getting lunch). As for SMS resets they also have the same flaw as email resets which make them highly susceptible to being compromised by the known adversary.

This motivated us to create models for quantifying the known adversary and providing empirical results as to whether or not this known adversary is indeed a threat. In this work we created three different models to quantify the known adversary, based on relationship closeness metrics from the field of Social Psychology [14, 16, 24].

In order to test these models, we performed a linear regression analysis corelating the scores obtained by our models and the results that the participants obtained during the GeoSQ pair guessing portion of our user study. We compared our models to the self-reported relationship characterizations which we asked the participants in our user study to provide at the beginning (during Session 1). In our study we attempt to determine whether or not our models to measure the known adversary are better suited for quantifying the known adversary than simple self-declaration by the pairs.

3.2 Designing Different Models

Our first model was based on the RCI [16]. We adapted the RCI to suit our needs because it was developed in 1989 and it was not geared towards measuring the known adversary, it was geared to measure relationship closeness broadly. In addition to that our way of life has changed since 1989 because of our increased use of technology. As a result, we added some more questions to reflect our current way of life to create the adapted RCI model for measuring the known adversary.

Refer to Appendix A for the original RCI, and refer to Appendix C for the adapted RCI that we utilized in our study, the appendices also contain all the information regarding the scoring of all the questions.

Our second model was based on the IoS Scale (Inclusion of the Other in the Self Scale) [14]. The IoS Scale is a simple pictorial tool that allows participants to measure their relationship closeness by selecting the image they believe best describes their relationship (see Appendix B).

We combined the IoS Scale with the We Scale [24]. The We Scale is a simple Likert Scale question with seven levels that asks "Please, select the appropriate number below to indicate to what extent you would use the term "WE" to characterize you and this individual" [24]. The average of the We Scale and the IoS Scale provides us with what is known as the Oneness score [24]. The average of the We Scale and the IoS Scale (Oneness) is what we utilize for our second model for a model to measure the known adversary.

The IoS Scale and the We Scale have both been positively corelated in providing accurate relationship closeness metrics to the RCI by Gachter et al. [28]. The IoS Scale and the We Scale provide the advantage of shortening the time it takes for researchers to conduct a survey since the IoS Scale is a simple pictorial tool and the We Scale is a simple Likert Scale question.

Our third model to measure and quantify the known adversary is what we call the Known Adversary Inventory (KAI). The KAI is composed of all the questions that were added to the RCI to create the adapted RCI model (see Appendix C). This model was created for the purpose of evaluating whether or not relationship closeness scales/inventories from the field of Social Psychology are needed at all, because what we are trying to measure is not only relationship closeness but proximity and threat posed to an authentication system in general, which can come from a variety of sources (e.g., acquaintances, co-workers, spouses)

We compare the models created to measure and quantify the known adversary against the self-declared relationship characterization provided by our participants. We perform this analysis in an effort to compare the benefits of self-declaration of relationship characterizations versus having a formal model that quantifies the known adversary.

Our approach for the statistical analysis is to perform a linear regression analysis to discern correlations between the models for measuring the known adversary and GeoSQ pair guessing scores. We cannot perform the same type of linear regression analysis on GeoPassHints pair guessing scores because, so few were successful therefore no positive correlation is possible. Our analysis begins by calculating the correlation coefficient between the model and the GeoSQ pair guessing scores. If the correlation coefficient was above 0.2 (the threshold for significance), we continued the linear regression analysis to find key statistical values that prove a significant relationship exists between the model and the GeoSQ pair guessing scores.

3.3 Results

In this section we will present the results of the linear regression analysis on each of our models, in addition to the self-reported relationship characterization. Followed by that we will compare the results of all the different models.

3.3.1 Adapted RCI (Model 1)

After performing a linear regression analysis on the guessing score of each participant and the corresponding adapted RCI score, where the adapted RCI score was the predictor (x) and the guessing score was what we were trying to predict (y); we attained a correlation of r=0.772 which is well beyond our threshold for significance (set at 0.2)

Figure 3.1 shows a scatter plot where there appears to be a linear relationship between the guessing score and the Adapted RCI score hence the positive correlation of r=0.772 attained.

Figure 3.2 features two boxplots that display the quartiles and the mean of both the GeoSQ guessing score, and the adapted RCI score.

Figure 3.3 showcases a density plot for the adapted RCI scores and the GeoSQ guessing scores

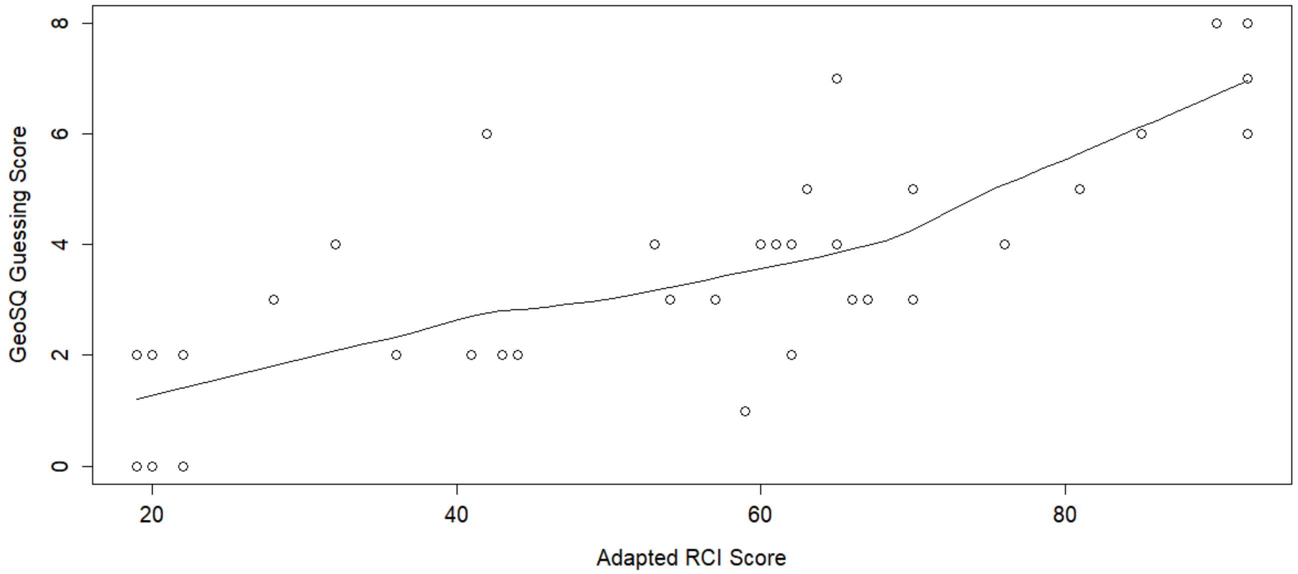


Figure 3.1: Scatter plot of guessing score/adapted RCI score that shows a linear relationship.

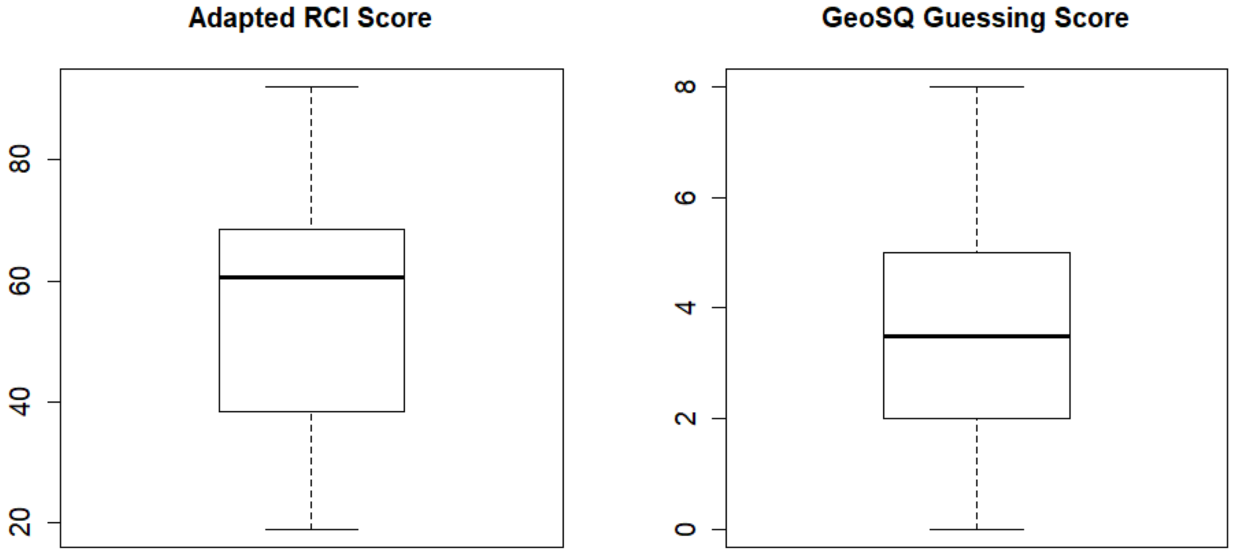


Figure 3.2: Boxplot for both the adapted RCI score and the GeoSQ guessing scores.

with n =36 instead of n=34 (which is the total number of participants for Session 2) because this graph includes one sided data were the participant's pair did not have enough location data to proceed with GeoSQ but was present for the guessing portion of the study, these two participants with not enough location data are only included and mentioned in the portions of the study that they participated in. We utilize the density plots in order to test for normality and the validity of the analysis below.

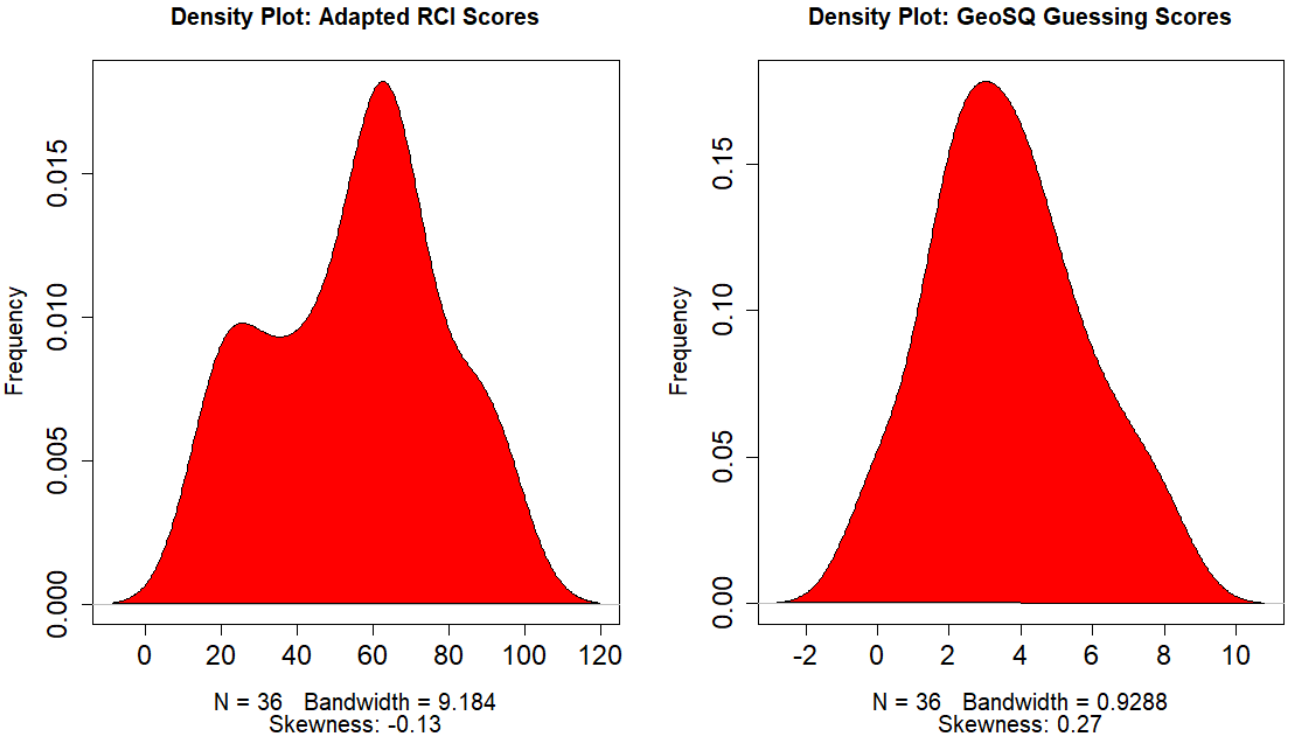


Figure 3.3: Density plot for the adapted RCI score and the GeoSQ guessing results.

Table 3.1 showcases all the relevant statistical values. The null hypothesis is that the p value will be zero which indicates no relationship between the two variables. Our alternative hypothesis is that there exists a relationship which ideally means our p value will be lower than our predetermined level of significant which is $p < 0.05$. Our attained p value using an independent t-test is $3.472e^{-8}$ which is significantly lower than 0.05 and is higher than zero therefore we can confidently reject the null hypothesis.

The t value is another important statistical figure because the higher the t value the less likely it is that the reported results are not zero by chance. Our reported t value as shown in Table 3.1 is 7.087 which indicates that it is not likely the p value and correlation coefficient is not zero by chance.

Our residual standard error for this analysis was 1.362 on 34 degrees of freedom. Our calculated multiple r^2 was 0.596 and our adjusted r^2 was 0.584. The r^2 value gives us the proportion of variation in the GeoSQ guessing result which in this case is y or dependent variable that we attempt to correlate with the adapted RCI result. The adjusted r^2 will penalize the analysis based on the total number of adapted RCI variables (independent variables). This is standard practice because larger sets can be set up to have large r^2 values, therefore we rely on the adjusted r^2.

Furthermore, we utilize two more measures of goodness fit in order to compare different models, because it is considered a best practice to not completely ignore models based on the r^2 and the adjusted r^2. The first is Akaike's information criterion (AIC) [10] and the second is Bayesian information criterion (BIC) [45]. The lower the AIC and BIC are the better, and both measures of good fit rely on a maximized value of likelihood function as shown in the formulas below. Our reported AIC and BIC are 128.368 and 133.119

AIC=$(-2) \times \ln(L) + (2 \times k)$, L is the likelihood function, and k is the number of parameters.

BIC= $(-2) \times \ln(L) + k \times \ln(n)$, n is the sample size.

Correlation Coefficient r	p Value	t Value	Standard Error	Degrees of Freedom	r^2	Adjusted r^2	AIC	BIC
0.772	$3.472e^{-8}$	7.087	1.362	34	0.596	0.584	128.368	133.119

Table 3.1: Relevant statistical figures for adapted RCI model.

3.3.2 Oneness Score (Model 2)

We performed the same analysis for the second model which is utilizing Oneness scores instead of the adapted RCI score. The correlation coefficient was weak at r=0.0966. This value does not merit further statistical analysis with the Oneness score being the independent (predictor) variable. Figure 3.4 shows the scatter plot of the Oneness scores and the GeoSQ guessing scores.

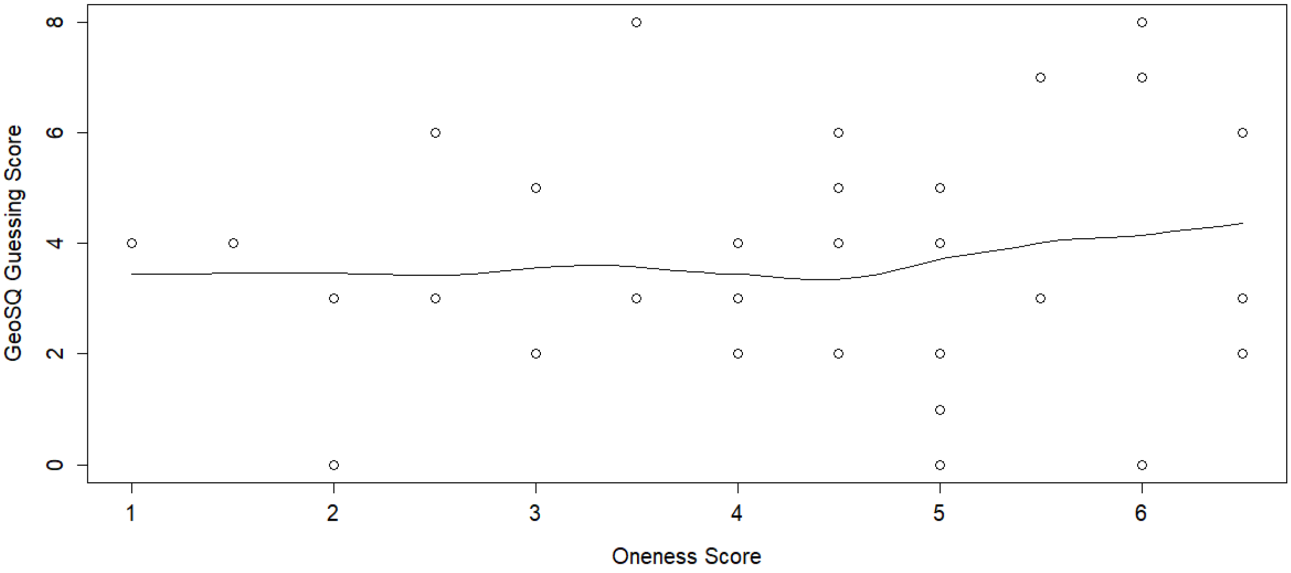


Figure 3.4: Scatter plot of GeoSQ guessing scores and Oneness Scores obtained by the participants.

3.3.3 KAI (Model 3)

After performing an initial analysis on the KAI we obtained a high correlation value of r= 0.729 which is statistically significant and warrants a full linear regression analysis. Figure 3.5 showcases the scatterplot of the guessing scores against the results of the KAI.

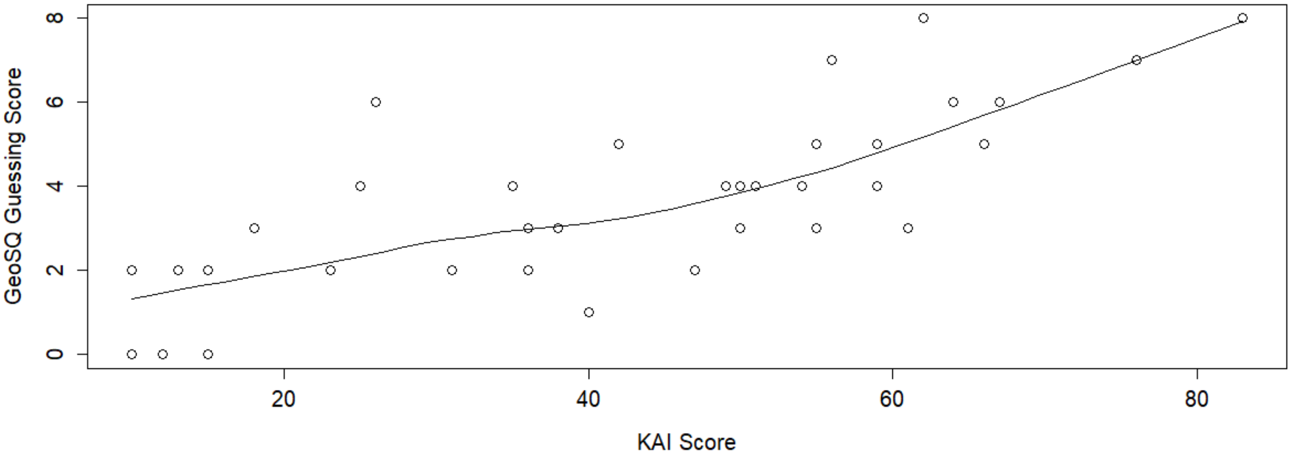


Figure 3.5: Scatterplot of the KAI model and GeoSQ guessing score that shows a linear relationship.

Figure 3.6 features two boxplots that displays the quartiles and the mean of both the guessing score and the KAI score.

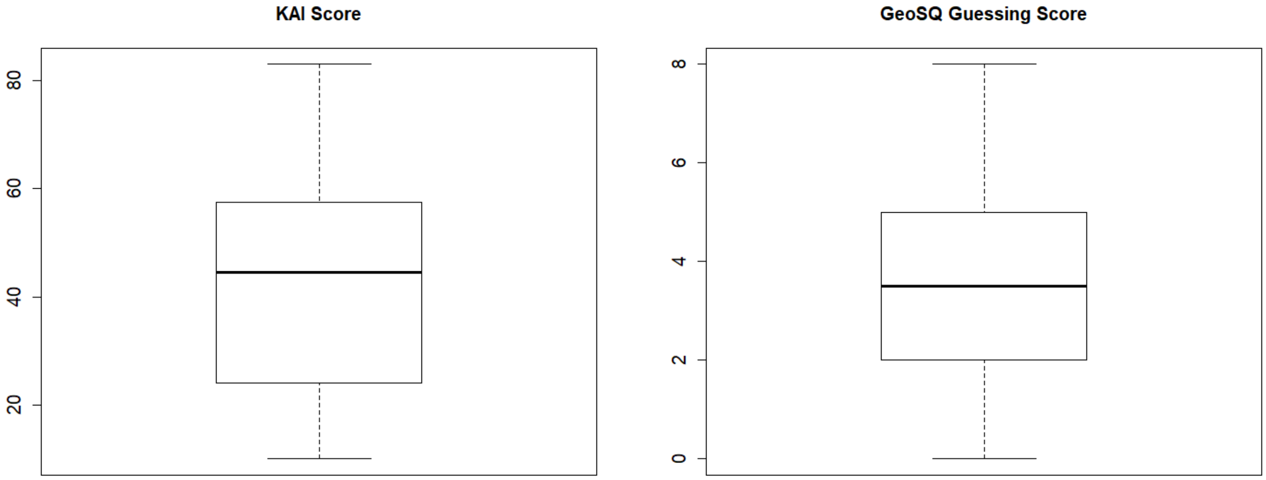


Figure 3.6: Boxplot for both the KAI score, and the GeoSQ guessing score.

Figure 3.7 showcases the density plot, as mentioned earlier the density plot is n=36 instead of n=34 because of two participants who did not have enough location data to continue with the

GeoSQ portion of the study.

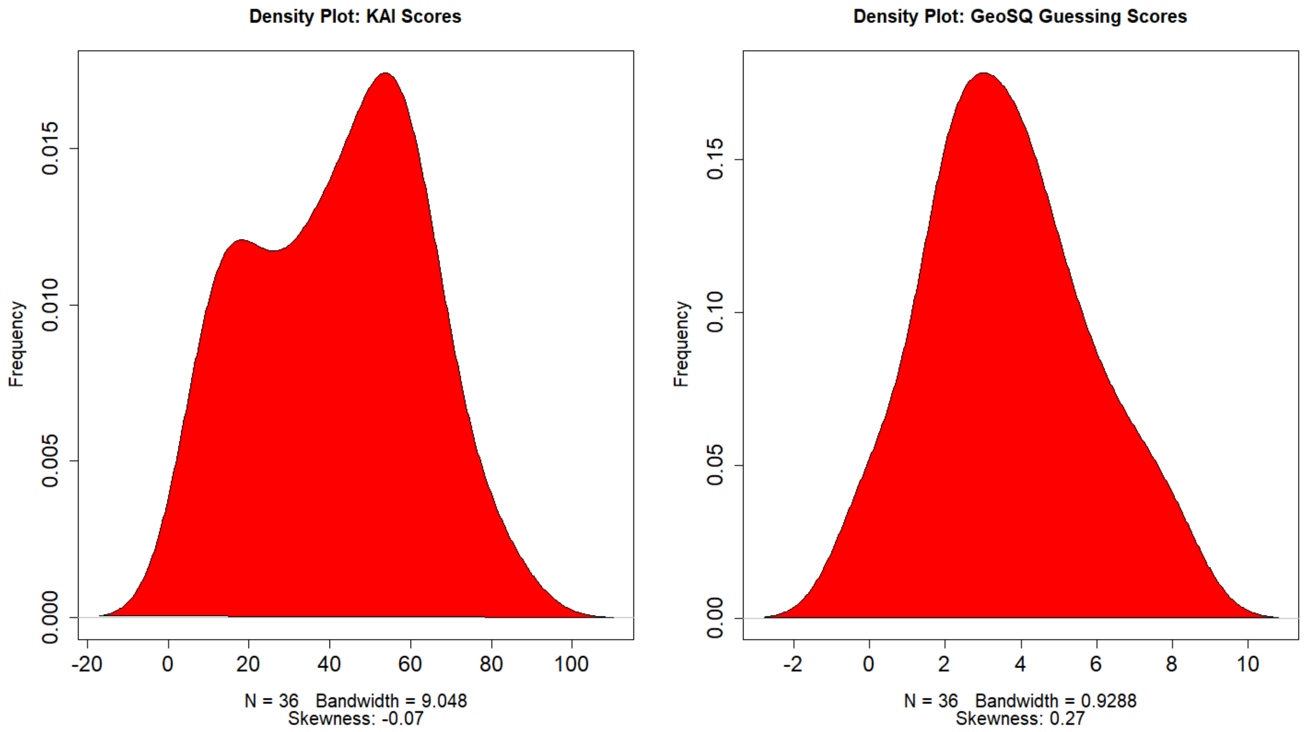


Figure 3.7: Density plot for KAI scores and GeoSQ guessing scores.

Correlation Coefficient r	p Value	t Value	Standard Error	Degrees of Freedom	r^2	Adjusted r^2	AIC	BIC
0.728	$4.703e^{-8}$	6.203	1.469	34	0.530	0.517	133.774	138.525

Table 3.2: Relevant statistical figures for KAI model.

The statistical figures in Table 3.2 show us that even without the RCI component, the questions we added to the adapted RCI are also good predictors for a participant's guessing ability.

3.3.4 Self-Reported Relationship Characterization

In order to prove that utilizing a model for measuring the known adversary is indeed more useful we also perform the same analysis on the self-reported relationship characterizations and

GeoSQ guessing scores. We obtained a negative correlation coefficient of -0.068, which is well below the threshold of significance that was set at 0.2.

The self-reported relationship characterization was split across four levels that participants could choose from. The first level was no relationship, the second level was not very close, the third level was close, the fourth level was very close. Of the 17 pairs that successfully completed all part of Session 1 and Session 2, there were seven instances in which the participant pairs self-reported their relationship characterization differently. All the differences were one level apart, meaning that it was typically a participant who picked a similar but different value from their pair.

As a result of the low correlation coefficient, further statistical analysis is not merited in this case, Figure 3.8 shows the relationship of the self-reported relationship with the GeoSQ scores.

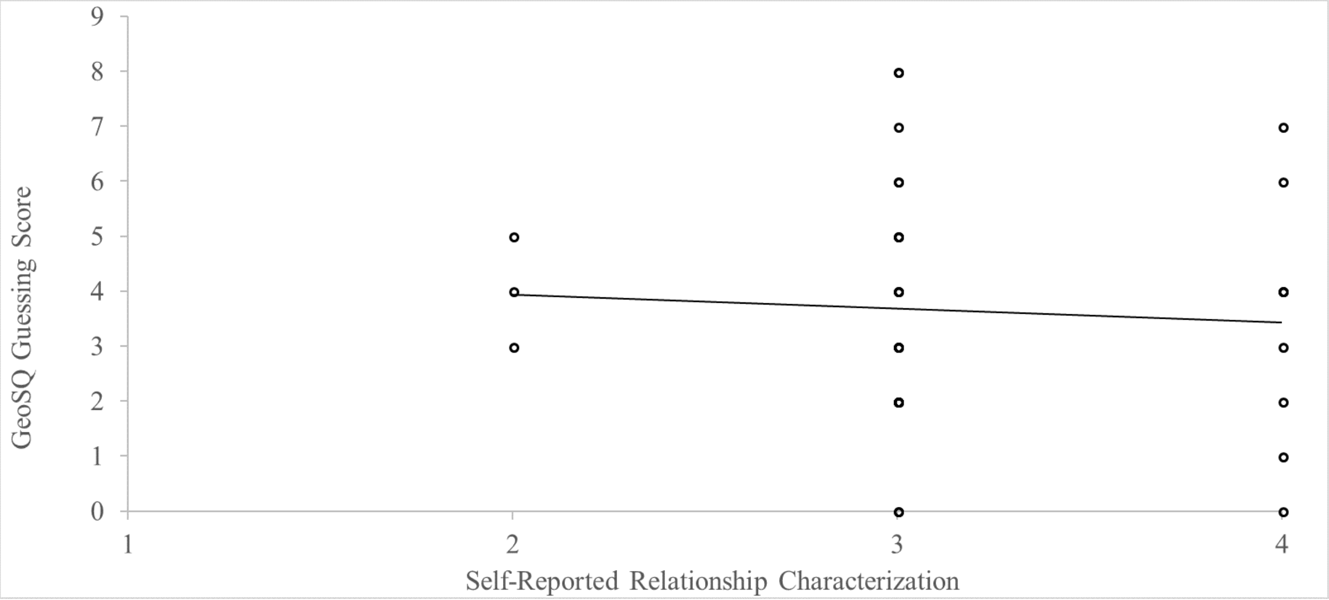


Figure 3.8: Scatterplot of the self-reported relationship characterization and the GeoSQ guessing scores that shows no clear relationship.

3.4 Comparison

Two models that quantify the known adversary stood out, the KAI and the adapted RCI. The self-reported relationship characterization was shown to have a negative correlation with the GeoSQ guessing scores, and the Oneness score has a low correlation with the GeoSQ guessing scores which makes both the Oneness score and the self-reported relationship characterization ineffective in quantifying the known adversary.

Out of the two successful models that have been proven to be adequate predictors of the ability of a pair to obtain high results on the GeoSQ guessing score, the adapted RCI seemed to be the best predictor. However, the margin between the adapted RCI and the KAI was very small. We attribute this closeness in correlation between those two models and the GeoSQ guessing scores to the fact that the original RCI portion of the adapted RCI model was geared towards relationship influence (see Appendix C), which is strictly a relationship closeness metric and not necessarily pertinent to measuring and quantifying the known adversary as proven by the analysis. That is why we did not discuss the analysis of the RCI portion of the adapted RCI like we did with the KAI. While the RCI portion of the adapted RCI provided a slight improvement on the original portion of the adapted RCI on its lonesome did not have a significant correlation.

When comparing the statistical figures obtained by the adapted RCI model and the KAI model, the difference is very small with the adapted RCI showing a slight advantage in terms of goodness of fit when we performed the full linear regression analysis. Table 3.3 compares the statistical figures obtained by the adapted RCI and the KAI. The AIC and BIC of the adapted RCI are only slightly lower (meaning they are a better fit) than the KAI.

This has ramifications because one of the issues with security research is the time burden placed on the participants by extremely long questionnaires therefore shortening them is always welcome. More importantly, this leads to the conclusion that perhaps social closeness metrics might not be pertinent when quantifying the known adversary, because the known adversary is not always an individual we consider as close to us from a social perspective.

Model	Adapted RCI	KAI
Correlation Coefficient	0.772	0.729
p Value	$3.472e^{-8}$	$4.703e^{-7}$
t Value	7.087	6.203
Standard Error	1.362	1.469
Degrees of Freedom	34	34
r^2	0.596	0.530
Adjusted r^2	0.584	0.517
AIC	128.368	133.774
BIC	133.118	138.525

Table 3.3: Side by side comparison of statistical figures obtained from the linear regression analysis of the adapted RCI model and the KAI model.

To account for the fact that we ran two tests on the same GeoSQ guessing scores data set which increases the chance that we obtained these results by coincidence; we perform a Bonferroni correction [9] which adjusts the level of significance from 0.05 to 0.025. Since the p values are considerably below that threshold, this does not affect the validity of our result.

3.5 Limitations and Future Work

The successful models for quantifying the known adversary are the KAI and the adapted-RCI. However, since they have only been tested using GeoPassHints and GeoSQ there is a question regarding their adaptability and the ability for us to claim that they can be utilized on any fallback authentication or primary authentication method.

The different models were designed in a way that should be more successful in measuring the ability of a known adversary to guess autobiographical authentication questions and geographical authentication questions. This is true because we are often measuring physical proximity by asking the questions in the model. Therefore, we cannot say that the models can be generalized for utilization in all research with a known adversary or insider threat aspect.

We stress that we are not only measuring relationship closeness, we are only trying to measure whether or not a pair in a research study is a truly strong known adversary to be able to evaluate our authentication systems to a more valid extent.

Future work should discuss the applicability of the two successful models to other methods of primary and fallback authentication. In addition to that another valid extension of this work would be altering the actual weights that have been utilized within the models. For example, we have a question on fettered and unfettered access to a pair's device. These questions are weighted more than other questions regarding proximity due to the fact that a simple answer to these questions is adequate in determining whether a pair is a true and strong known adversary that is able to guess and answer many verification questions successfully. Adjusting the weights and attempting to yield different results could alter the effectiveness of the models to either be even more successful in predicting whether a certain pair is a strong or weak known adversary and it could also alter the effectiveness of these models negatively.

Another extension of this work would be to break down the models into individual questions and take a closer look into the questions that are yielding the most accurate prediction results.

Chapter 4

GeoSQ

4.1 Introduction

In spite of their tremendous security flaws the most common fallback authentication mechanisms (security questions, Email resets, and SMS resets) continue to be used by many verifiers today. The three most common fallback authentication mechanisms are in use due to their perceived usability and less so for their proven security [18,31,34]. The more usable an authentication mechanism is, the more it tends to lack security. Highly secure authentication mechanisms on the other hand tend to be less usable. The perfect authentication mechanism is one that balances usability and security. Ideally, an authentication mechanism will bridge the chasm between usability and security.

Secure policies that make authentication mechanisms more secure are often not adopted to allow for increased usability. For example, we could set password policies that mandate a minimum length requirement of 16 characters. However, the longer the password, the less memorable it tends to be [46]. This is just an example of how usability and the relative security an authentication system provides, will ultimately decide whether or not an authentication mechanism is widely adopted.

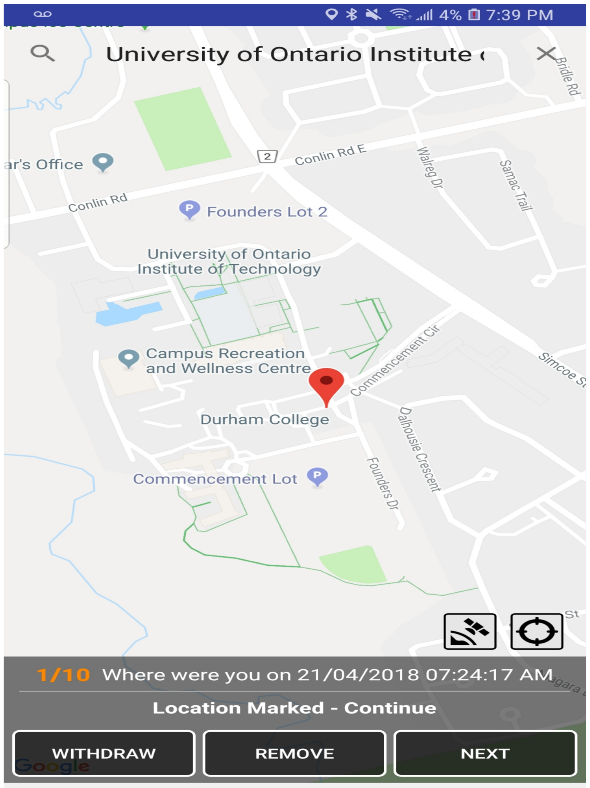
In recent years, we have seen a push towards creating alternative fallback authentication mechanisms that replace traditional commonly utilized fallback authentication mechanisms due to the security flaws present within security questions, email resets, and SMS resets. One such alternative fallback mechanism relies on autobiographical data in order to generate questions that the legitimate user can answer in order to gain access to an account/device [13,34]. Authentication systems utilizing autobiographical data tend to take longer to authenticate because more than one question is typically asked of the legitimate user. This makes systems utilizing autobiographical data unusable for primary authentication (i.e., entering a password) because primary authentication occurs far more often than fallback authentication [18]. As a result, autobiographical authentication systems are typically only thought of as replacements to current fallback authentication mechanisms.

In this work, we design and implement from scratch, an autobiographical authentication system called GeoSQ (geographical security questions), a variant of Albayram et al's and Hang et al's previously proposed autobiographical authentication systems [13, 34], and we test our system for security, usability, and deployability.

This chapter is organized as follows; we begin by introducing the system design. Followed by that we complete our security, usability, and deployability analysis utilizing data from our user study, and using the framework developed by Bonneau at al. [18] as a guide to our analysis. Lastly, we present our resulting system and policy recommendations, followed by a discussion and future work section.

4.2 GeoSQ System Design

GeoSQ is an Android application developed with the help of the Google Maps API [4] that utilizes autobiographical location data logged over 7 to 11 days. The user presses on the screen for 1.5 seconds in order to set a point when prompted with the question. Users of GeoSQ have the option of switching between satellite and terrain mode and have the ability to search the Google map in order to aid them in finding the location where they were present when prompted with the question. Users of GeoSQ also have the option of clicking on the current location button and are then zoomed in to their current location (requires location services to be activated). Lastly, the users can zoom in and zoom out as they please. Figure 4.1 features screenshots of the GeoSQ interface.


University of Ontario Institute o

1/10 Where were you on 21/04/2018 07:24:17 AM

Location Marked - Continue

Figure 4.1: GeoSQ Interface

4.2.1 GeoSQ System Design and Development Details

GeoSQ was designed in an agile development environment where development was simultaneously conducted with pilot testing, followed by more editing and development to come up with the final user interface and system settings. In the development of GeoSQ many usability/security tradeoffs had to be made, and many changes had to be made to account for the resource restricted environment an application like GeoSQ would run on (a smartphone with limited battery).

One of the biggest concerns faced when developing Android phone applications is ensuring that the application does not consume a great amount of battery. That is one of the first system design tradeoff between security and usability that we encountered. If we set the location accuracy to be too high and if we set it to refresh the location too often it would waste too much battery and hence make it less usable. On the other hand, it would make the location more accurate and hence the error margin would be less, and therefore it would be more secure because the sample space would be larger.

The accuracy of the Google Maps API does depend on how modern the phone is and what software is being run, but in the pilot testing with the location accuracy set to balanced power accuracy [4] we observed an accuracy similar to the documentation of about 2 blocks or roughly 200-400 meters. Balanced power accuracy is the middle ground in terms of accuracy; this setting does not completely favor saving battery power, but it balances saving battery power with obtaining the most accurate location possible. When testing on older smartphones running older versions of Android the same location accuracy setting gave us an accuracy of roughly 400 meters depending on whether we were connected to Wi-Fi or not. We decided to maintain the settings on balanced power accuracy and set the error margin to 400 meters which decreases the sample space, but saves power and makes the application more usable.

As mentioned earlier, the frequency of the refresh rate will also have an effect on the battery life. We set the refresh to be every one minute and we were able to make the battery life usage very efficient. Other system settings for GeoSQ include a minimum of a five-minute stay at any location in order to ensure that it would be memorable (and to ensure it was not a stop in traffic for example). In addition to that, we implemented a five-hour maximum stay at any one place to avoid obvious guesses such as work or home; this is another usability security trade off however, in this case we elected to have higher security and less usability.

We had several security, usability, and deployability goals in mind when creating GeoSQ. Our primary goals were enshrined in ensuring an authentication system that is more secure than current fallback authentication systems in use, in addition to ensuring that the system is highly memorable and usable. After the usability analysis was conducted it became obvious that the memorability of the system is a major issue as well as the time required to login. Both of these usability issues were not observed during pilot testing to the extent that they were in the actual user study. As for the deployability of the system, we were not as concerned about that because the deployability of an authentication system often comes after its adoption and maturity.

The greatest usability/security tradeoff came from our decision to only record unique locations. This greatly improves the security of the system because without this feature, adversaries can simply guess home or work depending on the timing and in less busy weeks, and especially in a university environment it would be an extremely successful guessing strategy. The usability tradeoff is that some participants did not have ten unique locations at the end of a week, so we extended the allowed timespan from anywhere between 7-11 days. Six individuals (2 of them are the researchers and four family members, and friends) pilot tested the application with and without the unique location requirement and we found that it made it more memorable to have the unique location requirement in 4/6 pilot testers. It had no negative impact on memorability for the other two pilot testers.

The default zoom level was zoom level 16 and it was concentrated over the current location; if the participant/user clicked on the current location button the map view would zoom in to level 14 and put the focus on the current location. The error margin was irrespective of the zoom level since we were using spherical distance to computer the distance between two coordinates that will mean the more zoomed in the more accurate the market placement. Participants were encouraged to zoom in and set the marker as accurately as they can remember.

Other settings and system design features that are worthy of mentioning are the minimum displacement which we set to be 200 meters before it is considered a unique location or before a location was recorded. This means that, for example, a movement from one class to another in the same building would not be recorded as a new unique location (unless it was a really large building). In order to calculate the distance from one pair of coordinates to another, we utilized spherical distance, which measures the most direct line irrespective of buildings and paths to actually get to a location.

The last system design specification that is worthy of mentioning is the order in which questions are presented. We utilized a last in first out algorithm to extract the questions from the database. This was done to measure the effect of time on memorability on location autobiographical data. We had no access to the actual locations just whether or not the answer was correct and the time it took to answer the question.

4.3 Results

In this section, we will present our analysis of our autobiographical authentication system GeoSQ. The analysis will be split into three different broad categories: 1- security, 2-usability, 3-deployability. Our analysis will be based on the criteria laid out by the Bonneau et al. framework for the comparative evaluation of authentication systems [18]. See Section 2.8 for a detailed description of each subcategory.

We added two more subcategories to the framework originally developed by Bonneau et al. [18]. The two categories that we added are *Resilient-to-Social-Media-Mining and Resilient-to-the-Known-Adversary.*

A system is awarded being *Resilient-to-Social-Media-Mining* if social media mining cannot provide an attacker with a leg up in guessing the user's authentication information. We award an authentication system this feature if social media mining does not improve the attacker's ability to guess by more than 1%. An authentication system is awarded the accolade of being Resilient-to-the-Known-Adversary if any attacker with elevated access to a user's devices and elevated knowledge of the user cannot compromise the system. To be more specific, a throttled known adversary should not be able to compromise more than 1% of accounts protected by the authentication system given 10 guesses a day for 365 days.

4.3.1 Security

Resilient-to-Physical-Observation

GeoSQ is considered to be Resilient-to-Physical-Observation since an attacker cannot reasonably successfully impersonate a user after observing them authenticate one or more times. The questions asked by GeoSQ are autobiographic which means that they will change over time. Also, once a question is asked, our system design marks the question as used and it cannot be asked again. In this case, physical observation of the user by the attacker would have to be constant over the past 7-11 days (the timeframe required to gather 10 unique locations).

Resilient-to-Targeted-Impersonation

Resilience to targeted impersonation involves being resilient to attacks that exploit personal details [18]. In the case of GeoSQ, which relies on location autobiographical information paired with our parameters put into place (the unique location requirement); an attacker with knowledge of personal details such is the place of work, and the home of the victim would only be able to score 2/10 within our system. However, more advanced targeted impersonation could reasonably attain high scores within GeoSQ therefore we award GeoSQ the perk of almost having the benefit of being *Resilient-to-Targeted-Impersonation.* We added another column to account for the known adversary specifically because targeted impersonation can be conducted by a stranger, while a known adversary is any individual who knows the potential victim on some level and has physical access to their devices and possibly accounts.

Resilient-to-Throttled-Guessing

According to Bonneau et al. [18], a system is awarded being Resilient-to-Throttled-Guessing if an attacker, while being constrained by a verifier, cannot compromise more than 1% of accounts a year given ten guesses a day. In our analysis, 34% of the location questions were answered correctly by the pairs when attempting to guess each other's locations. With our threshold of 7 correct answers out of 10, 11% of participants were able to successfully guess their pair's locations. Since the participant pairs knew each other it is expected that they would be able to guess more successfully than a typical throttled guessing attack that does not involve an individual that knows the potential victim. That is why this will be analyzed in the *Resilient-to-the-Known-Adversary* category that we added.

One feature that can assist against throttled guessing attacks by an unknown adversary on GeoSQ is the changing target, since autobiographical location data is dynamic, it would be difficult given a robust system design and a high enough threshold for a throttled attacker to be able to compromise an account, the attacker only gets one chance at answering questions correctly. Our GeoSQ system design does not allow for multiple guesses for each location question. Instead, the ten questions are asked, then each of these questions will never be reused again. In order for the attacker to try again, he/she would have to wait for the participant to generate new location data. This is a usability burden to the legitimate user who does not answer the questions correctly, therefore the legitimate user might need an alternative method of fallback authentication.

Therefore, since the criteria of being *Resilient-to-Throttled-Guessing* does not directly apply to GeoSQ (due to the fact that the location questions will change, and it is not possible to conduct that kind of guessing attack) we grant GeoSQ the accolade of being *Resilient-to-Throttled-Guessing.*

Resilient-to-Unthrottled-Guessing

According to Bonneau et al. [18] a system is *Resilient-to-Unthrottled-Guessing* if an attacker with the ability to guess 2^{40} per account is only able to compromise 1% or less of the accounts.

Based on that metric, GeoSQ would be *Resilient-to-Unthrottled-Guessing.*

The sample space for autobiographical location data paired with an error threshold of a 400-meter radius is very low (140,625). However, it is high when we factor in the threshold of 7 correct location answers. While calculating the sample space, we made the assumption that most participants will be within an hour and a half of a central home location at any time within the week. Therefore, we can assume a radius of 150 kilometers which is around 70,685 square kilometers. Our error radius is 400 meters which is 0.50265 km^2. Dividing 70,685 km^2 by 0.50265 km^2 yields us 140,625 unique locations.

Lastly, we assume that the threshold is 7 locations answered correctly out of 10. Therefore, our sample space is 140,625 choose 7 ($2^{107.411}$) which would not easily be guessable by an unthrottled attacker. We do not have access to the actual location therefore we cannot run an empirical test that guesses the locations logged on our participants' smartphones.

Resilient-to-Internal-Observation

According to Bonneau et al. [18], an authentication mechanism is Resilient-to-Internal-Observation if the attacker cannot impersonate the legitimate user by intercepting the legitimate user's input. An example of internal observation would be key logging or recording the screen of the legitimate user and relaying to an attacker's computer. Based on those parameters, GeoSQ is indeed Resilient-to-Internal-Observation to a high extent. GeoSQ relies on autobiographical location data, hence even if there was the threat of an internal observer, the questions would change over time and information gathered from one observation couldn't be reused in an attack.

Resilient-to-Leaks-from-other-Verifiers

A system is classified as *Resilient-to-Leaks-from-other-Verifiers*; if anything, another verifier leak could potentially help an attacker impersonate a legitimate user [18]. In the case of GeoSQ, even if there was a leak of a popular verifier such as Google Maps for example, assuming the leak is not ongoing; we would classify GeoSQ as being resilient to leaks from other verifiers.

However, if the leak is ongoing, GeoSQ would not be *Resilient-to-Leaks-from-other-Verifiers*. It is important to note that even with an ongoing leak it would be difficult to match the metrics since it is possible that not all the locations in your Google Timeline will be recorded by a system utilizing GeoSQ.

Resilient-to-Phishing

GeoSQ can be classified as *Resilient-to-Phishing* because it utilizes autobiographical location data which changes over time. In order for a classical phishing attack to be successful on the GeoSQ system, it would have to assume that a legitimate verifier utilizing GeoSQ as their authentication scheme will have the exact same questions as the phishing attack. The phishing attack would also have to happen right away and would have to have the exact timestamps utilized by the legitimate verifier. Since that is very unlikely, we assert that GeoSQ is resilient to phishing attacks. Please note that *Resilient-to-Phishing* does not consider relay style man in the middle attacks, which GeoSQ would be vulnerable to.

Resilient-to-Theft

This category is pertinent for authentication mechanisms that rely on a physical object for authentication, such as a magnetized card or a hardware token. In the case of GeoSQ, a physical object is required for authentication (a smartphone). Without the assumption that a smartphone is on the user's person at all times, it is infeasible to collect autobiographical location data. However, even if the smartphone was stolen, it would have to remain on for a great amount of time to collect location data. During this time a smartphone is usually traceable. Therefore, in the case of an untraceable smartphone, GeoSQ is not Resilient-to-Theft but is *Resilient-to-Theft* in the case of a traceable smartphone.

No-Trusted-Third-Party

This feature is awarded to systems that only rely on the user and verifier [18]. In the case of GeoSQ, it is unreasonable to assume that every verifier will have their own location logging software. Therefore, a third party would most likely be required. Hence, GeoSQ is not awarded this feature.

Requiring-Explicit-Consent

This feature is awarded to systems that require the user to initiate an authentication. It is regarded as a security and privacy feature [18]. GeoSQ requires explicit consent for an authentication process to begin, hence it is awarded this feature.

Unlinkable

An authentication system is classified as *Unlinkable* if a set of colluding systems cannot determine from the authenticator whether or not it is belongs to the same user on multiple different platforms [18]. In this case, since GeoSQ relies on a trusted third party, it would be classified as linkable. Hence it is not awarded this feature.

Resilient-to-Social-Media-Mining

This category is an addition to the original framework proposed by Bonneau et al. [18]. A system is awarded this feature if social media mining cannot provide an attacker with a leg up in guessing the user's authentication information. We award an authentication system this feature if social media mining does not improve the attacker's ability to guess by more than one percent.

In our study, we attempted to gauge the effect of social media mining on the participants ability to guess the pair's locations. In our adapted RCI framework (see Appendix C), we asked the participants if they follow their pairs on social media platforms, and whether or not they see frequent location posts. The results showed that out of the 38 participants (19 pairs) that conducted session 1, all of the pairs followed each other on social media. When asked on how many different mediums, the average was 2.67 per participant. Even though a pair might follow each other on social media that does not mean that they are active, therefore we asked how many posts per week do you does your pair make? The average was 3.05 posts per week. After that we asked how many location posts your pair makes per week the average for that question was 1.62.

This data is very important because it shows how pervasive social media mining can be when used as a tool by the attacker. Based on the threshold we set (7/10), and the average number of location posts per week (1.62), we estimate that social media mining will not provide an attacker a significant advantage. Hence GeoSQ is *Resilient-to-Social-Media-Mining.*

Resilient-to-the-Known-Adversary

Resilient-to-the-Known-Adversary is another category that we appended to the originally proposed Bonneau et al. Framework. An authentication system is awarded the accolade of being resilient to the known adversary if any attacker with elevated access to a user's devices and elevated knowledge of the user cannot compromise the system. To be more specific, a throttled known adversary should not be able to compromise more than 1% of accounts protected by the authentication system given 10 guesses a day for 365 days.

Since the known adversary is often a throttled adversary, that leads to the question, what are the optimal throttling settings? In addition to; what are the thresholds that we should use for authentication? When utilizing passwords, or security questions the answer is simply an exact character match. However, in more complex systems such as GeoSQ for example there are several thresholds that we could utilize for authentication.

As mentioned earlier, in our study we asked the participants to answer ten location questions gathered over a period of 7-11 days. In order to determine what the optimal number threshold for a successful authentication is (correct answers to autobiographical location questions out of 10) we conducted an analysis on the data to determine the true positive rate and the false positive rate in the form of a Receiving Operator Characteristics graph (see Figure 4.2).

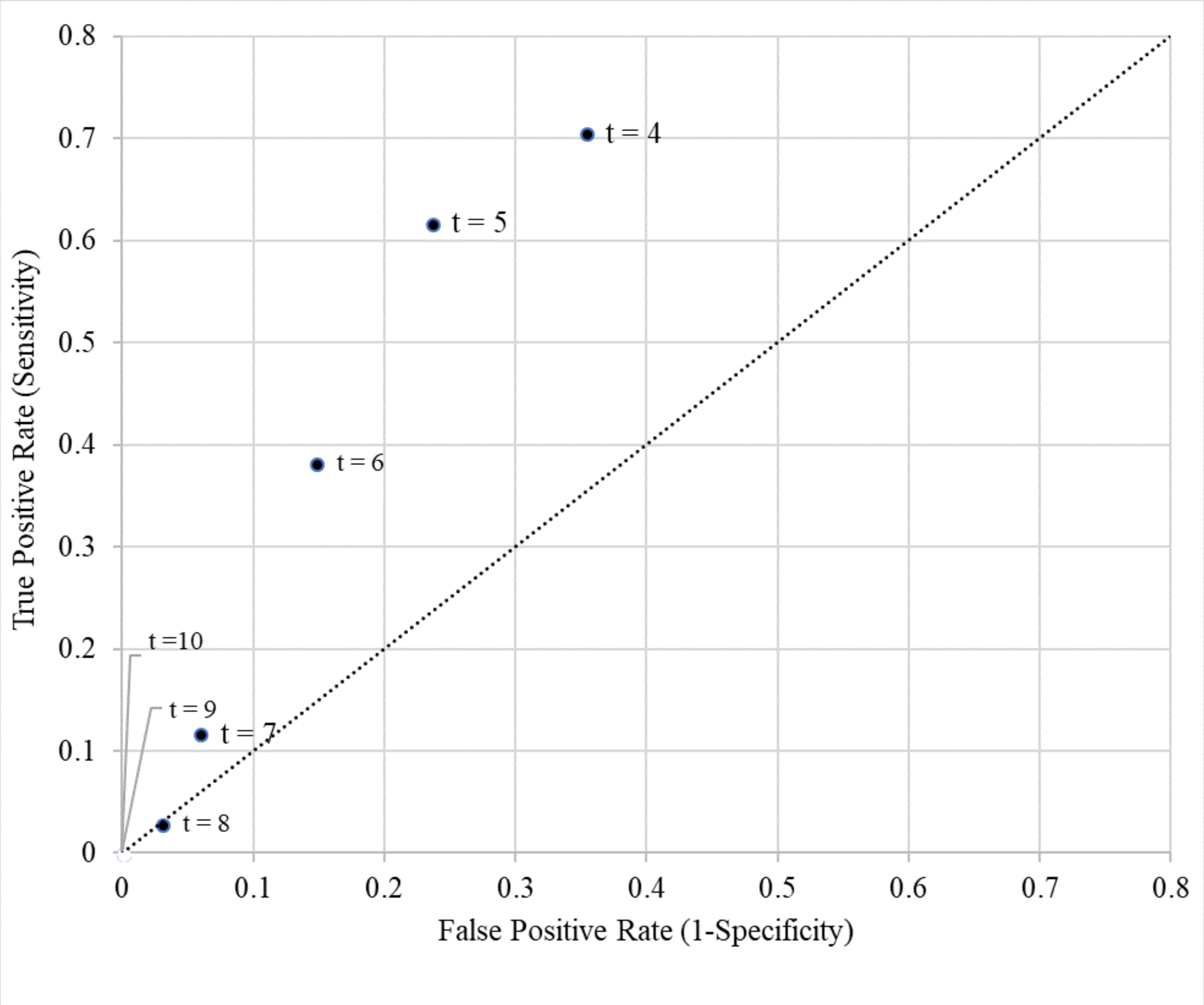


Figure 4.2: ROC graph showcasing the effect the threshold will have on the FPR and the TPR.

As mentioned earlier, according to the ROC graph the resilience to the known adversary depends solely on the threshold set. At a threshold of 7/10 correct questions, the false positive rate (FPR) is very low but so is the true positive rate (TPR), which is a preview on the memorability (to be discussed in the usability analysis). The TPR improves when the threshold is set to six, but the issue is that by then the FPR is close to 15% which is much higher than 1% of all accounts; the threshold we set for an authentication system to be considered resilient to the known adversary.

Therefore, GeoSQ is not *Resilient-to-the-Known-Adversary.*

4.3.2 Usability

Memorywise-Effortless

In order for an authentication system to be considered Memorywise-Effortless it must not require a user to remember any secret at all [18]. A system is rewarded with the status of being *Quasi-Memorywise-Effortless* if there is only one secret to remember for all verifiers [18]

Based on those parameters GeoSQ is considered to not be *Quasi-Memorywise-Effortless* or *Memorywise-Effortless* because more than one secret is required for all verifiers. The secret (the location) is also not stagnant and multiple questions are required based on our system design.

Scalable-for-Users

An authentication system is considered *Scalable-for-Users* if using the authentication system for many accounts does not put any extra burden on the user [18]. For example, if a user is asked to set multiple passwords across different verifiers, then as the number of accounts increase, the burden on the user is increasing.

Based on those parameters, GeoSQ is considered *Scalable-for-Users* because it does not increase the burden on the users as the number of accounts that utilize GeoSQ increase. For a specific timeframe, the same locations would have to be utilized. In addition to that, since the location being collected is autobiographic, all the user has to do is go on with normal life instead of having to register authentication credentials.

Nothing-to-Carry

This category is awarded to authentication systems that do not require the user to carry anything to be able to authenticate. Some systems require a physical card to be present for authentication, these systems would not be granted the Nothing-to-Carry accolade [18]. Moreover, an authentication system is given the status of Quasi-Nothing-to-Carry if the user will carry this device anyway (i.e., a smartphone) [18]. Based on these parameters GeoSQ would be granted the status of *Quasi-Nothing-to-Carry*.

Physically-Effortless

This category is granted to authentication systems that do not require the user to perform any action past the pressing of a button (to indicate authentication is occurring i.e., submit) [18].

Based on that definition, GeoSQ would not be considered physically effortless. The data collection process, which is equivalent to setting your credentials, is physically effortless (users are typically moving around with their smartphone anyway). However, the process of authentication requires the user to utilize the map and click to set a marker, rendering the system not *Physically-Effortless*.

Easy-to-Learn

In order to determine if the system was *Easy-to-Learn* and easy to use, we followed up the usage of the systems during Session 2 of our user study and asked a series of usability questions, including the following two Likert Scale questions: 1- I thought the system was easy to use, 2- I thought the system was easy to learn how to use. The users were asked to indicate the extent in which they agreed with those statements on a scale of 1 to 7.

Figure 4.3 shows the responses to the GeoSQ system. 65% agreed, strongly agreed, or very strongly agreed that the GeoSQ system was easy to use. 11% of our participants were neutral and 26% either disagreed, strongly disagreed, or very strongly disagreed. As for the second question 79% agreed, strongly agreed, or very strongly agreed that the system was easy to learn how to use. 6% of our participants were neutral and 26% very strongly disagreed). Based on that we grant GeoSQ the accolade of being *Easy to Use* and easy to learn how to use.

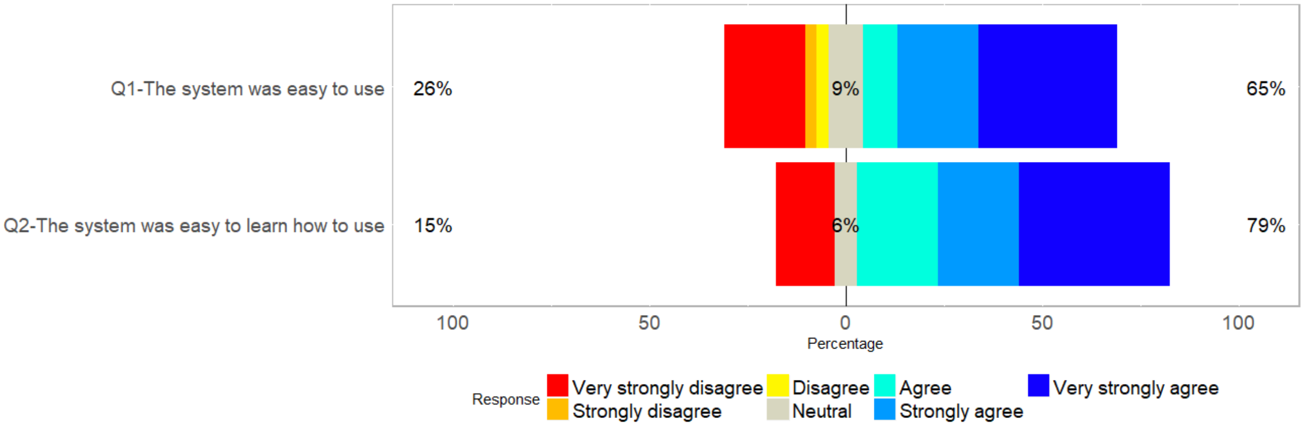


Figure 4.3: Usability Likert Scale questions - GeoSQ

Efficient-to-Use

In order to an authentication system to be considered *Efficient-to-Use* the time spent for each authentication must be acceptably short [18], the time required for a user to set up his/her credentials with a verifier should also be reasonable [18]. Bonneau et al. do not specify empirical values for this category which is appropriate because of the variety of authentication systems and the variety of intended use. For example, a smartphone would be more secure with an 8-character PIN; however, it is unusable for that target environment because a user typically wants to access the smartphone quickly.

Since GeoSQ does not really require the user to set their authentication credentials, it scores well in that half of this category. Figure 4.4 shows the login times per question, Figure 4.5 shows the average login time for all participants for all ten questions.

The most noticeable outlier is Q1 which has an average that is far higher than the other questions. Since we utilized a last in first out algorithm when deciding the order of the questions that are displayed to the participant, we know that Q1 was the most recent location that was logged. However, despite that the login time for that specific question appears to be a lot higher, we expected the login time for that specific question to be the lowest because it is the most recent. During our user study, we observed that the participants tended to fiddle around and get used to

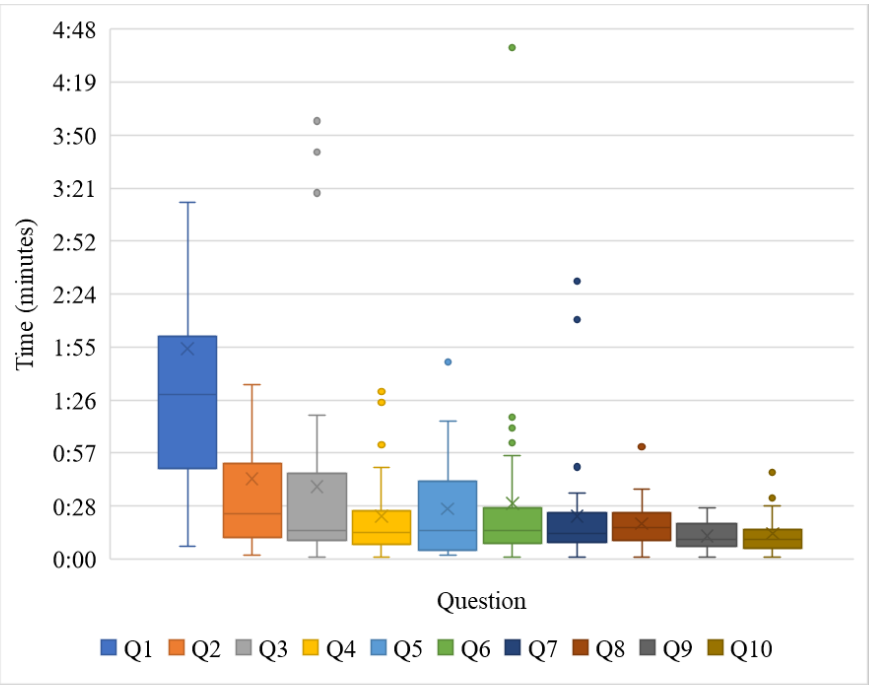


Figure 4.4: Average login time for GeoSQ for each question. Three outliers standing at 12:00

the

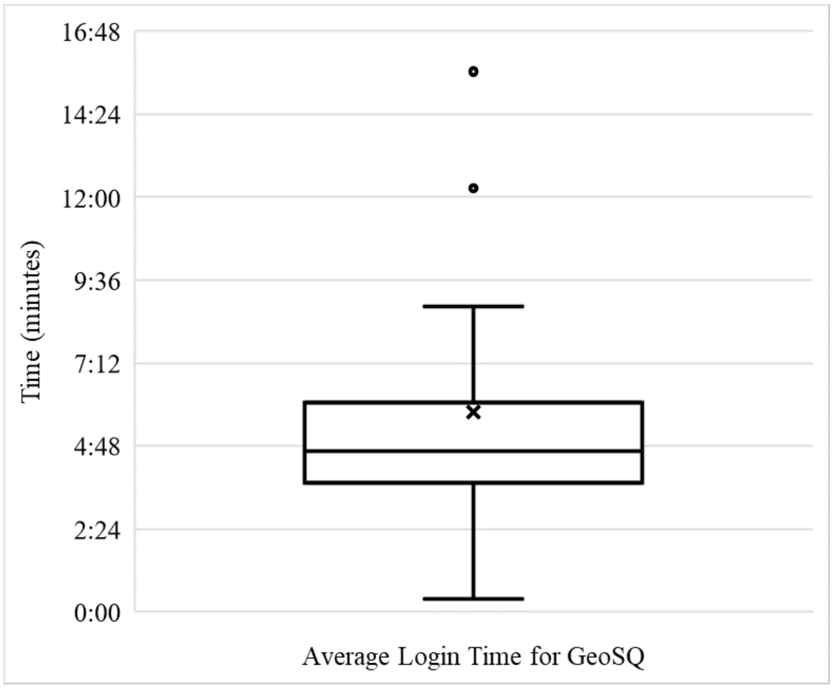


Figure 4.5: Average login time for GeoSQ for all ten questions.

interface in question one which explains the finding. The general trend after question one is a decreasing login time, this suggests that the users are becoming more familiar with the interface. GeoSQ cannot be classified as *Efficient-to-Use* due to the high login time.

Infrequent-Errors

An authentication system may be awarded in the *Infrequent-Errors* category if the login task typically succeeds when performed by the true user [18]. In the case of GeoSQ, since it is not only one question being answered, this category will depend on the user correctly answering a certain number of questions greater than the system threshold. Figure 4.6 showcases the number of correct/incorrect responses for all 34 participants of Session 2. Figure 4.7 showcases the average correct responses for each user.

According to the Figure 4.7, the average correct response is 5.06, which is just slightly higher than 50% of the questions asked. Recall that the participants were asked ten unique autobiographical location questions each. The participants were not allowed to attempt answering the same autobiographical location question if it was answered incorrectly, but they were simply moved to the next question.

The results indicate a dangerous trend to the security and usability of the system.

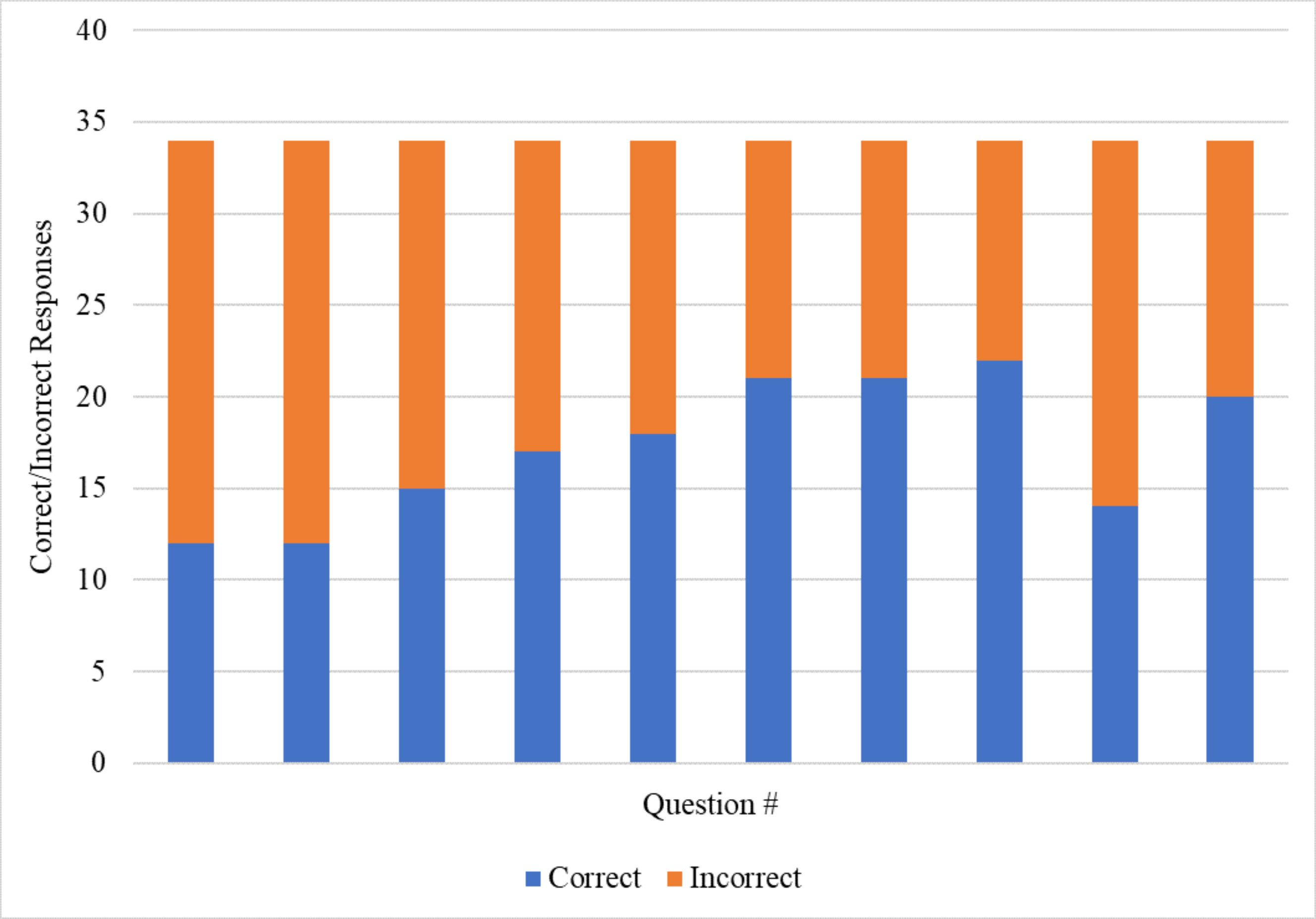


Figure 4.6: Total number of correct and incorrect responses per question, by legitimate users.

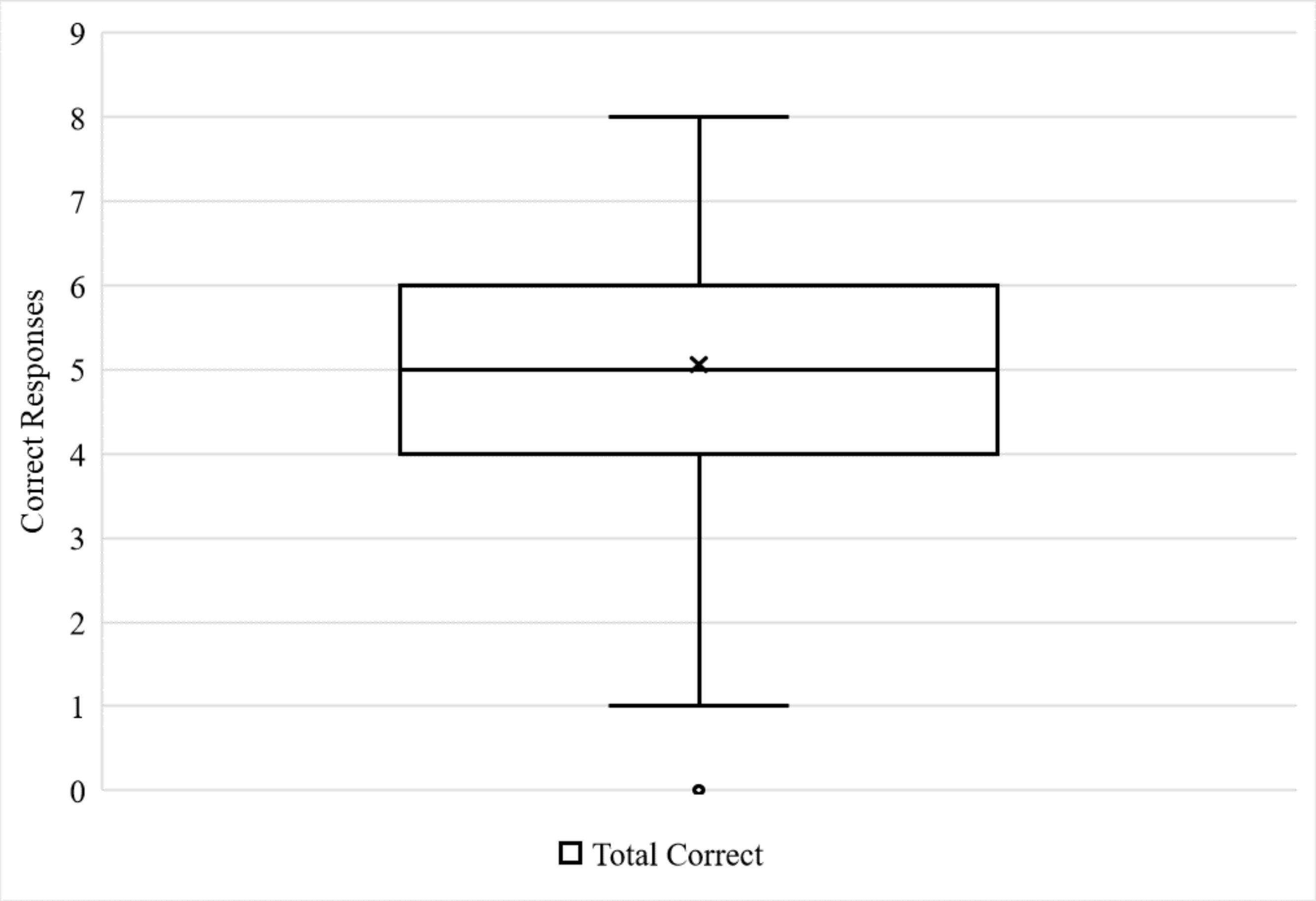


Figure 4.7: Correct responses by legitimate users.

Figure 4.2 showcases the ROC graph that determines the false positive rate and the true positive rate for each threshold. Given the number of incorrect answers in order to make this system usable for 61% of users, we must set the threshold to 5 correct answers out of 10. However, the risk becomes the security concern because at that threshold, the false positive rate is about 24% (n=34, 17 pairs).

As a result of the frequent errors and the need to decrease the threshold to an extent that would compromise GeoSQ's security, we do not consider GeoSQ to have *Infrequent-Errors*. Which was a surprising result given the promising results we obtained during pilot testing.

Easy-Recovery-from-Loss

This category rewards authentication systems that make it easy to recover from loss. However, in the case of GeoSQ, should it be deployed for the purpose of fallback authentication, then it would not be easy to recover from loss. Our implementation of this autobiographical location-based system is based upon the fact that the user will have at least ten unique locations to be queried about. That typically does take a significant period of time to accumulate (many of our participants did not have ten unique locations at the end of seven days so we had to extend the range to 11 days). Should GeoSQ fail as a fallback authentication system for a legitimate user, it would not be feasible to wait until the user has a set of new questions. Therefore, GeoSQ would not be easy to recover from loss.

4.3.3 Participant Preferences, Attitudes, and Concerns

This section will encompass other usability dimensions for GeoSQ that are not part for the Bonneau et al. Framework [18]. At the end of Session 2, we asked participants to fill out an exit survey for GeoSQ; which asked usability, preference, and concern questions in order to determine how the participants viewed the GeoSQ system. All of the questions were Likert Scale questions ranging from 1 to 7.

We begin by analyzing the data we collected for GeoSQ. After the two ease of use/ease of learning questions are asked (refer to Section 4.3.5 Figure 4.3 for the results), we asked the participants if they wrote anything down or recorded any information that helped them use this system. None of them indicated that they recorded anything manually/electronically or wrote anything down, but one participant did indicate that they used an online service that logs locations and associated times to attempt to remember the answers to the location questions asked.

Followed by that we ask: For a mobile device's forgotten password/PIN or failed biometric, to what extent would you prefer to use this system (GeoSQ) in lieu of any of the following fallback authentication methods? Figure 4.8 showcases the results.

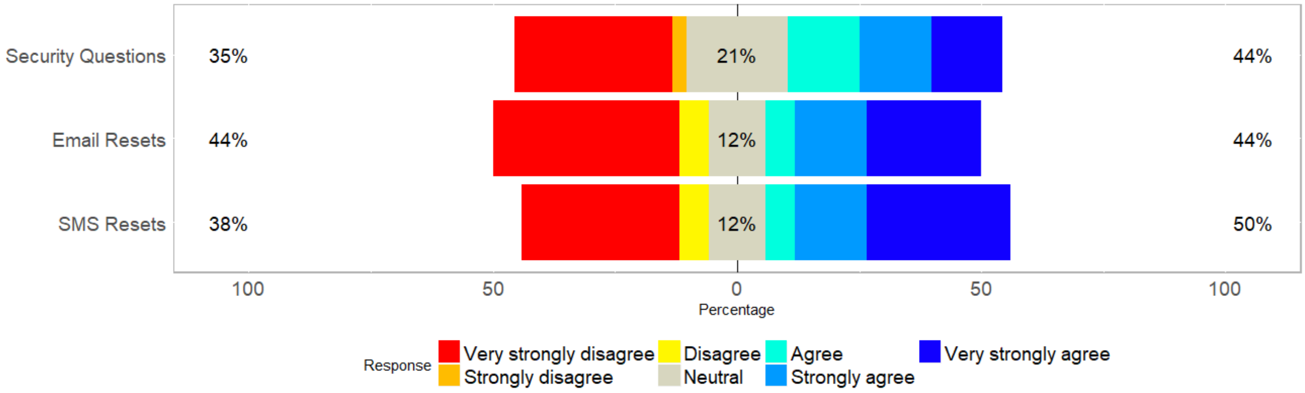


Figure 4.8: For a mobile device's forgotten password/PIN or failed biometric, to what extent would you prefer to use this system in lieu of any of the following fallback authentication methods?

Figure 4.8 shows that the participants were split on the issue of utilizing GeoSQ. With the majority leaning towards neutral or negative in lieu of any of the fallback authentication mechanisms mentioned above.

Next, we asked the participants: For an online account's forgotten password, to what extent would you prefer to use this system in lieu of any of the following fallback authentication methods? Figure 4.9 showcases the results.

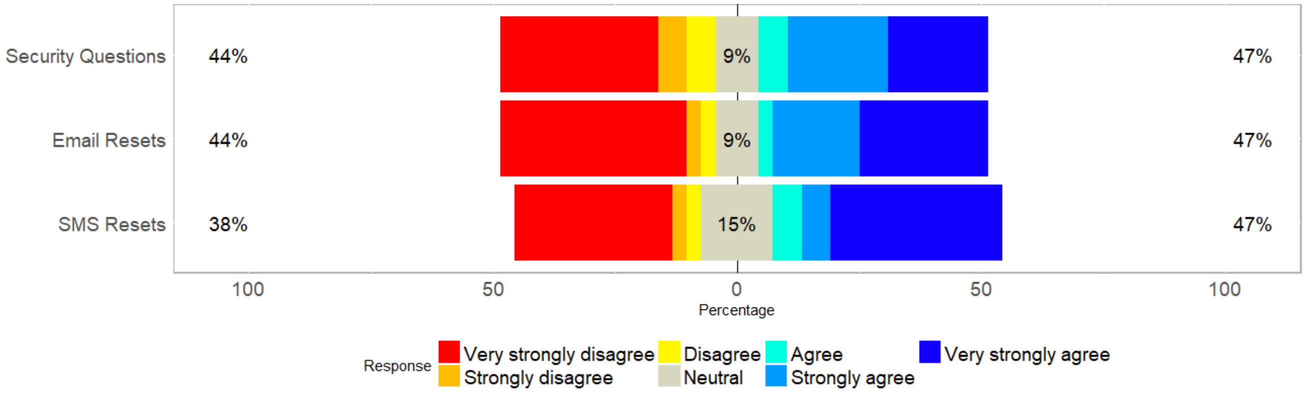


Figure 4.9: For an online account's forgotten password, to what extent would you prefer to use this system in lieu of any of the following fallback authentication methods?

Based on Figure 4.9, the results once again tend to favor the negative and neutral answers with an equal 47 percent of the answers being positive across all three types of popular fallback authentication techniques. This suggests a certain frustration with the time it took, during the user studies we observed that many participants would really try hard to remember their locations to no avail.

Lastly, we asked three more questions to gauge the concerns that the participants might have while utilizing an authentication system that utilizes sensitive location data. We asked three questions. 1 If you were to use this system, how likely do you think it is that someone you know could access your device/accounts without consent? 2- To what extent are you concerned of a privacy leak as a result of applications utilizing location services (e.g., Google Locations, GeoSQ, etc.)? 3- How concerned are you about a privacy leak on any medium (e.g. Facebook, contact information, pictures)? Figure 4.10 showcases the responses attained from the participants in the study.

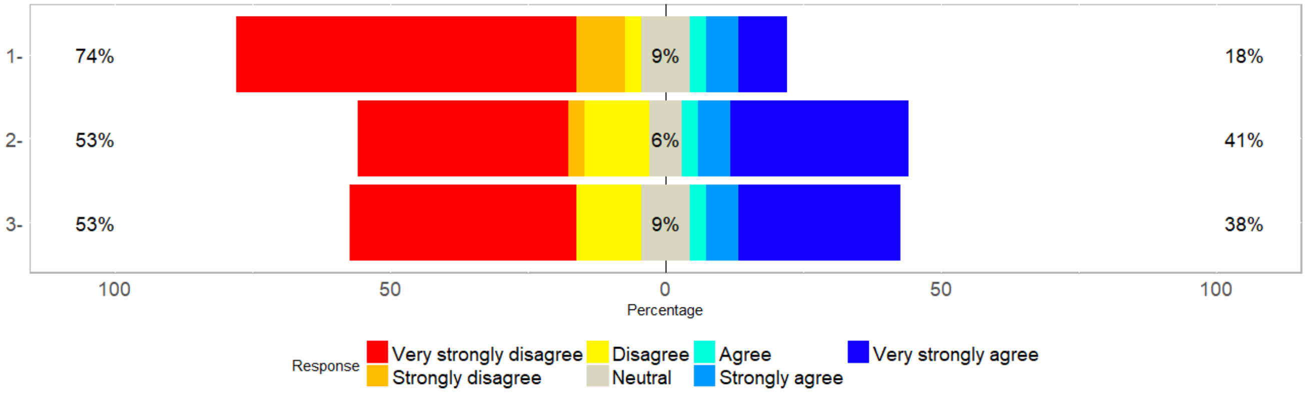


Figure 4.10: 1-If you were to use this system, how likely do you think it is that someone you know could access your device/accounts without consent? 2- To what extent are you concerned of a privacy leak as a result of applications utilizing location services (e.g., Google Locations, GeoSQ, etc.)? 3- How concerned are you about a privacy leak on any medium (e.g. Facebook, contact information, pictures)?

Figure 4.10 shows that most participants are not concerned about a breach caused by a known adversary, should GeoSQ be utilized to secure their devices/accounts. Only 18% of participants voiced a concern that GeoSQ could lead to a known adversary accessing their devices/accounts without consent. The concern for a privacy leak on any medium was also considerably low; only 41% of participants were either somewhat concerned, concerned, or highly concerned. Lastly, only 38% of participants were somewhat concerned, concerned, or highly concerned of a privacy leak as a result of application utilizing locations services. These figures are surprisingly low when taking into account the number of leaks and privacy breaches that have garnered media attention in the recent past.

Overall, there seems to be a reluctance by most participants when it comes to using GeoSQ in lieu of other forms of fallback authentication, and participants seem to oppose an authentication scheme such as GeoSQ to access their accounts/devices.

4.3.4 Deployability

Accessibility

The Android environment has several accessibility features that allow individuals who are differently able to use Google Maps [1]. We utilized the Google Maps API in order to develop GeoSQ. Therefore, the accessibility features offered are compatible with GeoSQ. On Android OS BrailleBack, TalkBack, Voice Access help are offered readily in many different distributions of the Android OS.

It is important to note that only certain devices offer the BrailleBack feature [1] and not all common devices offer this feature which is designed for individuals who have visual impairments and more specifically individuals that have visual and hearing impairments simultaneously. As a result of the Google Accessibility Suite we reward GeoSQ with the benefit of being accessible with one caveat regarding the speed of use. Individuals that are differently able will realistically require more time to use GeoSQ.

Negligible-Cost-per-User

GeoSQ is awarded in the *Negligible-Cost-per-User* category because implementing the system is possible with open and free to use APIs hence the burden on a verifier implementing such a system does not go up when the system is scaled to many users.

Server-Compatible

GeoSQ is not server compatible, because the definition of being server compatible in that category is not having the verifier significantly change their existing authentication settings, and the authentication system must be compatible with text-based passwords from the point of view of the verifier. GeoSQ would require significant edits made to the backend to make both these authentication systems work.

Browser-Compatible

GeoSQ is not designed for a browser environment, therefore this category does not pertain to it. It is however possible to implement a version of GeoSQ that is compatible with many modern browsers. Any modern browser that supports the Google Maps API would be able to run a browser based GeoSQ.

Mature

GeoSQ is not a mature authentication system, this category is gauged towards authentication systems that have already been implemented widely. GeoSQ is simply an authentication system we designed and implemented for testing purposes that has not been tested for a long period of time on a wide scale.

Nonproprietary

GeoSQ is not patented in an yway and the technology behind it (Google Maps API) is free and open to use with alternatives to this API being readily available (though not as good).

Table 4.1 shows the full system evaluation according to the Bonneau et al. framework [18] in addition to our two added categories, and compares GeoSQ to passwords, security questions, Email resets, and SMS resets.

Table 4.1: Adapted Bonneau Framework [18]. Analysis of Security Questions [44], analysis of web passwords [19]. Please note that the two extra security categories that we added were not evaluated in the original framework proposal [19].

	Memorywise-Effortless	Scalable for Users	Nothing to Carry	Physically Effortless	Easy to Learn	Efficient to Use	Infrequent Errors	Easy Recovery from Loss	Accessible	Negligible Cost per User	Server Compatible	Browser Compatible	Mature	Non-Proprietary	Res. to Physical Observation	Res. to Targeted Impersonation	Res. to Throttled Guessing	Res. to Unthrottled Guessing	Res. to Internal Observation	Res. to Leaks from Other Verifiers	Res. to Phishing	Res. to Theft	No Trusted Third Party	Requiring Explicit Consent	Unlinkable	Res. to Social Media Mining	Res. to the Known Adversary
Passwords (Baseline)			●		●	●	○	●	●	●	●	●	●	●		○						●	●	●	●	▲	▲
Security Questions	○ ⦀		●		●	●	○	●	●	●	≡	●	●	●	≡	≡				≡		●	●	●	● ≡	≡	≡
Email Resets	● ⦀	● ⦀	●		●	●	● ⦀	●	●	●	≡	●	●	●		○	● ⦀		⦀	≡	≡	●	≡	●	≡	△	△
SMS Resets	● ⦀	● ⦀	≡		●	●	● ⦀	●	●	●	≡	●	●	●		○	● ⦀	△	⦀	≡	≡	○ ≡	≡	●	≡	△	△
GeoSQ	≡	● ⦀	○ ≡	≡	●	≡	≡	≡	●	●	≡	●	≡	●	● ⦀	○ ≡	● ⦀	● ⦀	⦀	● ⦀	● ⦀	● ≡	○ ≡	●	≡	● ⦀	≡

● – offers the benefit, ○ – almost offers the benefit, no circle – does not offer the benefit, ▲ – potential to have the benefit , △ – potential to almost offer the benefit, ⦀ – better than passwords, ≡ – worse than passwords

4.4 Resulting System and Policy Recommendations

Based on the findings from our user study we suggested several system and policy recommendations for GeoSQ.

- As newer smartphones and more accurate localization comes along the error margin of 400 meters must be decreased.

This recommendation would greatly increase the sample space of the locations possible during a guessing attack hence increasing the security of GeoSQ. Our error margin was devised based on low end devices with lower than normal accuracy, as time passes by and localization improves we expect that the error margin can be safely decreased without any negative effect on usability.

- As newer smartphones and more accurate localization comes along the unique location margin of 200 meters can also be decreased.

The Google Maps API is undergoing constant updates that improves localization, this can be observed by the fact that many different services that utilize the Google Maps API have become more successful in pinpointing the exact establishment a user is in to a high degree of accuracy. Our demographic (undergraduate university students) move around very little. Therefore, oftentimes after the seven-day marker they did not have ten unique locations. Movements from building to building on our university campus were often not registered due to the 200-meter margin set for counting a unique location. Decreasing this margin for counting a unique location will greatly improve usability and will make it more likely for users in the future to obtain 10 unique locations in a shorter period of time. At the same time, the sample space will increase, hence improving the security of GeoSQ.

- The optimal system setting occurs when the error margin and the margin for a unique location is set to be equal.

The margin for error currently set at 400 meters and the margin for a unique location which is currently set to 200 meters is not the optimal configuration. Ideally, the margin for error would be equal to the margin for a unique location. We were bottlenecked by the fact that older devices had such a high error margin, following the same theme as previous recommendations that more modern devices are capable of more accurate localization and very little cost to battery life. Therefore, when those two system settings are set to be equal, the sample space greatly increases at no foreseen cost to usability and an increase in the security of GeoSQ.

- Ensure that the timestamp labels the date and time (MM-DD) (HH:MM) as well as the day of the week.

Throughout our user study, we observed that many participants had to look up the day of the week for a certain date to try to remember where they were. They usually used their calendar either on their smartphone or on their laptops. This increased the time it took for login and made GeoSQ less usable. Therefore, we recommend adding the day of the week to the location question for GeoSQ instead of simply have the date and the time as noted above.

4.5 Discussion and Future Work

Perfect authentication systems are ones that managed to bridge the security usability chasm effectively. When designing and implementing GeoSQ, several decisions that affect usability, but benefit security had to be made and vice versa as is the norm for any authentication system being developed.

Similar work in the realm of autobiographical authentication made the same compromises. For example, in the work done by Albayram et al. [13] they only utilized locations from the past 24 hours. Their goal was to improve memorability, but this affects security because it makes monitoring the potential victim much easier, and the diversity of locations in only one day in a real-world environment is very low.

In GeoSQ we opted for more security by embedding a ten unique location requirement in our authentication system which takes longer to collect and according to our results is much less memorable than AlBayram et al's. system [13]. AlBayram et al. observed a success rate of 69% with their autobiographical location questions, while we only observed an average accuracy of 5.2/10 questions (roughly a 52% accuracy). Albayram et al. only asked one location question from data gathered within the past 24 hours, while we asked 10 questions from data collected from the past 7-11 days. It is important to note that our system was significantly different than AlBayram et al's. autobiographical authentication system [13] because a mixture of autobiographical data was utilized in their system. That is why we only focus on the results of their autobiographical location data.

The timespan to obtain ten unique locations in our demographic was typically 7-11 days as observed in our study. This greatly affected the memorability as noted in our security and our usability analysis of GeoSQ. Decreasing the timespan for collecting location and decreasing the 10 unique location requirement should lead us to observe an increase in memorability but a corresponding decrease in the sample space which would negatively affect our security. The first question asked (which should have been from the previous 24 hours) was the least memorable according to our user study we believe that it was an anomaly because AlBayram et al's. work [13] suggests that autobiographical location data from the past 24 hours is memorable 69% of the time.

In future work it would be interesting to alter GeoSQ so that the number of questions are decreased with the timespan of data collected also being decreased. Based on the results of similar work [13, 35], we should observe an increase in the memorability which would make the system more usable.

Our results indicated that many participants would not prefer to utilize GeoSQ over other authentication schemes (for fallback authentication or primary authentication). The average time it took for participants to answer all 10 questions was 5 minutes and 40 seconds, which is considerably more than it would take them to fallback authenticate using security questions, Email resets, and

SMS resets. We also observed a great amount of audible frustration ("I don't remember where I was") regarding the lack of memorability. Despite the security of GeoSQ it is highly unusable because of high login time. It may be appropriate for fallback authentication but the frequency of errors (due to lack of memorability) makes it highly unusable on a large scale.

Our threshold for counting a login attempt as successful was 7/10 correct questions answered. By decreasing this threshold, we present a security challenge especially when it comes to the known adversary (refer to Section 4.3.1, Figure 4.2 for the ROC graph). As a result, we do not recommend that GeoSQ be utilized on a larger scale without further testing and tweaking the system settings in order to increase the usability.

4.6 Conclusion

In conclusion we designed, implemented, and tested an autobiographical location authentication system called GeoSQ. Our analysis leads us to the conclusion that GeoSQ is highly unusable given the current system settings. While the observed security meets several requirements, the usability in terms of two categories; login time (*Efficient-to-Use*) and accuracy (*Infrequent-Errors*) was abysmal. The average login time was 5 minutes and 40 seconds, while the average accuracy of the legitimate users was around 5.2/10 (52%). GeoSQ did not show much resistance to the known adversary as 7% of the accounts would be guessed by the known adversary. Should we decrease the threshold for a successful login from7/10 correct answers to 6/10 or 5/10 correct answers for improved usability, then the known adversary would be a significant threat to the system. Unfortunately, it appears that this system is unable to bridge the security-usability chasm. While the target environment for GeoSQ was fallback authentication for smartphones, we do not recommend the system for such an environment using the current settings due to its low usability.

Chapter 5

GeoPassHints

5.1 Introduction

Today, the most commonly utilized fallback authentication mechanisms are security questions, Email resets, and SMS resets, but unfortunately they have many security [18,31,34]. Despite that, due to their usability, these fallback authentication mechanisms are still in use [18]. The flaws in the current most popular fallback authentication mechanisms motivate our work to attempt to find alternate replacement authentication mechanisms.

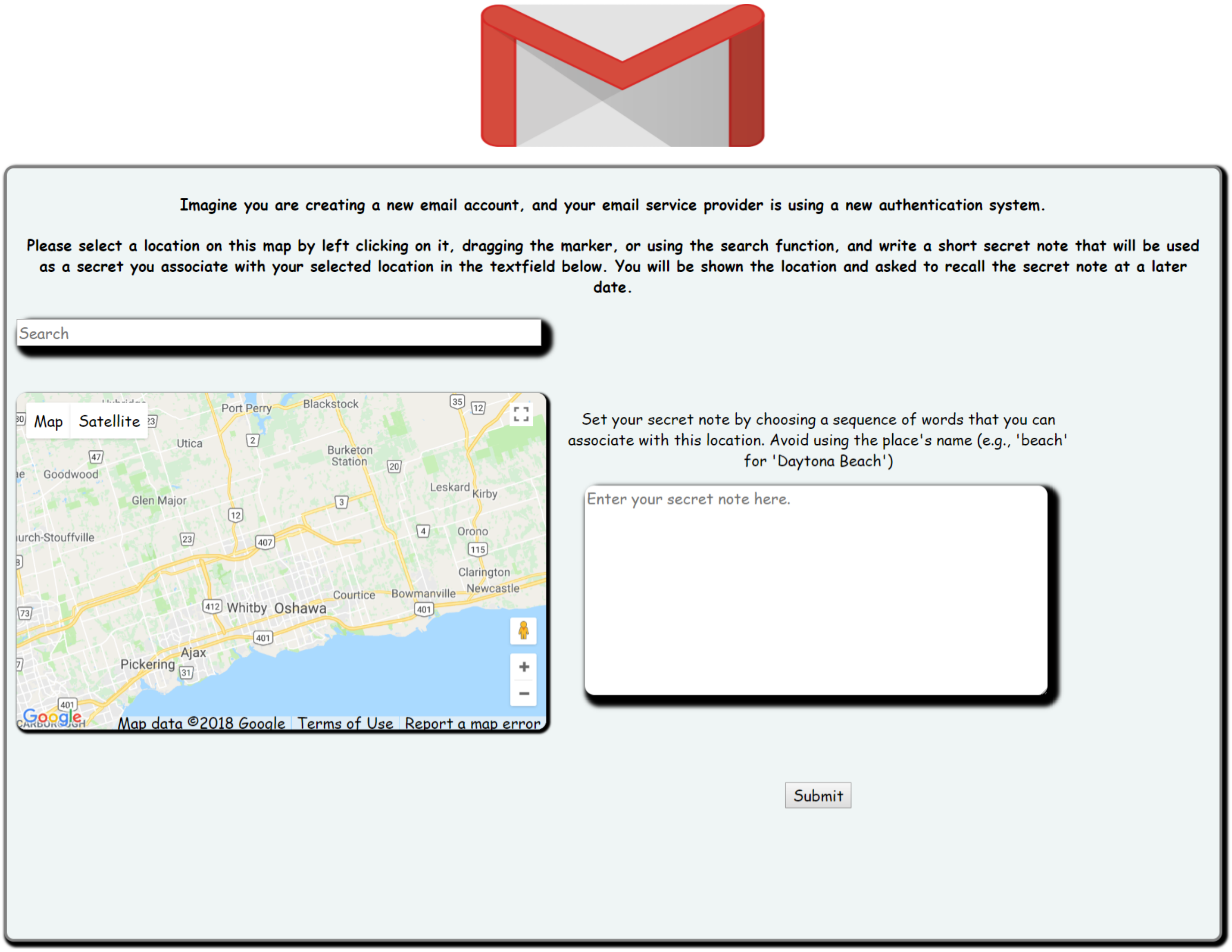
One such attempt at bridging the security-usability chasm is the utilization of geographic authentication systems. Geographic authentication systems rely (at least) in part on a map, or a location on a map, for authentication. Geographic authentication systems are distinct from autobiographical location authentication systems (such as GeoSQ) that use a map as geographic authentication systems do not rely on dynamic location data.

We attempt to bridge the security-usability chasm by designing, implementing and testing an alternative authentication system called GeoPassHints. GeoPassHints is a geographic authentication system that can be used for fallback authentication or primary authentication. However, we focus on GeoPassHints being deployed in a fallback authentication environment because of its use of a hint to hopefully help trigger long-term memory.

This chapter will be organized as follows; we will begin by introducing the system design. Followed by that, we will complete our security, usability, and deployability analysis using data we collected from our user study. Similar to GeoSQ, we will utilize the Bonneau et al. [18] framework, which was developed for the comparative evaluation of web authentication systems. Lastly, we will present our resulting system and policy recommendations, followed by a discussion and future work section.

5.2 GeoPassHints System Design

GeoPassHints is a web-based authentication system that allows the user to set a location and attach a secret note they associate with that location. The user is not required to remember the location but is required to remember the secret note. Users have the option of searching to get to a location or they can alternatively click, drag, or even touch on the screen if they have a touchscreen laptop device. Users can also zoom in and zoom out as they please. GeoPassHints was also developed with the Google Maps API [4]. Figure 5.1 shows the GeoPassHints interface.



Imagine you are creating a new email account, and your email service provider is using a new authentication system.

Please select a location on this map by left clicking on it, dragging the marker, or using the search function, and write a short secret note that will be used as a secret you associate with your selected location in the textfield below. You will be shown the location and asked to recall the secret note at a later date.

Search

Set your secret note by choosing a sequence of words that you can associate with this location. Avoid using the place's name (e.g., 'beach' for 'Daytona Beach')

Enter your secret note here.

Submit

Figure 5.1: GeoPassHints interface.

5.2.1 GeoPassHints System Design and Development Details

The design of GeoPassHints had two challenges:

1. For improved security, we wanted to influence the participants to create longer secret notes

2. For security reasons, we wanted to make sure that the participants do not simply label the location they have selected, as the hint would be useful to the attacker.

To address the latter challenge, as done by MacRae et al. [39] [48] we added the following instructions during the secret note creation phase “Set your secret note by choosing a sequence of words that you can associate with this location. Avoid using the place's name (e.g., 'beach' for 'Daytona Beach')". To address the former challenge, we made sure that the input text area was large enough to let the user know that this is meant to be a longer secret note. All of this can be seen in Figure 5.1.

The solutions to the challenges seem to have worked. We had set a policy of at least five characters for the secret note. According to the logs, only two participants ran into that error when setting their GeoPassHints authentication credentials.

Much like GeoSQ the security goals we wanted to achieve were related to improving the overall security and usability of fallback authentication systems.

As with GeoSQ, we also had to make some security concessions in order to ensure increased usability. As mentioned earlier in Section 2.6, longer passphrases, user assigned passwords, and system assigned passwords all tend to be unmemorable and annoying to use because of input error. In order to account for that, we ensured that we had an error distance that was acceptable from a security and usability point of view. We utilized the Levenshtein distance in order to achieve that. The Levenshtein distance is a level of similarity between two strings [53]. We set the threshold to be 80%; therefore, if a string was 80% similar to another string, we accepted that as correct input. This could make it easier to guess, but it definitely makes it easier to input a longer secret note because of the error margin allowed. We will analyze the usability and the security concerns in Section 6.3.

After users set their location and associated secret note, they were prompted to confirm their secret note. This is similar to how passwords are confirmed to ensure that there was no input error. During the confirmation phase, we did not utilize the Levenshtein distance; the users had to have an exact character match (including whitespace). The Levenshtein distance is only utilized during subsequent authentication attempts, not when the users are setting their credentials.

5.3 Results

In this section we will present our analysis of GeoPassHints. The analysis will be split into three different broad categories 1-security, 2-usability, 3- deployability. Our analysis will be based on the criteria laid out by the Bonneau et al. framework for the comparative evaluation of web-based authentication schemes [18]. See Section 2.8 for a detailed description of each criteria.

Two criteria were added to the original Bonneau et al. framework, the first criteria added is *Resilient-to-Social-Media-Mining* and the second criteria added is *Resilient-to-the-Known-Adversary*. We added these two additional criteria for evaluation because of the increased risk of the known adversary, and the increased risk posed by the information we post on our social media account which is often open to mining.

A system is awarded being *Resilient-to-Social-Media-Mining* if social media mining cannot provide an attacker with a leg up in guessing the user's authentication information. We award an authentication system this feature if social media mining does not improve the attacker's ability to guess by more than 1%. An authentication system is awarded the accolade of being *Resilient-to-the-Known-Adversary* if any attacker with elevated access to a user's devices and elevated knowledge of the user cannot compromise the system. To be more specific, a throttled known adversary should not be able to compromise more than 1% of accounts protected by the authentication system given 10 guesses a day for 365 days..

5.3.1 Security

Resilient-to-Physical-Observation

GeoPassHints is not considered to *Resilient-to-Physical-Observation*, While passwords are also not *Resilient-to-Physical-Observation* [18], GeoPassHints is considered more resilient than passwords because of the longer average character length which stands at 17.3. However, GeoPassHints is still not resilient to physical observation it is still conceivable that an attacker could impersonate a legitimate user after observing a user authenticate a few times.

Resilient-to-Targeted-Impersonation

GeoPassHints does not offer the benefit of being *Resilient to Targeted Impersonation.* We examined 152 secret notes set and observed that many individuals set locations such as a previous home or a current home and the associated secret note was descriptive, along the lines of “this is my home”, “I live here”, “I used to live here”, “my future house”, “my home”. In addition, many secret notes included references to where participants’ parents live. In our manual analysis of the secret notes set, roughly 21% (32/152) of secret notes set were of that nature. Another common trend we noticed was labelling where the participant works. In our manual analysis of the secret notes set, roughly 4.6% (7/152) secret notes set were of that nature.

7.8% (3/38) participants reused notes for more than one account, and 10.5% (4/38) participants labelled at least one of the locations despite instructions not to do so (i.e., the location set was a restaurant called xyzrestaurant, and the secret note was “xyzrestaurant”). Lastly, there was a common trend of setting locations and secret notes to schools and universities (i.e, set a university location, and the secret note was “my university”, or “I go to school here”); roughly 7.2% (11/152) secret notes had some variation of that trend. Overall, 3.2% (5/152) secret notes were direct labels, 32.9% (50/152) were personal descriptive labels (e.g., involving home, work, or school), and 5.9% (9/152) were reused. An in depth analysis will be conducted in Section 5.4.

We evaluate whether or not the common trends in labelling and setting commonly known locations such as home or work made it easier for the known adversary to guess in Section 6.3.1 - Resilient-to-the Known Adversary.

Resilient-to-Throttled-Guessing

According to Bonneau et al. [18], a system is awarded being *Resilient-to-Throttled-Guessing* if an attacker, while being constrained by a verifier, cannot compromise more than 1% of accounts a year given ten guesses a day.

We utilized three different methods for password cracking. The first method we utilized was a probabilistic context-free grammar [54]. A probabilistic context free grammar is considered to be a more sophisticated attack against passwords that generates password guesses in descending probability order. A probabilistic context-free grammar will generate password guesses in descending order based on a training set. The training set will first be parsed into base structures for example "ILoveKarina123." would be LLLLLLLLLLDDDS where L is alphabetic based string, D is a digit string, and S is a special character string. Once a training set (or several training sets) are input and parsed using the probabilistic context-free grammar, then a list of password guesses in descending probability of occurrence is generated based on the most probable base structures and terminals. In our work, we train using the RockYou dataset [6], a leaked dataset of 32 million passwords that is publicly available for analysis.

Figure 5.2 showcases the number of secret notes guessed for each account type utilizing Weir's probabilistic context-free grammar [54]. Approximately 2.2% (3/136) of banking account secret notes, and e-commerce account secret notes were guessed after 2^{30} guesses. This means that secret notes are Resilient-to-Throttled-Guessing. Because 3650 guessing is 211.85 according to Figure 5.2 only 0% of secret notes were guessed utilizing the PCFG at $2^{11.85}$ guesses. No social media secret notes were guessed using the PCFG.

The second method we utilized for password cracking is the semantic cracker [51]. The semantic cracker segments a set of training passwords into their core components (alphabetic, numeric, special symbol, and then further characterizes the alphabetic components using part-of-speech and then Wordnet [8] which is a database for English language vocabulary, to determine their semantic categories). The core components are used in a PCFG to generate guesses from highest to lowest probability. We used the RockYou [6] dataset for training. Figure 5.3 showcases the results of the semantic cracker, which was the most successful at guessing secret notes. Banking account secret notes were the most successfully guessed (5.1% or 7/136)). Again, based on the results obtained by the semantic cracker, GeoPassHints is *Resilient-to-Throttled-Guessing*. Due to the fact that no secret notes were guessed by the semantic cracker at 211.85 guesses.

The third and last method we utilized for password cracking was John the Ripper [3]. John the Ripper is a popular password cracking tool that has many different modes. We utilized the incremental mode. The incremental mode within John the Ripper utilizes frequency tables in order to crack passwords. Frequency tables rely on letters that occur most in the English dictionary. For example, the letter e is highly likely to come after the letter h. JtR incremental will utilize trigram frequencies for each character position within a password. Trigrams are sequences of three letters, meaning that John the Ripper in incremental mode will utilize frequency tables for sequences of three letters to determine which letter is most likely to occur next. This is a slow cracking method that has the capability of guessing passwords that are not in the English dictionary.

Figure 5.4 showcases the password guessing results utilizing John the Ripper in incremental mode. John the Ripper was most successful in guessing e-commerce account secret notes and bank account secret notes(1.4% or 2/136). Based on the definition of *Resilient-to-Throttled-Guessing* GeoPassHints is also *Resilient-to-Throttled-Guessing* when utilizing John the Ripper in incremental mode.

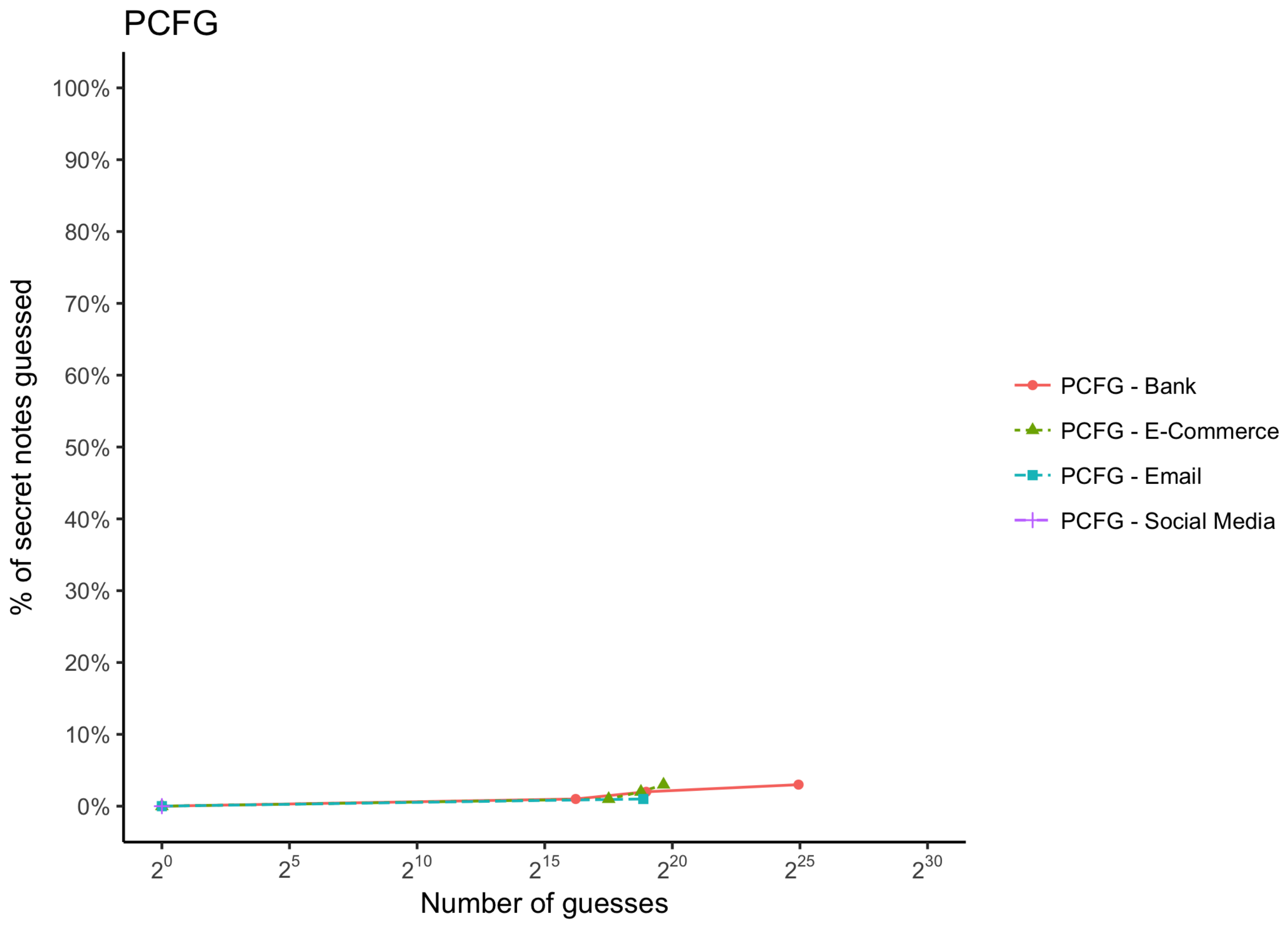


Figure 5.2: Percentage of secret notes guessed utilizing Weir's Probabilistic Context Free Grammar (PCFG) [54].

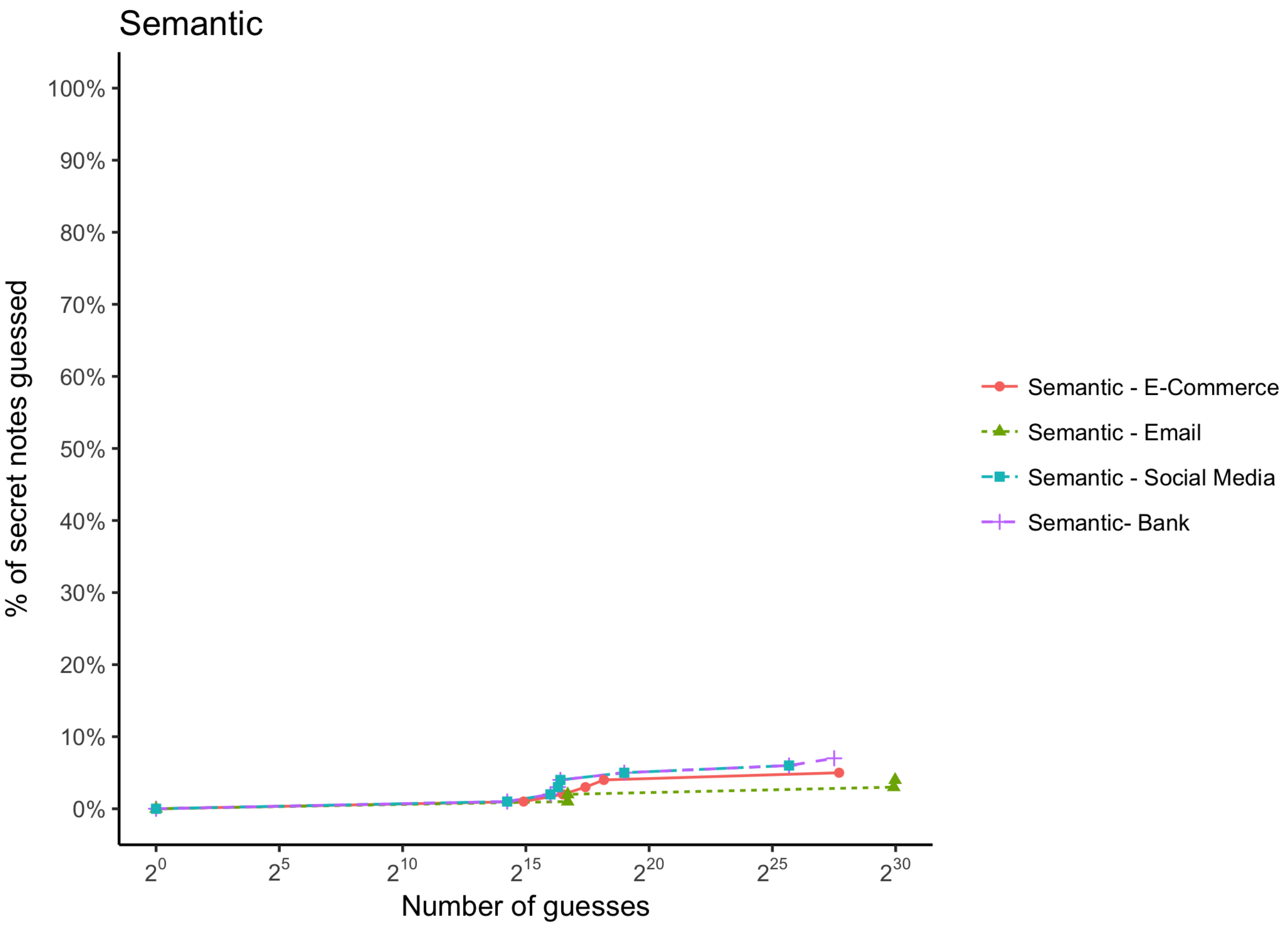


Figure 5.3: Percentage of secret notes guessed using a semantic cracker.

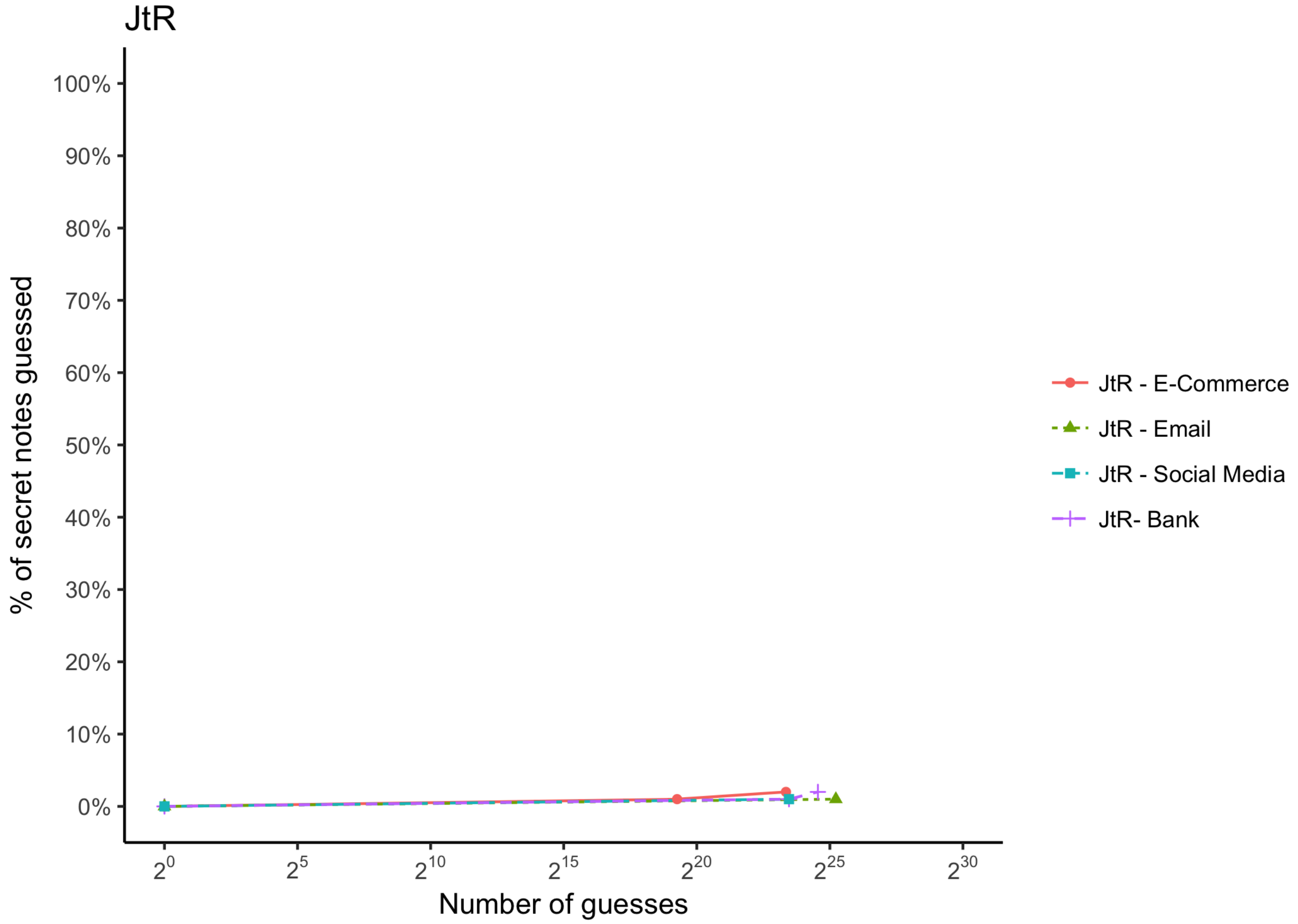


Figure 5.4: Percentage of secret notes guessing utilizing John the Ripper in incremental mode.

Resilient-to-Unthrottled-Guessing

GeoPassHints does not provide the feature of being *Resilient-to-Unthrottled-Guessing*. The 17.3 average character length of a secret note provides additional security than a typical password because it is much larger than the space of a typical password as evaluated by Bonneau et al. [18]. Therefore, at minimum GeoPassHints offers more resilience than the typical password despite the fact that our minimum character policy was set to 5. However, more than 1% of secret notes were guessed by our guessing algorithms in only 1 billion guesses (Weir's PCFG [54], John the Ripper (incremental mode) [3], and a semantic cracker [51]) as shown in Figures 5.2,5.3, and 5.4.

Resilient-to-Internal-Observation

According to Bonneau et al. [18], an authentication mechanism is *Resilient-to-Internal-Observation* if the attacker cannot impersonate the legitimate user by intercepting the legitimate user's input. An example of internal observation would be key logging or recording the screen of the legitimate user and relaying to an attacker's computer.

GeoPassHints is not *Resilient-to-Internal-Observation*. That is due to the fact that GeoPassHints relies on previously set credentials. Therefore, on an infected computer, GeoPassHints is not effective because the secret note can be compromised.

Resilient-to-Leaks-from-other-Verifiers

A system is classified as *Resilient-to-Leaks-from-other-Verifiers*; if anything, another verifier leak could potentially help an attacker impersonate a legitimate user [18].

GeoPassHints does not offer any protection when it comes to leaks from other verifiers. If the user decides to reuse a GeoPassHints secret note paired with the same location, then it would be easy for an attacker to impersonate a legitimate user. This would mimic the reuse of passwords that is typical of the majority of users [33]. Our user study showed 7.9% (3/38) of participants reused some note/location combinations, despite instructions to choose unique locations and unique secret notes.

Resilient-to-Phishing

GeoPassHints is considered *Resilient to Phishing*, against classical phishing attacks, because the location is set by the user. Therefore, one piece of information is being provided by the verifier, which the user can detect as being incorrect. A successful classical phishing attack can only be conducted if there is a leak from a verifier. Of course, it is not resilient to more advanced phishing attacks, such as targeted attacks, or man-in-the-middle attacks.

Resilient-to-Theft

This category is pertinent for authentication mechanisms that rely on a physical object for authentication. *Resilient-to-Theft* is not a pertinent issue for GeoPassHints, because there is no physical object to steal.

No-Trusted-Third-Party

This feature is awarded to systems that only rely on the user and verifier [18]. GeoPassHints does not require a trusted third party. In our implementation we utilized the Google Maps API for convenience; however, there are open source Map APIs that can be utilized [5]. These open source map APIs are not considered a third party as they would run locally, hence GeoPassHints does not rely on a trusted third party.

Requiring-Explicit-Consent

This feature is awarded to systems that require the user to initiate an authentication. It is regarded as a security and privacy feature [18]. GeoPassHints requires explicit consent for an authentication process to begin, so it is also awarded this feature.

Unlinkable

An authentication system is classified as *Unlinkable* if a set of colluding systems cannot determine from the authenticator whether or not it is belongs to the same user on multiple different platforms [18].

Except for a user that utilizes the same GeoPassHints secret note across multiple different verifiers, colluding verifiers cannot determine whether the same user is authenticating to different platforms. Therefore, unlike GeoSQ, GeoPassHints is classified as *Unlinkable*.

Resilient-to-Social-Media-Mining

This category is an addition to the original framework proposed by Bonneau et al. [18]. A system is awarded this feature if social media mining cannot provide an attacker with a leg up in guessing the user's authentication information. We award an authentication system this feature if social media mining does not improve the attacker's ability to guess by more than 1%.

GeoPassHints is *Resilient-to-Social-Media-Mining*. While location posts could uncover that a potential victim lives at a certain address, which could give the attacker a leg up in terms of a targeted guessing attack with that information, the large sample space (average character length was 17.3) even with an error margin of 20% (Levenshtein distance) we suggest that it would not be a significant advantage. In addition to that, we offer that GeoPassHints is *Resilient-to-Social-Media-Mining* because the all the participants indicated that they follow each other on social media, on an average of 2.67 different social media mediums (platforms), with an average of 3.05 social media per week, of which only 1.62 were location posts. Despite the fact that all participants followed each other on social media, only 2.9% (4/136) secret notes were guessed with a Levenshtein distance of 80%.

As mentioned earlier, there was a significant amount of descriptive labelling of previous, current, and future addresses, in addition to labelling universities and places where the participants work. A total of 32.9% (50/152) of secret notes followed that trend, not to mention direct labelling of locations, a total of 3.2%(5/152) of notes followed that trend. None of them were successfully guessed by known adversaries, despite the fact that they follow each other on social media hence GeoPassHints is *Resilient-to-Social-Media-Mining.*

Resilient-to-the-Known-Adversary

Resilient-to-the-Known-Adversary is another category that we appended to the originally proposed Bonneau et al. framework [18]. An authentication system is awarded the accolade of being resilient to the known adversary if any attacker with elevated access to a user's devices and elevated knowledge of the user cannot compromise the system. To be more specific, a throttled known adversary should not be able to compromise more than 1% of accounts protected by the authentication system given 10 guesses a day for 365 days.

Based on those parameters, GeoPassHints is not *Resilient-to-the-Known-Adversary.* While conducting a manual analysis on the secret notes created by our participants, certain trends were clear. A great number of people would set the location to their home and associate a secret note that eludes to that fact such as "this is my home" or "this is where I live". We also noticed a significant labelling of geographic locations and businesses.

We also found that the account type did have an effect on the secret note/locations that the participants set. For example, for the hypothetical bank account we found that many individuals selected their bank or associated a secret note with that hypothetical account with something to with "money" or "my bank". The logo we used in order to cue the participants that this was a bank account was a logo for the Bank of Montreal, a popular bank with a well-known logo, one user simply labelled the closest BMO branch with "bmo bank".

In addition to that, we found that four participants reused secret notes across the hypothetical accounts. It is clear that this trend is following the user's tendencies to pick the same password across multiple different accounts [40]. All of these factors would seem to indicate that GeoPassHints is not resilient to the known adversary.

Contrary to these initial patterns found in the manual analysis, GeoPassHints appeared to be strong in its resilience to the known adversary and the secret notes and the patterns mentioned above were not the secret notes that were guessed successfully the pairs. Despite that, a total of 4 notes were guessed by four participants, out of a total of 136 notes set (2.9%), which is higher than the threshold set out by this category. Please note that this number is not out of 152 as indicated in the note patterns mentioned in previous sections because only 17 pairs came back for Session 2. To the best of our ability, we ensured that the participants did not communicate when they were attempting to guess each other's secret notes to preserve the integrity of the data and to mimic a real-life scenario with a known adversary attempting to access an account or device without authorization; however, we were not able to prevent communication between sessions.

GeoPassHints is a system where the settings and thresholds matter to a great extent. In our case we were using a Levenshtein distance metric to give the participants and error margin (since the idea was to have longer secret notes and reduce the annoyance caused by typos affecting large strings for passwords). The question then becomes what is the ideal error margin? In order to answer that question, we created a ROC graph which is featured in Figure 5.5.

As shown in the ROC graph (Figure 5.5), it is clear that a threshold of 1 (an exact character match) is not usable; roughly 15 percent of the participants managed to remember a secret note correctly with this threshold after a period of 7-11 days. Our target threshold for success in our study was a Levenshtein distance of 0.8 which is a 20% error margin between the target string (secret note set during Session 1) and the input string (secret note recall during Session 2).

However, after conducting this analysis, we see that the false positive rate does not improve when the threshold is decreased to 0.7 (30% different than the target string), while at the same time the true positive rate does increase significantly when the threshold is set lower. The same trend follows when we decrease the threshold to 0.6. Only 2.9% (4/136) of accounts were compromised with a threshold of 0.8, 0.7, and 0.6. This data suggests that GeoPassHints is not *Resilient-to-the-Known-Adversary*, and it also indicates that the correct usability/security tradeoff in terms of the

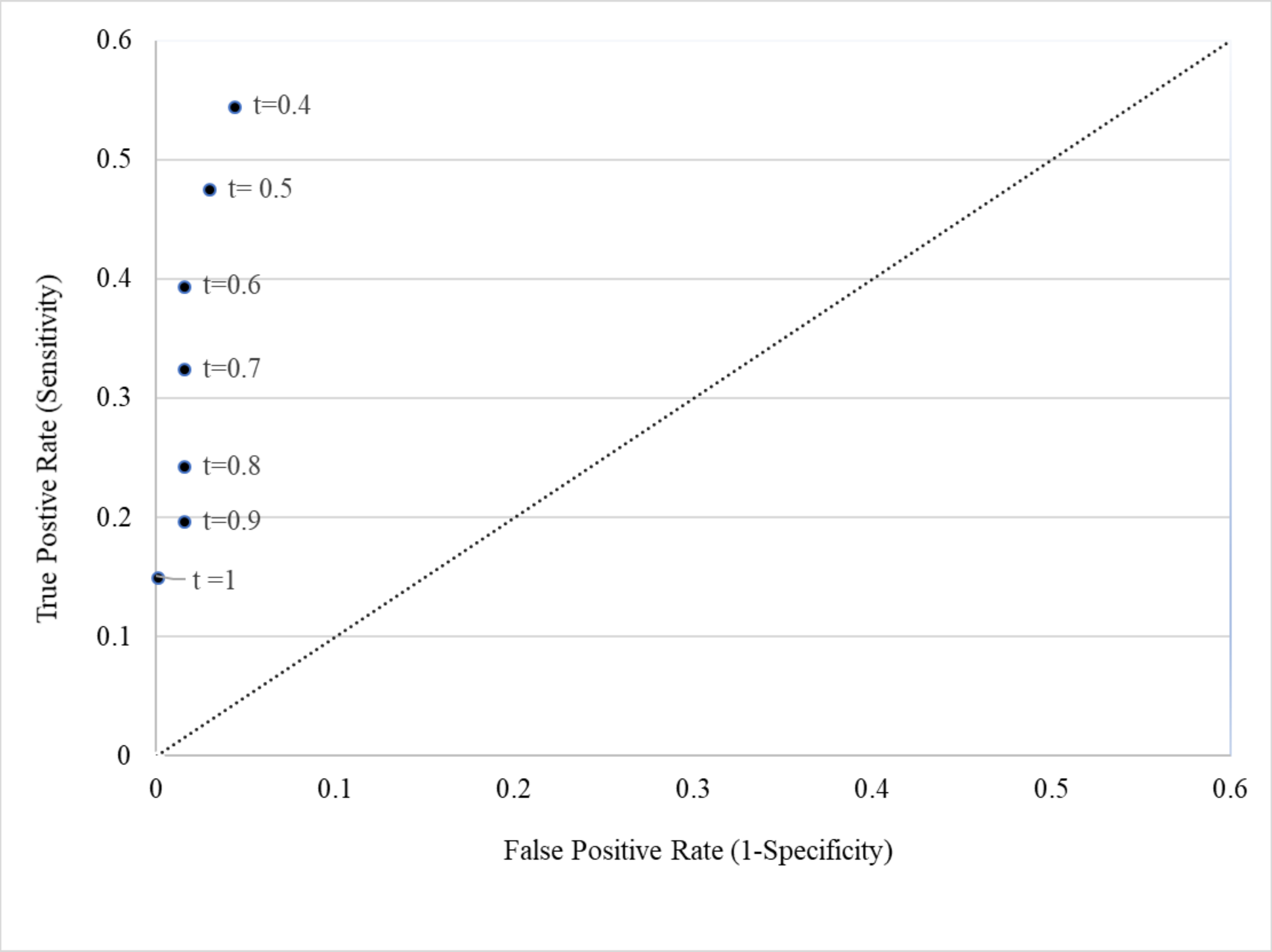


Figure 5.5: ROC graph showing the effect the threshold will have on the false positive rate and the true positive rate.

threshold is anywhere between 0.6-0.8, with the system becoming highly unusable with thresholds that are above that limit.

In our study, the pairs only had 5 mandatory guesses for each account, we did not make it 10 mandatory guesses to match this framework criteria due to time constraints. However, the Levenshtein Distance for the five guess attempts for each account were often too far below the threshold for that to matter. The Levenshtein distance average of all guesses was around 0.02.

5.3.2 Usability

Memorywise-Effortless

In order for an authentication system to be considered *Memorywise-Effortless* it must not require a user to remember any secret at all [18]. A system is rewarded with the status of being *Quasi-Memorywise-Effortless* if there is only one secret to remember for all verifiers [18].

Based on the parameters laid out above, GeoPassHints is also not *Memorywise-Effortless.* Technically, the users could have one location and one secret note across all verifiers; however, that is not recommended for the same reason it is not recommended for passwords due to the avalanche effect, which in this scenario means a single leak or compromise causing the multiple accounts and devices to be compromised.

Scalable- for-Users

An authentication system is considered *Scalable for Users* if using the authentication system for many accounts does not put any extra burden on the user [18]. For example, if a user is asked to set multiple passwords across different verifiers, then as the number of accounts increase, the burden on the user is increasing.

GeoPassHints is not *Scalable for Users* because as the number of accounts increase, the user has to spend time registering authentication credentials. In addition, with the assumption that the users will (hopefully) utilize different locations and different secret notes for each account, the memory burden increases on the user

Nothing-to-Carry

This category is awarded to authentication systems that do not require the user to carry anything to be able to authenticate. Some systems require a physical card to be present for authentication, these systems would not be granted the *Nothing to Carry* accolade [18]. Since GeoPassHints only requires the memorability of the secret note and does not require any physical object, it would be considered a system with *Nothing to Carry*.

Physically-Effortless

This category is granted to authentication systems that do not require the user to perform any action past the pressing of a button (to indicate authentication is occurring i.e., submit) [18].

GeoPassHints requires the user to register the authentication credentials and requires the user to type and click submit in order to complete an authentication sequence correctly. Hence it is not *Physically Effortless*.

Easy-to-Learn

In order to determine whether or not GeoPassHints is easy to learn how to use, and easy to use, we followed up the usage of GeoPassHints with two usability questions: 1-I thought the system [GeoPassHints] was easy to use, 2- I thought the system [GeoPassHints] was easy to learn how to use. The users were asked to indicate the extent in which they agreed with those statements on a scale of 1 to 7. Figure 5.6 showcases the results.

76% of our participants agreed, strongly agreed, or very strongly agreed that GeoPassHints was easy to use. 11% of our participants were neutral and only 13% either disagreed or strongly disagreed with the statement. As for the second question regarding ease of learning, GeoPassHints did even better with 87% of participants agreeing, strongly agreeing, or very strongly agreeing that GeoPassHints was easy to learn how to use. 11% of our participants were neutral and 3% disagreed. Based on those results we grant GeoPassHints the accolade of being easy to learn how to use and *Easy to Learn*.

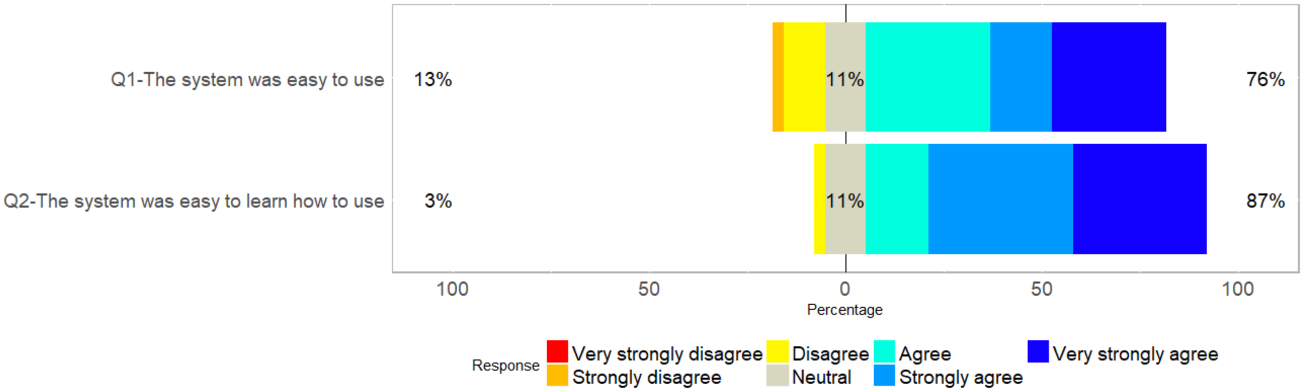


Figure 5.6: GeoPassHints usability Likert Scale questions.

Efficient-to-Use

In order to an authentication system to be considered *Efficient to Use* the time spent for each authentication must be acceptably short [18], the time required for a user to set up his/her credentials with a verifier should also be reasonable [18]. Bonneau et al. do not specify empirical values for this category which is appropriate because of the variety of authentication systems and the variety of intended use. For example, a smartphone would be more secure with an 8-character password; however, it is unusable for that target environment because a user typically wants to access the smartphone quickly.

GeoPassHints requires the user to register the location and secret note. Figure 5.7 shows the registration time for each account, in order of registration, in the user study. One thing automatically pops out; unlike the expected gradual decrease in registration time as people became more familiar with the interface, GeoPassHints actually had a gradual increase in the registration time for each account.

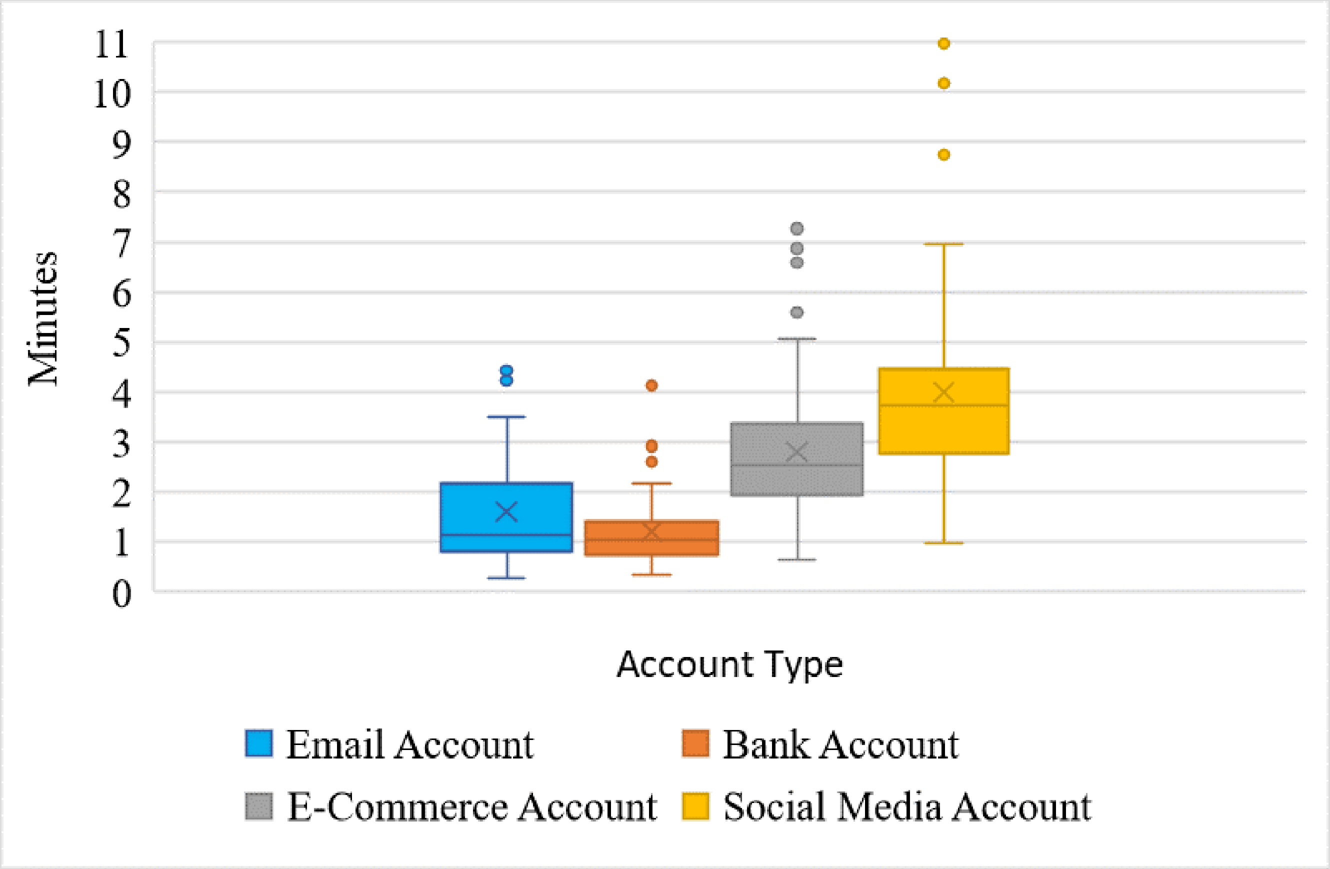


Figure 5.7: Boxplot for setting and confirming account credentials.

The average credential set time (set + confirm) across all four accounts (n=38) was 2.39 minutes. That is considered inefficient when compared to equivalent geographical authentication schemes such as GeoPassNotes which had a reported 31 second set time on average [39] [48], and a 24 second confirm time on average (total = 55 seconds).

As for the average login time, GeoPassHints does not perform very well since the average number of attempts until a successful login was 3.26 (std= ± 0.52). Figure 5.8 showcases the average login time for successful attempts and Figure 5.9 showcases the average login time for all the attempts until successful authentication was achieved.

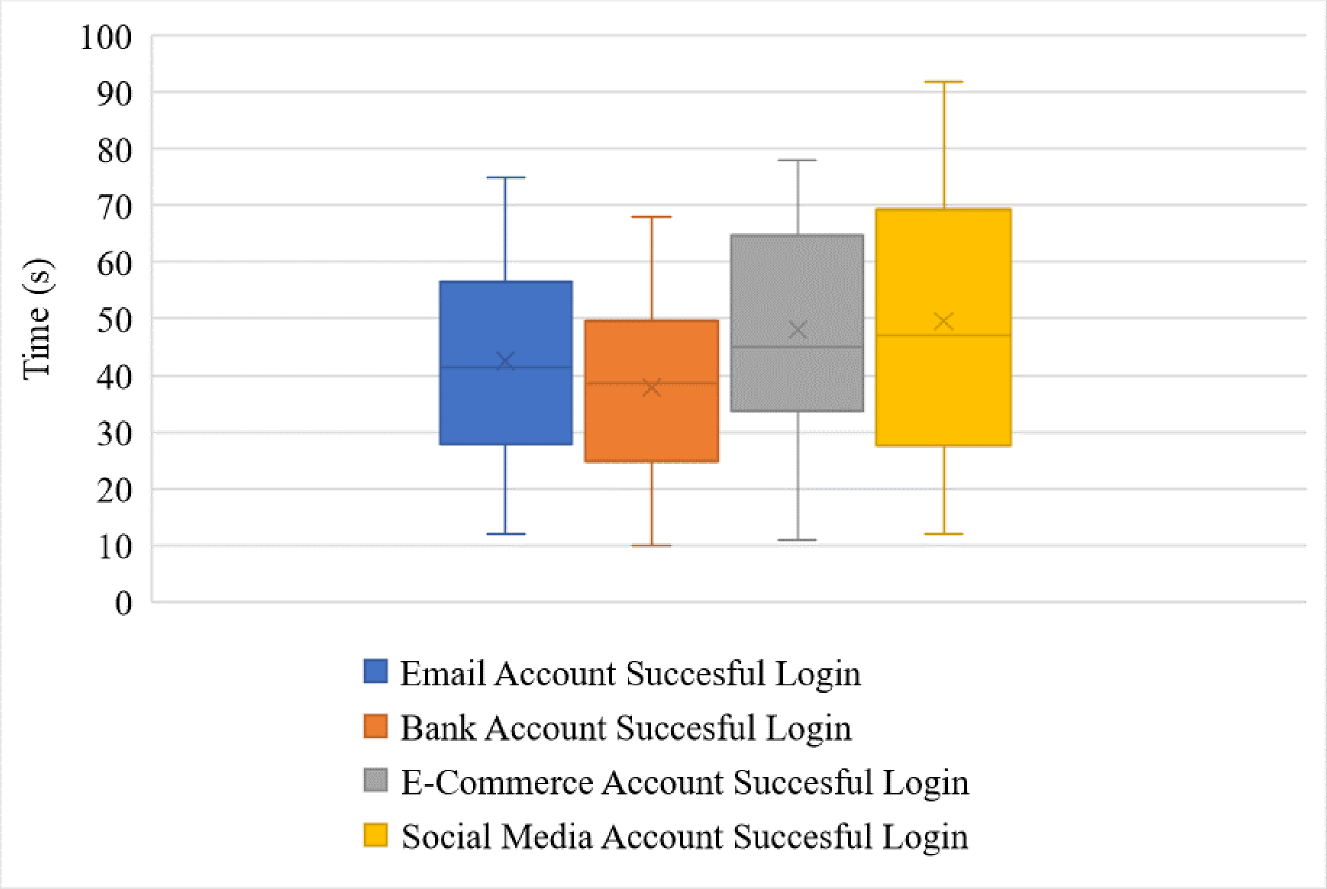


Figure 5.8: Successful login attempt time.

As shown in Figure 5.8 and 5.9, on average there were significant delays in successful logins as a result of failed attempts. The average login time on a successful attempt for email accounts,

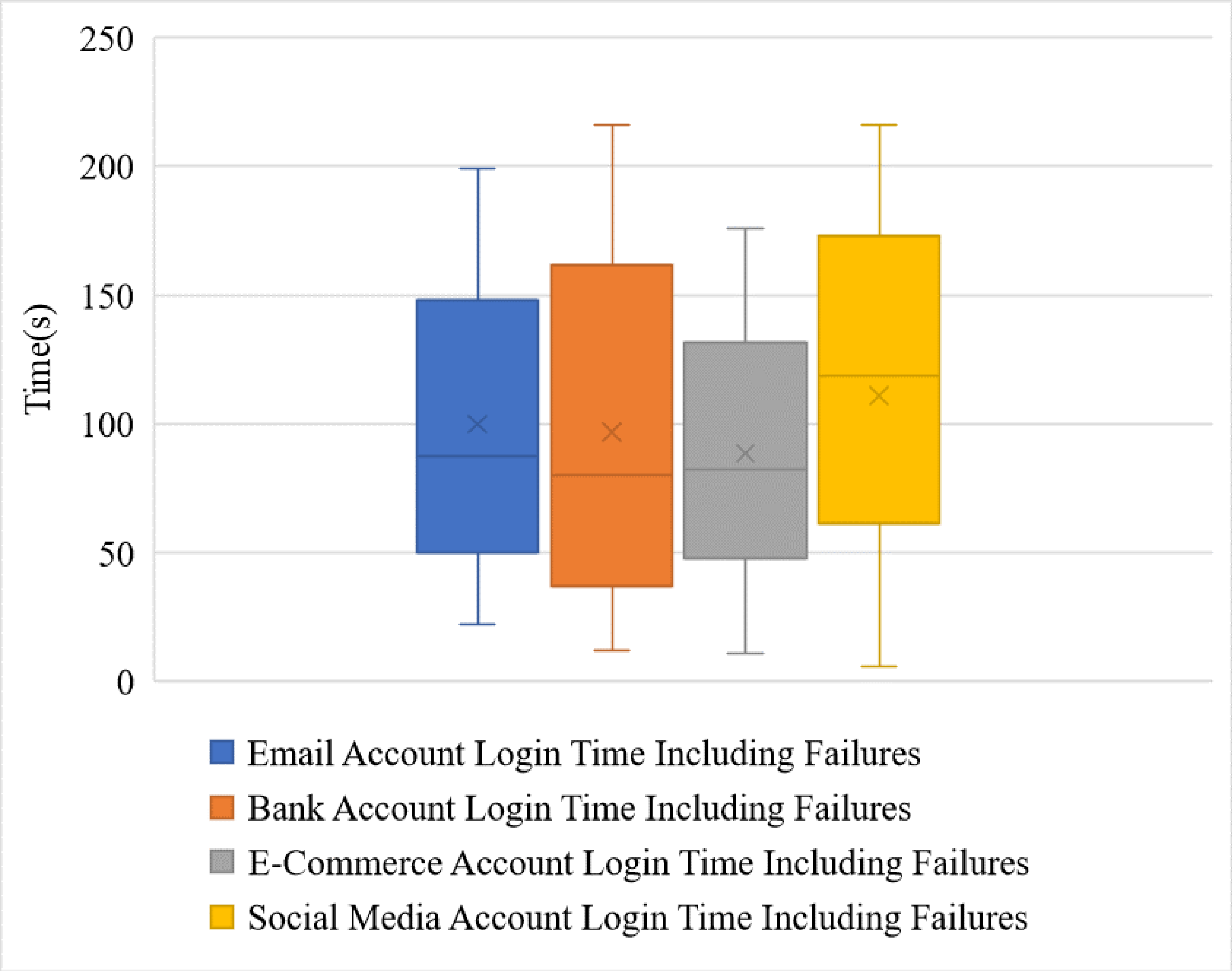


Figure 5.9: Login time including failures

bank accounts, e-commerce accounts, and social media accounts were 42.6s, 37.9s, 48s, and 49.6s respectively. If the unsuccessful attempts are included, then the times shoot up to 100s, 96.9s, 88.5s, and 111s respectively. To compare it to a similar geographic authentication system GeoPassNotes [39] contained far fewer failures only 4 participants in Session 2 of the study conducted by MacRae et al. (n=35) failed to login on their first attempt, all participants in that study eventually managed to login. The average login time in Session 2 of that study which was 7-8 days after Session 1 was around 30s which is much lower than GeoPassHints. We conduct further analysis on this matter in the *Infrequent-Errors* Section. Because of the high login times GeoPassHints will not be classified as *Efficient to Use*.

Infrequent-Errors

An authentication system may be awarded in the *Infrequent Errors* category if the login task typically succeeds when performed by the true user [18].

GeoPassHints is also prone to user input error despite having an error margin of 20% utilizing Levenshtein's Distance. Participants had an average of 3.26 (std= ±0.52) failed attempts until a successful login. Only 9% (3/34) of participants managed to remember all four hypothetical account secret notes, one of which utilized the same secret note for each account. Figure 5.10 showcases the correct/incorrect responses for all participants for each account type, and Figure 5.11 showcases the average for each participant for every account type. A detailed secret note analysis in Section 5.4 will go into detail regarding an in-depth analysis by categorizing each failure (or successful) attempt.

At this error threshold (0.8 Levenshtein's Distance), we expected that more participants would be able to login successfully and without that many errors. The long character length that the participants set tended to make the secret note harder to remember, even with the error margin allotted. However, we did observe that the true positive rate increased as the threshold was decreased, which means the participants were at times close. Figure 5.5 showcases how the true

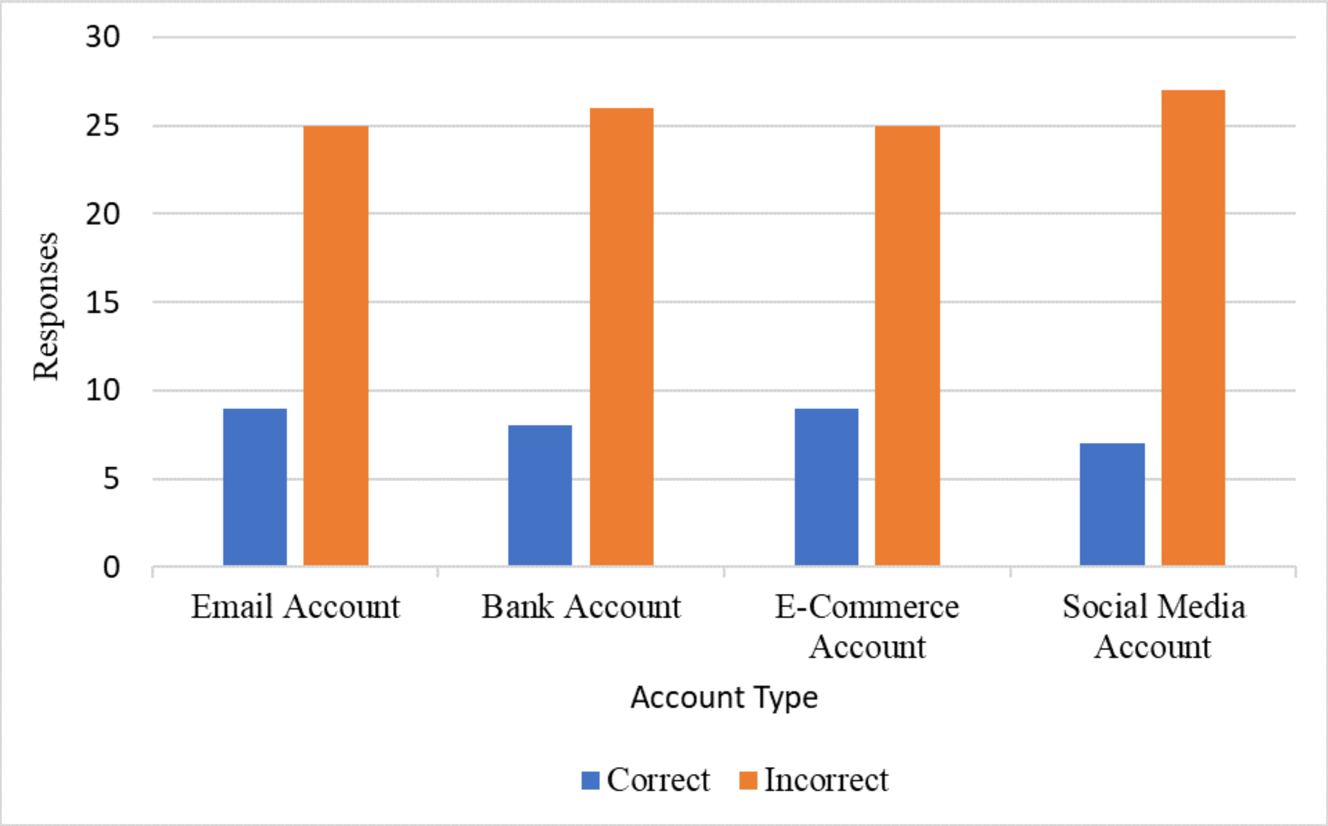


Figure 5.10: Comparison of correct and incorrect responses for all users across all four different account types (n=34).

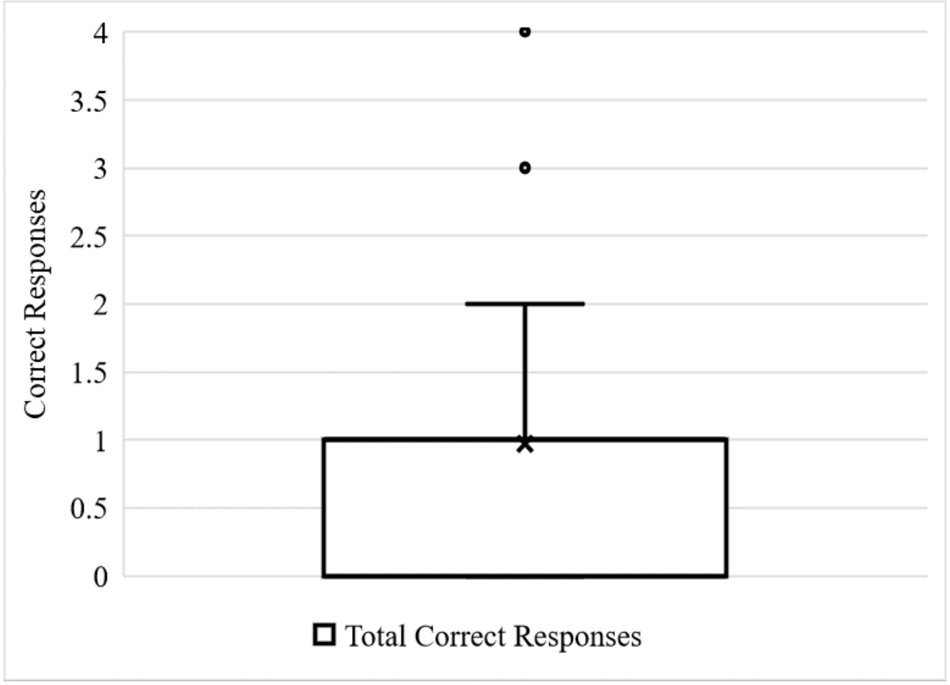


Figure 5.11: Boxplot of the total correct responses per participant for all four accounts. Participants received five attempts before having to reset that account.

positive rate increases steeply to 0.39 when the error margin is set to 0.6 instead of 0.8, without increasing the false positive rate. This is a promising feature that could be investigated further; however, any improvements to the system need to go beyond this one thing as a true positive rate of 0.39 is too low.

Overall, due to the fact that at a threshold of 0.8 this authentication system was not sufficient in increasing the true positive rate to usable levels we do not award GeoPassHints as having *Infrequent Errors.*

Easy-Recovery-from-Loss

This category rewards authentication systems that make it easy to recover from loss. it is very easy to recover from loss in the GeoPassHints authentication system. In our study, users could simply reset the location and the associated secret note. In terms of reset time, the reset + reset confirmation time was much shorter (average = 1m:15s std = ±20s) than the time it took participants to set and confirm their credentials in the first place (refer to Figure 5.7). This is probably due to the fact that at this stage, they no longer had to ask questions and they remembered how to use the system from Session 1.

As a result of the ease of reset, the recovery from loss in all scenarios should not be time consuming, hence GeoPassHints is considered to be easy to recover from loss.

5.3.3 Participant Preferences, Attitudes, and Concerns

This section will encompass other usability dimensions for GeoPassHints that are not part for the Bonneau et al. Framework [18]. At the end of Session 2, we asked participants to fill out an exit survey for GeoPassHints; which asked usability, preference, and concern questions in order to determine how the participants viewed the GeoPassHints system. All of the questions were Likert Scale questions ranging from 1 to 7.

We began by asking participants if GeoPassHints was easy to use, and if GeoPassHints was easy to learn how to use. The results are showcased in Figure 5.6, most participants indicated that GeoPassHints was easy to use and easy to learn how to use. After asking the participants whether or not the system was easy to use and easy to learn how to use, we asked the participants whether or not they wrote anything down to assist them in recalling the secret note. Only one participant indicated keeping a record of one secret note.

Followed by that we asked: For a mobile device's forgotten password/PIN or failed biometric, to what extent would you prefer to use this system [GeoPassHints] in lieu of any of the following fallback authentication methods? The fallback authentication methods were security questions, Email resets, and SMS resets. Figure 5.12 showcases the results.

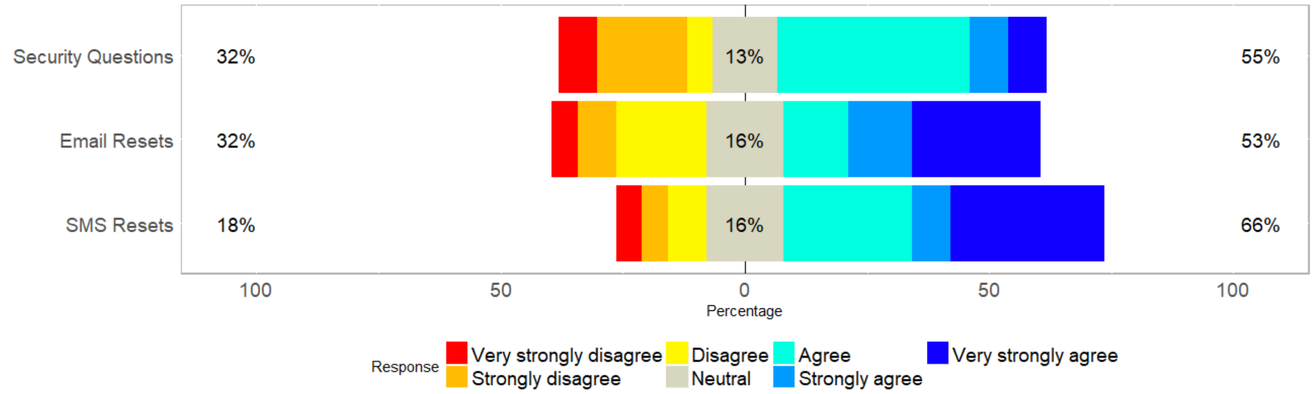


Figure 5.12: For a mobile device's forgotten password/PIN or failed biometric, to what extent would you prefer to use this system (GeoPassHints) in lieu of any of the following fallback authentication methods?

GeoPassHints seems to be appealing to the participants, the majority of the participants appeared to favor GeoPassHints in lieu of Security Questions, Email Resets, and SMS Resets, when it comes to fallback authentication for mobile devices, as shown in Figure 5.12.

After asking whether or not participants would prefer to utilize GeoPassHints in lieu of the commonly utilized fallback authentication methods we asked: For an online account's forgotten password, to what extent would you prefer to use this system in lieu of any of the following fallback authentication methods? Figure 5.13 showcases the results.

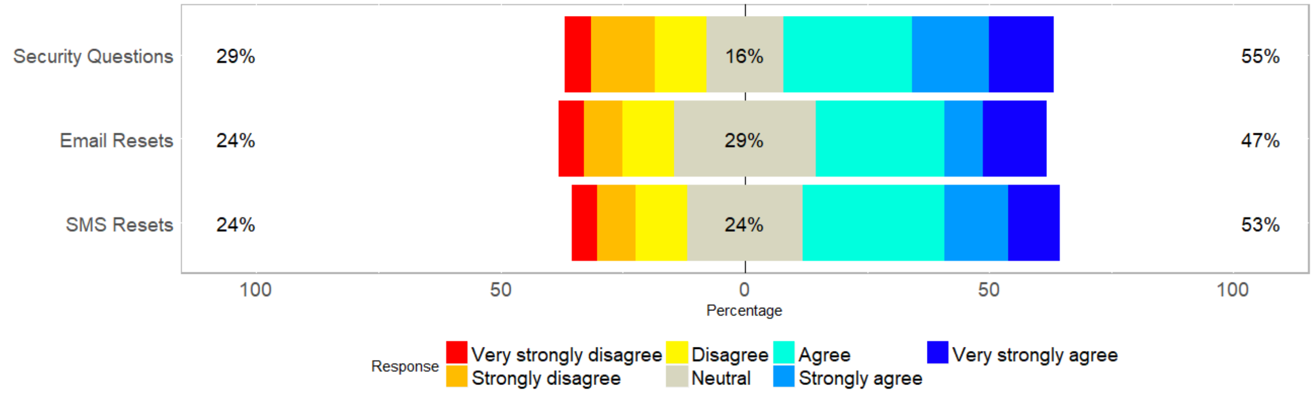


Figure 5.13: For an online account's forgotten password, to what extent would you prefer to use this system in lieu of any of the following fallback authentication methods?

The participants also appeared to respond positively to utilizing GeoPassHints instead of common fallback authentication mechanisms for the recovery of online accounts as shown in Figure 5.13.

Next, we move on to the preferences, we asked the participants whether not they prefer to utilize GeoPassHints over a password for a series of different account types as shown in Figure 5.14.

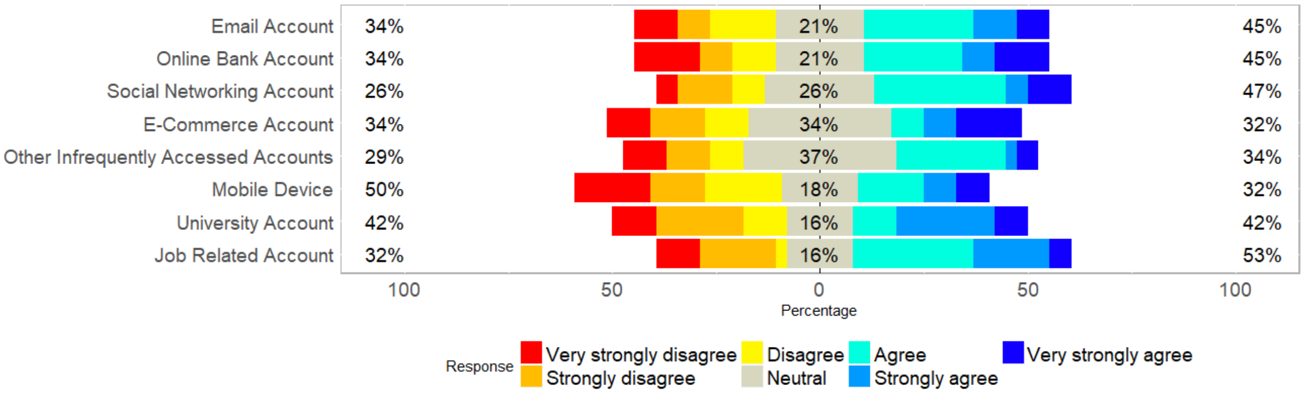


Figure 5.14: I would prefer to use the GeoPassHints system to login to any of the following accounts, instead of entering a password, for these account types.

Figure 5.14 paints a diverse range of opinions regarding the utilization of GeoPassHints for different account types. We observe a great amount of neutral responses for the GeoPassHints. We suspect it is because GeoPassHints was greatly similar to entering a password. Overall, participants are more likely to prefer to utilize GeoPassHints over entering a password or be neutral regarding the issue.

Furthermore, we asked the participants two questions about their concerns regarding GeoPassHints. 1- If you use this system, how likely do you think it is that someone you know could access your device/accounts without consent? 2- How concerned are you about a privacy leak on any medium? Please note that for GeoPassHints we did not ask the participants about their concerns over a location service application leaking their information because it is not pertinent. The results are displayed in Figure 5.15.

Overall, participants were not concerned that utilizing GeoPassHints would in anyway lead to

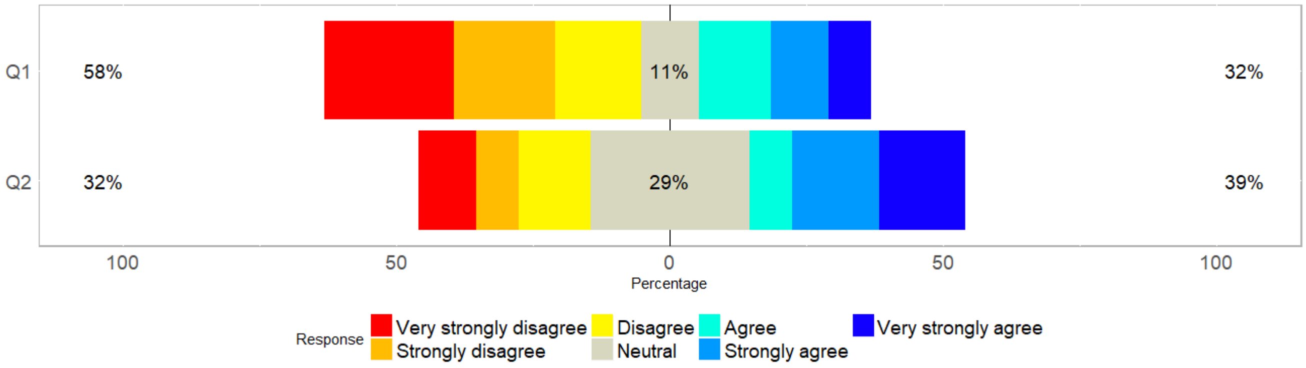


Figure 5.15: Participant security/privacy concerns regarding GeoPassHints. 1- If you use this system, how likely do you think it is that someone you know could access your device/accounts without consent? 2- How concerned are you about a privacy leak on any medium?

the known adversary being able to access their accounts/devices without consent. Only 32% of participants indicated any concern that GeoPassHints would could lead to the known adversary being able to access their accounts/devices. Participants did not seem too concerned regarding a privacy leak on any medium with only 32% of participants indicating that they were concerned about a privacy leak on any medium.

Lastly, we asked the participants to give us some feedback regarding the system, only two people gave negative regarding the responsiveness of the system (they felt like it was too slow in some instances), there are a variety of factors for that including the connection speed, implementation errors that lead to four participants not being able to load the location properly (the map was not rendering correctly for 5 participants, but it was resolved). In two of those instances, a reset was forced.

5.3.4 Deployability

Accessibility

We did some research regarding the accessibility tools provided by the Google Maps API specifically tailored towards browsers. We found that Google does provide accessibility features such as screen reader compatibility on browsers, the ability to explore a Google Map using a keyboard (for individuals with problems relating to motor control who cannot utilize a mouse), and keyboard shortcuts that can aid individuals with problems relating to motor control. These accessibility features are available on desktop/laptop browsers [1]. Hence, we award GeoPassHints in the *Accessibility* category.

Negligible-Cost-per-User

GeoPassHints is awarded in the *Negligible-Cost-per-User* category because implementing the system is possible with open and free to use APIs hence the burden on a verifier implementing such a system does not go up when the system is scaled to many users.

Server-Compatible

In order for a system to be awarded the accolade of being server compatible, verifiers must not have to significantly change their existing authentication settings, and the authentication system must be compatible with text-based passwords from the point of view of the verifier. GeoPassHints would require significant edits made to the backend to make both this authentication system work.

Browser-Compatible

The Google Maps API is what we used to develop the geographic portion of GeoPassHints. That API which renders a map and provides us with several different functions including setting a marker on a map is compatible with most modern browsers. Versions of popular browser vendors up to three years of age such as Safari, Google Chrome, and Firefox support every feature we have implemented according to our testing on BroswerStack and personal browsers. Therefore, GeoPassHints is *Browser-Compatible.*

Mature

GeoPassHints is not a mature authentication system, this category is gauged towards authentication systems that have already been implemented widely. GeoPassHints is simply an authentication system we designed and implemented for testing purposes that has not been tested for a long period of time on a wide scale.

Nonproprietary

GeoPassHints is not patented in anyway and the technology behind it (Google Maps API) is free and open to use with alternatives to this API being readily available (though not as good). Table 6.1 shows the full system evaluation according to the Bonneau et al. framework [18] in addition to our two added categories, and compares GeoPassHints to passwords, security questions, Email resets, and SMS resets.

5.4 Secret Note Recall Analysis

In order to further understand the usability issues generally, and more specifically the memorability issues, we conducted a semantic analysis on all the secret notes set and input during login by our participants. We eliminated some notes and recall attempts based on some criteria such as technical issues while attempting to remember the note, repetitiveness, or failure to adhere to the instructions. Participants could fail to adhere to instructions for a variety of reasons, for example. we provided instructions regarding labelling in the oral instructions portion of the study specifically in Session 1 where we said direct labelling is barred (e.g., setting Adam Scott High School as the location and labelling it as such), another common failure to adhere to instructions would be setting the same note for all four hypothetical accounts. Repetitiveness was not expressly barred during the instructions, however, any participant that utilized the same location/secret note combination for all four of their hypothetical accounts were excluded. Lastly, technical difficulties that

Table 5.1: Adapted Bonneau et al. framework [18]. Analysis of Security Questions [44], analysis of web passwords [19]. Please note that the two extra categories that we added were not evaluated in the original framework proposal [18].

	Memorywise-Effortless	Scalable for Users	Nothing to Carry	Physically Effortless	Easy to Learn	Efficient to Use	Infrequent Errors	Easy Recovery from Loss	Accessible	Negligible Cost per User	Server Compatible	Browser Compatible	Mature	Non-Proprietary	Res. to Physical Observation	Res. to Targeted Impersonation	Res. to Throttled Guessing	Res. to Unthrottled Guessing	Res. to Internal Observation	Res. to Leaks from Other Verifiers	Res. to Phishing	Res. to Theft	No Trusted Third Party	Requiring Explicit Consent	Unlinkable	Res. to Social Media Mining	Res. to the Known Adversary
Passwords (Baseline)			●		●	●	○	●	●	●	●	●	●	●		○						●	●	●	●	▲	▲
Security Questions	○ ⦀		●		●	●	○	●	●	●	≡	●	●	●	≡	≡				≡		●	●	●	● ≡	≡	≡
Email Resets	● ⦀	● ⦀	●		●	●	● ⦀	●	●	●	≡	●	●	●		○	● ⦀		⦀	≡	≡	●	≡	●	≡	△	△
SMS Resets	● ⦀	● ⦀	≡		●	●	● ⦀	●	●	●	≡	●	●	●		○	● ⦀	△	⦀	≡	≡	○ ≡	≡	●	≡	△	△
GeoPassHints	≡		●		●	≡	≡	●	●	●	≡	●	≡	●	⦀	≡		⦀		● ⦀	● ⦀	●	●	●	●	● ⦀	≡

● – offers the benefit, ○ – almost offers the benefit, no circle – does not offer the benefit, ▲ – potential to have the benefit , △ – potential to almost offer the benefit, ⦀ – better than passwords, ≡ – worse than passwords we excluded from the analysis were ones relating to not seeing the location. This is an implementation issue that was subsequently fixed. In the end we had 124 valid secret notes out of a total of 136 secret notes from a total of 16 pairs, only 17 pairs returned for Session 2 of our study. One participant was disqualified due to repetitiveness and direct labelling. Four hypothetical accounts (spread out across different participants were also excluded due to technical errors.

First we begin by analyzing and classifying the type of secret notes that the participants set. Before we delve into the breakdown it is important to explain the categories we sorted the secret notes set into and explain how we came up with them.

During the manual observation analysis of the notes set, we noticed that a great deal of them were labels of some sort (e.g., this is my home, a place where I attended a lot of concerts etc..). This lead to us noticing a large scale pattern that a great percentage of the notes revolved around labelling a place descriptively, directly, or at a higher level. Therefore we sorted the notes set by our participants into four categories.

The first category is a general direct label of some sort (e.g., this is my home). The second category is a descriptive label (e.g., first house in Canada). The third category is a direct naming. Lastly, the fourth category is a more general category encompassing secret notes set that did not fall within the general labelling trend, the fourth category is not a label or other. The results are displayed in Table 5.2 below showcases the results.

Category	Number of Notes
Descriptive Label	31 (25%)
Label	47 (37.9%)
Name	5 (4%)
Other	41 (33%)

Table 5.2: Breakdown of secret notes into categories.

Since we consider naming to be labelling 83 (67%) of the 124 secret notes set by our participants was some kind of labelling. However, the failure rates were very high as noted previously. As a result, we conducted a further analysis on each recall attempt in order to ascertain what might be the real issue. We would like to find out whether or not the labelling was an effective strategy to help the user remember the secret note.

In order to categorize the failures we created four general categories. The first category was memory loss, the second category was semantic similarity, and the third category was rewording, the last category was interference. Semantic similarity is a category that we created in order to categorize recall attempts that were similar in meaning to the set secret note but incorrect because it was not an 80% match to the set secret note (e.g., the set note being "this is my home" while the recall attempt was something relating to the home). Memory loss is a category for failed recall attempts that were not close to the original set note. Rewording is a category that encompasses recall attempts that were similar to the originally set secret note but were reworded in some way that led to a failure (e.g., "I saw the raptors play here" as the set secret note and the recall attempt being "This is where I saw the raptors play"). Lastly, we include a category for the interference effect which occurs as a result of not knowing which secret note belongs to an account, since they set four locations and associated secret notes (e.g., a user sets "This is my home" is a the secret note for their hypothetical bank account but instead of using that secret note for their hypothetical bank account they utilize that secret note for their hypothetical email account). This analysis includes all failed attempts until success or all failed attempts until 5 failure attempts at which point the participants were asked to reset the note. Table 5.3 showcases the results of our analysis.

Category	Number of Notes
Semantic Similarity	74 (26.8%)
Memory Loss	100 (36.2%)
Interference	21 (7.6%)
Rewording	81 (29.3%)

Table 5.3: Categorization of failed attempts.

The semantic similarity issues appeared to occur as a result of the participant remembering that this note was set as a result of a location they were associating with a memory. However, they could not remember exactly how they wrote the secret note in Session 1, as a result many attempts with a semantic similarity ended in failure as shown in the failure rates. Memory loss was the primary source of failed attempts. Please note that some of these failed attempts ended in success eventually. Labelling did present itself as a viable method of increasing memorability. Combined, the different types of labelling compose 66.9% of the total set secret notes (83/124). The participants that labelled did have a higher success rate than those who did not. Of those 83 notes we categorized as some sort of label, the eventual success rate (despite some failures in the beginning) is 74.6% (62/83). This compares well to the success rate of the secret notes that were not categorized as labels which had a 48.7% (20/41) success rate.

5.5 Resulting System and Policy Recommendations

Based on the findings from our user study we suggested several system and policy recommendations for GeoPassHints.

- Blacklisting certain keywords is necessary, to prevent the secret note from being descriptive.

During our analysis of the secret notes set by the participants, several trends were noticed such as some variation of "I live here" or "where I work". This is very dangerous because a known adversary with the right motivation is more likely to know current and previous work and home addresses. That in turn makes it easier for a known adversary to guess the secret note should this trend continue. Which leads to the necessity of blacklisting certain common secret notes that tend to be highly descriptive.

Another type of blacklisting that is necessary is to avoid directly labelling an institution/business that has been set as a location. That can easily be done by utilizing Google Places to get a list of businesses/institutions within a certain range of the location set. Verbatim labelling of a location within the secret note is very dangerous because it makes the secret notes highly vulnerable to a guessing attack that involves a list of places.

Blacklisting should improve how GeoPassHints fares in the *Resilient-to-Unthrottled-Guessing* criteria. In future work, it would be interesting to implement these blacklists and perform the secret note guessing analysis on notes created under both those criteria.

- Lowering the Levenshtein distance threshold for a successful login is important for usability.

Currently the threshold for a successful login utilizing GeoPassHints is a 0.8 string similarity utilizing the Levenshtein distance. Based on the ROC graph (Figure 5.5) the true positive rate increases significantly with a Levenshtein distance of 0.7 or even a Levenshtein distance of 0.6 without affecting the false positive rate. That means even if the Levenshtein distance threshold for a successful login was lowered all the way to 0.6, we did not observe a security implication, while the usability greatly increased.

- Setting the secret note, and confirming the secret note can occur on the same page.

In our system design of GeoPassHints the participant had to set the location and the secret note. After they clicked submit, the secret note had to be confirmed. However, in lieu of that configuration, we propose that the secret note set, and secret note confirmation occur on the same page much like how passwords are set and confirmed on the same page. We make this recommendation in an effort to decrease the time required to set GeoPassHints credentials which would in turn make it more usable.

- Set a policy for whitespace and punctuation.

During our analysis of the results, we noticed that some participants were confused whether they should include whitespace and punctuation because they usually do not do so with passwords. Secret notes are incredibly similar to passwords which causes that confusion. Therefore, it would be helpful to include a note while the users are setting their credentials that whitespaces and punctuation are okay.

- Encourage Association

As mentioned earlier labelling of any kind does improve the chance that a secret note will be successfully remembered 7-11 days after it was set. While direct labelling and direct naming should be discouraged for fear of a security compromise, it is important to associate the secret note and the selected location with a memory of some sort. Many labels analyzed were a type of association, and that yielded a higher success rate. Therefore, changing the instructions to include an emphasis on association with some sort of memory could enhance the memorability of the secret note to a great extent.

5.6 Discussion and Future Work

A perfect authentication system is one that balances security and usability. In order for an authentication system to bridge the security-usability chasm, it must offer quick credential setting and quick authentication paired with durable security to a whole variety of threats.

Unfortunately, GeoPassHints does not bridge the security-usability chasm. The login time, which stands at an average of 100 seconds (refer to Figure 5.8), appears to be too lengthy especially when counting in the number of failures until a successful login.

Given the lengthy login time, GeoPassHints would not be well placed in a primary authentication setting. However, as a fallback authentication system GeoPassHints would be comparable with security questions, Email resets, and SMS resets [17]. Fallback authentication occurs much less often than primary authentication [17, 34]. Therefore, it is more acceptable for an authentication system to take longer to authenticate if it provides a security improvement. In the case of GeoPassHints which provides many advantages over security questions, Email resets, and SMS resets, the login time which stands at an average of roughly 100 seconds including all the failed attempts is deemed acceptable.

Where GeoPassHints lacks is its success rate. Figure 5.5 showcases the ROC graph on how altering the Levenshtein distance threshold for a successful login will affect the true positive rate and the false positive rate. Even with the Levenshtein distance threshold set to be 0.6 (60% string similarity), the true positive rate is still very low standing at roughly 39%. That means even with that low threshold, 61% of participants would fail to login using GeoPassHints 7-11 days after setting that credential.

It is important to note that fallback authentication typically occurs over a much longer timespan after setting the fallback authentication credentials. The typical average time from fallback authentication credential setting to utilization is around 6 months [17]. Session 3 of our user study has not been conducted yet but based on the abysmal results from Session 2 of our user study in terms of GeoPassHints secret note memorability, we can assume that this 39% true positive rate even at a Levenshtein distance of 60% will drop. Hence, making GeoPassHints highly unmemorable for as a fallback authentication system.

To compare, security questions require you to remember a piece of information, much like GeoPassHints requires you to remember the secret note. Security questions succeed roughly 60% of the time [17] after a six-month interval, and even that is considered a low score for memorability.

According to our analysis of GeoPassHints secret notes and their resilience to throttled guessing, we found that the secret notes showed a great amount of resilience to all three password guessing methods that we utilized (Weir's probabilistic context-free grammar [54], a semantic cracker [51], John the Ripper (incremental mode) [3]). When compared to the resilience of passwords attained from popular password leaks, the secret notes set by the participants appear to offer much more resistance to password guessing attacks. This is an important security feature, but due to the high failure rate this system still cannot be recommended as a fallback authentication system.

We also observed an interesting anomaly when analyzing the data obtained from the user study.

Credential setting time tended to increase as the participants were going through different accounts. Figure 5.7 showcases how the credential set time tended to increase. We designed the user study so that the participants would have to set credentials for four different accounts to test for multiple password interference [11, 22]. During our physical observation throughout Session 1, we noticed that participants did take their time deciding which location to select; as they progressed through the accounts they had to set credentials for, they were running out of places to set their location markers.

As part of our multiple password interference analysis, we observed that only 8.8% (3/34) participants input a secret note for one account that was meant for another account (e.g., they input their hypothetical Email account secret note in their hypothetical banking account secret note). All three of those participants input the secret note incorrectly for all 5 attempts. One participant who made that mistake had input the same secret note for more than one hypothetical account. We suspect that this low multiple password interference effect is due to the location cue provided, which helped guide their secret note input.

In future work, it would be interesting to implement a blacklist such as the one recommended in Section 7.4, and attempt the same security and usability analysis. We suspect that the security should improve. In addition to that, it would be interesting to observe the effect on the credential setting time for GeoPassHints should the setting and confirmation of the secret note appear on the same page. Lowering the credential setting time will improve the overall usability of GeoPassHints.

5.7 Conclusion

In conclusion, we designed, implemented, and tested a new geographic authentication system called GeoPassHints. Our security, usability, and deloyability analysis lead us to the conclusion that GeoPassHints is not particularly usable in a primary authentication environment or a fallback authentication environment. While the attained average login time is acceptable in a fallback authentication environment, it is extremely long for use in a primary authentication environment (i.e., password replacement).

In addition to that, while GeoPassHints did provide some benefits in relation to security especially when compared to other commonly utilized methods for fallback authentication (see Table 9 for a comparison), it proved very poor in terms of usability. The long login time might be acceptable in a fallback authentication environment, but the failure rate of 61% after 7-11 days of setting is too high . The increased security but poor usability ultimately means that GeoPassHints does not bridge the security-usability chasm that exists in many authentication systems. As a result, we do not recommend GeoPassHints as a replacement to the currently used fallback authentication mechanisms. Certain improvements to the system might lead to better results, but as it stands GeoPassHints would not fare well on a larger scale implementation by a popular verifier.

Chapter 6

User Study

We evaluated GeoPassHints and GeoSQ for security and usability, in addition to the viability of our known adversary frameworks through a 38-participant (19 pairs) user study that was approved by our university's Research Ethics Board. Before the studies proceeded both of our systems (GeoPassHins, and GeoSQ) were pilot tested by 6 colleagues, friends, and family. Three pilot testers were experienced computer users with degrees in computing science/networking, the remaining three were casual computer users. Our pilot testing allowed us to discover usability flaws within our system, in addition to discovering any bugs or development related issues that could hinder our user study.

Participants for this study were recruited using a broadcast email in addition to posters around campus. Participation was limited to students, visitors, and staff of our university. In order to participate, all participants must have met the following criteria:

1. 18 years of age or above.

2. Participants must bring a pair.

3. Participants must have an Android smartphone.

4. Participants must be willing and able to turn on location services throughout the week.

6.1 Study Sessions

Our user study contained three sessions spanning 6 months and 2 weeks. Session 3, was scheduled to be 6 months after Session 2, however due to the abysmal memorability and due to the fact that GeoPassHints did not meet the minimum viability for memorability we elected to not have a long term memorability study. The pairs completed the exact same steps; we did not have a main participant and a pair, both pairs performed all parts of this study.

6.1.1 Session 1

Session 1 was an in-lab session that lasted an average of 35 minutes. Each participant was compensated $8 for their participation. We held Session 1 on several different dates, and different time slots throughout the day. Each time slot had a maximum of 2 pairs (4 participants). At the beginning of each iteration of Session 1 we asked pairs to sit across from each other to avoid any type of contamination of the results (e.g., participant seeing their pair's secret note or feeling pressured to answer questions about their relationship differently). Followed by that we read the instructions to the participants from a prewritten script in order to ensure that there was no variation in the instructions between the iterations that could affect our results. We also did a practice run of the system to ensure that the participants knew how to use our systems.

After the instructions portion of Session 1 was complete, the participants were then asked to open their laptops and login to the GeoPassHints system. After reading the consent form and agreeing to it, they proceeded to complete the entry survey. The entry survey was a mix of background questions such as age, gender, academic background, as well as the We Scale, IoS Scale, and the adapted RCI scale (see Appendix A, B, and C). After the participants completed the entry survey they would then set a location and a secret note corresponding to four hypothetical accounts.

The first hypothetical account they enroll a location and a secret note for was their Email account, the second hypothetical account they would enroll their credentials for was their bank account, the third and fourth hypothetical accounts were e-commerce accounts and social media accounts respectively.

By asking the participants to set their credentials for more than one hypothetical account we aimed to investigate the effect of multiple password interference [11]. We also wanted to observe whether or not participants would set their secret note differently based on account type. We cued the participants as to the account type with a logo representing a major service provider, for example an Email account would have a Gmail logo because that is a popular Email service provider (it is the official Email provider for our university).

After each location and secret note enrolled (for each account type) the participants are asked to confirm their secret note. The location is given to them during the confirmation, but they must have an exact character match for their secret note to confirm enrollment of their credentials. After all four hypothetical accounts were enrolled, they have completed the GeoPassHints portion of Session 1.

At that time, we would instruct participants to take out their Android Smartphones and download the application with the previously emailed download link. After they install the GeoSQ application they are asked to read the instructions relating to the app, enter their assigned ID for the study, and accept the terms and conditions. Lastly, they would be reminded to keep location services on and we reminded them that GeoSQ is logging location information locally in the background. We also informed them that location services could be turned off at any time and turned back on at any time if they did not feel comfortable with their locations being logged at a certain time. That is the end of Session 1.

6.1.2 Session 2

Session 2 was scheduled 7-11 days after Session 1 in the lab. Session 2 lasted an average of 45 minutes and participants were compensated $10 for their participation. In addition to that the participating pairs were entered into a $100 draw if they were both present for Session 2.

The purpose of Session 2 being scheduled 7-11 days after Session 1 was to assess memorability as well as allowing time for the GeoSQ application to log location information (the requirement was ten unique locations before the application unlocks). Session 2 also started with the instructions being read off a prewritten script to avoid variations in the instructions having an effect on the study results. We also demonstrated the systems in Session 2 to ensure all participants knew how to use both our systems and what was expected of them.

After the participants were given time to read the consent form, the participants were then asked to login to the GeoPassHints system again. They would then begin by trying to remember the credentials they set 7-11 days ago. They were given the location as a cue and they just had to remember the secret note associated with that location. For each account type (hypothetical email, bank, e-commerce, and social account) the participants were given five tries. If they are unsuccessful after five tries they were asked to reset that account's login credentials.

After the participants completed the recall portion of Session 2 for GeoPassHints, they were asked to guess their pair's secret note (again given the location on a Google Map as a cue). The pairs had five guess attempts for each hypothetical account type (hypothetical email, bank, e-commerce, and social media). However, the pairs were not told if they guessed the secret note correctly or incorrectly. The pairs were simply given five attempts and were not given any feedback to preserve the privacy of their pair.

Following the guessing portion of Session 2 for GeoPassHints, the pairs were then asked usability questions regarding GeoPassHints. At that point we were done with GeoPassHints for Session 2 and we switched the participants' attention to GeoSQ (the smartphone-based autobiographical authentication system).

The participants were prompted with ten questions regarding their whereabouts throughout the previous 7-11 days. After that was complete the pairs were asked to switch phones, at this time the participants were asked to attempt to guess their pairs' location questions throughout previous 7-11 days, these were the exact same questions asked of their pair.

The participants were asked to switch phones because the autobiographical location data was logged locally, and to further protect the participant's privacy we did not want to relay that information on a network of any kind to avoid a leak scenario. They were asked to sit next to each other to avoid any kind of discomfort from having their smartphone with their pair without the ability to see what their pair was doing with it. We also kindly requested that the participants not communicate at all during this period to avoid the contamination of results.

During the location recall and location guessing phases of Session 2 the users were actively encouraged to use the Internet for research. Some sources such as their Google Timelines or transportation system cards can give them information regarding their whereabouts at a certain point in time (we observed those two sources being used).

After the pairs had completed guessing their pair's locations throughout the past 7-11 days they would return each other's smartphones and they would begin the exit survey which is composed of similar usability questions that were asked after the GeoPassHints system. However, this time the usability questions were obviously geared towards GeoSQ. The exit survey marked the termination of Session 2.

6.2 Participants

Our recruited participants were all undergraduate students between the ages of 18-30. The average age of our participants was 21.3. Out of our 38 initial participants, 13 were female (34.2%) and 25 were male (65.7%). As mentioned earlier, in order to meet the criteria for participation our participants had to be a minimum of 18 years of age (age of majority in our jurisdiction), and they were required to come in pairs. Each participant had to have his/her own Android smartphone (Android 3 or above). 15 of our participants (39%) had taken some computer security/IT course before.

6.3 Study Environment

The first two sessions of our user study were conducted in person. Participants were required to bring their own smartphones and they were also required to bring their own laptops to be able to participate in this study. The sessions were held in a secluded room, and a maximum of two pairs were scheduled for each iteration of Session 1 and Session 2. This environment was carefully designed to allow us to ensure participant comfort, and ensure no communication was occurring between participant pairs.

6.4 Limitations and Ecological Validity

Our initial two sessions are completed in person in a lab environment, as opposed to being conducted online on a platform such as Amazon's Mechanic Turk. This is a limitation because our data obtained is from a small set of participants from a similar demographic (relatively similar age, university students). The user study was conducted in this fashion to allow us to ensure participant comfort, and ensure no communication was occurring between participant pairs. In addition to that, it would be difficult to know if the online participants were truly pairs.

We recognize the fact that GeoPassHints and GeoSQ would be utilized by a more diverse demographic in a real-world scenario. We also recognize the fact that different patterns in the secret notes being set, and different patterns in answers to GeoSQ's autobiographical location questions might emerge given a more diverse demographic. However, this is the method currently used for studies that evaluate the known adversary, for the reasons mentioned above. A larger online study (with a different user study design) to determine the usability and security of our systems is not in order until results from a study involving a smaller set of participants such as ours has already provided us with some preliminary results.

Another limitation is presented to us by the "lab effect" [27]. Conducting studies in a lab environment is proven to have a significant impact on the results obtained. For example, it was shown that participants in formal lab environments typically act and perform a certain way that they would not in their everyday lives. Additional patterns might be observed should this user study be conducted online without a researcher present; this could include a higher occurrence of reused notes. Throughout the lab, participants were being given instructions that aided in their utilization of the system. This does not mimic a real-world environment.

Moreover, the time for each Session was long and a financial incentive was presented. That could also be another factor related to the "lab effect" that challenges the real world ecological validity. Participants might be pressured to complete the study quicker in order to be financially rewarded quicker.

In addition to that, we cannot control what the pairs communicate to each other after the study, they could let each other know their locations/secret notes anytime throughout the week.

In addition to that, there was no reward for guessing locations correctly or guessing secret notes correctly; in fact, the participants never knew those scores (for privacy reasons). Therefore, participants did not have a real incentive to guess correctly (they were compensated no matter how well they guessed). This does not imitate a motivated known adversary in a real-world environment that might be more motivated to gain access to an account or a device.

Furthermore, we would like to comment on the order of systems in which the study was completed. As mentioned above, GeoPassHints was always the first system that participants interacted with. In hindsight that could have created a usability issue due to the fact that the study sessions were quite long. When participants were completing the GeoPassHints exist survey they were typically twenty minutes into the second session. However, when they were completing the GeoSQ exit survey they were typically 35 to 40 minutes into their second session of the study. In hindsight the order in which GeoPassHints and GeoSQ were tested in the study sessions should have been switched. We cannot say anything for certain because there is no way to measure this effect after the fact, but it could have had a negative impact on the usability and results of GeoSQ due to the long study sessions.

Lastly, the exit survey for each GeoPassHints and GeoSQ were highly similar (almost identical questions). However, there are more effective tools to gauge usability. We designed the exit survey questions from scratch, and we modelled these questions using a comparative approach were we asked whether or not they would prefer to use this system over another, typically in the same domain (fallback authentication in the case of GeoSQ and fallback or primary authentication in the case of GeoPassHints). In the exit survey we also included questions to gauge participant concerns regarding privacy leaks as a result of utilizing GeoPassHints and GeoSQ. In hindsight we could have asked those questions at the end of Session 1 to decrease the length of the exit survey (the longer the session the more we could observe that the participants just wanted to leave).

The diagram below illustrates the exact sequence of events that occurred during the Sessions.

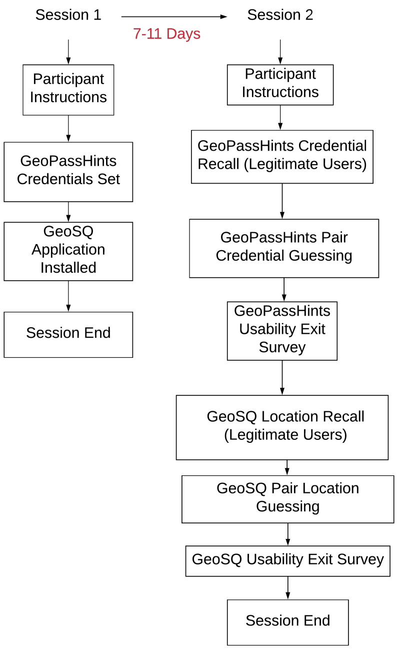

Session 1
Session 2
7-11 Days
Participant Instructions
GeoPassHints Credentials Set
GeoSQ Application Installed
Session End
Participant Instructions
GeoPassHints Credential Recall (Legitimate Users)
GeoPassHints Pair Credential Guessing
GeoPassHints Usability Exit Survey
GeoSQ Location Recall (Legitimate Users)
GeoSQ Pair Location Guessing
GeoSQ Usability Exit Survey
Session End

Figure 6.1: Sequence of events during Session 1 and Session 2.

Chapter 7

Discussion

7.1 Interpreting Results from the Known Adversary Models

In this work, we created three models, all of which proved better at predicting the extent to which a participant would be able to guess his/her pair's location question from the past 7-11 days than simply relying on self-reported relationship characterization. Two of the models we used to quantify the known adversary (the KAI and the adapted RCI) proved to be highly correlated with the GeoSQ pair guessing scores, indicating their promise.

Our primary goal was to quantify the known adversary more accurately than self-reported relationship closeness as used in existing security research. However, the general utilization of the KAI and the adapted RCI comes with some reservations. Both the KAI and the adapted RCI are heavily geared towards the concept of physical proximity, due to the definition of the known adversary which is any individual with elevated knowledge of a potential victim and/or elevated access to a potential victim's devices. Since GeoSQ is an autobiographical location authentication system, and the KAI and the adapted RCI have many questions regarding physical proximity, it is possible that high scores from the KAI and the adapted RCI could be better suited to autobiographical data for authentication than other systems. Since the KAI and the adapted RCI were only tested against GeoSQ, in future work it will be important to use them when performing user studies with different authentication systems as further proof of concept.

Another important question to address is the known adversary classification point. In other words, at what point do we classify an attacker to be a known adversary based on the results in the KAI or the adapted RCI? Based on our results, any participant score above 40 points typically yielded great guessing results. However, we caution that this score should not be used as a general cutoff point. Consider for example, given a different authentication system, participants with scores above 30 on the adapted RCI or the KAI could have great guessing scores. Based on that particular analysis, the cut off for the known adversary should be 30. Hence, we do not recommend a generalized cut off score, instead we recommend that security researchers evaluate whether a system is *Vulnerable-to-the-Known-Adversary*, by detecting a correlation between the guessing scores and the model scores. Should a high correlation exist then their authentication system would be declared vulnerable to the known adversary, if the guessing scores lack a correlation with the model scores then the authentication system is deemed not vulnerable to the known adversary.

Certain authentication systems will not yield a significant correlation with the KAI or the adapted RCI scores because there is a low probability that the authentication credential can be guessed. For example, GeoPassHints guessing scores would naturally not lead to a high correlation with the KAI score or the adapted RCI score because so few of them were guessed successfully. Certain authentication systems will have a very low probability of being guessed, and GeoPassHints was one of them, that is why no correlation analysis was conducted with GeoPassHints and the KAI or the adapted RCI. Only 2.9% (4/136) secret notes were guessed therefore all the correlation tests we ran were negative.

Furthermore, another important point to discuss is what are we truly measuring. In our research we began by investigating relationship closeness. However, as noted earlier, the known adversary does not necessarily have to be an individual that we consider to be socially close. The known adversary can be a co-worker, physical proximity to a co-worker can be more pertinent than physical proximity to a spouse in some cases. That is what motivated the design of the adapted RCI and the KAI. If we were simply measuring relationship closeness, we would have simply utilized the original RCI or the Oneness score, but our objective was different. When designing the adapted RCI and the KAI, we kept in mind that the known adversary does not necessarily have to be socially close. That is why they are not geared towards measuring relationship closeness, instead they are geared towards measuring factors such as proximity and influence which are indicators of an individual with elevated access to a potential victim's devices and/or elevated knowledge of a potential victim.

Lastly, the KAI and the adapted RCI scores obtained highly similar correlations to the GeoSQ pair guessing scores. While the adapted RCI did obtain slightly better results, the difference was miniscule. The adapted RCI contains 7 questions from the original RCI that measure the influence that one person has over the other. Since the results from the adapted RCI were only slightly better than the correlation results obtained utilizing the KAI, it is recommended that the adapted RCI be the standard model utilized to measure the known adversary going forward. The justification behind that lies in the 7 questions from the original RCI that attempt gauge influence. Since the KAI does not contain these 7 influence gauging questions, and is more geared towards quantifying the known adversary by using proximity questions, we expect that it will not perform well in establishing a correlation between the guessing scores of pairs in studies evaluating authentication frameworks that have nothing to do with autobiographical data.

We stress that the results obtained from the KAI and the adapted RCI are preliminary. Testing on a wider scale with a more diverse demographic is necessary to provide further proof of concept that the utilization of the KAI or the adapted RCI is indeed more fruitful for security research.

7.2 Comparison of Results GeoSQ/GeoPassHints

This section will be dedicated to comparing the results obtained from the security, usability, and deployability analysis of GeoPassHints and GeoSQ. We will utilize the Bonneau et al. [18] framework in order to guide this comparison, while focusing on the important metrics and categories where one authentication system was superior to the other. Table 7.1 showcases the comparison of GeoSQ, GeoPassHints as well as commonly utilized fallback authentication techniques.

Before we delve into the comparison it is important to note the target environment. GeoSQ is more suited towards a mobile fallback authentication environment, wheras GeoPassHints is more versatile and can be utilized in any primary or fallback authentication environment. Both of the authentication systems we devised failed to meet several important metrics in terms of usability, therefore we would not currently recommend them to be replacements to current primary/fallback authentication mechanisms.

7.2.1 Security

Both GeoSQ and GeoPassHints failed to meet the threshold to be *Resilient-to-the-Known-Adversary.* Despite that, GeoPassHints performed considerably better than GeoSQ in being able to repel the known adversary with only 2.9% of secret notes being guessed by the pairs. On the other hand, roughly 11% of pairs were successful in guessing 7/10 locations which was the threshold we set for GeoSQ.

Both systems appeared to be *Resilient-to-Social-Media-Mining.* The average number of location posts per week was 1.26 which is not significant enough to completely compromise a system such as GeoSQ due to the high number of questions asked. Since only 4 secret notes were guessed by pairs in our user study, we conclude that following each other on social media could did not greatly contribute to a pair's ability to guess secret notes.

Lastly, there are two security metrics that GeoSQ is superior to GeoPassHints in. GeoSQ has a significantly larger sample space because 7 correct answers out of 10 are required. Therefore, it is *Resilient-to-Unthrottled-Guessing*. While we could not run a similar password cracking analysis on GeoSQ that we performed on GeoPassHints (because we did not have access to the locations), the sample space is large enough that a password cracker attempting guessing would fail due to the system settings (you do not get more than one shot at answering a location questions). On the other hand, our password cracking analysis performed on GeoPassHints secret notes yielded a higher end of 20% of secret notes being guessed by utilizing the semantic cracker [51]. The second security metric that GeoSQ is superior to GeoPassHints in is the resilience to physical observation. GeoPassHints offers similar protection against physical observation as passwords. However, in order to compromise GeoSQ, a prolonged period of physical observation is required, hence making it more immune to physical observation attacks.

According to our analysis, GeoSQ appears to be the most secure authentication system in principal. However, it is extremely vulnerable o the known adversary while GeoPassHints is less so. Refer to Table 7.1 for a side by side comparison.

7.2.2 Usability

Both GeoSQ and GeoPassHints obtained terrible results in our usability analysis. There is only one category that GeoSQ outperforms GeoPassHints in, which is *Scalable-for-Users*. The burden on the user does not increase when the number of verifiers that utilize GeoSQ increases, because it is the same location data that will be utilized. In the case of GeoPassHints however, the number of secret notes will greatly increase the memory burden on the user should many verifiers adopt GeoPassHints.

Both GeoPassHints and GeoSQ are extremely inefficient to use and have frequent errors due to the high login time and the high error rate in both systems. In addition to that, GeoSQ performs very poorly in terms of credential setting time. While the credential setting is happening automatically through the user's natural movements, it still does take 7-11 days to gather 10 unique locations. So theoretically GeoSQ needs anywhere between 7-11 days to set credentials on average.

GeoSQ does not offer the perk of being easy to recover from loss. Should the user fail to authenticate when asked ten questions, it would be difficult to wait for a new set of questions to be generated (because the user would need to move around). Alternative system designs could reuse location questions, but the current configuration of GeoSQ does not. When compared to GeoPassHints, that is not very efficient, as the reset + reconfirm time for GeoPassHints is comparable to the initial set + confirm time at around one minute on average for each account type.

7.2.3 Deployability

In terms of deployability, both GeoSQ and GeoPassHints scored identically on the Bonneau et al. framework [18]. The accessibility of authentication schemes that rely on maps is very important. We believe the Google Maps API provides a great set of accessibility features [1] in a mobile environment or in a browser environment.

Table 7.1: Adapted Bonneau et al. framework [18]. Analysis of Security Questions [44], analysis of web passwords [19]. Please note that the two extra security categories that we added were not evaluated in the original framework proposal [18].

	Memorywise-Effortless	Scalable for Users	Nothing to Carry	Physically Effortless	Easy to Learn	Efficient to Use	Infrequent Errors	Easy Recovery from Loss	Accessible	Negligible Cost per User	Server Compatible	Browser Compatible	Mature	Non-Proprietary	Res. to Physical Observation	Res. to Targeted Impersonation	Res. to Throttled Guessing	Res. to Unthrottled Guessing	Res. to Internal Observation	Res. to Leaks from Other Verifiers	Res. to Phishing	Res. to Theft	No Trusted Third Party	Requiring Explicit Consent	Unlinkable	Res. to Social Media Mining	Res. to the Known Adversary
Passwords (Baseline)			●		●	●	○	●	●	●	●	●	●	●		○						●	●	●	●	▲	▲
Security Questions	○ ▥		●		●	●	○	●	●	●	▤	●	●	●	▤	▤				▤		●	●	●	● ▤	▤	▤
Email Resets	● ▥	● ▥	●		●	●	● ▥	●	●	●	▤	●	●	●		○	● ▥		▥	▤	▤	●	▤	●	▤	△	△
SMS Resets	● ▥	● ▥	▤		●	●	● ▥	●	●	●	▤	●	●	●		○	● ▥	△	▥	▤	▤	○ ▤	▤	●	▤	△	△
GeoSQ	▤	● ▥	○ ▤	▤	●	▤	▤	▤	●	●	▤	●	▤	●	● ▥	○ ▤	● ▥	● ▥	▥	● ▥	● ▥	● ▤	○ ▤	●	▤	● ▥	▤
GeoPassHints	▤		●		●	▤	▤	●	●	●	▤	●	▤	●	▥	▤		▥		● ▥	● ▥	●	●	●	●	● ▥	▤

● – offers the benefit, ○ – almost offers the benefit, no circle – does not offer the benefit, ▲ – potential to have the benefit , △ – potential to almost offer the benefit, ▥ – better than passwords, ▤ – worse than passwords

Chapter 8

Conclusion

In conclusion, we designed, implemented, and tested two authentication systems for security, usability, and deployability. The results of our analysis concluded that GeoPassHints and GeoSQ would not be usable in a primary authentication or fallback authentication environment due to the high error rate and the long login time. According to the Bonneau et al. [18] framework analysis (refer to Table 7.1) the relative security that both GeoSQ and GeoPassHints offer is stronger than commonly utilized methods of fallback authentication such as security questions, Email resets, and SMS resets. However, the usability of GeoPassHints and GeoSQ is the real drawback against recommending these authentication systems according to the Bonneau et al. [18] framework analysis.

Moreover, we developed three different models in an attempt to quantify the known adversary. This was done in the hopes of addressing a flaw in current security research where known adversaries are classified as either strong or weak based on self-reported closeness measures. Two of our models (the KAI and the adapted RCI) showed promising correlation results with GeoSQ pair guessing scores. As a result, we recommend that future work should utilize the KAI and the adapted RCI with different types of authentication systems (since we only performed the analysis with an autobiographical authentication system) for further proof of concept.

References

[1] Accessibility products and features, `https://www.google.ca/accessibility/products-features.html`, site accessed June 2018.

[2] The common cold project, `https://www.cmu.edu/common-cold-project/measures-by-study/psychological-and-social-constructs/marital-quality-measures/relationship-closeness.html`, site accessed June 2018.

[3] John the ripper password cracker, `http://www.openwall.com/john/`, site accessed June 2018.

[4] Location request, `https://developers.google.com/android/reference/com/google/android/gms/location/LocationRequest`, site accessed June 2018.

[5] Open layers api , `https://openlayers.org/`, site accessed June 2018.

[6] Passwords, `https://wiki.skullsecurity.org/Passwords`, site accessed June 2018.

[7] Ss7 hack tutorial, `https://fedotov.co/ss7-hack-tutorial-software-video`, site accessed June 2018.

[8] Wordnet, `https://wordnet.princeton.edu/`, site accessed June 2018.

[9] Hervé Abdi. Bonferroni and šidák, corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics*, 3:103–107, 2007.

[10] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[11] Mahdi Nasrullah Al-Ameen and Matthew Wright. Multiple-password interference in the geopass user authentication scheme. In *Proceedings of the Proc. Workshop Usable Secur.(USEC)*, pages 1–6, 2015.

[12] Yusuf Albayram and Mohammad Maifi Hasan Khan. Evaluating the effectiveness of using hints for autobiographical authentication: A field study. In *Symposium on Usable Privacy and Security*, pages 211–224, 2015.

[13] Yusuf Albayram and Mohammad Maifi Hasan Khan. Evaluating smartphone-based dynamic security questions for fallback authentication: a field study. *Human-Centric Computing and Information Sciences*, 6(1):16, 2016.

[14] Arthur Aron, Elaine N Aron, and Danny Smollan. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4):596, 1992.

[15] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of Learning and Motivation*, volume 2, pages 89–195. Elsevier, 1968.

[16] Ellen Berscheid, Mark Snyder, and Allen M Omoto. The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology*, 57(5):792, 1989.

[17] Joseph Bonneau, Elie Bursztein, Ilan Caron, Rob Jackson, and Mike Williamson. Secrets, lies, and account recovery: Lessons from the use of personal knowledge questions at google. In *Proceedings of the 24th International Conference on World Wide Web*, pages 141–150. International World Wide Web Conferences Steering Committee, 2015.

[18] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 553–567. IEEE, 2012.

[19] Joseph Bonneau and Sören Preibusch. The password thicket: Technical and market failures in human authentication on the web. In *WEIS*, 2010.

[20] Paul C. van Oorschot, Amirali Salehi-Abari, and Julie Thorpe. Purely automated attacks on passpoints-style graphical passwords. *IEEE Transactions on Information Forensics and Security*, 5(3):393–405, 2010.

[21] Claude Castelluccia, Chaabane Abdelberi, Markus Dürmuth, and Daniele Perito. When privacy meets security: Leveraging personal information for password cracking. *CoRR*, abs/1304.6584, 2013.

[22] Sonia Chiasson, Alain Forget, Elizabeth Stobert, Paul C van Oorschot, and Robert Biddle. Multiple password interference in text passwords and click-based graphical passwords. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pages 500–511, 2009.

[23] Soumyadeb Chowdhury, Ron Poet, and Lewis Mackenzie. Passhint: memorable and secure authentication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2917–2926, 2014.

[24] Robert B Cialdini, Stephanie L Brown, Brian P Lewis, Carol Luce, and Steven L Neuberg. Reinterpreting the empathy–altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, 73(3):481, 1997.

[25] Martin A Conway. Episodic memories. *Neuropsychologia*, 47(11):2305–2313, 2009.

[26] Fergus IM Craik and Endel Tulving. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104(3):268, 1975.

[27] Armin Falk and James J Heckman. Lab experiments are a major source of knowledge in the social sciences. *science*, 326(5952):535–538, 2009.

[28] Simon Gächter, Chris Starmer, and Fabio Tufano. Measuring the closeness of relationships: a comprehensive evaluation of the inclusion of the other in the self scale. *PloS one*, 10(6):e0129478, 2015.

[29] Simson L Garfinkel. Email-based identification and authentication: An alternative to pki? *IEEE Security & Privacy*, 99(6):20–26, 2003.

[30] Nethanel Gelernter, Senia Kalma, Bar Magnezi, and Hen Porcilan. The password reset mitm attack. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, pages 251–267, 2017.

[31] Maximilian Golla and Markus Dürmuth. Analyzing 4 million real-world personal knowledge questions (short paper). In *Proceedings of the International Conference on Passwords*, pages 39–44. Springer, 2015.

[32] Mordechai Guri, Eyal Shemer, Dov Shirtz, and Yuval Elovici. Personal information leakage during password recovery of internet services. In *Proceedings of the Intelligence and Security Informatics Conference*, pages 136–139, 2016.

[33] Ameya Hanamsagar, Simon S Woo, Chris Kanich, and Jelena Mirkovic. Leveraging semantic transformation to investigate password habits and their causes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 570, 2018.

[34] Alina Hang, Alexander De Luca, Matthew Smith, Michael Richter, and Heinrich Hussmann. Where have you been? using location-based security questions for fallback authentication. In *Proceedings of the 11th Symposium On Usable Privacy and Security*, pages 169–183, 2015.

[35] Alina Hang, Alexander De Luca, Emanuel Von Zezschwitz, Manuel Demmler, and Heinrich Hussmann. Locked your phone? buy a new one? from tales of fallback authentication on smartphones to actual concepts. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 295–305, 2015.

[36] Shouling Ji, Shoukon Yang, Xin Hu, Weili Han, Li Zhigong, and Raheem Beyah. Zero-sum password cracking game: A large-scale empirical study on the crackability, correlation, and security of passwords. *IEEE Transactions on Dependable and Secure Computing*, 14(5):550–564, 2017.

[37] Mike Just and David Aspinall. Personal choice and challenge questions: a security and usability assessment. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 8, 2009.

[38] Andy Lilly. Imsi catchers: Hacking mobile communications. *Network Security*, 2017(2):5–7, 2017.

[39] Brent MacRae, Amirali Salehi-Abari, and Julie Thorpe. An exploration of geographic authentication schemes. *IEEE Transactions on Information Forensics and Security*, 11(9):1997–2012, 2016.

[40] Fatma Al Maqbali and Chris J Mitchell. Web password recovery—a necessary evil? *CoRR*, 1801.06730, 2018.

[41] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. Usability

and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 527–539, 2016.

[42] Ildar Muslukhov, Yazan Boshmaf, Cynthia Kuo, Jonathan Lester, and Konstantin Beznosov. Know your enemy: the risk of unauthorized access in smartphones by insiders. In *Proceedings of the 15th international Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 271–280, 2013.

[43] Amirali Salehi-Abari, Julie Thorpe, and P. C. van Oorschot. On purely automated attacks and click-based graphical passwords. In *Proceedings of the 2008 Annual Computer Security Applications Conference*, pages 111–120, 2008.

[44] Stuart Schechter, AJ Bernheim Brush, and Serge Egelman. It's no secret. measuring the security and reliability of authentication via "secret" questions. In *Proceedings 30th IEEE Symposium on Security and Privacy*, pages 375–390, 2009.

[45] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[46] Richard Shay, Patrick Gage Kelley, Saranga Komanduri, Michelle L Mazurek, Blase Ur, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proceedings of the 8th Symposium on Usable Privacy and Security*, page 7, 2012.

[47] Julie Thorpe, Muath Al-Badawi, Brent MacRae, and Amirali Salehi-Abari. The presentation effect on graphical passwords. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 2947–2950, 2014.

[48] Julie Thorpe, Brent MacRae, and Amirali Salehi-Abari. Usability and security evaluation of geopass: A geographic location-password scheme. In *Proceedings of the 9th Symposium on Usable Privacy and Security*, pages 14:1–14:14, 2013.

[49] Julie Thorpe, Amirali Salehi-Abari, and Robert Burden. Video-passwords: Advertising while authenticating. In *Proceedings of the 2012 New Security Paradigms Workshop*, pages 127–140, 2012.

[50] Julie Thorpe and P. C. van Oorschot. Human-seeded attacks and exploiting hot-spots in graphical passwords. In *Proceedings of 16th USENIX Security Symposium*, pages 8:1–8:16, 2007.

[51] Rafael Veras, Christopher Collins, and Julie Thorpe. On semantic patterns of passwords and their security impact. In *NDSS*, 2014.

[52] Merrill Warkentin and Robert Willison. Behavioral and policy issues in information systems security: the insider threat. *European Journal of Information Systems*, 18(2):101–105, 2009.

[53] Jie Wei. Markov edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):311–321, 2004.

[54] Matt Weir, Sudhir Aggarwal, Breno De Medeiros, and Bill Glodek. Password cracking using probabilistic context-free grammars. In *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pages 391–405, 2009.

[55] Bill Welch. Exploiting the weaknesses of ss7. *Network Security*, 2017(1):17–19, 2017.

[56] Susan Wiedenbeck, Jim Waters, Jean-Camille Birget, Alex Brodskiy, and Nasir Memon. Passpoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies*, 63(1):102–127, 2005.

Appendices

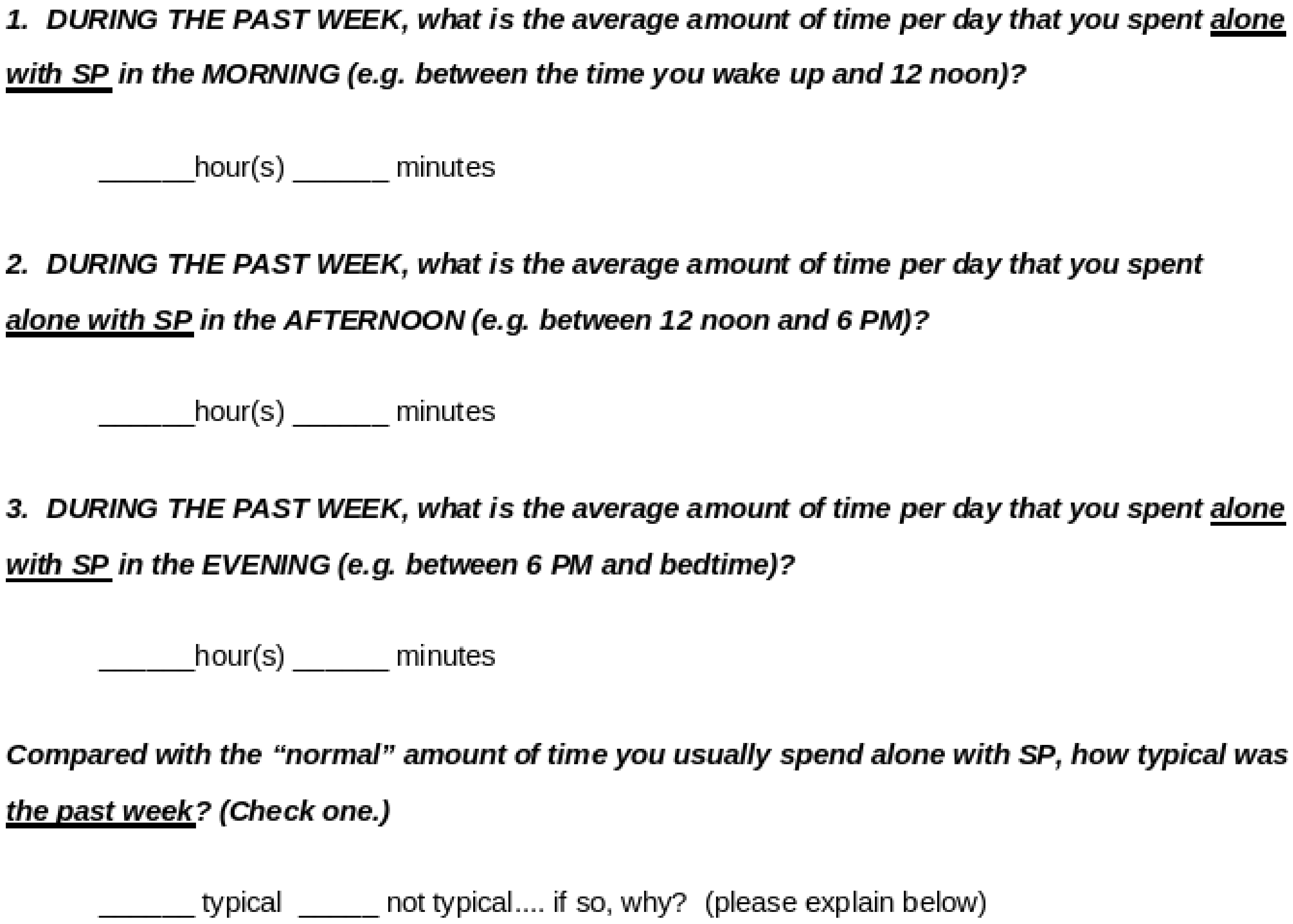
Appendix A

Original RCI

Digital remake of the original is a courtesy of: The Common Cold Project [2]

INSTRUCTIONS:

We would like you to estimate the amount of time you typically spend alone with your spouse/partner (referred to as SP below) during the day. We would like you to make these time estimates by breaking the day into morning, afternoon, and evening, although you should interpret each of these time periods in terms of your own typical daily schedule. (For example, if you work a night shift, "morning" may actually reflect time in the afternoon, but is nevertheless time immediately after waking.) Think back over the past week and write in the average amount of time, per day, that you spent alone with your SP, with no one else around, during each time period. If you did not spend any time with SP in some time periods, write 0 hour(s) and 0 minutes.



1. DURING THE PAST WEEK, what is the average amount of time per day that you spent alone with SP in the MORNING (e.g. between the time you wake up and 12 noon)?

______hour(s) ______ minutes

2. DURING THE PAST WEEK, what is the average amount of time per day that you spent alone with SP in the AFTERNOON (e.g. between 12 noon and 6 PM)?

______hour(s) ______ minutes

3. DURING THE PAST WEEK, what is the average amount of time per day that you spent alone with SP in the EVENING (e.g. between 6 PM and bedtime)?

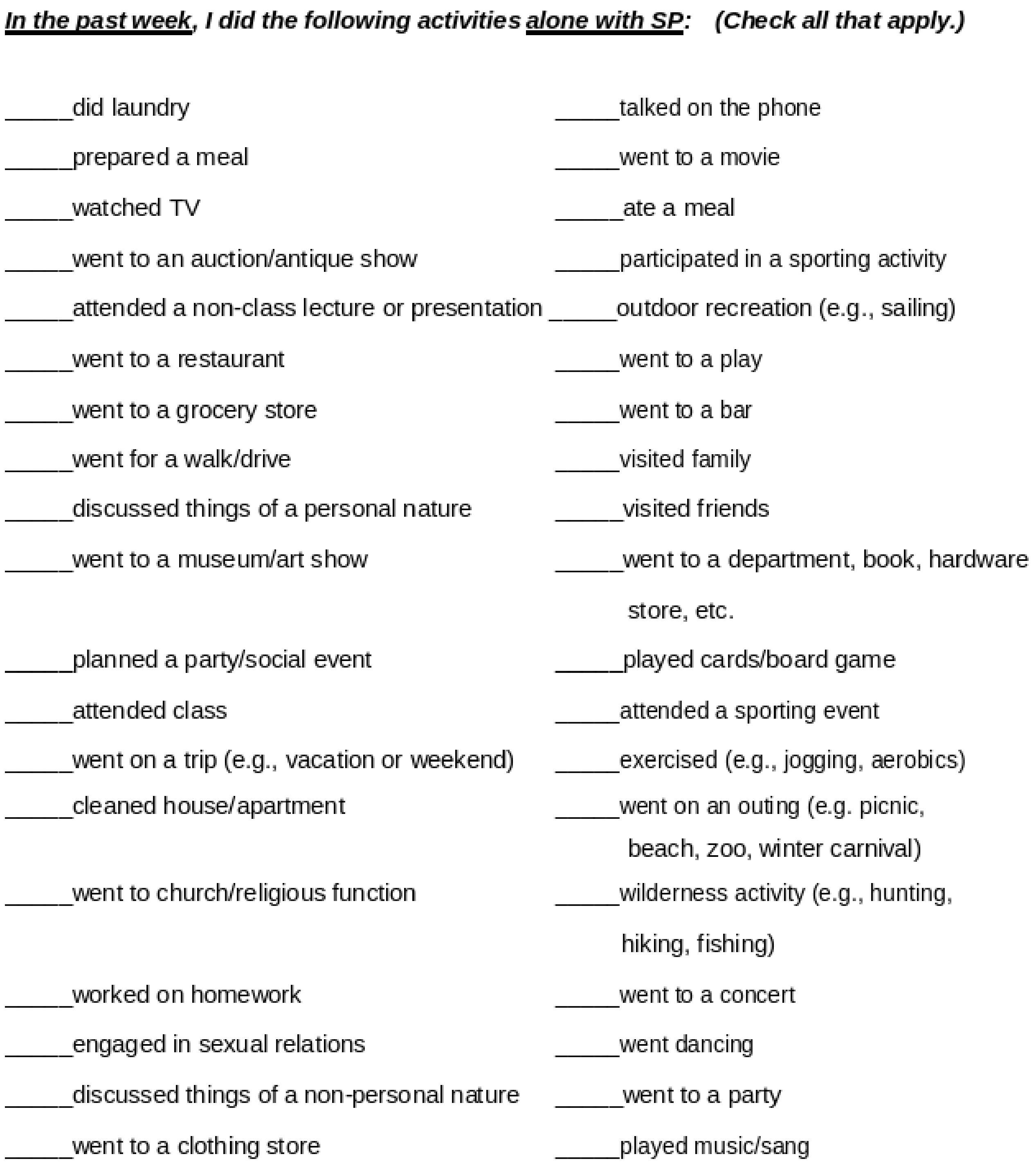
______hour(s) ______ minutes

Compared with the "normal" amount of time you usually spend alone with SP, how typical was the past week? (Check one.)

______ typical _____ not typical.... if so, why? (please explain below)

INSTRUCTIONS:

The following is a list of different activities that people may engage in over the course of one week. For each of the activities listed, please check all of those that you have engaged in alone with SP in the past week. Check only those activities that were done alone with SP and not done with SP in the presence of others.



***In the past week*, I did the following activities *alone with SP*: (Check all that apply.)**

_____did laundry

_____prepared a meal

_____watched TV

_____went to an auction/antique show

_____attended a non-class lecture or presentation

_____went to a restaurant

_____went to a grocery store

_____went for a walk/drive

_____discussed things of a personal nature

_____went to a museum/art show

_____planned a party/social event

_____attended class

_____went on a trip (e.g., vacation or weekend)

_____cleaned house/apartment

_____went to church/religious function

_____worked on homework

_____engaged in sexual relations

_____discussed things of a non-personal nature

_____went to a clothing store

_____talked on the phone

_____went to a movie

_____ate a meal

_____participated in a sporting activity

_____outdoor recreation (e.g., sailing)

_____went to a play

_____went to a bar

_____visited family

_____visited friends

_____went to a department, book, hardware store, etc.

_____played cards/board game

_____attended a sporting event

_____exercised (e.g., jogging, aerobics)

_____went on an outing (e.g. picnic, beach, zoo, winter carnival)

_____wilderness activity (e.g., hunting, hiking, fishing)

_____went to a concert

_____went dancing

_____went to a party

_____played music/sang

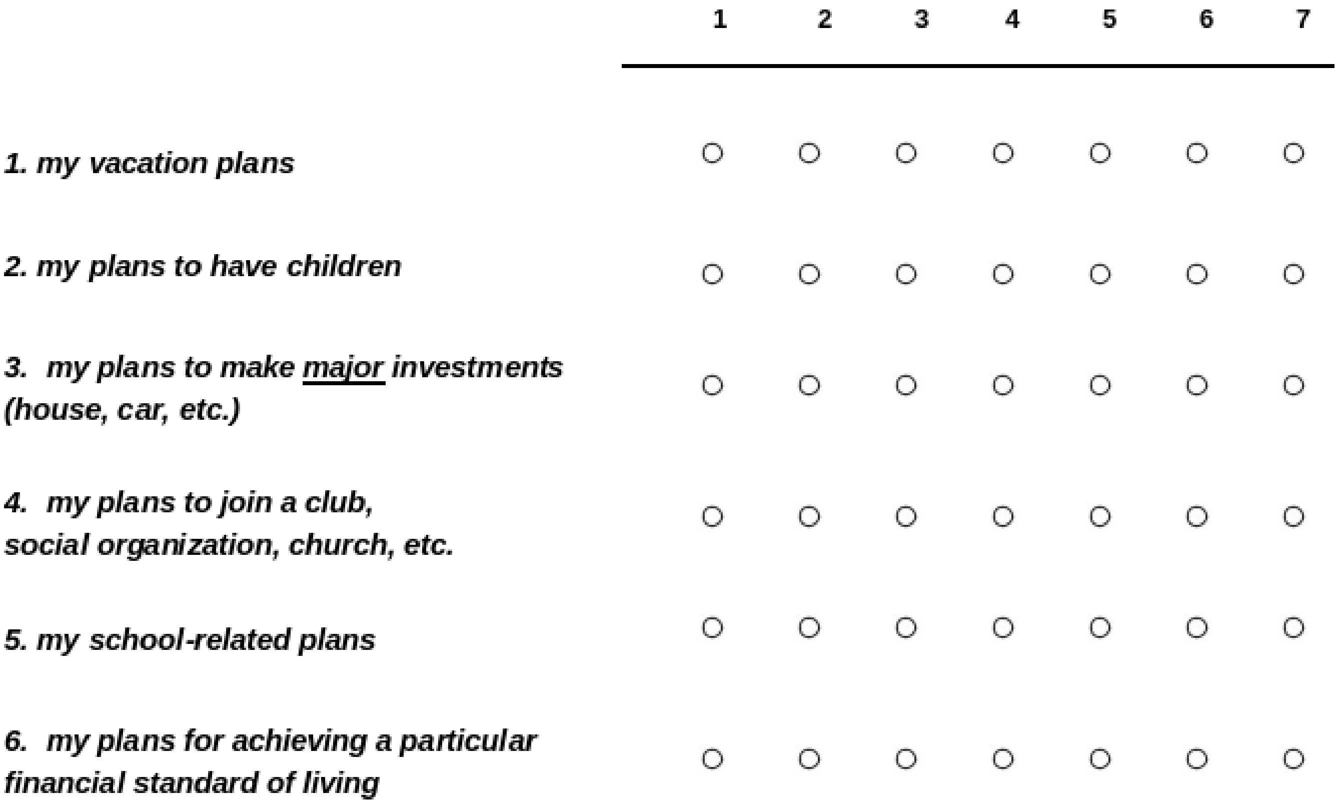
INSTRUCTIONS:

The following questions concern the amount of influence SP has on your thoughts, feelings, and behavior. Using the 7-point scale below, please indicate the extent

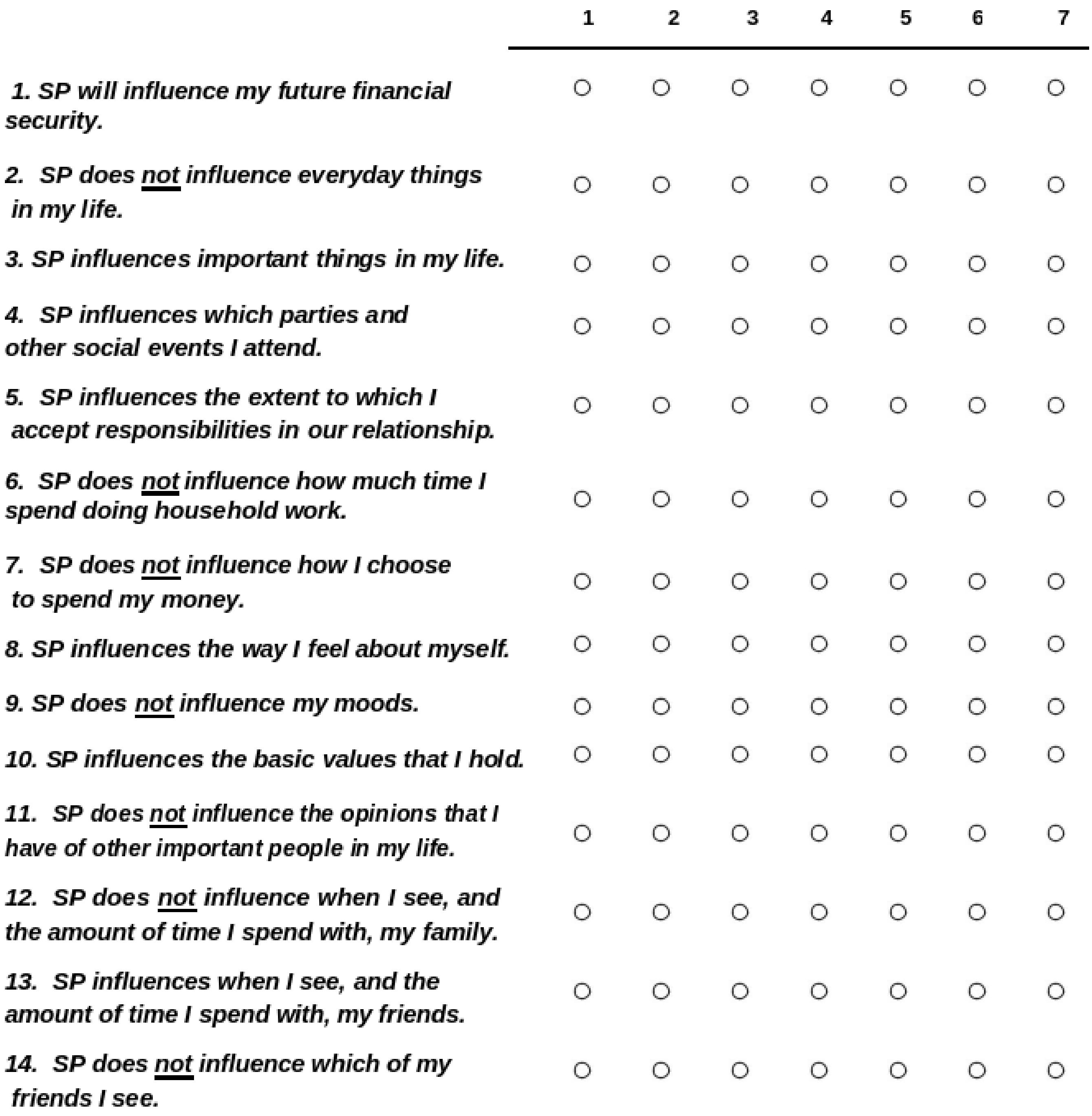
to which you agree or disagree with each statement by placing an “X” over the appropriate circle.

INSTRUCTIONS:

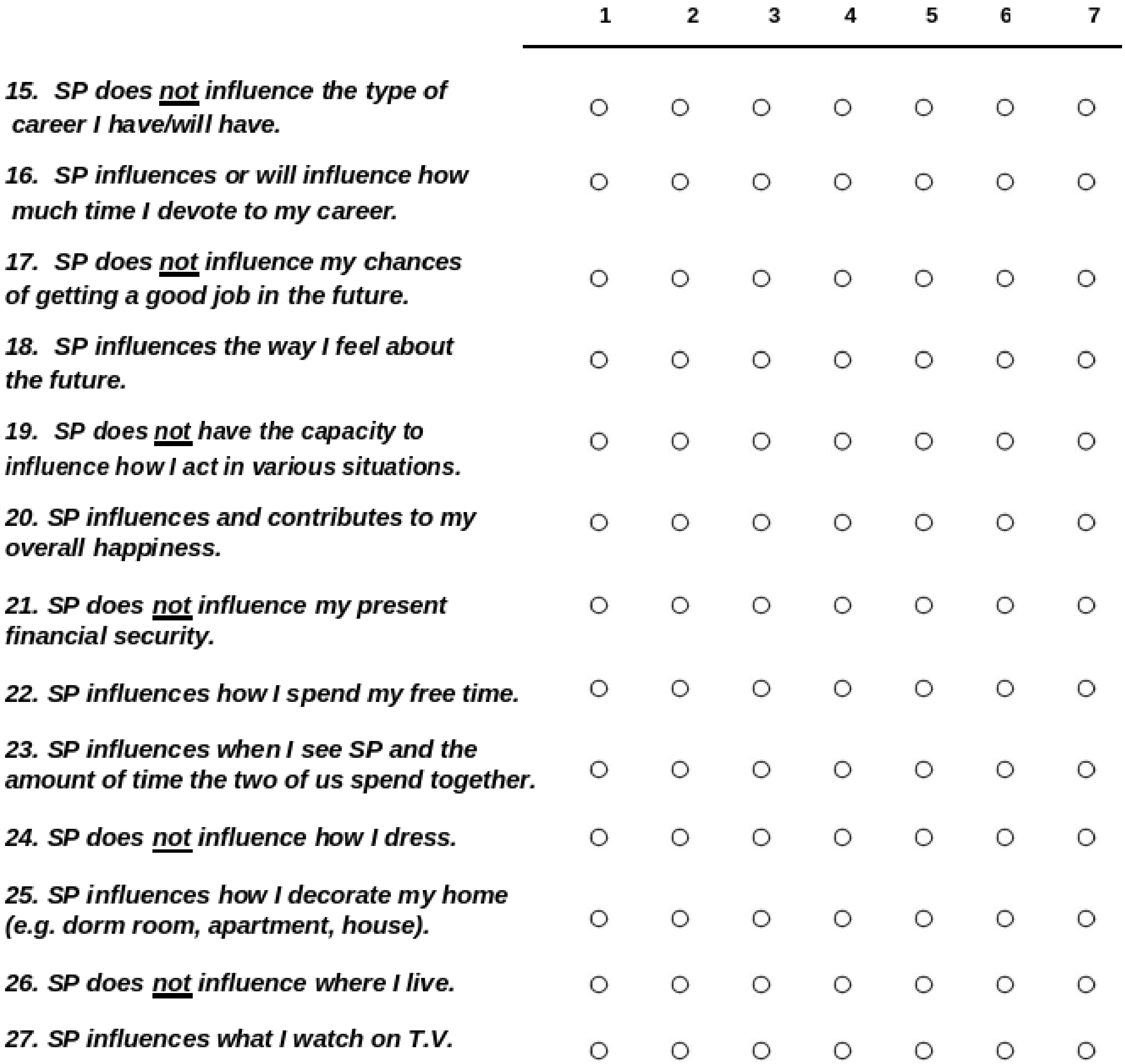
Now we would like you to tell us how much SP affects your future plans and goals. Using the 7-point scale below, please indicate the degree to which your future plans and goals are affected by SP by placing an “X” over the appropriate circle for each item. If an area does not apply to you (e.g. you have no plans or goals in that area), put an “X” over the circle for “1” (not at all).



	1	2	3	4	5	6	7
1. my vacation plans	○	○	○	○	○	○	○
2. my plans to have children	○	○	○	○	○	○	○
3. my plans to make major investments (house, car, etc.)	○	○	○	○	○	○	○
4. my plans to join a club, social organization, church, etc.	○	○	○	○	○	○	○
5. my school-related plans	○	○	○	○	○	○	○
6. my plans for achieving a particular financial standard of living	○	○	○	○	○	○	○



	1	2	3	4	5	6	7
1. SP will influence my future financial security.	○	○	○	○	○	○	○
2. SP does not influence everyday things in my life.	○	○	○	○	○	○	○
3. SP influences important things in my life.	○	○	○	○	○	○	○
4. SP influences which parties and other social events I attend.	○	○	○	○	○	○	○
5. SP influences the extent to which I accept responsibilities in our relationship.	○	○	○	○	○	○	○
6. SP does not influence how much time I spend doing household work.	○	○	○	○	○	○	○
7. SP does not influence how I choose to spend my money.	○	○	○	○	○	○	○
8. SP influences the way I feel about myself.	○	○	○	○	○	○	○
9. SP does not influence my moods.	○	○	○	○	○	○	○
10. SP influences the basic values that I hold.	○	○	○	○	○	○	○
11. SP does not influence the opinions that I have of other important people in my life.	○	○	○	○	○	○	○
12. SP does not influence when I see, and the amount of time I spend with, my family.	○	○	○	○	○	○	○
13. SP influences when I see, and the amount of time I spend with, my friends.	○	○	○	○	○	○	○
14. SP does not influence which of my friends I see.	○	○	○	○	○	○	○



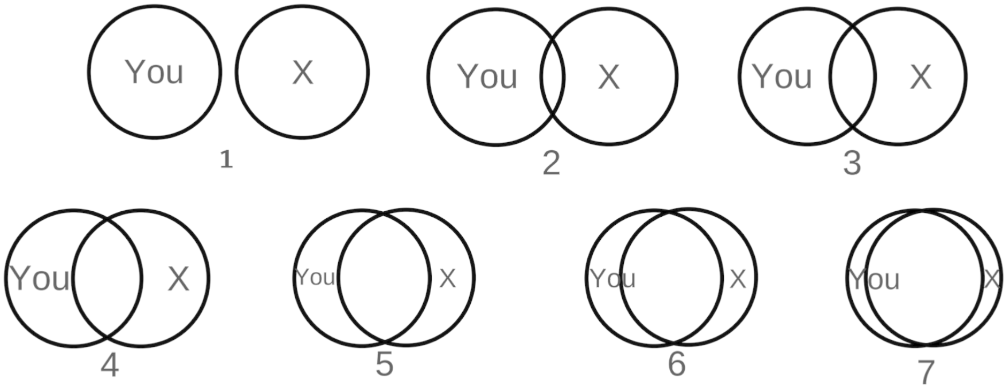
	1	2	3	4	5	6	7
15. SP does not influence the type of career I have/will have.	○	○	○	○	○	○	○
16. SP influences or will influence how much time I devote to my career.	○	○	○	○	○	○	○
17. SP does not influence my chances of getting a good job in the future.	○	○	○	○	○	○	○
18. SP influences the way I feel about the future.	○	○	○	○	○	○	○
19. SP does not have the capacity to influence how I act in various situations.	○	○	○	○	○	○	○
20. SP influences and contributes to my overall happiness.	○	○	○	○	○	○	○
21. SP does not influence my present financial security.	○	○	○	○	○	○	○
22. SP influences how I spend my free time.	○	○	○	○	○	○	○
23. SP influences when I see SP and the amount of time the two of us spend together.	○	○	○	○	○	○	○
24. SP does not influence how I dress.	○	○	○	○	○	○	○
25. SP influences how I decorate my home (e.g. dorm room, apartment, house).	○	○	○	○	○	○	○
26. SP does not influence where I live.	○	○	○	○	○	○	○
27. SP influences what I watch on T.V.	○	○	○	○	○	○	○

Appendix B

IoS Scale & We Scale

IoS Scale:

Please note that this is a remake of the original for web viewing purposes.



We Scale:

“Please, select the appropriate number below to indicate to what extent you would use the term “WE” to characterize you and [This Individual]. Answers were on a 7-point scale (1 = “not at all”; 7 = “very much so ”)” [24]

This is the exact wording for the We Scale, we “[This Individual]” with “your pair”.

The average of the We Scale and the IoS Scale is the oneness score [24].

Appendix C

Adpated RCI

Questions in the original RCI are marked with a red asterisk.

1. What is the approximate number of hours in physical proximity to your pair per week? This refers to the total number of hours you and your pair are in the same house, in the same room, same cafeteria, or same workspace etc...

   Scoring Range:

Answer	Score
0-1 hours	1
2-7 hours	2
8-10 hours	3
10-15 hours	4
16-20 hours	5
21-25 hours	6
26-30 hours	7
31-40 hours	8
41-45 hours	9
45+ hours	10

2. Approximately how many meals do you share with your pair per week?

   Scoring Range:

Answer	Score
0	0
1-3	1
4-7	2
8-11	3
12-15	4
16-19	5
21-24	6
24+	7

3. Does your pair have access do any of your social media accounts? This could mean a social media account password.

   Scoring:

   Yes = 10 points

   No = 0 points

4. Does your pair have restricted physical access to your devices/accounts? (i.e. , does your pair have access to your locked device?)

   Yes = 10 points

   No = 0 points

5. Does your pair have unrestricted access to your devices/accounts? (i.e. , do they know one or more of your passwords?)

   Yes = 15 points

   No = 0 points

6. In the past week, approximately how many hours have you spent with your pair awake? (i.e. , at home together, in the same work space, or in the same social gathering).

   Scoring Range:

Answer	Score
0	0
1-10 hours	1
11-20 hours	2
21-30 hours	3
31-40 hours	4
41-50 hours	5
51-60 hours	6
61-70 hours	7
70+ hours	8

7. Does your pair affect your vacation plans? *

   Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your vacation plans).

8. Does your pair have any affect/influence on your marriage plans? *

   Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your marriage plans).

9. Does your pair have any affect/influence on your plans to have children? *

   Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your plans to have children).

10. Does your pair have any affect/influence on your plans to make major investments? *

    Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your major investment plans).

11. Does your pair have any affect/influence on your plans to join a club/social organization? *

    Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your plans to join a club/social organization).

12. Does your pair have any affect/influence on your school related plans? *

Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your school related plans).

13. Does your pair have any affect/influence on your financial standing/wellbeing? * Likert Scale question with seven levels. The score the level chosen (i.e, 1 = pair has no power, 7 = pair has significant power to affect your financial standing/wellbeing).

14. Do you and your pair follow each other on social media?

    Yes = 10 points

    No = 0 points

    If the answer was no questions 15-17 were not asked.

15. On how many different mediums do you follow your pair on social media?

    Scoring Range:

Answer	Score
0 mediums	0
1 medium	1
2 mediums	2
3 mediums	3
4 mediums	4
5 mediums	5
6 mediums	6
7 mediums	7
8+ mediums	8

16. How often do you see your pair's social media posts per week?

    Scoring Range:

17. How many times per week do see your pair's social media posts relating to his/her location? (This could be a check in , a snapchat post indicative of location etc...).

    Scoring Range:

Answer	Score
0 posts	0
1 post	1
2 posts	2
3 posts	3
4 posts	4
5 posts	5
6 posts	6
7+ posts	7

Answer	Score
0 posts	0
1 post	1
2 posts	2
3 posts	3
4 posts	4
5 posts	5
6 posts	6
7+ posts	7

Appendix D

GeoPassHints Exit Survey Questions

1. Was GeoPassHints easy to use? Likert Scale with seven levels.

2. Was GeoPassHints easy to learn how to use? Likert Scale with seven levels.

3. For a mobile device's forgotten password/PIN or failed biometric, to what extent would you prefer to use this system (GeoPassHints) in lieu of any of the following fallback authentication methods?

   i. Security Questions

   ii. Email Resets

   iii. SMS Resets

4. For an online account's forgotten password, to what extent would you prefer to use this system in lieu of any of the following fallback authentication methods?

   i. Security Questions

   ii. Email Resets

   iii. SMS Resets

Likert Scale with seven levels for each fallback authentication method.

5. I would prefer to use the GeoPassHints system to login to any of the following accounts, instead of entering a password, for these account types.

   i. Email Account

   ii. Online Bank Account

   iii. Social Networking Account

   iv. E-Commerce Account

   v. Other Infrequently Accessed Accounts

   vi. Mobile Device

   vii. University Account

   viii. Job Related Account

   Likert Scale with seven levels for each account type.

6. If you use this system (GeoPassHints), how likely do you think it is that someone you know could access your devices/accounts with your consent.

   Likert Scale with seven levels.

7. How concerned are you about a privacy leak on any medium.

   Likert Scale with seven levels.

Appendix E

GeoSQ Exit Survey Questions

There is only one different question from the GeoPassHints exist survey that is marked clearly with a red aseterisk.

1. Was GeoSQ easy to use? Likert Scale with seven levels.

2. Was GeoSQ easy to learn how to use? Likert Scale with seven levels.

3. For a mobile device's forgotten password/PIN or failed biometric, to what extent would you prefer to use this system (GeoSQ) in lieu of any of the following fallback authentication methods?

   i. Security Questions

   ii. Email Resets

   iii. SMS Resets

   Likert Scale with seven levels for each fallback authentication method.

4. For an online account's forgotten password, to what extent would you prefer to use this system (GeoSQ) in lieu of any of the following fallback authentication methods?

i. Security Questions

ii. Email Resets

iii. SMS Resets

Likert Scale with seven levels for each fallback authentication method.

5. I would prefer to use the GeoSQ system to login to any of the following accounts, instead of entering a password, for these account types.

   i. Email Account

   ii. Online Bank Account

   iii. Social Networking Account

   iv. E-Commerce Account

   v. Other Infrequently Accessed Accounts

   vi. Mobile Device

   vii. University Account

   viii. Job Related Account

   Likert Scale with seven levels for each account type.

6. If you use this system (GeoSQ), how likely do you think it is that someone you know could access your devices/accounts with your consent.

   Likert Scale with seven levels.

7. How concerned are you about a privacy leak on any medium. Likert Scale with seven levels.

8. To what extent are you concerned of a privacy leak as a result of applications utilizing location services? *

   Likert Scale with seven levels.

Appendix F

Bonneau et al. Framework

Each category within usability, security, and deployability are explained in detail in this section.

Usability sub categories:

Subcategory	Explanation
Memorywise-Effortless	In order for an authentication system to be considered *Memorywise-Effortless* it must not require a user to remember any secret at all [18]. A system is rewarded with the status of being *Quasi-Memorywise-Effortless* if there is only one secret to remember for all verifiers [18].
Scalable-for-Users	An authentication system is considered *Scalable-for-Users* if using the authentication system for many accounts does not put any extra burden on the user [18]. For example, if a user is asked to set multiple passwords across different verifiers, then as the number of accounts increase, the burden on the user is increasing.

Nothing-to-Carry	This category is awarded to authentication systems that do not require the user to carry anything to be able to authenticate. Some systems require a physical card to be present for authentication, these systems would not be granted the *Nothing to Carry* accolade [18]. Moreover, an authentication system is given the status of *Quasi Nothing to Carry* if the user will carry this device anyway (i.e., a smartphone) [18].
Physically-Effortless	This category is granted to authentication systems that do not require the user to perform any action past the pressing of a button (to indicate authentication is occurring i.e., submit) [18].
Easy-to-Learn	In order to a system to be classified *Easy-to-Learn* users must be able to learn how to use it without too much effort and attention required, and must be able to remember how to use it with ease [18]
Efficient-to-Use	In order to an authentication system to be considered *Efficient to Use* the time spent for each authentication must be acceptably short [18], the time required for a user to set up his/her credentials with a verifier should also be reasonable [18]. Bonneau et al. do not specify empirical values for this subcategory which is appropriate because of the variety of authentication systems and the variety of intended uses.
Infrequent Errors	An authentication system may be awarded in the *Infrequent Errors* subcategory if the login task typically succeeds when performed by the true user [18].
Easy-Recovery-from-Loss	This subcategory rewards authentication systems that make it easy to recover from loss.

Deployability subcategories:

Subcategory	Explanation
Accessible	This subcategory is awarded to authentication schemes that can be utilized by user with disabilities who can utilize passwords [18]
Negligible-Cost-per-User	This subcategory is awarded to authentication schemes that do not increase the burden on the verifier or the authentication schemes that do not user in terms of cost. Startups with revenue per user should be able to reasonably utilize this authentication scheme [18]
Server-Compatible	This subcategory is awarded to authentication schemes that are compatible with text-based passwords and do not require any changes in the backend [18]
Browser-Compatible	This subcategory is awarded to authentications schemes with broad compatibility to the browsers of the current time [18]
Mature	This subcategory is awarded to authentication schemes that have been tried and tested on a large scale in a real-world environment [18]

Security subcategories:

Subcategory	Explanation
Resilient-to-Physical-Observation	An authentication scheme is considered *Resilient-to-Physical-Observation* when an attacker cannot reasonably successfully impersonate a user after observing them authenticate one or more times. An authentication system is considered *Quasi-resilient-to-physical-observation*, when the authentication scheme can only be compromised by physically observing a user authenticate more than ten times [18].
Resilient-to-Targeted-Impersonation	An authentication scheme is considered *Resilient-to-Targeted-Impersonation* if it is resilient to attacks that exploit personal details [18].
Resilient-to-Throttled-Guessing	An authentication system is awarded this benefit if an attacker, while being constrained by a verifier, cannot compromise more than 1% of accounts a year given ten guesses a day [18]
Resilient-to-Unthrottled-Guessing	An authentication system is *Resilient-to-Unthrottled-Guessing* if an attacker with the ability to guess 2^{40} per account is only able to compromise 1% or less of the accounts [18]
Resilient-to-Internal-Observation	An authentication scheme is *Resilient-to-Internal-Observation* if the attacker cannot impersonate the legitimate user by intercepting the legitimate user's input. An example of internal observation would be key logging or recording the screen of the legitimate user and relaying to an attacker's computer [18].
Resilient-to-Leaks-from-other-Verifiers	An authentication scheme is classified as *Resilient-to-Leaks-from-other-Verifiers*; if anything, another verifier leaks could potentially help an attacker impersonate a legitimate user [18].

Resilient-to-Phishing	An authentication scheme is *Resilient-to-Phishing* if an attacker posing as a verifier cannot fool the user into giving up authentication credentials/personal information [18]. Please note that this subcategory only refers to classical phishing attacks.
Resilient-to-Theft	This subcategory is geared towards authentication schemes that require some sort of physical object or hardware token for verification. authentication *Resilient-to-Theft* is awarded to schemes that cannot be compromised by stealing a physical object or hardware token [18].
No-Trusted-Third-Party	This subcategory rewards authentication schemes that only rely on the user and verifier [18].
Requiring-Explicit-Consent	This subcategory is awarded to authentication schemes that require the user to initiate an authentication. It is regarded as a security and privacy feature [18].
Unlinkable	An authentication scheme is classified as *Unlinkable* if a set of colluding verifiers cannot determine from the authentication credentials alone, whether or not it belongs to the same user on multiple different platforms [18].