

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Xinran Li

2023/03/30

Signaface: Face generation for SSL/TLS certificate change detection

By

Xinran Li

Ymir Vigfusson, Ph.D.
Advisor

Computer Science

Ymir Vigfusson, Ph.D.
Advisor

Ramnath K Chellappa, Ph.D.
Committee Member

Emily Wall, Ph.D.
Committee Member

2023

Signaface: Face generation for SSL/TLS certificate change detection

By

Xinran Li

Ymir Vigfusson, Ph.D.
Advisor

An abstract of
A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Bachelor of Science with Honors

Computer Science

2023

Abstract

Signaface: Face generation for SSL/TLS certificate change detection

By Xinran Li

Cybercriminals and nation-states have a history of targeting individuals by luring them onto legitimate-looking websites to compromise them. The protection of modern websites, SSL/TLS, relies on a cryptographic certificate typically consisting of hard-to-parse sequences of numbers. Users, therefore, often fail to notice a change in the certificate and are tricked into visiting rogue websites that appear legitimate. Browsers are often incapable of detecting such attacks, such as redirection to rogue servers via misspelled URLs. This paper proposes Signaface as a method to help users curb potential security risks by representing SSL certificates with computer-generated faces and relying on the human eyes' familiarity with faces to detect SSL changes. Signaface inputs a website's SSL certificate signature to a generative adversarial neural network (GAN), which produces a synthetic and unique human face. We developed Chrome and Firefox extensions that show a face 'mascot' in the corner of each web page every time the user visits so that they can familiarize themselves with the face mascot. We hypothesize that when a familiar face changes, a sufficient number of users will notice and report the change, lowering the attackers' benefit-to-cost ratio. To test our theory, we devised a gamified experiment to evaluate people's ability to detect face changes. This experiment investigates the effectiveness of using synthetic face mascots to visualize the web's security information and how these results vary depending on whether participants and faces are from the same demographic group.

Signaface: Face generation for SSL/TLS certificate change detection

By

Xinran Li

Ymir Vigfusson, Ph.D.
Advisor

An abstract of
A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Bachelor of Science with Honors

Computer Science

2023

Acknowledgments

I would like to first thank Dr. Ymir Vigfusson for coming up with this research idea presented in this thesis and for bringing me into the world of research. I've been working under Dr. Vigfusson's guidance for the past one and a half years and really appreciate his patience when answering all my questions and helping me to improve and refine my research skills.

I want to also thank Dr. Emily Wall for all her responsiveness and extremely hands-on feedback during the research I conducted with her during my undergraduate senior year. I've been working closely with Etna Ozakara, a current undergraduate junior at Emory University, on the research presented in this thesis since the spring semester of my junior, to which she has contributed a lot, and I appreciate her support.

Many thanks to Egill Valdimarsson, who has offered a lot of help on the psychological components of our research, and Dr. Ramnath K Chellappa, for being my committee member and providing valuable feedback. My friend and roommate Yunjie(Ruby) Wu has been offering me all kinds of support both in regard to research and life, so I hope to give her a special thank-you.

Lastly, I want to thank all my family and friends for being supportive during my four years at Emory.

Contents

1	Introduction	1
2	Background and Related Work	5
2.1	SSL/TLS and Web Certificates	5
2.2	Attacks on SSL/TLS & CA	6
2.3	Hash Visualization	8
2.4	Face Recognition via Human Eyes	11
2.4.1	Face vs. Non-face Objects	11
2.4.2	Familiar vs. Unfamiliar Faces	12
2.4.3	Own-group bias in face recognition	12
3	Extension Design	13
3.1	Pipeline and StyleGAN3	13
3.1.1	Current Prototype	13
3.1.2	Alias-Free Generative Adversarial Networks	14
3.2	Formative Evaluation	16
4	Experiment	19
4.1	Motivation and Assumption	19
4.2	Hypothesis	20
4.3	Participants	21

4.4	Method	22
4.4.1	Gamification	22
4.4.2	User Study	26
4.4.3	Planned Analysis	30
5	Discussion	32
6	Conclusion and Future Works	35
	Appendix A	37
	Bibliography	38

List of Figures

1.1	The lock button used by Firefox.	2
2.1	SSL handshake example.	7
2.2	An example of Passface [42]	10
2.3	Two fingerprints visualized [34]	11
3.1	Face for www.emory.edu	14
3.2	Extension pipeline	15
3.3	Fingerprint used as seed for library.emory.edu	16
3.4	Faces generated by Signaface for Stanford, FJTV, and Washington Post clockwise	17
3.5	3 Faces generated by Signaface that look abnormal	17
4.1	Sample user experience.	23
4.2	Gamified Extension	24
4.3	Sample Rounds	25
4.4	Project Website	27
4.5	Login Page	27
4.6	Scoreboard	29
5.1	Face for emory.edu VS. face for biology.emory.edu	33
5.2	Face for emory.edu in Dec 2022(bottom) and Mar 2023(top)	34

List of Tables

4.1	Sample User Score	22
4.2	Users' Requests vs Domains	26
4.3	Participants Info	29
4.4	Confusion matrix	31

Chapter 1

Introduction

The trustworthiness of websites' capability to protect user privacy is a critical aspect of online browsing, and it directly impacts users' willingness to interact with them. To gain users' trust, popular browsers such as Chrome and Firefox employ a variety of privacy protection mechanisms. Typically, when visiting a website, users are presented with a lock button, indicating that the connection is secure (Figure 1.1). However, this message only shows that the browser trusts the server's presented certificate and does not guarantee the website's actual security. In technical terms, if an HTTPS connection, which is an encrypted "TLS/SSL" version of the standard HTTP web-query protocol, has been established between the browser and web server, users are led to believe that a secure session has been established with the website [18]. In order to verify this secure HTTPS connection that provides transport-level encryption of web traffic, servers purchase an SSL/TLS certificate from a certificate authority (CA) that identifies them as the domain owner. Certificate authorities serve as trustworthy third-party entities that verify the validity of certificates by confirming the identity of those requesting a certificate and digitally signing the certificate [3]. When a CA issues a certificate for a particular website, no content is actually examined, resulting in a business relationship rather than a trust-based one between

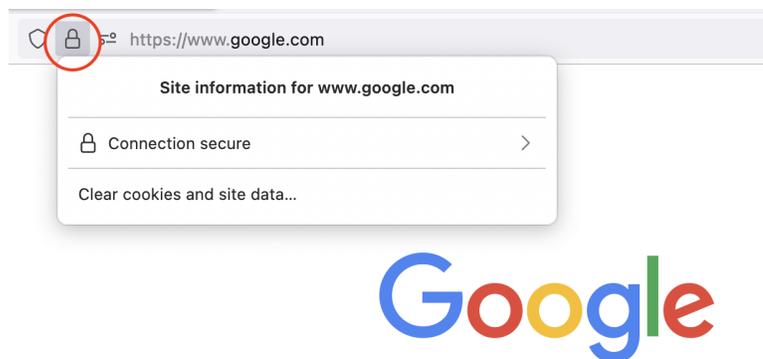


Figure 1.1: The lock button used by Firefox.

the website's operator and the CA. This verification mechanism also suggests that as long as the certificate is valid, which is very much achievable by attackers, the lock icon will continue to label the HTTPS connection as secure even if the certificate is no longer the original one. After observing the lock being displayed, ordinary users often consider their sensitive data safe with the website.

Unfortunately, cybercriminals and nation-states have a history of targeting individuals by luring them into visiting legitimate-looking websites, which are in reality equivalent to sites with valid SSL/TLS certificates, to compromise them. Snowden disclosed, for instance, that the National Security Agency (NSA) of the United States side-loads malware onto targets' computers when they visit a website, thereby undermining security protocols and posing threats to the internet's integrity [32]. These attacks are automated by intercepting users' connection to their intended websites and directing them to the NSA's server, where their computers are infected with malware, and their information is exploited [2]. Similarly, targeted Border Gateway Protocol (BGP) attacks allow hackers to request CAs to obtain their own letsencrypt.org credentials for the website [4]. In this case, attacks occur during the domain verification process, and certificate authorities often lack a way to identify attackers using intercepted routes resulting in the issuance of certificates to individuals who are not the domain's owners [6]. Lastly, there are countless typo-squatting sites, which are web-

sites that appear identical to their targets (for example, "bankofamerlca.com instead of bankofamerica.com") but are designed to trick users into divulging sensitive information or downloading malware. Current browser security mechanisms are powerless against these attacks, which we detail in Chapter 2.

The crux of the problem lies in the inability of a computer to know that a legitimate user is being fooled, whether it's through a hijacked domain, a misspelling, a DNS attack, or other forms of showing a website, even one that supports HTTPS, that imitates another one in a manner that's dangerous for the user. Hence, it is essential to devise an approach that is user-friendly for informing common users of the potential threat they face. One way to warn users that their connection may be insecure is to notify them once the site's security information contained in the certificates has been changed, i.e., having a completely different certificate. Our research not only attempts to identify possible directions to rogue servers but also proposes hash visualization in the form of a fake face as a novel method to alert users of this situation so they can report such attacks. Our core hypothesis is that for websites that they frequently visit, such as online banking, social media, or simply the Google search index page, users could notice certain differences regarding the page's display if its identity can be presented in a novel way, which is a GAN-generated human face. Unlike fingerprints or other cryptographical information included in a web certificate, human eyes are more reactive toward images, and users will not be pressured into trying to discern the potential differences between current and past certificates [7] [33] [36] [38].

Chapter 2 of this thesis provides an overview of the literature on attacks related to SSL/TLS certificates and hash visualization works that inspired the creation of the "Signaface" extension and human brains' functioning in terms of face recognition. Chapters 3 and 4 detail the pipeline and implementation of the "Signaface" extension, followed by a concrete plan for the user study, in which we will use gamification to

conduct an experiment to examine the human capacity for recalling faces. In the final chapters, I discuss the limitations of our extension and methodology, leading to possible future directions for this research.

Chapter 2

Background and Related Work

This section provides background information on SSL/TLS handshakes and certificate authorities, followed by an explanation of the possible attacks against them. The final section provides an overview of previous research conducted primarily in two fields: hash visualization and the human capacity for face recognition.

2.1 SSL/TLS and Web Certificates

Secure Socket Layer(SSL) and its newer version Transport Layer Security (TLS), are cryptographic protocols that operate on the Transport layer to ensure the secure transformation of data during end-to-end communication [35]. The SSL/TLS handshake protocol consists mainly of negotiating which secret key will be used to encrypt future communications between web servers and clients to prevent eavesdropping [26] [44]. HTTPS is one type of application protocol that SSL and TLS provide cryptographic security for [18]. To establish a secure HTTPS connection, the SSL/TLS handshake starts with the client sending a "hello" message, to which the server responds with a "hello" [13]. The server will then send a certificate to the client issued by a trusted certificate authority and containing the server's public key [13] [35]. Certificate authorities (CA) are a chain of CAs consisting of a root CA

typically self-signed and pre-trusted by end users and CAs signed by other CAs [3]. The certificate at the bottom of the chain is the one used to generate a face. If a server wishes to obtain a certificate from the CA, it must send the CA its public key. The CA will then create a certificate containing this information, the server's ID, and other parameters [13]. A message digest algorithm will be utilized to generate a certificate fingerprint, which will then be encrypted with the CA's private key to produce a signature [13]. The seed used as input for GAN in this research is a part of the mentioned fingerprint.

In order to authenticate the server by verifying the validity of the certificate received, the client has to take several steps. First, they must use the CA's public key, obtained by maintaining a list of trusted CAs and their respective public keys, to decipher the signature on the certificate [13]. The fingerprint obtained from the deciphered signature must then be compared to that independently computed by the client, with a match indicating the validity of the certificate [13]. After determining a secret key using a public-key algorithm to conclude the handshake, an HTTPS connection is secured [13]. The steps of the SSL handshake are depicted in Figure 2.1.

2.2 Attacks on SSL/TLS & CA

Many types of attacks can occur when users access seemingly safe websites and secured connections. According to National Security Agency (NSA)'s documents, a program with the name FOXACID enables the NSA to place its malware between targeted computers and the Internet, redirecting any signals the malware intercepted to rogue servers operated by the NSA called "FoxAcid" servers [32] [19]. Rogue webpages will be hosted on these servers impersonating those of legitimate websites such as LinkedIn and Gmail, thus tricking targeted users into entering their account info and

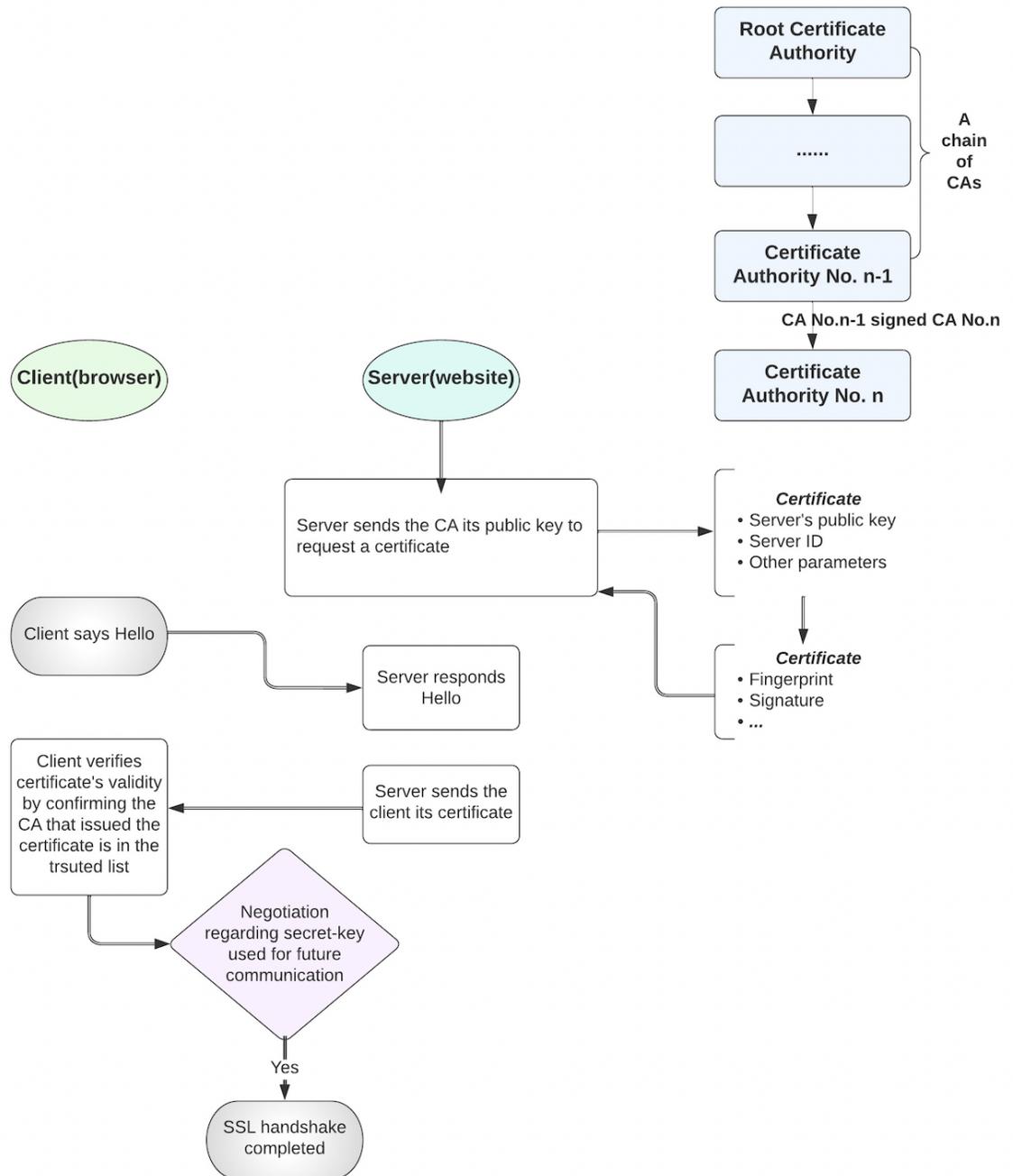


Figure 2.1: SSL handshake example.

collecting personal data [19]. The fact that users are oblivious to the fact that they have been redirected to alternative websites makes it difficult to defend against such attacks.

Border Gateway Protocol (BGP) hijacking is an attack that occurs during digital certificate issuance and domain owner verification [5]. Attackers typically try to impersonate the victim and obtain a valid TLS certificate for the victim domain while taking the victim's network prefix(which is an aggregation of IP addresses) [5] [23]. Since the IP address failed to be routed to the correct server, CA cannot verify the identity of the party that submits a Certificate Signing Request(CSR), giving attackers a chance to get a certificate not belonging to their own servers [23]. Therefore, when users are redirected to the attackers' servers, they will still see a valid SSL/TLS certificate for the intended domain.

Typosquatting is another form of attack that seeks to profit from users' inability to distinguish the original website from its imitation. Many typosquatting sites are advertisement platforms, while others function as phishing sites or malware servers [40]. All of these sites rely on users' misspellings, typing errors, and the popularity of websites with similar names [40]. Typosquatting domains are often registered in bulk, aiming to capitalize on users' mistakes and their inability to detect the difference between the two URLs. Our research aims to simplify this challenging comparison of two complex strings to two faces.

2.3 Hash Visualization

Upon realizing the importance of human factors in promoting security, existing studies have worked to spare users the tedious task of memorizing or manually comparing complex strings by converting the latter to structured images, a technique known as hash visualization [34] [16].

This replacement of traditional alphanumeric passwords with graphical passwords has been the focus of several studies [39]. The usage of hash visualization has frequently been applied to the user authentication process, given its usability in addressing users' difficulty remembering complicated passwords [34] [17]. Users frequently choose passwords that are either short and easy to remember or the same for multiple accounts [1] [17]. These choices render users vulnerable to all kinds of security risks, such as dictionary attacks, where attackers search for possible candidates for passwords in a small dictionary [31] [49] [41]. Graphical passwords can be effective in dealing with both brute-force search and dictionary attacks because a large number of possible images can make attackers' attempts to find the correct password much more expensive [41]. In terms of recognition and recall, it is easier for humans to memorize pictures than strings, according to psychological studies [36] [7] [33] [38]. More importantly, people often remember their passwords approximately instead of exactly, so if a system can convert the portion of a password that is typed in so far to an image, users will be able to recognize the image that corresponds to the correct password, without having to memorize the exact password string [34]. With all the advantages discussed above, the efficacy of graphical passwords has also been proved by existing studies, with a clear advantage compared to traditional password and PIN authentication [17] [46].

Some study concerning graphical passwords has been particularly relevant to our research due to their adoption of human faces as passwords called "passface" [39]. The main idea of passface is for users to prove their identity by identifying a sequence of faces [20] [42]. Each of these faces will represent a short string like in Figure 2.2 [42], and users will initially choose their own sequence of faces. Though passface has not demonstrated sufficiently regarding its effectiveness, some user studies have indicated that users were capable of memorizing their passfaces for a long period of time [39] [43].



Figure 2.2: An example of Passface [42]

Besides user authentication, hash visualization also applies to the root key validation during public key infrastructure(PKI) [34]. Root key validation mainly involves users verifying the validity of a locally stored root key, by checking if a specific CA issues it [34]. Just like the problem we are trying to solve, this validation process requires a not-user-friendly comparison between two fingerprints, and a conversion from the two fingerprints to two images no doubt makes the comparison a lot easier(Figure 2.3 [34]).

The design of the "Signaface" extension combines the above two ideas of graphical passwords with faces and visualization of fingerprints and aims at making the detection of changes in SSL/TLS certificates more accessible and user-friendly.

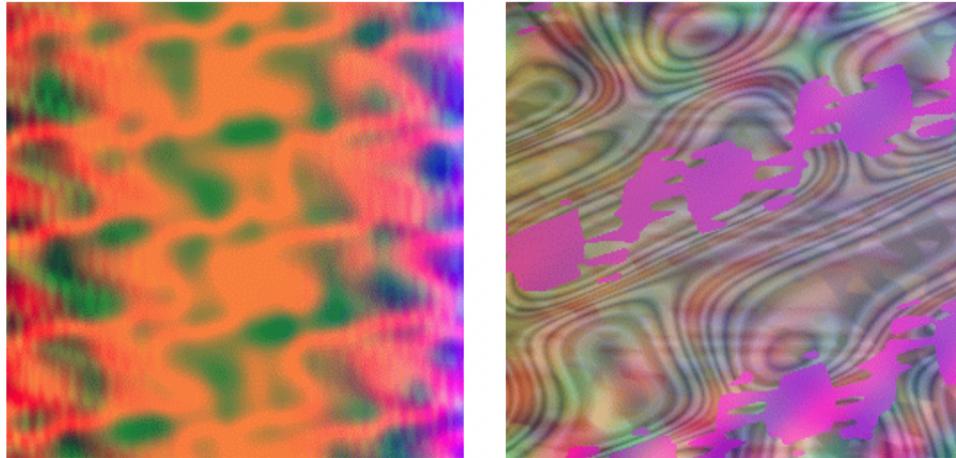
(a) F_1 (b) F_2

Figure 2.3: Two fingerprints visualized [34]

2.4 Face Recognition via Human Eyes

2.4.1 Face vs. Non-face Objects

As discussed in the preceding section, psychologists have discovered that humans find it easier to recall and recognize images than meaningless strings [7] [33] [36] [38]. Face recognition, as a particular case of image memorization, has frequently been studied compared to human brains' performance on other non-face objects. Several studies have shown that face recognition and common-object recognition may depend on functionally independent and anatomically separate systems [21]. Evidence indicated that some patients who experience a neuropsychological impairment called prosopagnosia have difficulty recognizing human faces even though animal faces are pretty recognizable for them [29]. Other studies have discovered significant differences in prosopagnosia patients' ability to recognize human faces and everyday objects. [22]. It can thereby be deduced that there may be an area within the human brain that is specialized for human face recognition and less critical for recognition of other visual images [21].

2.4.2 Familiar vs. Unfamiliar Faces

The extension developed by our research group relies heavily on users' ability to recall whether they have seen a certain face before, which is, in other words, a decision regarding whether users consider a face to be familiar. Researchers from all fields have studied this face familiarity task, and some found making this decision much easier than having to recall any specific information [9] [11] [25] [50]. After a face has been judged as familiar, people achieve higher accuracy in identifying them, even if the circumstances under which they have to make that identification are much less optimal than when they are asked to match unfamiliar faces [10].

2.4.3 Own-group bias in face recognition

One of the vital factors that we attempt to include in our research is whether users and faces being memorized belonging to the same social group will make a difference in users' ability to judge the familiarity of those faces. One concept especially relevant to this issue is the own-race bias (ORB) phenomenon, where findings suggest own-race faces to be better remembered [8] [30]. Cross-race recognition errors are common, and many people have experienced the feeling of "they all look the same" [27]. One possible explanation for ORB is the contact hypothesis, which attributes ORB to limited contact with members from other races [12]. Studies have shown that although subjects identify their own-race faces more accurately and confidently, ORB is less prevalent among groups with high contact with other races [12]. Other types of face recognition biases resulting from social categorization, such as own-sex bias [47], own-age bias [48], and own-species bias [37] has also been documented in earlier literature, suggesting a need to take users' and faces' social group into account when conducting research related to face recognition [15] [27].

Chapter 3

Extension Design

This section describes the pipeline of the Signaface extension, which is currently implemented as a temporal add-on for Mozilla Firefox. A description of the generative adversarial network(GAN) used for our extension styleGAN3 is also included [28]. I then move to the formative evaluation section, where I elaborate on the preliminary user experience of the "Signaface" extension and some refinements that have been or could be made.

3.1 Pipeline and StyleGAN3

3.1.1 Current Prototype

Our current prototype is capable of generating a unique synthetic face for most domains and display it in the upper left corner. Figure 3.1 is the face generated by www.emory.edu's SSL/TLS certificate. Our implementation mainly involves the communication between three parties: users/browsers, the "Signaface" extension, and Emory's Titan GPU server. As shown in Figure 3.2, the pipeline begins with the user visiting webpages under the domain name abc.com. Once the webpage finishes loading, "Signaface" will request a certificate from the server. If abc.com supports

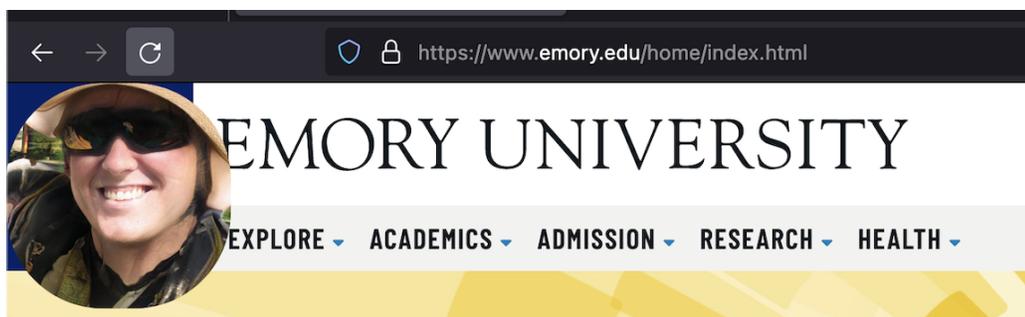


Figure 3.1: Face for www.emory.edu

and uses HTTPS and a certificate is successfully retrieved, part of the fingerprint is retrieved and passed to the GPU server Titan as a seed 3.3. Titan will generate a face using styleGAN3, an alias-free generative adversarial network, and send it back to the extension after receiving a request for a face and the seed. The extension will insert the face into the HTML of the website, allowing users to view the face in the upper left corner of the page. At this point, they will have to try to recall if they have seen this face before and draw conclusions about the security of their information based on the outcome.

3.1.2 Alias-Free Generative Adversarial Networks

Generative Adversarial Networks (GAN) is an artificial intelligence algorithm used for generative modeling, intending to learn the probability distribution of training samples and generate new samples based on this distribution [24] [45]. By putting two networks through competition training, GAN excels in its ability to output realistic-looking, high-resolution images [14] [24]. StyleGAN3 was chosen as our face generation algorithm due to StyleGAN's exceptional ability to generate natural human faces. The StyleGAN has been regarded as useful for generating realistic artificial faces, and StyleGAN3 is an alias-free GAN that attempts to address unnatural visual flaws such as texture sticking, where substance like hair appears to be stuck on the screen [28].

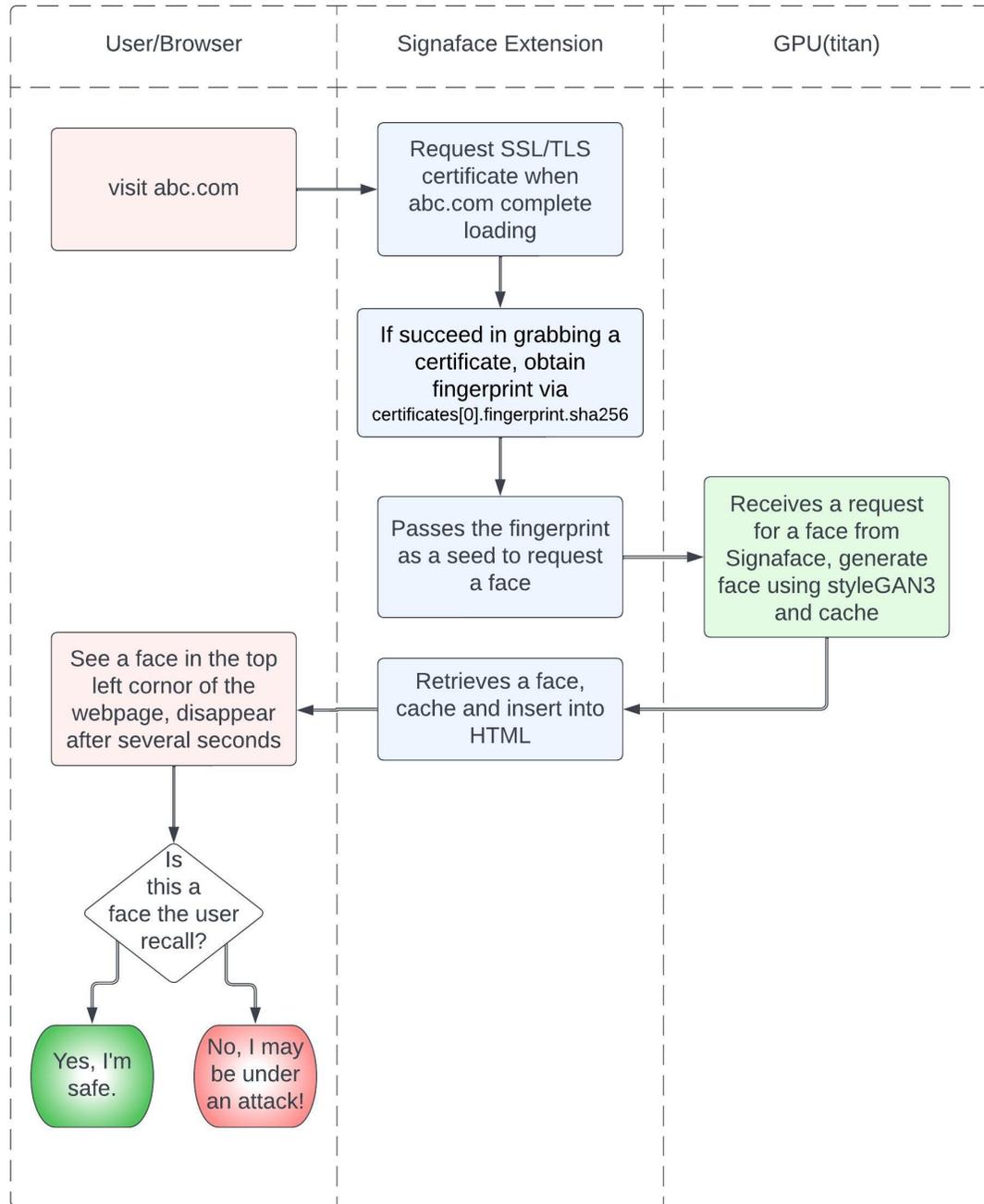


Figure 3.2: Extension pipeline

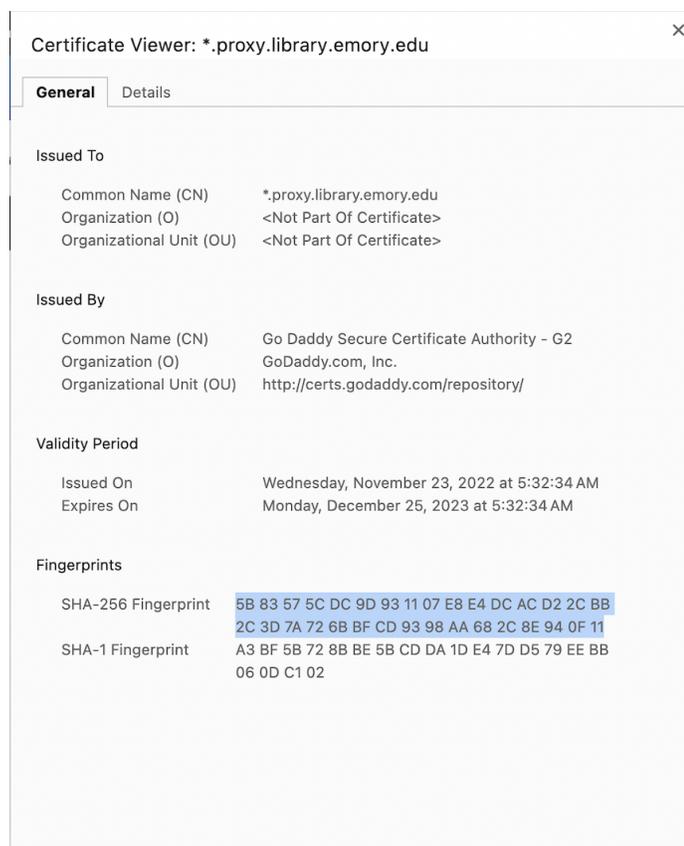


Figure 3.3: Fingerprint used as seed for library.emory.edu

3.2 Formative Evaluation

To test the usability of our extension, we have asked four current students at Emory University to try using "Signaface" and provide any feedback. Their usage is approximately three weeks, and they are all frequent computer users. At the conclusion of the third week, we asked the four students 1) if they could recognize a familiar face from a website they frequently visit and 2) if they had observed any "strange" display of faces or had any suggestions for the extension design. We are told that the extension, in general, functions well in terms of displaying a face and allowing for a certain amount of exposure to the face. Students who have tried the extension also mention that it is helpful for different web pages of the same domain to show the same face, which is also reloaded when clicking onto a different web page and drawing the user's

attention to the face. Based on their feedback and testing on our own, we adjusted the face to make it fade away after several seconds and also ensured that other web page HTML elements would not cover the image. Additionally, we made a few



Figure 3.4: Faces generated by Signaface for Stanford, FJTV, and Washington Post clockwise



Figure 3.5: 3 Faces generated by Signaface that look abnormal

discoveries that may be useful in refining "Signaface." First, a minority of websites alternate between two distinct certificates. Google's search index page, for instance, utilizes either the `www.google.com` or `*.google.com` certificate, thus displaying one of the two faces, yet the latter with a much higher frequency. Therefore, we may have to devise a way to prevent users from becoming alarmed when they see a "stranger's" face on the website's other certificate. Second, our research team and current Emory students have reported that a disproportionate number of faces displayed are wearing

sunglasses, hats, or both, making it more difficult for them to memorize the face due to a lack of distinguishing facial features (Figure 3.4). To resolve this issue, examining real faces used during StyleGAN3's training procedure may be necessary to generate faces with fewer accessories.

Many generated faces are also deformed (Figure 3.5), so we may need to investigate the cause of these deformations or consider alternative face generation models. However, a face with an abnormal appearance does not necessarily contradict our current goal of having users remember them.

Chapter 4

Experiment

In this chapter, I discuss the design of the user experiment, including major hypotheses, the gamification in the form of a web browser extension that we will be using for our study, a typical user experience participating in the experiment, and a hypothetical evaluation of the result.

4.1 Motivation and Assumption

A crucial part of our research is to conduct a scalable user study online in which participants will be asked to install and use the extension over the course of several months. We will randomly alter the face displayed for each domain. The participants' primary task will be to indicate whether they believe a different face is displayed for a specific website, presumably one they visit frequently enough. Instead of conducting a psychologically controlled study, we will focus on ecological validity, which more closely aligns with our objectives. To accomplish this goal, we will ask participants to do no more than their usual routine for internet browsing. To ensure that their experience during the user study most closely resembles that with the extension in the real world, we designed the study to be longitudinal and span six months rather than a short-term one. This allows for the simulation of seeing the same face, memorizing

it, and noticing a change through frequent visits to a website.

One of our assumptions for this study was that we could conduct multiple rounds of gamification over six months. Depending on the preliminary results, each round lasts between one week and one month and begins with presenting a different face. We hypothesize that the duration of each gamification round is sufficient to ensure adequate exposure for the frequently visited websites of users. To clarify, we assume the practicability of the previously mentioned cycle of memorizing the face well enough to detect a change within the time constraints of each round.

4.2 Hypothesis

We hope to derive from this experiment the degree of detection accuracy users can attain after repeatedly encountering the same face on the website's pages. We are interested in both the possibility of this detection and the amount of exposure required for a significant percentage of users to detect the change. In order to better tailor our extension to the needs of our users, we also plan to collect the demographic information of our participants to better understand how people of different genders, races, and ethnicities differ in their ability to recognize synthetic faces from various demographic groups. Following are our hypotheses for this experiment.

H1 To become familiar with a human face, individuals must be exposed to it for at least days if they see it for several minutes per day.

H2 After a certain amount of exposure, the average user can memorize a face to the extent that they can decide whether a face is familiar.

H3 It is easier for users to determine whether a face is recognizable if it is of the same race, gender, or age group.

4.3 Participants

Besides familiarity with online settings, frequent web-browsing activities are also vital for evaluating the usability of our extension due to the fact that the latter relies heavily on the accumulated time individuals see displayed faces for each website they visit. In order to fulfill this requirement, we sought users with a certain level of computer literacy who are capable of installing a web browser extension with the help of the guides we provide and who also use browsers on a daily basis to visit a variety of web pages, so that using our extension can become a normal part of their routine, as well as a simulation of users' actual experience with the "Signaface" extension. We will conduct an initial screening based on this criterion during the recruitment phase. Since the ultimate goal of this study is to promote user privacy by detecting possible hacker activities, we will recruit participants using any of the following methods: (1) direct email recruitment or email listservs, (2) social media advertisement (e.g., Facebook, Twitter) and HackerNews or related news aggregators that link to our website where they will find a page with complete study details and eligibility requirements. In addition, we will likely use cybersecurity workshops, conferences, meetups, websites, and chat groups to recruit participants. We plan to recruit at least 500 participants who are at least 18 years and capable of speaking English to conduct a scalable experiment. All communications with participants will take place virtually, and the only requirement for this experiment is that participants use the gamified extension after it has been installed.

4.4 Method

4.4.1 Gamification

A gamified version of the "Signaface" extension was developed for this user experiment. Figure 4.1 illustrates in detail how the gamified extension functions, from a user clicking on a website to our data collection. The extension will request a face once the user visits a website to our database, from which a face will be displayed on the screen for several seconds as long as the user is still viewing the page. In each round, the face retrieved will be the same for the same user and domain. Suppose the extension receives a request for a face from the user's account. In that case, we assume the user has seen the face and marks the identification attempt as either correct or incorrect, depending on whether the user has clicked the button and whether it is a different face. If the user correctly clicks the button, they receive 1 point, and 0.5 if they correctly choose not to click the button. They will get -0.5 points if they incorrectly choose not to click the button and -1 if they incorrectly choose to do a click. Table 4.1 is a sample score record for a user. It can be observed from Table 4.1 that user A participated in rounds 1 and 3, meaning the extension has received requests from user A's account during those two rounds asking for a face to be displayed.

User A	Round 1	Round 2	Round 3
Participation	Y	N	Y
Number of requests made	126	N/A	304
Correct Detection(by clicking the button)	6	N/A	5
Correct Detection(by not clicking the button)	100	N/A	151
Incorrect Detection(by clicking the button)	4	N/A	3
Incorrect Detection(by not clicking the button)	16	N/A	145
Score	44	N/A	5

Table 4.1: Sample User Score

This gamified version of the extension has a similar structure to that of the original version except for the part where faces are generated. As seen in Figure 4.1, instead of

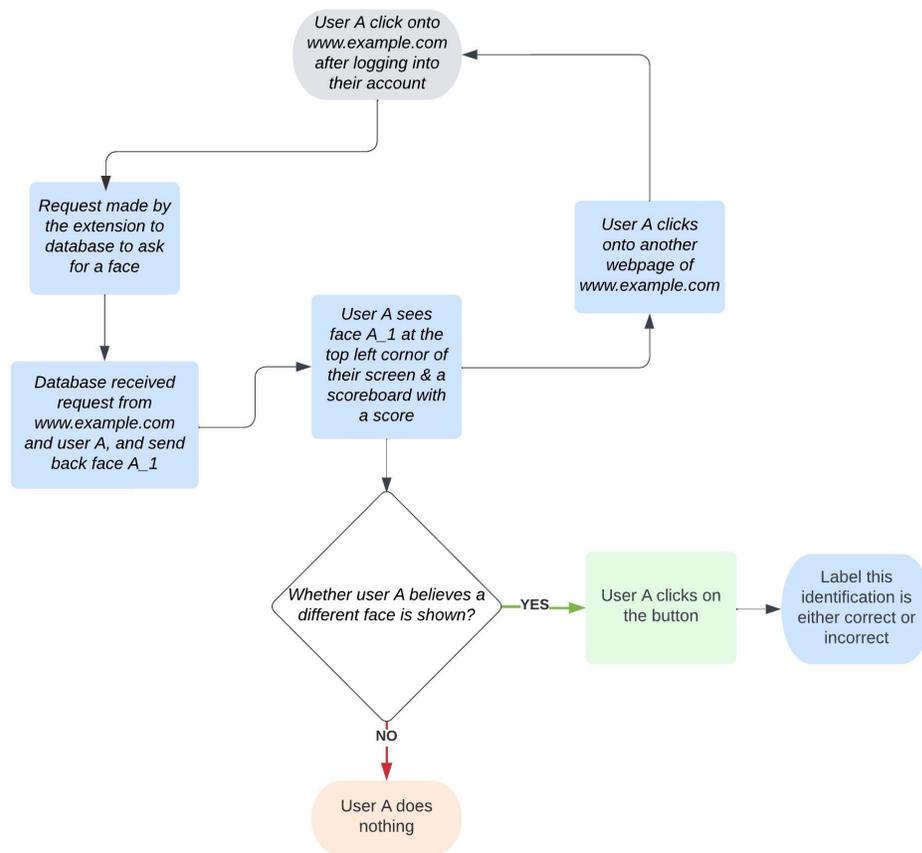


Figure 4.1: Sample user experience.

requesting a part of the SSL certificate to use as seed for the generation of synthetic faces for each domain, we will pre-generate faces using either styleGAN3 or other generative adversarial networks and store them in a database prior to the start of the experiment. This is due to the fact that the “Signaface” extension converts part of the site’s certificate’s fingerprint to a human face, which renders it difficult for us to manipulate the face in terms of the frequency we would like to change it. Using pre-generated faces, on the other hand, allows us to decide the face we want to display for each website and the length of this display. We will generate approximately 100,000 faces using styleGAN3 to ensure we have a sufficient number of faces to allocate for each website each participant visits and also to apply changes to test users’ “different-face detection” capacity.

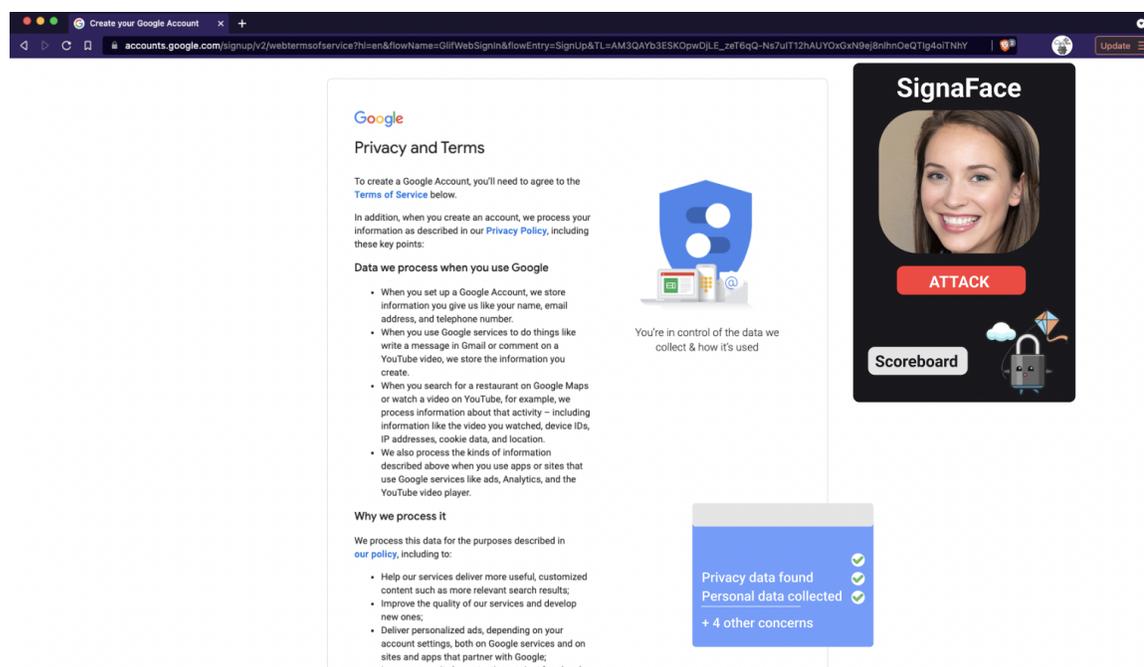


Figure 4.2: Gamified Extension

The user interface of this extension includes three components: the pre-generated face, a button for users to click on if they believe a different face has been shown, and a mini scoreboard that displays users’ overall score (Figure 4.2). We can see website requests on our end to our extension asking for a face as soon as a webpage has finished

loading so that we can still detect whether participants have seen the face when no button has been clicked. The gamification will be carried out using rounds as units, of which the number of days within each round will range from a week to one month. Each round of the experiment is defined by our research team changing the face shown for all participants. Specifically, a new face will be displayed at the beginning of each round for each website that each participant visits. Each round's conclusion will be chosen at random to prevent participants from concluding the change in faces from sources other than their own observations of the faces. For instance, if we set the duration of one round to 9 days, we may wish to set the duration of the next round to 20 days and the third round to 15 days rather than setting all three rounds to 9 days. During the preliminary phase of the study, we will increase or decrease the number of days included in a round based on the performance of the participants in order to determine a lower threshold for the number of days required for a certain percentage of individuals to correctly identify a change of face. After determining the "minimum days," the length of one round will be chosen at random as long as the threshold is met. To set the timeout for the display of a face, we will take a similar approach, initially setting the display time to 10 seconds and then adjusting it throughout the experiment. Figure 4.3 demonstrates a sample of rounds.

No. Round	Date	Number of Days	Display time for faces(seconds)	Faces used
3rd	2023/05/01-2023/05/09	9	10	No.25001-40000
...
7th	2023/06/17-2023/06/29	13	8	No.100000-110000
8th	2023/06/30-2023/07/16	17	13	No.110000-125000

Figure 4.3: Sample Rounds

Round No. 7			
User	Domain	Number of Requests	Face No.
B	www.a1.com	24	No. 72
	www.a2.edu	10	No. 837
	www.a3.com	15	No. 1123

Table 4.2: Users' Requests vs Domains

We will keep a record of users' reports regarding the change of face in terms of the score for each round. We will also keep track of the number of requests made to our extension to ask for a face for a website to obtain a more comprehensive evaluation of the amount of exposure users have with regard to the same face. Due to the importance of this exposure, we are likely to discard the results collected from websites where few requests are made. To protect user privacy, we will hash the domain names that appear in their requests and assign each of them site numbers. Table 4.2 is an example of how our record of the result for one round may look like for a single participant.

4.4.2 User Study

This section will detail participants' experiences, from being interested in our study to officially joining and participating and ending the study via self-withdraw or upon completion. When potential participants click on our website's homepage, they can find a link to download the gamified extension from our website and install it for their browsers(Figure 4.4).

After successful installation, potential participants will be directed to create their accounts with a valid email address and receive a confirmation link and an electronic form they will be required to complete to officialize their participation(Figure 4.5). In this form, they will be asked to answer some basic questions, including their age group, their usual frequency using that browser, their intended times of participation in terms of weeks, and a short answer question asking why they want to participate

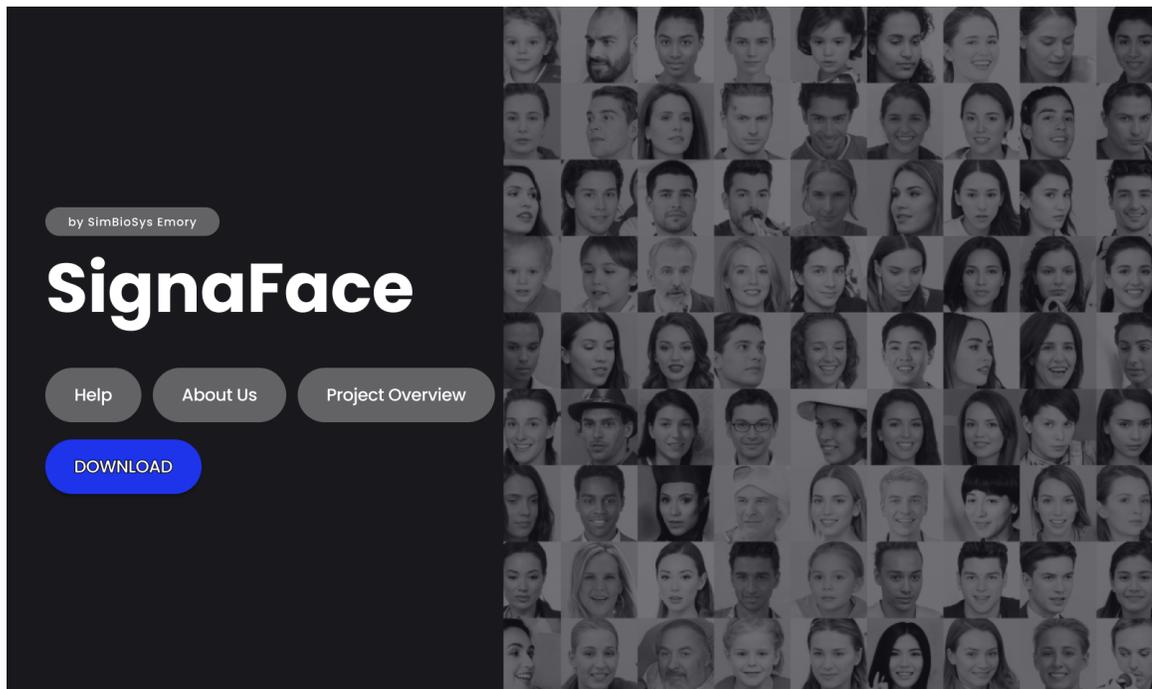


Figure 4.4: Project Website

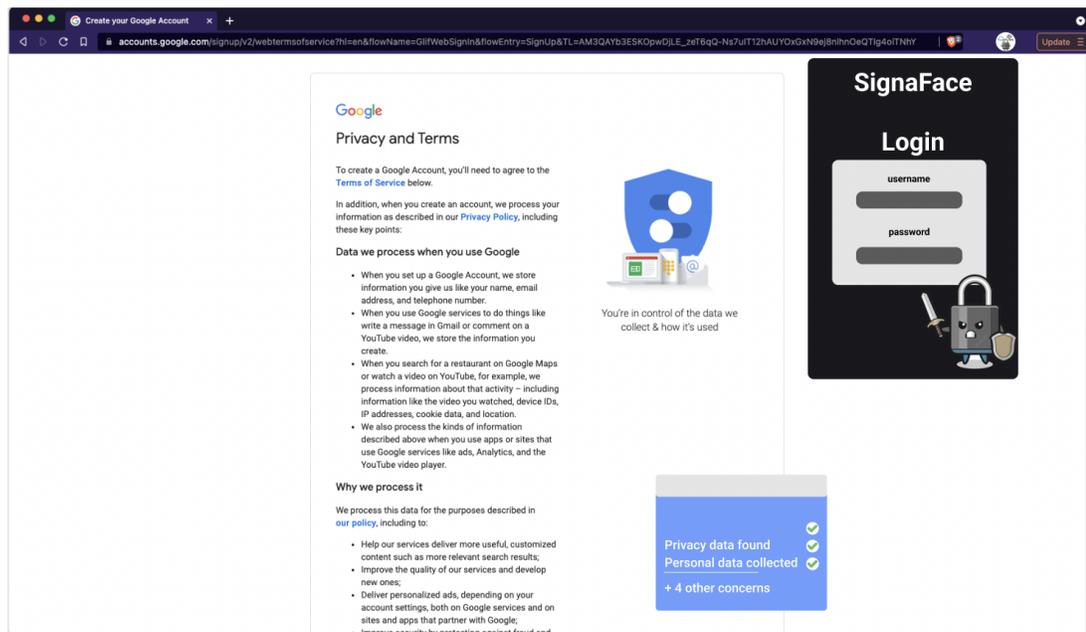


Figure 4.5: Login Page

and if they can click a "yes" button included in the extension if they believe the face "mascot" generated for the website has changed. By asking these questions, we hope to filter out potential participants who are minors, more frequently use browsers that do not support our Signaface extension, cannot participate long enough to allow for sufficient exposure, or are not committed enough to actually use the extension. Based on the participants' answers, we will decide whether we believe they satisfy the basic condition to become official participants in our study and email the latter with the experiment's consent forms.

After participants send back the consent forms via email, their participation will be officialized, and they will receive an email with a guide on how to get started with using the extension and a second online questionnaire(see Appendix A) that will request information such as gender, race, age bracket, and the browser they will be using during the experiment. Participants' main task will also be conveyed through this email: to surf the web as normally as possible and click on a button in the browser extension to indicate if they believe the face shown for the current webpage has been changed. We will recommend that the participants not pay too much extra attention to the face, which may render our result inaccurate. One way to avoid this is to refrain from refreshing the page to allow for longer exposure time or take screenshots or pictures that may exceed natural browsing. Participants will be told that constant refreshing of the same page will appear as continuous requests from the same domain and may render their result of that round of the game to be discarded. In the last part of the email, we will ask participants to click on the button several times to confirm the receipt of requests on our end.

During the duration of the study, if participants have any questions or want to withdraw from the study, they can either write to the email address we use to communicate with them to indicate their intentions, or they can also find the email addresses of the researchers on our website. If participants decide to leave the study, the scores

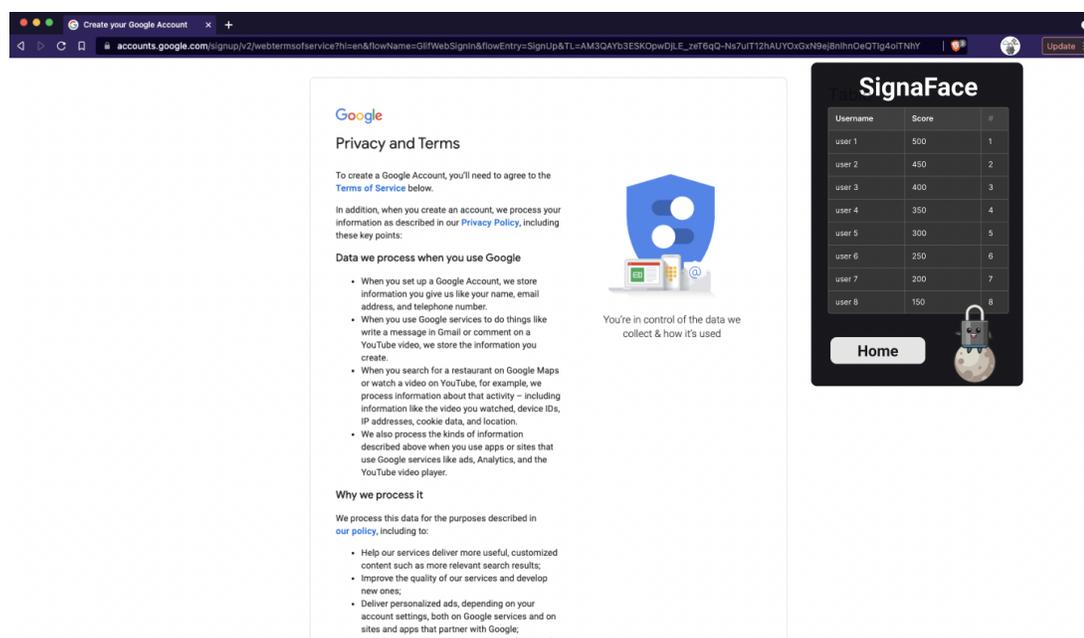


Figure 4.6: Scoreboard

they receive for the last round will not be recorded. Our group keeps track of all the information collected from participants, including everything collected from online forms and their accumulative scores. Table x is an example of information in our database for one participant (Table 4.3). More details about a specific participant's performance have been recorded in Table 4.1 and Table 4.2 shown earlier. Although the number of days in a single round will be made random to avoid participants figuring out potential patterns when we change faces, we generate an up-to-date scoreboard by the end of each week displaying the top five participants in terms of points gained (Figure 4.6).

Participant account	Most-used browser	Average usage time per day	Gender	Race	Age Group	Overall Score
abcd@gmail.com	Chrome	1h	F	Black or African American	18-22	9.5
acde@outlook.com	Firefox	1.5h	M	Asian	23-27	6

Table 4.3: Participants Info

The scoreboard is accessible through our website and also via the extension, and individual participants will also receive an email containing their own score for the given week. This email will include a weekly check-in questionnaire about partici-

participants' experience with the game, which will help us in terms of adjusting the extension to provide a better user experience. For each round of the game, if the participants' browsers make few requests to the extension, indicating limited exposure and lack of usage of the extension, their participation for that specific round will be recorded as inactive and the result discarded. At the end of the study, participants with top scores will be given the chance of being selected to receive gift cards of varying amounts based on performance.

4.4.3 Planned Analysis

Over the span of the study, we collect data regarding users' exposure time to faces via the number of requests made to the extension and the number of days in each round. We will examine if there exists a correlation between this exposure and the accuracy rate regarding noticing the change of face, by developing regression and models, deriving correlation scores such as Pearson coefficients, and generating a confusion matrix with the actual change/no change of face as columns and perceived as rows as shown in Table 4.4. The accuracy rates will be calculated from the below results:

1. True positive rate(TP): The percentage of average accurate detection for all rounds
2. True negative rate(TN): The percentage of "no-button-clicked" when the face remains the same
3. False positive rate(Type I error): The percentage of the incorrect clicks of buttons when the face remains the same
4. False negative rate(Type II error): The percentage of "no-button-clicked" when the face changes

	Face actually changed	Face hasn't changed
Button clicked	TP	Type I
No button clicked	Type II	TN

Table 4.4: Confusion matrix

Whether participants view faces from their own social group is also likely to be treated as a relevant factor when building the model and evaluated with regards to their importance to accurate detection rate.

Chapter 5

Discussion

Since the efficacy of our extension depends on whether users can familiarize themselves enough with faces, our method is more applicable to websites users frequently visit. For websites that users visit less regularly, however, our faces serve more as a warning that they are entering an unfamiliar zone and a reminder to be cautious when entering personal information. The same applies to users visiting frequently-used websites during the initial period of using our extension: they have to take time to familiarize themselves with the face presented. Another issue our extension may present is "fake-warning" users by displaying a different face upon entering a new domain, when the latter may believe they are still on the same website and should thus see the same face. When visiting `biology.emory.edu`, for example (Figure 5.1), a different face will be shown than that of `emory.edu`, even though the former is often clicked onto from the latter.

There are also some websites that are not using HTTPS, meaning they do not have valid SSL/TLS certificates, so our extension is not able to generate a face. This absence of a face may be useful in offering users an extra reminder that their connection may be insecure.

We are also aware that participants who perform poorly during the initial phase

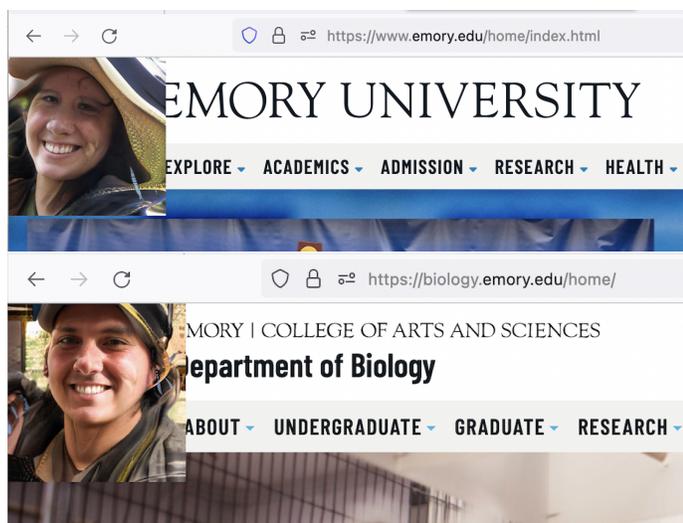


Figure 5.1: Face for emory.edu VS. face for biology.emory.edu

of the experiment may leave, whereas those who achieve great performance may tend to stay. As a result, the experiment's outcome may be biased in favor of persons skilled at detecting facial change.

Additionally, people also concentrate more during studies than in real life by paying a lot more attention when they know they are doing an experiment with the incentive of receiving a financial reward. When browsing in real-life settings, they may care less about what is shown on the screen beside the content they view. Despite all of this, users are still likely to pay some attention to a displayed face, even if they do not intend to do so.

One of our extension's most severe issues is the expiration issue of the SSL/TLS certificate. A certificate's validity is typically around six months, with the exact time varying depending on the website. Based on our observation, however, the fingerprint of a certificate tends to change prior to its expiration date; therefore, the time that the same face will be displayed may be shorter than anticipated. For instance, Figure 5.2 shows the face generated by the certificate of www.emory.edu in December 2022 and March 2023, respectively, and it is evident that the face has changed. One possible way to deal with this issue is to alternate between the face generated by

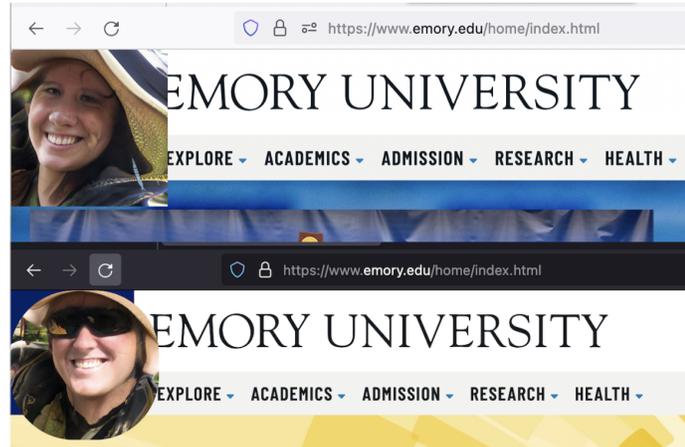


Figure 5.2: Face for emory.edu in Dec 2022(bottom) and Mar 2023(top)

the old certificate and the new certificate during the first or two weeks of adopting a new one. In this way, users can become acquainted with the new face while still confirming the security of their connection by viewing the old face.

Chapter 6

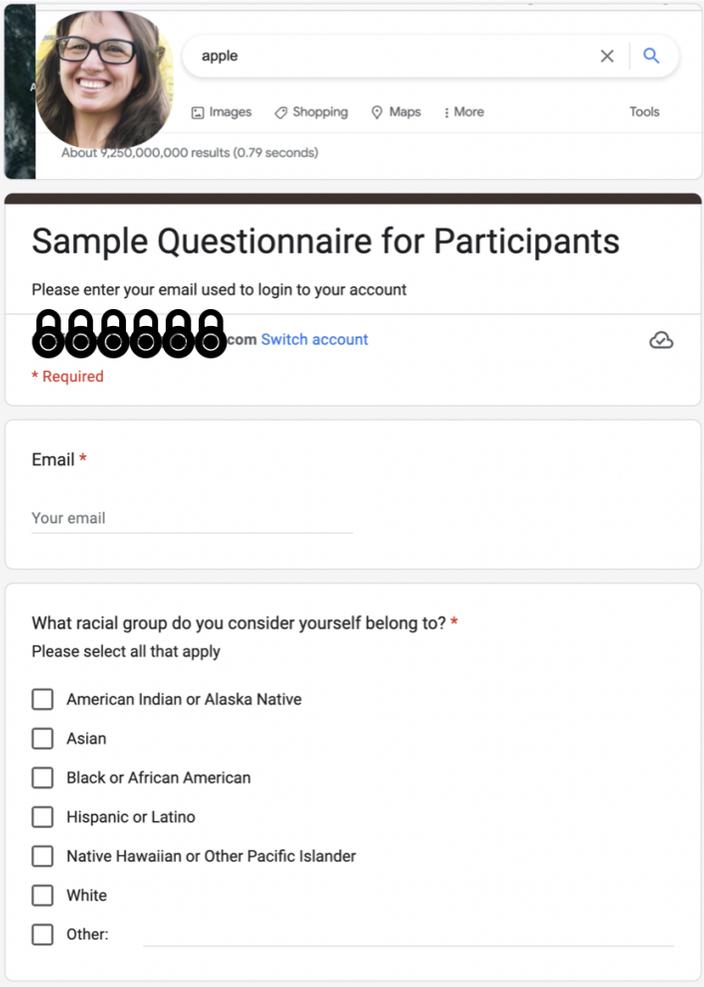
Conclusion and Future Works

Our research proposes a novel method in the form of a web browser extension that uses synthetic human faces to warn users of possible directions to malicious servers. Existing research has suggested the uniqueness of face recognition in terms of how it functions in human brains, and we hope to test the efficacy of this method through a scalable user study in which we ask the user to click a button provided by our gamified extension. If optimal results are obtained, the subsequent step will be to make the temporal add-on a formal web extension that can be downloaded from the web extension store. As previously mentioned, the human brain functions differently when memorizing and recognizing human faces versus non-face objects, a distinction we have not attempted to verify in this case. Thus, we may attempt to incorporate this variable into our experiment and compare the effectiveness of faces and familiar objects in assisting individuals in detecting changes.

The use of face in hash visualization likely has broader implications outside of cyber security and user privacy. One example of this application can be making a library or Python package called "Hashface," which displays a face to replay random hashes like passwords, private keys in cryptocurrency wallets, data anonymization, or hashes of any kind. The visual primitive we will be offering is to provide ways

to help people quickly discern between two things in a more user-friendly way, not only to address the insufficient consideration for human factors in security but also to take advantage of the specialty of face recognition when it comes to representing less straightforward information.

Appendix A



apple

Images Shopping Maps More Tools

About 7,250,000,000 results (0.79 seconds)

What is your gender? Select the one below

Female

Male

Other: _____

What is your age? Select the one below th

18-27 years old

28-37 years old

38-47 years old

48-57 years old

58-67 years old

68+ years old

What is the browser you are going to use

Google Chrome

Mozilla Firefox

Safari

Microsoft Edge

Other: _____

Sample Questionnaire for Participants

Please enter your email used to login to your account

○○○○○○○○ com [Switch account](#)

* Required

Email *

Your email _____

What racial group do you consider yourself belong to? *

Please select all that apply

American Indian or Alaska Native

Asian

Black or African American

Hispanic or Latino

Native Hawaiian or Other Pacific Islander

White

Other: _____

Bibliography

- [1] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.
- [2] Jacob Appelbaum. Jacob Appelbaum: To Protect And Infect, Part 2. <https://www.youtube.com/watch?v=vtQ7LNeC8Cs>, 2013. [Online; accessed 21-March-2014].
- [3] Jake A Berkowsky and Thaier Hayajneh. Security issues with certificate authorities. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, pages 449–455. IEEE, 2017.
- [4] Henry Birge-Lee and Liang Wang. Experiences deploying multi-vantage-point domain validation at let’s encrypt. In *USENIX Security*, 2021.
- [5] Henry Birge-Lee, Yixin Sun, Annie Edmundson, Jennifer Rexford, and Prateek Mittal. Using bgp to acquire bogus tls certificates. In *Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2017), Minneapolis, MN, USA*, 2017.
- [6] Henry Birge-Lee, Yixin Sun, Anne Edmundson, Jennifer Rexford, and Prateek Mittal. Bamboozling certificate authorities with bgp. In *27th USENIX Security Symposium*, 2018.

- [7] Gordon H Bower, Martin B Karlin, and Alvin Dueck. Comprehension and memory for pictures. *Memory & cognition*, 3:216–220, 1975.
- [8] John C Brigham and Roy S Malpass. The role of experience and contact in the recognition of faces of own-and other-race persons. *Journal of social issues*, 41(3):139–155, 1985.
- [9] Vicki Bruce. Recognizing faces. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 302(1110):423–436, 1983.
- [10] Vicki Bruce, Zoë Henderson, Craig Newman, and A Mike Burton. Matching identities of familiar and unfamiliar faces caught on cctv images. *Journal of Experimental Psychology: Applied*, 7(3):207, 2001.
- [11] A Mike Burton, Vicki Bruce, and Peter JB Hancock. From pixels to people: A model of familiar face recognition. *Cognitive science*, 23(1):1–31, 1999.
- [12] Patrick Chiroro and Tim Valentine. An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology*, 48(4):879–894, 1995.
- [13] Wesley Chou. Inside ssl: the secure sockets layer protocol. *IT professional*, 4(4):47–52, 2002.
- [14] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.
- [15] John F Cross, Jane Cross, and James Daly. Sex, race, age, and beauty as factors in recognition of faces. *Perception & psychophysics*, 10(6):393–396, 1971.
- [16] Rachna Dhamija. Hash visualization in user authentication. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, pages 279–280, 2000.

- [17] Rachna Dhamija, Adrian Perrig, et al. Deja vu-a user study: Using images for authentication. In *USENIX Security Symposium*, volume 9, pages 4–4, 2000.
- [18] Zakir Durumeric, James Kasten, Michael Bailey, and J Alex Halderman. Analysis of the https certificate ecosystem. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 291–304, 2013.
- [19] Hugh Eakin. The swedish kings of cyberwar. *The New York Review of Books*, 64(1):56–58, 2017.
- [20] Ali Mohamed Eljetlawi and Norafida Ithnin. Existing recognition base usability features of the graphical password. In *Recent Trends in Wireless and Mobile Networks: Second International Conference, WiMo 2010, Ankara, Turkey, June 26-28, 2010. Proceedings*, pages 379–388. Springer, 2010.
- [21] Martha J Farah. Is face recognition ‘special’? evidence from neuropsychology. *Behavioural brain research*, 76(1-2):181–189, 1996.
- [22] Martha J Farah, Karen L Levinson, and Karen L Klein. Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33(6):661–674, 1995.
- [23] Artyom Gavrichenkov. Breaking https with bgp hijacking. *Black Hat. Briefings*, 2015.
- [24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [25] Dennis C Hay, Andrew W Young, and Andrew W Ellis. Routes through the face recognition system. *The Quarterly Journal of Experimental Psychology Section A*, 43(4):761–791, 1991.

- [26] Kipp Hickman and Taher Elgamal. The ssl protocol. 1995.
- [27] Kurt Hugenberg, John Paul Wilson, Pirita E See, and Steven G Young. Towards a synthetic model of own group biases in face memory. *Visual Cognition*, 21(9-10):1392–1417, 2013.
- [28] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.
- [29] Jane E McNeil and Elizabeth K Warrington. Prosopagnosia: A face-specific disorder. *The Quarterly Journal of Experimental Psychology*, 46(1):1–10, 1993.
- [30] Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3, 2001.
- [31] Robert Morris and Ken Thompson. Password security: A case history. *Communications of the ACM*, 22(11):594–597, 1979.
- [32] Jared Naude and Lynette Drevin. The adversarial threat posed by the nsa to the integrity of the internet. In *2015 Information Security for South Africa (ISSA)*, pages 1–7. IEEE, 2015.
- [33] Allan Paivio, Timothy B Rogers, and Padric C Smythe. Why are pictures easier to recall than words? *Psychonomic Science*, 11(4):137–138, 1968.
- [34] Adrian Perrig and Dawn Song. Hash visualization: A new technique to improve real-world security. In *International Workshop on Cryptographic Techniques and E-Commerce*, volume 25, 1999.
- [35] Ashutosh Satapathy, Jenila Livingston, et al. A comprehensive survey on ssl/tls

- and their vulnerabilities. *International Journal of Computer Applications*, 153(5):31–38, 2016.
- [36] Roger N Shepard. Recognition memory for words, sentences, and pictures. *Journal of verbal Learning and verbal Behavior*, 6(1):156–163, 1967.
- [37] Rodrigo Sigala, Nikos K Logothetis, and Gregor Rainer. Own-species bias in the representations of monkey and human face categories in the primate temporal lobe. *Journal of Neurophysiology*, 105(6):2740–2752, 2011.
- [38] Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25(2):207–222, 1973.
- [39] Xiaoyuan Suo, Ying Zhu, and G Scott Owen. Graphical passwords: A survey. In *21st Annual Computer Security Applications Conference (ACSAC'05)*, pages 10–pp. IEEE, 2005.
- [40] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. The long “taile” of typosquatting domain names. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 191–206, 2014.
- [41] Hai Tao and Carlisle Adams. Pass-go: A proposal to improve the usability of graphical passwords. *Int. J. Netw. Secur.*, 7(2):273–292, 2008.
- [42] Farnaz Towhidi, Maslin Masrom, and Azizah Abdul Manaf. *An enhancement on Passface graphical password authentication*. PhD thesis, Universiti Teknologi Malaysia, 2010.
- [43] Tim Valentine. Memory for passfaces after a long delay. Technical report, Technical Report, Goldsmiths College, University of London, 1999.

- [44] David Wagner, Bruce Schneier, et al. Analysis of the ssl 3.0 protocol. In *The Second USENIX Workshop on Electronic Commerce Proceedings*, volume 1, pages 29–40, 1996.
- [45] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhua Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017.
- [46] Susan Wiedenbeck, Jim Waters, Jean-Camille Birget, Alex Brodskiy, and Nasir Memon. Authentication using graphical passwords: Effects of tolerance and image choice. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 1–12, 2005.
- [47] Daniel B Wright and Benjamin Sladden. An own gender bias and the importance of hair in face recognition. *Acta psychologica*, 114(1):101–114, 2003.
- [48] Daniel B Wright and Joanne N Stroud. Age differences in lineup identification accuracy: People are better with their own age. *Law and human behavior*, 26: 641–654, 2002.
- [49] Thomas D Wu. A real-world analysis of kerberos password security. In *Ndss*, 1999.
- [50] Andrew W Young, Dennis C Hay, and Andrew W Ellis. The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology*, 76(4):495–523, 1985.