Department of Computer Science

Submitted in part fulfilment for the degree of
MSc in Cybersecurity.

Investigating the Security of p≡p 's Trustword PGP Fingerprint Encoding

Aidan Fray

DRAFT PROCESSED 6th August 2019

Supervisor: Siamak Fayyaz Shahandashti

Contents

List of Figures

List of Tables

1 Introduction

The increasing use of public key cryptography by instant messaging and secure email means key fingerprint verification is an ever more important task. One of the most significant risks to the security of the communication channel is a Man-in-the-middle (MiTM) attack. A successful MiTM attack can circumvent the encryption as it allows an attacker to read all encrypted data. A countermeasure for this is the verification of each parties' fingerprint. Fingerprints can come in a number

TODO: Listing of chapters and how the report is to be laid out

2 Background

The chapter explains aspects required to fully comprehend the project's achievements alongside an evaluation of current literature and the respective gaps.

The chapter will contain the following sections:

- **Project context**
  Will provide the user with the base knowledge required to understand the context and purpose of the following literature review.
- **Definition of terms**
  Specific definitions relevant to the literature review will initially be defined.
- **Literature Review**
  A review of currently available literature on relevant topics. Recommendations and research gaps will be identified and discussed.

2.1 Project context

To fully understand the literature review it is necessary to introduce and explain the base knowledge before continuing.

Public-key Cryptography and Key Exchange

Asymmetric cryptography facilitates the secure encryption of messages in end-to-end encryption (E2EE), verification of digital signatures and sharing of pre-communication secrets, among others. The use-case of E2EE messages will be the primary focus of this paper.

The asymmetry stems from the use of "Public" and "Private" keys. The Public key is used to encrypt data that only the respective Private key can decrypt. This means the keys required to encrypt can be easily shared across insecure channels.

To construct an E2EE connection the initial stage will be a key exchange using the key types discussed previously. This is the sharing

of a pre-shared secret between two verified parties, this will then be used to encrypt subsequent messages with symmetric encryption[1]. This, however, hinges on the verification of the initial parties, if one party impersonates another and gets to view the pre-shared secret all communication becomes decryptable. Therefore, the correct identification of parties is crucial to maintaining secure communication channels. The exploitation of this is known as a Man-in-the-middle (MiTM) attack. This can be bidirectional where the attacker can sit in the middle of a communication channel and decrypt all correspondence. Figure 2.1 contains a visual representation of this attack.

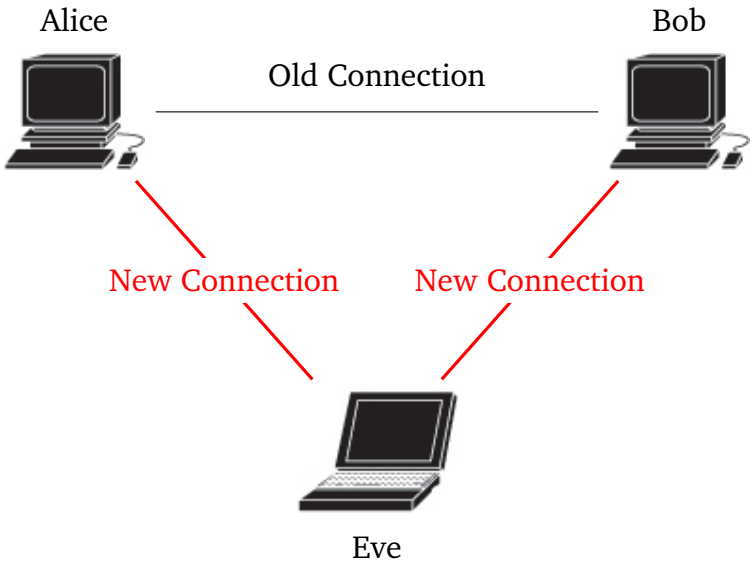

Figure 2.1: Photo depicting a MITM attack

Due to the assumption that the attacking party (Eve) does not have the same public-private key pair as either party (Alice or Bob) the fingerprint of the public key can be used for identification. There are, however, issues in how a user should link a key's fingerprint to real world entity. This paper will, however, assume that users know the respective parties' correct fingerprint, as the protocols used to achieve this are outside the project's scope.

The fingerprint of the key is generated by running the main key components through a secure one way hash function such as SHA-1. This process produces a digest of fixed length that can be used to compare keys. Therefore, the comparison of expected and actual fingerprints can be used too detect MiTM attacks. Historically these have been represented as a hexadecimal string whereon verification fingerprints are

[1]This is due to the speed increase of symmetrical over asymmetrical encryption

compared between two substrates, for example, a monitor screen and a business card. This process is known as the *"authentication ceremony"*. However, system designers are moving away from this structure and are now creating a combined key from each sides' public key. In the case of WhatsApp for example, the connection fingerprint of two parties keys is the first 30 bytes of a SHA-512[2] hash of each parties' identity key, these are then both concatenated together to form a single key[1]. This produces a unique key communication pair.

Due to the manual comparison of the fingerprint, length and format is a consideration. Prior research has shown that the average human can only hold around 7-digits worth of data in their working memory[2]. This rules out the possibility of comparing complete digests. For example, SHA-1 is 40 hex digits (160-bit) and, therefore, difficult to effectively compare. Therefore, if human interaction is required; there is a need for schemes that work effectively with consideration to individual limitations.

Research in this area has taken various encoding schemes and compared their fallibility to impersonated attacks. The main elements of comparison have been the "accuracy of attack detection" and "time to compare". In this paper, these are considered the metrics of "security" and "usability", respectively.

2.2 Definition of terms

Throughout the paper, various terms are used to define elements of the authentication ceremony relevant to fingerprints verification. Explicit descriptions are, therefore, stated in this section. These are used to remove ambiguity and underlying associations readers may have for the words chosen.

Encoding schemes - This is the physical method of encoding used to represent the fingerprint. For example, "Hexadecimal" or "Words" are a unique way to represent a fingerprint and thus, are encoding schemes.

Method of comparison - This is how the user assesses the encoding scheme. The most common example and the one used as default is "Compare-and-Confirm" (CaC). CaC, as the name suggests, is the *comparison* of two fingerprints on different devices or mediums that are then *confirmed*. The following sections explain alternative methods of comparisons.

[2] 5200 iterations

2.3 Literature Review

This section will now discuss the current research around the proposed topic.

The section is organized as follows:

- **Authentication ceremony performance**
  This section explains the current research on performance of various encoding schemes and methods of comparison. This is then complemented with an assessment of the experimentation design for the respective studies.
- **Encoding schemes**
  The following section then discusses research around the design of an encoding scheme. Focus in this section will be allocated to the technical details around the creation and design of relevant schemes.
- **Attacks on encoding schemes**
  The final section will then discuss literature that investigates the creation of attacks against encoding schemes in order to deceive the respective users.

2.3.1 Authentication ceremony performance

Encoding scheme performance

Results from the literature consistently show the effectiveness of language-based encodings such as Words or Sentences with accuracies ranging from 94% [3][4][5]. In all cases, these were the best schemes from the sets assessed. The exception to this is the work performed by **Hsiao, *et al.*** [6] in 2009 with Words achieving an abnormal accuracy of 63.00%.

Aside from textual representations were graphical schemes. Examples of schemes assessed were: Random Art[7], Flag[8], T-Flag[9], Vash[3], OpenSSH visual host key and Unicorns[4], among others. These schemes had mixed accuracy with ranges as large as 50% - 94% in work by **Hsiao, *et al.*** [6]. The only other paper assessing graphical representations was the work of **Tan *et al.*** [4] where they also achieved mixed results with accuracies ranging from 46% to 90%.

In terms of usability of graphical schemes, the literature concurred on

[3] https://github.com/thevash/vash
[4] https://unicornify.pictures/

Scheme	Hsu-Chun[6]	Kainda[5]	Dechand[3]	Tan[4]
Hexadecimal			11.20	9.00
Numerical		6.00	10.60	9.00
Base32	3.51	6.00	10.20	
Words	4.63	7.00	8.70	7.00
Scentences		11.00	12.30	8.00
Chinese Symbols	5.01			
Japanese Symbols	5.07			
Korean Symbols	4.92			
Random Art	3.21			
Flag	4.28			
T-Flag	4.00			
Flag Ext.	4.02			
OpenSSH‘				5.00
Unicorns				3.00
Vash				3.00

Table 2.1: Timing results in seconds for the related schemes

their high usability. The comparison speed of these schemes were all among the quickest (See Table 2.1 for an overview of timings). Other work also indirectly concurred with graphical encodings having significantly quicker comparison times compared to non-graphical schemes [3][5]. In terms of research into the performance of graphical schemes, the literature does not contain an extensive review with literature only containing two papers. There is also no overlap in the schemes assessed with each paper reviewing a unique set. This is, therefore, a promising candidate for further research.

One unique paper in this research area was the work by **M. Shirvanian *et al.*** [10] produced in the context of secure messaging pairing. This paper was unique for several reasons. First was to consideration for “remote-vs-proximity” pairing where this is the first consideration of this aspect found in the literature. There is room for further research to compare encoding schemes in the context of “remote-vs-proximity.” Another unique aspect was the end-to-end encryption context of the study.

The findings from the paper showed a high false negative rate for all the schemes; this is a consideration also missing from the literature. Alongside this, results for usability were lower in a remote setting for all the schemes. The author, however, comments on the expected nature of this result.
Images were assessed as being the most secure method of authentication in the remote setting but voted as the method with the worst usability.

This conclusion is highly inconsistent with all other work in this area. However, this could be due to the unique setting of remote verification, resulting in distinct results from user studies. Without further work in this area, it is difficult to validate these results conclusively.

Alongside this, it was shown in work by **Hsiao, *et al.*** [6] that age and gender do not affect the accuracy of the scheme. However, younger participants were considerably faster. Furthermore, findings also showed that language comprehension helped in discerning small differences between schemes encoded in Chinese, Japanese, Korean or English. Subsequently, knowledge of the language did not assist in differentiating more significant changes in the schemes as these had high accuracy regardless. These were interesting and unique considerations. Further work could aim to corroborate these conclusions.

Scheme	Hsu-Chun[6]	Kainda[5]	Dechand[3]	Tan[4]	M. Shirvanian[10]
Hexadecimal			89.56%	79.00%	
Numerical		100.00%	93.66%	65.00%	97.33%
Base32	86.00%	90.00%	91.50%		
Words	63.00%	100.00%	94.25%	94.00%	
Scentences		100.00%	97.01%	94.00%	
Chinese Symbols	59.00%				
Japanese Symbols	57.00%				
Korean Symbols	54.00%				
Random Art	94.00%				
Flag	50.00%				
T-Flag	85.00%				
Flag Ext.	88.00%				
OpenSSH‘				90.00%	
Unicorns				46.00%	
Vash				88.00%	

Table 2.2: Accuracy of correct comparison for the encoding schemes assessed

Tables 2.1 & 2.2 contain the accuracy results of all papers assessed; this is to aid in visual comparison. Each paper used a different metric for measuring accuracy. Therefore, all results have been translated into “overall accuracy.”

Methods of comparison performance

Aside from “Compare-and-Confirm” (CaC), there is “Compare-and-Select” (CaS) and “Compare-and-Enter” (CaE). CaS is the method where one

device displays the fingerprint, and the other user is provided with several options. The user then has to choose the correct value from the list of candidates. If there is no match, the user must deny the connection attempt. The creation of CaS was due to concerns that CaC would be "too easy" for users leading to complacency and errors[11]. CaE is designed for scenarios where both devices might not have a display, i.e. pairing between a phone and a keyboard. One device displays the checksum. This checksum is entered into the other device. The first device then compares the entered string and checks for a match.

Research has been performed assessing the performance of these schemes and how they affect the ultimate security of the authentication ceremony. The literature agrees on CaC being the best overall scheme to compare fingerprints [4][11] with CaS being highlighted for its poor security and usability. CaE has had contrasting results. [11] discarded it after one round due to "poor usability." However, [4] considered it the best method overall for usability and security. These conflicting results, therefore, show polarisation in the results of CaE. However, this could be due to the different overall use-cases of the studies. Validation of results from either study would, therefore, be an area of additional study.

Experimentation methodology comparison

To further look into the validity of previously discussed results, it is necessary to asses how the respective studies reached these conclusions. Areas for consideration are scheme entropy, attacker strength and participant attributes.

The starkest limitations of the literature are the range of participants and encoding entropy. The worst studies tested only 22-bits of entropy. This makes it difficult to compare results directly. One of the papers this most effects is the early work by **Hsiao, *et al.*** [6] where their highest entropy is 28-bits. This level of entropy was inadequate even at the time of publication. There is an attempt by the authors' to address this issue in the later stages of the paper where they write *"[...] increasing entropy is not a solution because it sacrifices usability and accuracy. With more entropy, representations will contain longer sequences of characters or more minute details which will lead to increased time and errors during comparisons."*. This statement is backed up with no empirical evidence, and the authors' fail to consider how the low entropy would affect the overall security of the schemes.

Attacker strength is also another metric used to compare results concluded by each paper. Some papers failed to address their attacker strength consideration directly, but the overall strength has been inferred from the changes made to their schemes. For example, if they

decided to change a single character in a 40-digit (160-bit) SHA-1 hex digest, they are indirectly stating that the attacker can control 39-digits (156-bit). To achieve this, the attacker would have to compute 2^{156} SHA-1 compressions to find a key match. In the literature, this element ranged from 2^{28} to around 2^{242}; however, this has some relation to the size of the encoding schemes used. This is a substantial range that makes it ultimately challenging to compare and confer results confidently. Further work could exclusively look into the affects attacker strength has on the success of attacks. Moreover, all the papers assessed failed to fully consider the feasibility of attacks in terms of computer and storage requirements. This again, shows gaps in the literature.

All of the studies considered demographical data when presenting their results. Their average ages were all around ~ 35 years old with the majority of participants educated with at least a bachelor degree. This was alongside the equal split between male and female participants.

One consideration of note is that made by Dechand *et al.* [3] where they briefly consider medical conditions such as ADHD and reading or visual disorders and the way they affect the comparison's effectiveness. They highlight a slight reduction in overall accuracy, although due to their small sample size, they cannot conclusively validate these results. This is an aspect unique to all literature. Further work would be required to produce conclusive results. Therefore, highlighting a gap in the literature.

One glaring issue with the demographical health of **Ersin Uzun *et al.*** [11] study is the use of two entirely different groups of participants. Not only were their demographics different, but they were from different countries (America and Finland). Different cultures contain inherent biases and assumptions. Moreover, more concerns are raised around the dual study design of this paper with two completely different set of participants. Changes were made pragmatically regarding the results from the first round of 40 participants. These were then re-assessed and the results were directly compared. This puts huge doubts on the validity of the results with no consideration made by the authors to control external factors that may affect the performance of the method of comparison.

Hexadecimal and numerical schemes were not included as encoding schemes in work by **Hsiao, *et al.*** [6] (some of the most widespread encoding schemes). This decision was based on their claims on similarities to Base32 and "well-known deficiencies" in the excluded schemes. This point was provided with no further justification or quantified in any way. It is also not consistent with available research, for example, in *"Empirical Study of Textual Key-Fingerprint Representations"* [3] it was shown numerical representations performed significantly better than

that of Base32.

	Hsu-Chun[6]	Kainda[5]	Dechand[3]	Tan[4]	M. Shirvanian[10]
Attacker Strength[5]	$\sim 2^{28}$	$\sim 2^{40}$	2^{80}	2^{60}	$\sim 2^{242}$
Entropy Range	22-28 bits	20-40 bits	122 bits	128 bits	160-256 bits
No° Participants	436	30	1001	661	25

Table 2.3: Paper attribute comparison

Table 2.3 has been provided to visually compare the differing aspects of the papers' parameters. Clearly it can be seen from the table the large ranges in participant size, entropy and attacker strength.

Topic conclusion

Overall, this review has identified several key areas suitable for further work. The first is the performance assessment of graphical encoding schemes. Recreation of pre-existing results from [6][4] is required to corroborate current conclusions and validate results. Another gap in the research is the consideration into the utilization of encoding schemes in realistic conditions, i.e. “remove vs. proximity." This topic was initially covered by [10] but their scope was limited. Further work, therefore, could increased the scope and touch upon a large number of schemes in these settings. The final aspect for further work is the limited consideration into the feasibility of attacks on encoding schemes. All of the papers assessed simulated attacks and had minimal consideration for the execution of these attacks. Therefore, further work could delve into the implementation of such attacks and their feasibility in terms of compute and storage complexity.

2.3.2 Encoding schemes

Another area of research is investigations into the actual physical encodings of the hash digest. This section will briefly discuss the current research available on the creation and security of actual encoding schemes. The actual details of the operation of the schemes are outside the scope of this literature review, therefore, minimal attention will be allocated to these details.

Some of the oldest preliminary work into visual encoding schemes was performed by **Adrian Perrig *et al***[7]. in the creation of their scheme “Random Art" in 1999. The motivation for creating such a scheme was the perceived flaws in the ways humans verify and compare written

information. As mentioned in previous sections visual encoding schemes have been shown to have mixed success, with low security being one of their most alarming flaws. This research laid the foundation for further work in analysing the security of visual encoding schemes.

Further research into the creation of unique visual hash schemes have been performed by **C Ellison *et al.*** [8] (Flag), **Yue-Hsun Lin *et al.*** [9] (T-Flag) and work by **M. Olembo *et al.*** [12]. Each publication has provided a new way to visually represent a key fingerprint. Alongside the academic literature, there are more informally presented methods of visual fingerprints such as Unicorns[6] and Robots[7]. This list is by no means exhaustive but is used to depict the amount of research and work invested into graphical hash representations.

One paper of note is the preliminary work performed by **D Loss *et al.*** [13] in their *"An analysis of the OpenSSH fingerprint visualization algorithm"* where their aim was to spur on further research with their initial findings into the security of the OpenSSH scheme. The authors claim that the use of the algorithm in OpenSSH is only heuristically defined and there is a need for a formal proof of its security.
The paper proposed a number of ways to generate similar fingerprints. The methods proposed were: Naive brute force, Graph Theory, and brute force of a full visual set. They were only able to produce only very basic results and have proposed a large amount of potential further work. Since the paper's publication in 2009, there seems to have been no research building on the work of the authors. This highlights a current gap in the available literature.

Minimal research has also focused on basic textual fingerprint representations and their respective security. Work by **A. Karole** and **N. Saxena**[14] looked into ways to improve the security of a textual representation. This research aim was to improve the secure device pairing process of comparing two numerical values. The devices used (Nokia 6030b; Mid-range devices at the time of publication) and the SAS compared results in findings that are not directly applicable in a fingerprint comparison context.

A more specific subsection of textual fingerprints is the use of words and sentences to encode hash digests. Some of the first work in this area was produced by **Juola** and **Zimmermann** [15] and their work in generating a word list where phonetic distinctiveness was prioritised. Each word is mapped to a single byte. The unique aspect of the word list is the separation of "even" and "odd" words where "even" byte positions are sample from the even-list and "odd" from the odd-list. This effectively creates two sub-word lists. The maximisation of linguistic

[6] https://unicornify.pictures/
[7] https://github.com/e1ven/Robohash

distinctiveness of these word lists were maximised through the use of a Genetic algorithm. The paper also includes a study on effective measures of "linguistic distances" of words and provided an in-depth discussion into these areas.

Overall the paper provides a foundation for formalising the creation of effective wordlists. A limitation is the lack of empirical data gathered on the performance. However, this was later evaluated in work by Dechand *et al.* [3] and shown to be an effective encoding scheme.

Other research of note is work by **M. Goodrich *et al.*** [16] called *Loud and Clear: Human-Verifiable Authentication Based on Audio*. As the name suggests the authors were researching ways to improve current methods of secure device pairing. The unique aspect of this work is the use of a Text-to-Speech system reading out syntactically correct English sentences. The sentences are based on a MadLibs[8] where static placeholders were replaced with potential words.
The work into a potential wordlist can be seen as an extension to the work performed by Juola and Zimmermann[15] as they aimed to emulate the techniques used in PGPfone. The paper's finding are limited by the lack of empirical data backing up claims made by the author as the systems performance and security are only theoretically assessed.

Aside from this research, there have been further informal implementations of fingerprint encodings. The first being by **Michael Rogers**[9]. Rogers' implementation is a program designed to map fingerprints to pseudo-random poems. This implementation was again, empirically evaluated by Dechand *et al.*[3]. Older work by **N. Haller** with the S/KEY[17] shows the implementation of a system designed to represent a hash as a series of six short words. However, this system is designed for a one-time-password purpose and only provides word mappings for basic human usability of the password and not within a fingerprint verification context. Therefore, the wordlist has not been designed with pronounceability in mind.

Pretty Easy Privacy

A very recent implementation of a word list can be found in Pretty Easy Privacy (p≡p) implementation of TrustWords[10]. p≡p is a data encryption system that utilises PGP encryption to provide E2EE on all common channels of communication such as email or SMS. The imbedded design principles state that above all the systems should be

[8] https://en.wikipedia.org/wiki/Mad_Libs
[9] https://github.com/akwizgran/basic-english
[10] https://tools.ietf.org/html/draft-birk-pep-trustwords-03

easy to install, use and understand.

p≡p deals with the threat of MiTM attacks by having users compare the respective key fingerprints encoded as a set of words. Figure 2.2 shows the p≡p Android implementation of Trustwords. The users then authenticate the words on an OOB (Out of Band) channel such as a phone call or in-person communication. If both users decide the words match they will accept or decline respectively.

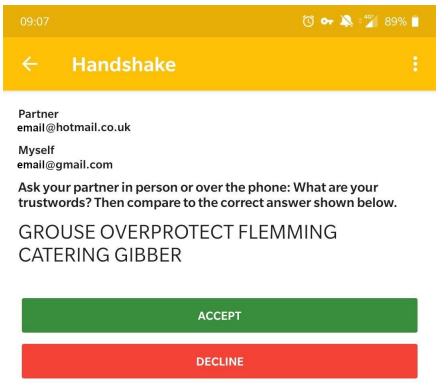

Figure 2.2: Trustword fingerprint verification

The unique aspect of TrustWords is its mapping of a single word to 16-bits. In comparison to other literature, this is the highest number of bits-per-word seen. Full mappings (no duplication of words) would, therefore, require 2^{16} words in the dictionary where arguably the dictionary size is higher than most users' vocabulary. This deviation from the norm has not been currently backed up by research.

```
    [...]
52127 ZYGOTE
52128 ZYGOTIC
52129 ZYMURGY
52130 AACHEN
52131 AARDVARK
52132 AAREN
    [...]
```

Figure 2.3: Re-mapping position in Trustword dictionary

Alongside the abnormally high number of words the design choice to exclude slang and profanities from the English Trustword dictionary requires dual-mapping of a section of words. Approximately 13633/65536 (20.8%) of words are re-mapped in the dictionary leaving 51903 unique words. The re-mapping is also done on a loop with it remaining alphabetical. Figure 2.3 shows the position in the dictionary where this occurs. This predictability within the dictionary will be explored later in the paper.

Moreover the main RFC documentation still remains in a draft stage and states *"It is for further study, what minimal number of words (or entropy) should be required."*. These aspects clearly highlight on a gap in the current literature.

Topic conclusion

In conclusion to this topic, the current research has primarily focused on the research and creation of visual representations. Research for textual fingerprints is fragmented and incomplete with work Juola and Zimmermann [15] and M. Goodrich *et al.*[16] providing meaningful research to build upon in terms of word a sentence based encodings. The fragmentation of this research leaves room for further work into this topic area. Alongside this, findings from the previous sections research shows that human language based encodings provided the best usability and, therefore, should be a target for further research looking to improve upon their security and usability.

2.3.3 Attacks on encoding schemes

This area of research studies ways to physically execute attacks on fingerprint encoding schemes. This differs from previously examined work due to papers discussing the performance and fallibility of encoding schemes simulated the attack without consideration for how the attack would be performed. Research in this area is scant, with lots of research attention being directed towards the security of Man-in-the-Middle (MITM) attacks and not the encoding schemes themselves.

Research in 2002 by **Konrad Rieck**[18] is the first formalisation of attacks on fingerprint representations. The paper titled *"Fuzzy Fingerprints Attacking Vulnerabilities in the Human Brain"* aimed to look into ways users check hexadecimal encoded OpenSSH fingerprint representations. The author created an elegant way to 'weight' more important chunks of the digest. The bytes furthest to the right and left of the digests provided the highest weight. The weight was the smallest in the centre of the digest. This provides a way to score digests and determine the best partial collisions found. For example with the target fingerprint: `9F23` a partial match `9313` is given a score of 45% even though only two characters were matching. This is due to the weightings.

The paper contains an implementation with a "1.2GHz CPU" being able to obtain 130,000 H/s (With MD5). In comparison to this, a mid-range Intel i5-3320M CPU can today obtain 111,700,000 MD5 H/s. This shows that the results obtained from the paper are significantly outdated. However, even with the low hash rate, the author was able to obtain some promising results. Figure 2.4 contains the best example used.

Overall the paper shows an interesting way to create partial fingerprint matches but is not quantified by any empirical evidence gathered on real world users. This, therefore, highlights on gaps in the coverage of

```
TARGET: d6:b7:df:31:aa:55:d2:56:9b:32:71:61:24:08:44:87
MATCH:  d6:b7:8f:a6:fa:21:0c:0d:7d:0a:fb:9d:30:90:4a:87
```

Figure 2.4: Best match obtained after a few minutes of hashing

this literature.

The only other relevant research on this topic is the work by **M Shirvanian *et al.*** [19] and their paper *"Wiretapping via Mimicry: Short Voice Imitation Man-in-the-Middle Attacks on Crypto Phones"*. Further research in the area of "human voice impersonation" has received lots of attention [20][21][22]. This paper was chosen over other alternatives due to is specific use of encoding schemes in its evaluation.

In this paper, the authors develop a way to impersonate users when authenticating Short-authentication-Strings (SAS) in pairing of Cryptophones. To achieve this impersonating they propose two methods: "Short voice reordering attack" where an arbitrary SAS string is recreated by re-ordering snipped obtained from eavesdropping a previous connection and "Short voice morphing attacks" whereby the use of previously eavesdropped audio snippets the attacker can morph their own voice to match that of the victim. With these methods, they aimed to attack encodings of Numbers, PGP word list (previously discussed work by Juola and Zimmermann [15]) and MadLib (M. Goodrich *et al.* [16] work also previously discussed). The effectiveness of these attacks were evaluated with a study involving 30 participants.

Results from the paper show the effectiveness of these methods. Compared to the baseline of the attacker's voice replacing the victim where this performed with a $\sim$18% success rate. Morphing gained an overall success rate of 50.58% and Reordering a very impressive 78.23% success rate. Showing that these attacks provide an improvement on top of the naive implementation.

One of the biggest limitations addressed by the authors was the reduction in success rates as the size of the authentication string grew. The morphing and reordering attacks become increasingly ineffective as the user has more time to detect imperfections. This is not quantified by the author and the extent of this degradation is never empirically discussed. Therefore, the results from this study are only effective and applicable in a SAS context.

Topic Conclusion

Overall the literature for this subtopic remains sparse and incomplete. Further suggested work could look into the feasibility of generating partial collisions for all textual representations alongside quantified effectiveness on users. With the possibility to concentrate on a few selected implementations. The work would aim to focus on the various physical methods used and their feasibility. This is one area the previous literature has failed to cover and has only theoretically quantified attacker strength without consideration for the actual real-world cost of these attacks.

2.4 Overall Summary

In summary of the literature gaps areas discovered; encoding scheme performance requires further work in the performance of all graphical schemes to back up the results made by other work. Furthermore, there is a gap with the assessment of encoding schemes in the context of "remote-vs-proximity" first proposed by M. Shirvanian et al. as it would arguably provide a better simulation of real world scenarios.

In the context of human attributes and the way they alter the performance of the schemes; further research is required into how the fluency of languages affects authentication ceremony. This is a continuation of the initial work by Hsiao et al and could allow the creation of schemes that are better suited to certain languages. Alongside this, there is a research gap in the investigation of how mental impairments affect the performance. This is useful due to the number of potential users being affected due to systems not being designed with them in mind. For example, 7% of the population has been identified as having dyslexic tendencies[23]. This means with a UK population of around 63,000,000[11] 4,410,000 people could benefit from encoding schemes including dyslexic oriented benefits. This is also only one of many visual impairments that could affect a scheme's performance, thus further contributing to the substantial affect further work could have on this area.

Other possible large areas for consideration is the lacks of direct consideration for attack strength when assessing the vulnerability of encoding schemes. Issues included the wide range of attack strength (2^{28} - 2^{242}) and the lack of consideration at all from certain papers. This alongside the complete lack of computation and storage complexity considerations means this is a prime area for further research as it would improve the applicability of results.

[11] From the 2011 Census collated by the Office for National Statistics

The final major research gap identified is the lack of justification into the newly created p≡p Trustwords. Abnormal design choices (16-bit per word) and its recent creation makes it another area for a wide variety of further research. This is also combined with the conclusive results from the literature on the effectiveness of language based encodings such as words. This would suggest that research into the improvement of Trustword encoding would be highly beneficial.

3 Research Questions

Following on from the previous chapter, this chapter will define the selected gaps and the subsequent research questions and aims.

The chosen project aims to asses the security of p≡p 's minimum recommend number of four Trustwords (As stated in Figure 3.1). As discussed in the previous chapter, Trustwords aims to sacrifice security for increased usability. The encoding scheme has been designed to assist this by having the user compare a reduced set of words. Moreover, issues with the dictionary's design such as the presence of homophones and dual-mapping of words shows the potential for possible vulnerabilities.

As discussed in the previous chapter, the identified research gap regarding the minimal consideration of attack complexity would be a suitable addition for this project. This is due to it providing the ability to asses actual performance and, thus, the actual real-world strength of Trustwords while providing research to a scant area. Therefore, the security assessment will be supplemented with complexity considerations and an actual implementation of the attack proposed.

```
"Short Trustword Mapping (S-TWM) requires a number of
Trustwords that MUST retain at least 64 bits of entropy.
Thus, S-TWM results into at least four Trustwords to be
compared by the user."
```

Figure 3.1: Most recent RFC security recommendation

By sampling the p≡p documentation they have defined the use case of the Trustword handshake: *A handshake is done by comparing the Trustwords between two users through a separate communication channel (e.g. in person or by phone)*.[1]. It can be assumed due the increased globalization of the planet that the most common handshake occurrence will be over the phone, and, therefore, will not be in person. This is, therefore, the assumed context the handshake will be occurring in. This means the similarity of words will need to be quantified phonetically instead of visually.

[1]https://www.pep.security/docs/general_information.html#handshake

3.1 Questions

With the previously discussed points in mind the following research questions have been proposed.

1. Is the recommended minimum number of four Trustwords enough to provide a basic level of security?
2. What are the different ways phonetic similarity can be quantified?
3. Out of a chosen set of metrics, which are the most effective?
4. What is the time and computation complexity required to generate a 'similar' keys for a targeted key pair?
5. What kind of hardware is required to compute a matching key?
6. What percentage of attacks successfully deceive a user?

With these research questions defined, the project will now discuss the design of the proposed attack.

4 Design

This chapter will discuss and overview of the proposed attack and the elements that are required.

4.1 Overall attack design

The attack on Trustwords involves generating "near-collision" keys. This will attempt to provide the base to answer Research Questions 1, 4 and 5.

Near-collision keys are keys that are deemed a match by the phonetic similarity metric (More on this aspect in the following sections). A similarity metric is an algorithm designed to determine if two words are phonetically a match. For example, the words "THEIR" and "THERE" will be a match whereas the words "DARK" and "PRINCIPLE" are not phonetically matching.

As each combined key in Trustwords is an exclusive-or of both sides public key (Figure 4.1 shows this process) the attack is target to a single pair of users and will require recomputation for every attack target. This will be considered when discussing the attack feasibility. Each pair is also split into a "Uncontrolled" and "Controlled" key. Uncontrolled is the reciever of the communication, and, thus, we cannot control their key. The Controlled key is the one we are attempting to impersonate, and it is assumed that we have the ability to replace the Controlled key with our malicious option. These descriptions will be the terminology used throughout the paper. However, it should also be considered that the uncontrolled and controlled can be swapped around with the ability to compute both directions. Thus, resulting in the possibility to intercept both directions of communication. This, however, will require performing the attack separately for both directions.

$$KeyFingerprint_1 \oplus KeyFingerprint_2 = TrustwordsFingerprint$$

Figure 4.1: Creation of the combined Trustword fingerprint

When attacking the respective similarity metric will be used to compute a list of possibilities for each position in the combined fingerprint.

Figure 4.2 shows the process of generating a sub section of the combinations. A list of each words matches are generated and then the total number of permutations are generated overall.

```
CHOKE BLUSHING FRIGHTENING HAND
COKE
SMOKE

CHOKE
COKE BLUSHING FRIGHTENING HAND
SMOKE

CHOKE
COKE
SMOKE BLUSHING FRIGHTENING HAND
```

Figure 4.2: Visualization of the generation of near matches

To generate an actual list of fingerprints to search for, the near collision words are converted back into hexadecimal and XORed with the uncontrolled key. The provides the impersonated key that will produce the desired combined near-collision. Completing this for all of the near-collision permutations will produce a list of fingerprints that can be inserted into a tool designed to hash keys and search for targets. This aspect of using a large list to search for keys massively reduces the complexity of the search.

In summary the, attack steps are:

1. Compute all possible matches using a similarity metric on all words in a dictionary (Only needs performing once).
2. Select a target and allocate "Uncontrolled" and "Controlled" key identification.
3. Calculate all permutations of near-collisions for the key pair and produce a list of similar key fingerprints.
4. Use list of similar keys in mass computation of keys to find near-collision keys.

4.2 Similarity metrics

One of the first requirements for the attack is quantifying "phonetic similarity". This section is looking to provide an answer for the Research Question 2. There are many algorithms currently available to provide

this functionality. This section will describe algorithms alongside the reasons they were selected for assessment later in the project.

4.2.1 Soundex

One of the earliest example of a phonetic algorithm is known as Soundex. It is one of the most famous algorithm due in part to its implementation into major database clients like MySQL[24], Oracle[25] and PostgreSQL[26]. Originally designed for indexing names by phonetics alongside spelling mistakes due to transpositions in letters. Therefore, due to it being based on the phonetic features, it is highly suitable for this project's use-case.

Soundex produces a four digit code for each word assessed. The first letter of the word is retained alongside the removal of all of a, e, i, o, u, y, h and w. The remaining letters are then mapped to numbers[1]. These mappings are displayed in Figure 4.3.

```
b, f, p, v                1
c, g, j, k, q, s, x, z    2
d, t                      3
l                         4
m, n                      5
r                         6
```

Figure 4.3: Soundex mappings of letters to numbers

4.2.2 Soundex Issues

Due to the fixed length and limited digit set the initial concerns from this design is the limited number of combinations. There are a total of 5616 codes due to 26 initial letters and three digits of 6 values ($26 * 6^3$). The limited combinations will result in matches of a limited quality. This will require consideration later in the project.

Further issues posed in [27] are discussed below and show the further deficiencies of Soundex in a dictionary word matching context.

1. **Dependency on the first letter:** Soundex cannot match words together if their first letters are different, meaning for example the words "KORBIN" and "CORBIN" will never be matched.

[1] Further steps are performed on the code, as the technical details do not add anything to the discussion, the steps of the algorithm can be found in Appendix TODO.

2. **Silent consonants.** Soundex does not have logic embedded to deal with silent consonants.
3. **Poor precision.** Due to the previously discussed point of a small code space. [27] re-iterates this point but in the context of name matching where Soundex's poor performance was demonstrated. Soundex only gained and overall accuracy 36.37% when matching names within a provided database.

However, even with Soundex's fallbacks its algorithmic simplicity and popularity allows it to still remain relevant in this application.

4.2.3 NYSIIS

The New York State Identification and Intelligence System (NYSIIS) phonetic code was created for use matching the phonetics of American names. It was created due to the presence of hispanic names in the American based databases (this was an aspect Soundex was known to have low accuracy with). However, due to it having embedded rules to handle word phonetics it would, again be applicable in this application.

It also allows for variable length codes and, thus, allows the applicability of the application to increase due to it not confronting the limited code issue of Soundex. It was in use right up the end of 1998 within various US Government departments. Due to this prolific stature and proven track record it was deemed suitable as one of the selected similarity metrics.

4.2.4 Metaphone

Metaphone was invented by Lawrence Philips in 1990[28] in response to the deficiencies in Soundex. It improves on Soundex by including information around inconsistency and variation in English spelling in an attempt to create a more accurate phonetic representation. Metaphone is arguably on the same level of ubiquity as that of Soundex with it finding itself implemented in languages such as PHP[29].

Further work would involve the implementation of newer versions of Metaphone. Double Metaphone (2000) and Metaphone 3 (2009) that claim to improve over the original version due to further research performed by Philips. The original version was chosen due to its historical and widespread usage.

4.2.5 Levenshtien Distance

Levenshtein distance is a string metric designed to measure the 'distance' between two strings. It is simply the number of single-character edits (insertions, deletions or substitutions) required to reach the other string. Edit distance is not technically designed as a phonetic algorithm, but due to similar-sounding words often being spelt in similar ways[30] Levenshtien distance was deemed another suitable metric.

An example distance between `trace` and `place` would be the substitutions of the first to letters, from `tr` to `pl` meaning the two strings have a Levenshtein distance of 2.

4.2.6 Phonetic Vectors

Phonetic Vectors is the unique addition to the chosen set. Created by Allison Parrish in 2017[31], Phonetic vectors is as the name suggests the vectorization of a words phonetics. This allows a words phonetics to be represented in vector space.

Phonetic features are used in this work as a way to compare the similarity of a words phonemes. Phonemes are the phonetic elements that construct a word. For example the word "RING" translated into the phonemes `/R IH NG/`.

Extensive prior work has gone into producing models of features that map to phonemes. [32][33][34]. Features, therefore, are an attempt at mapping the varying and inconsistent rules around pronunciation of the English language. The vectors were created using lists phonemes from the CMU Pronouncing Dictionary and mapping them to possible features.

Phone	Features	Phone	Features	Phone	Features
AA	bck, low, unr, vwl	F	frc, lbd, vls	P	blb, stp, vls
AE	fnt, low, unr, vwl	G	stp, vcd, vel	R	alv, apr
AH	cnt, mid, unr, vwl	HH	apr, glt	S	alv, frc, vls
AO	bck, lmd, rnd, vwl	IH	fnt, smh, unr, vwl	SH	frc, pla, vls
AW	bck, cnt, low, rnd, smh, unr, vwl	IY	fnt, hgh, unr, vwl	T	alv, stp, vls
AY	cnt, fnt, low, smh, unr, vwl	JH	alv, frc, stp, vcd	TH	dnt, frc, vls
B	blb, stp, vcd	K	stp, vel, vls	UH	bck, rnd, smh, vwl
CH	alv, frc, stp, vls	L	alv, lat	UW	bck, hgh, rnd, vwl
D	alv, stp, vcd	M	blb, nas	V	frc, lbd, vcd
DH	dnt, frc, vcd	N	alv, nas	W	apr, lbv
EH	fnt, lmd, unr, vwl	NG	nas, vel	Y	apr, pal
ER	cnt, rzd, umd, vwl	OW	bck, rnd, smh, umd, vwl	Z	alv, frc, vcd
EY	fnt, lmd, smh, unr, vwl	OY	bck, fnt, lmd, rnd, smh, unr, vwl	ZH	frc, pla, vcd

Table 4.1: Phonemes to feature mapping table

Table 4.1 contains the mappings used in [31] to create the phonetic feature lists. Using this with all 133,852 entries in version 0.7b of the CMU Pronouncing Dictionary, 949 unique properties were produced

overall. The author then performed principal components analysis[2] on the unique properties to reduce them down to 50.

This metric allows for a unique set of actions to be performed on the phonetic output. Not only does this metric allow the user to measure *dissimilarity* (opposed to the similar-or-not method of the alternatives) the continuous nature of the value allows mathematical operations to be performed on the output. An example shown in [31] was the addition of word vectors.

No	Operation	Result
1	$Vec(\texttt{sub}) + Vec(\texttt{marine})$	`submarine`
2	$Vec(\texttt{miss}) + Vec(\texttt{sieve})$	`missive`
3	$Vec(\texttt{fizz}) + Vec(\texttt{theology})$	`physiology`

Table 4.2: Examples of vector addition

For example, the addition of vectors can be seen in Table 4.2. This works for any mathematical operation with multiplication allowing the 'tinting' words with a theme.

The ability to perform operations and measure dissimilarity allows for a plethora of applications. One possible application of note would be the creation of a wordlist where the phonetic difference (vector distance) is maximized. This, if the vector mappings are an accurate representation could allow a creation of a phonetically distinctive wordlist. This will be discussed further at a later stage of the report.

4.3 Alternative Similarity Metrics

4.3.1 Match Rating Approach

Matching Rating Approach (MRA) is another algorithm designed to match names within a database, therefore, its operation can be grouped with that of NYSIIS and Soundex. MRA was discarded as an option. This is because of the lack of known utilization in any substantial real world use-cases alongside its similarity to more well established algorithms of Soundex and NYSIIS.

Further work could include the comparison of this algorithm to similar alternatives in phonetically matching of words to quantify performance. This will be discussed further in the evaluative sections of the report.

[2]Details regarding this process is outside the scope of the project. Please, however, if interested please refer to this resource for more information: http://setosa.io/ev/principal-component-analysis/

4.3.2 Caverphone

Another notable alternative solution is that of Caverphone that was designed in New Zealand. Caverphone as is the case with the vast majority of phonetic algorithms was designed for use with name matching. Caverphone was not chosen to similar reasons to that of MRA alongside its optimization for accents in location of New Zealand it was conceived in. Therefore, the low level of utilization alongside the unique design features that suggest it may not be suited for this application. This has, however, not be assessed empirically and thus would be a candidate for further work.

Metric	Output
Soundex	T614
NYSIIS	TRAVAL
Metaphone	TRFL
Caverphone	TRF111
MRA	TRVL

Table 4.3: The various phonetic encodings of the word "Travel"

4.4 Design of GreenOnion

To answer Research Questions 4 and 5 fully, it is necessary to implement an actual code base that will be used to generate actual keys.

Inspiration of the design of this tool was taken from a tool called Scallion[3]. Scallion was designed by Richard Klafter and Eric Swanson and used to demonstrate that 32-bit PGP key ids were insufficient. Keeping with the onion-based theme the proposed tool is known as GreenOnion and is a re-write of the tools structure in C++. This language was chosen due to the well understood efficiency benefits. The proposed tool differs from Scallion most notably in its ability to concurrently search for a large number of keys, GreenOnion improves on this substantially. More implementation and experimental details will be discussed in later chapters.

The tool should take two keys as parameters (Uncontrolled/Controlled) and a chosen similarity metric and produce a list of target keys fingerprints. This list is then used as a search criteria when hashing a checking a large number of keys. In order to utilise the high parallel nature of the GPU to compute the hash of a large number of keys the tool will utilise a GPGPU (General-purpose computing on graphics processing units)

[3] https://github.com/lachesis/scallion

framework. The chosen framework was OpenCL due to its support for the chosen language (C++) and platform (Linux). OpenCL allows the creation of code chunks refereed to as "kernels" to be executed concurrently, this will provide a massive speed increase compare to the sequential nature of the CPU. Technical details will be explained further in Chapter 5.

4.5 Experiment Design

In order to answer some of the research questions empirical evidence is required and, thus, experiments are required. In this section the designs and considerations of the chosen experiments will be discussed.

4.5.1 Metric performance

In order to answer Research Question 3 and reduce the number of metrics to asses in the later rounds, the performance of similarity metrics required comparison. The thinning out of metrics is required due to limited resources and, therefore, possible further work could involve repeating this work with a much more varied selection of metrics. As explained in a previous section (See Section 4.2) the chosen metrics up for assessment are Soundex, Metaphone, NYSIIS, Levenshtein and Phonetic vectors.

The design for the experiment involves assessing the quality of the metrics matches by having participants rate them on a scale of 1 to 5.

Matches

A match for each code based metric (Soundex, Metaphone and NYSIIS) occurs when the codes are identical. However, for the other cases where the difference is variable other ways of defining a match are required.

In the case of Levenshtein values of similarity are discrete due to it being the number of single character edits. Therefore, this requires less deliberation. Match size was used as a way to decide on a cut-off point.

As it can be seen from Table 4.4 the number of matches for L2 is massively larger than all metrics excluding Soundex. As the issues with Soundex have been discussed in previous sections (See Section 4.2.2) it can be excluded as an abnormality in this context. Therefore, due to the excessively high value for a L2, L1 was chosen as the Levenshtein cap for defining a match.

Metric	Matches
Soundex	1,527,554
Metaphone	412,916
NYSIIS	188,474
Levenshtein (Distance 1) [L1]	**97,730**
Levenshtein (Distance 2) [L2]	1,070,656

Table 4.4: Levenshtein number of matches comparison

Match size was also used to define the cap for the phonetic vectors. This is less distinct that that of Levenshtein due to the continuous nature of the distance between two vectors, however the complexity was reduced by limiting it to increments of 0.5.

Metric	Matches
Soundex	1,527,554
Metaphone	412,916
NYSIIS	188,474
Levenshtein	97,730
Phonetic Vector (Distance 3.0)	14,550
Phonetic Vector (Distance 4.0)	73,962
Phonetic Vector (Distance 4.5)	**216,156**
Phonetic Vector (Distance 5.0)	685,516

Table 4.5: Phonetic vector number of matches comparison

Table 4.5 contains the match size comparison for Phonetic vectors distance. Distance 4.5 was chosen as the suitable size due to the total matches fitting well within the other metrics. Distance 4.0 was the other alternative, this was discarded due to its small size. Any metric with a number of matches below 100,000 results in an inadequate number of possible near-collision keys later in the process, a lower number of matches may imply better quality but a balance is required between computational cost and attack quality. This will be discussed in more detail in later chapters.

Comparisons

Each participant is asked to rate the similarity of a match on a scale of 1 to 5. Figure 4.4 shows an example match and the connected scale. Users will complete 5 of these comparisons per metric.

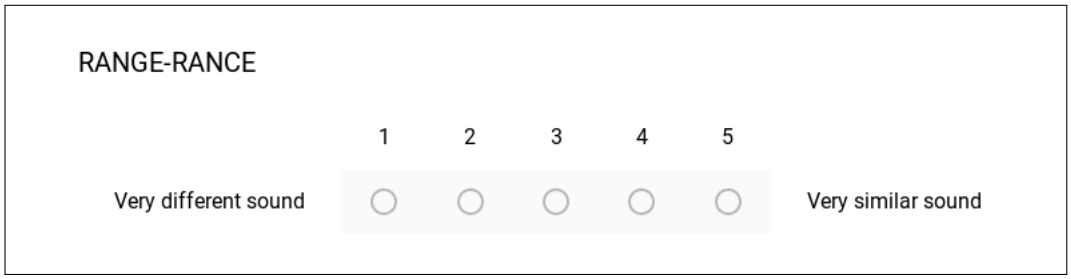

Figure 4.4: Example experiment question

The experiment randomizes the order of these comparisons per session alongside a complete refreshing of matches once per submission. This makes sure selections from the samples are fair and consistent as each user will have a different selection of matches.

Quality control

As the study was being outsourced to Amazon's Mechanical Turk the requirement to check the quality of results is important. Therefore, a couple of additions were provided to check a result's validity.

The first was the addition of 5 "Random matches" questions. These were two random words selected from the dictionary. Therefore, the overall rating of these words should be close to an average of 1. This allows a simple check to see if the user is providing valid results. If their average is too high, the results will be discarded.

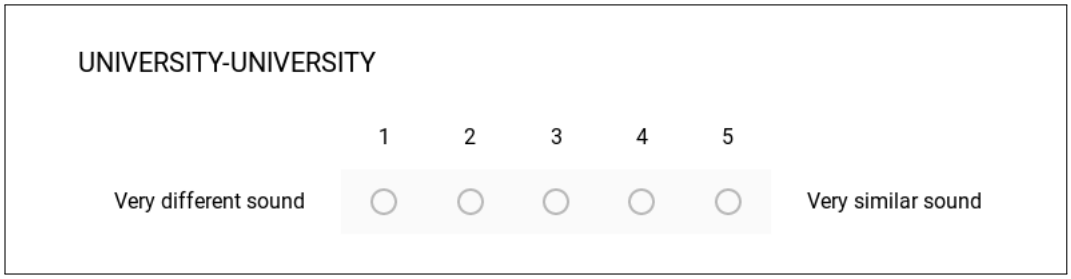

Figure 4.5: Exact experiment attention question

Alongside this, was the addition of two questions comparing exactly the same word. As both words are the same the result should always be a full 5/5 rating, any results without a full rating will be discarded. Figure 4.5 contains one example of the attention questions used to filter inaccurate results.

The final check to ensure the validity of results is a check for native English fluency. Having non-native English speakers complete the experiment has the possibility to introduce inconsistencies into the data and, therefore, needs to remain controlled. To achieve this a preliminary MTurk study will be run to ask users for their perceived English fluency on a scale of 1 (Basic Understanding) to 5 (Native). Workers with an answer of 5 will be provided with a Qualification[4] that tags them as defining themselves as native speakers. This is then added as a prerequisite when running the experiment. This will insure only fully native speakers are being assessed. The possibility of the worker falsy stating their level of fluency has also been considered. However, as the worker does not know the purpose for this initial study, thereby providing inaccurate results has no real valid motive for the worker.

Statistical

TODO

Considerations

This section will discuss the considerations required when interpreting the results of the study.

The first consideration is the way the words are compared. Due to the channel of authentication being phone based the comparison of words will be a mixture of auditory and textual. This is because a user needs to match a words sound to the one displayed on their device. In the case of this experiment the users are asked to compare the sound of two word that are visually displayed to them. Therefore, this experiment does not fully capture the targeted scenario. This was decided partly to aforementioned lack of resources. The main design of the experiment was to provide a simple way to cull the large number of metrics. The initial choice of five metrics discussed in Section 4.2 was to be followed by a more empirical decision. Further work, therefore, could improve by producing more conclusive results on the performance of the metrics.

Another consideration is the demographics of people accessible on

[4] https://blog.mturk.com/tutorial-understanding-requirements-and-qualifications-99a26069fba2

Mechanical Turk. A comprehensive and still active survey run by D Difallah *et al.* [35] shows that MTurk workers are younger and have a larger household income than that of the US population. This has to be taken into consideration when interpreting the results because it will inherently introduce a bias. This is due to the assumptions that an increase in age provides a better understanding of more obscure words, and, therefore, a better representation of the metrics performance over the entire dictionary.

4.5.2 Trustword Attacks

The final experiment required to answer the Research Question 6 will simulate attacks on participants. The user will be presented with the same design as presented in the p≡p Android application (See Section 2.3.2).

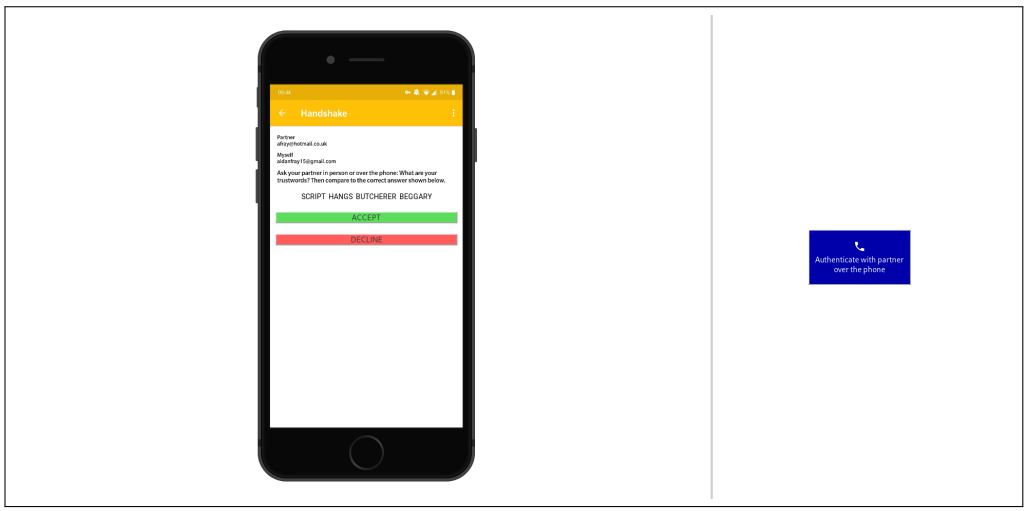

Figure 4.6: Experiment UI

Figure 4.6 shows the design of the experiments front end. As can be seen by comparing it to Figure 2.2 it has been designed to be a close as possible to the actual application. Users will click the blue button to simulate the authentication ceremony over the phone. A text-to-speech system will then read a set of words and the user should accept if they match and decline if they don't.

Design

Due to the scarcity of attacks in a real-world setting users will are often complacent about their occurrence. Therefore, in this experiment attacks only occur 30% of the time. This is a much higher value than in a real-world setting, but it is a balance between resources and realistic design. Alongside, this the first 5 trials of a new experiment will always be benign. This is to ensure users are lulled into a level of complacency regarding the possibility for an attack.

Certain keys have higher levels of near-collision possibilities than others due to how certain words are deemed similar by the chosen metric. Therefore, this presents the possibility to have keys with a very low number of similar matches. The distribution of these keys will be

explained further in Chapter 6. Therefore, if random sets of words were chosen with no restraints there are attacks presented to the user that in a real setting would be infeasible in this context as they are close to a 2^{64} attack (4 Words * 16-bits). Therefore, the experiment will sample from a list of "vulnerable" keys. These are keys that have a number of combinations that allow an attack in a certain timeframe. Furthermore, certain levels of attacks are also simulated. Below are the list of possible attacks:

- **Zero static words** - All words in the set can vary.
- **One static word** - All but the first word can vary.
- **Two static words** - All but the first and last word can vary.

The start and ends were chosen as the highest priority words to keep static. This is due to research highlighting on the common habit of users to only check the start and end of a checksum. [36] shows this by measuring user eye movement and displaying it as a heat map.

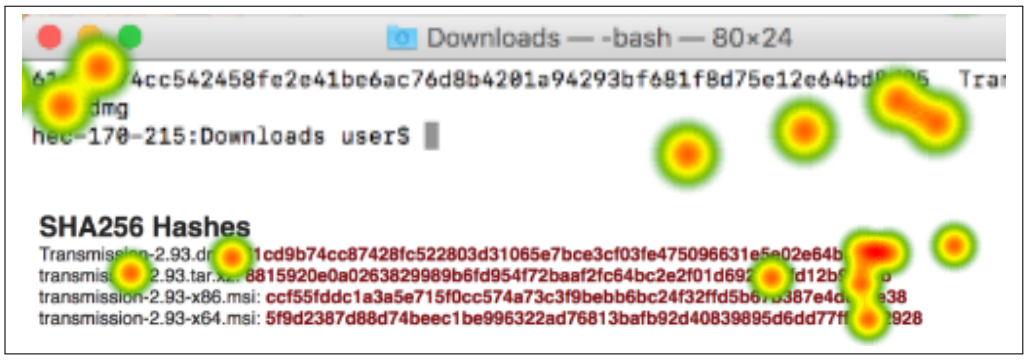

Figure 4.7: Failed verification of an incorrect checksum[36]

Figure 4.7 shows the results of a eye tracking experiment where the user failed to detect a mismatching digest. As it can be clearly seen the users never compared the center of the digest.

However, this may not be applicable in this context due to the user being linearly fed an audio stream that they cannot skip to the start or end. However, we believe that attention will also play a part in defining the most important areas of comparison as the hypothesis is that the users attention will be lower in the middle sections of comparison where it peak at the start and end. All these aspects, however, will require empirical evidence to arrive at a conclusion. Therefore, this consideration will need to be made when assessing the results of the experiment.

These attacks are ascending in complexity. The timeframe for finding a match was decided as 7-days on with attack strengths of 1 GPU/day, 10

GPU/days and 100 GPU/days on a mid range GPU (AMD RX 480 was the example in experiments).

GPU/days	No of combinations required	Attack Type
1	15250	Zero static
10	1525	One static
100	152	Two static

Table 4.6: Summary of attack requirements

Table 4.6 contains a summary of the different level of attacks and their respective restrictions. More combinations makes the attack search quicker as there are more possibilities.

Any key that exceeds this criteria is deemed as vulnerable is one of the three attack contexts. A list of these vulnerable keys are sampled from when an attack is simulated. The percentage of vulnerable keys will be explored in Chapter 6.

Quality control

As the previous experiment, participants will be sourced from MTurk. Therefore, again, quality control is and aspect that requires consideration. As with the previous study, initially screened native speakers will be used. The same process will be performed to recruit suitable workers. Two metrics will be used to detect invalid results. Audio button clicks and overall time taken.

Audio buttons clicks is the number of times the blue "*Authenticate with partner over the phone*" button in Figure 4.6. If there are rounds without button clicks, this is a sign of non-attentive users. The design choice was made to keep this as a result filter, as preventing non-clicks would be a trivial task. This provides a way to detect and discard click-throughs[5].

Overall time taken is as the name suggest a recording of the time taken to complete the entire experiment. If users complete the experiment in an abnormally small amount of time it can be used as another detection for invalid responses. Anything below a certain threshold will be discarded.

Statistical

TODO

[5]Workers that aim to complete the task as fast as possible, with no regard for the quality of responses

Considerations

TODO

Demographics of MTurk workers (Discussed in the previous section)

5 Implementation

TODO: Talk here about the interesting and challenges that were overcome

5.1 GreenOnion

General design

This section will discuss the overall technical implementation of GreenOnion alongside a deeper discussion into the unique aspect of the tool.

The tool starts by generating a 2048-bit RSA key through GPG[1]. This key is then used to create a hash of all but the final block of the key. The final block of the key is left un-hashed due to the presence of the exponent (e). This exponent is incremented by one to provide a new key. This is the aspect provides rapid generation of unique key as work for the GPU. This prevents issues in such as entropy starvation when generating a large number of valid RSA keys. This produces valid keys but with abnormally large exponents, this was deemed as suitable due to the short term use-case for these keys. 3 bytes are used to represent the exponent giving the potential to create $2^{24} - 1$ extra keys for one expensive key generation.

All keys plus the intermediate hash are loaded on the GPU via an OpenCL kernel. This kernel is designed with hashing the final block, obtaining the final fingerprint and checking if the fingerprint is present in the provided list.

Bloom filter

Mass checking of fingerprints is the tools main functionality as it allows for millions of keys to be checked with a very small amount of overhead. This is achieved through the use of a bloom filter. A bloom filter is a probabilistic data structure that allows efficient checking for present of

[1] GNU Privacy Guard

an element in a set. It is effectively a very large array of booleans that state wether an element is present. A position in the array is decided by a hashing algorithm. This is repeated a number of times with a number of hashing algorithms (k) to populate the array of length (m). This then means checking the presence of an element in the set will just mean hashing the target and checking the elements returned. This data structure therefore has a complexity of $O(k)$ regardless of the number of elements in the set.

However, due to the random distribution of the hashing algorithms commonly there is the possibility for collisions and, thus, the possibility of false-positives. The data structure also does not produce any false-negatives. This is fully suited to this use-case as any fingerprints tagged as "possibility" being present in the bloom filter can be followed up with a more expensive hash-table check to determine the actual presence. Therefore, if the levels of false-positives are controlled (by altering k and m) the tool can search through huge number of potential keys without any decrease in speed.

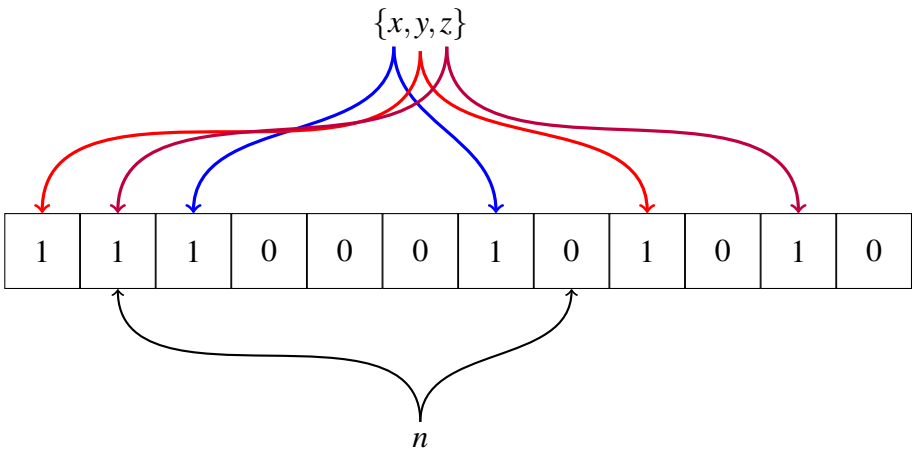

Figure 5.1: Bloom filter example

Figure 5.1 shows the operation of a simple bloom filter. As it can be seen the element n definitely does not exist in the set due to both array indices not being set. The performance boost when comparing large numbers of keys is substantial when compared to the similar tool Scallion mentioned in Section 4.4. The results from the comparison will be quantified in Chapter 6.

5.2 First Experiment

Explain how Google App Script was used?

5.3 MainExperiment

Design of the application etc

6 Experiments

6.1 Scallion vs GreenOnion

6.2 Metric Performance - Results

As discussed in Section 4.5.1 the goal of the experiment was the cull the number of metrics to be assessed in the following experiment. This section will discuss the demographics of participants alongside the subsequent results.

Gender	Male:	46.2%
	Female:	53.8%
Age:	18-24:	10.6%
	25-29:	20.2%
	30-39:	30.8%
	40-49:	22.1%
	50-59:	11.5%
	60-69:	3.8%
	70-79:	1.0%
Highest Education:	Bachelor's degree:	51.0%
	A-Level/O-Level:	18.3%
	GCSE:	15.4%
	Master's degree:	13.5%
	PhD:	1.9%

Table 6.1: Participant demographics

Metric	Average Rating
Leven	3.66
NYSIIS	2.92
Metaphone	2.56
Phonetic Vec	2.50
Soundex	2.08
Random	1.16

Figure 6.1: Average metric performance

Table 6.1 contains the demographical breakdown of the participants assessed. As it can be seen over 60% of participants can be considered highly educated (Bachelor's and up). This is not fully reflective of the general population and therefore, has to be considered when interpreting the results. All participants were sourced

from the US, this again requires consideration due to the large range of dialects present that may bias the results. Further work could investigate the affect of location and dialect on similar results.

Overall, 104 participants were assessed in this study. Five results were discarded from the set due to either failing the attention questions (See in Section 4.5.1) or having having too low of a fluency rating. This was a necessary process to improve the health of the results.

Figure 6.1 shows the average results for the metrics. It can be seen that Levenshtein came out substantially above the rest. The breakdown of the ratings in Figure 6.7 also shows Levenshtein's dominance. Levenshtein has a much larger proportion of 4 and 5 ratings than the alternatives alongside a very low level of low ratings. This performance may, however, be due to the visual way the comparisons are being performed. (Discussed in detail in Section 4.5.1).

The visual comparison will be the first point of contact between the question and the participant, then followed by the mental comparison of the phonetics. Therefore, this initial visual contact has the potential to bias the results of the study. This would affect Levenshtein due to the matches never being more than one character different, therefore, resulting in very similar looking words. The alternative is to run the study with only audio based comparisons between words, this will force the user to compare just the phonetics of the words and not be influenced by the visuals of the pair. This, however, adds cost onto the time and execution of the study. Therefore, even with the discussed issues, the aim of the experiment was to promptly reduce the number of metrics for use later in the project due to a lack of project resources. This experiment, therefore, has achieved that goal of providing three metrics for the subsequent experiment while balancing between accuracy and expenditure. Further work could aim to reproduce this study with the proposed audio based design.

6.3 Experiment 2

6.4 Distribution of keys permutations (Vuln-keys)

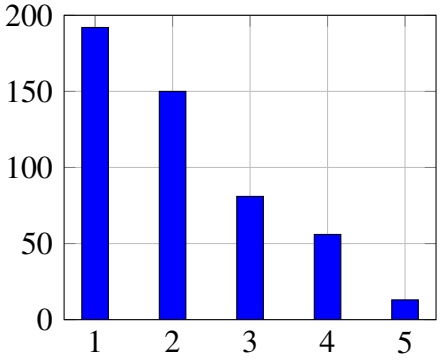

Figure 6.2: Soundex

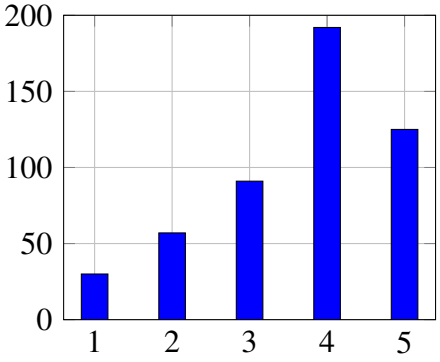

Figure 6.3: Levenshtein

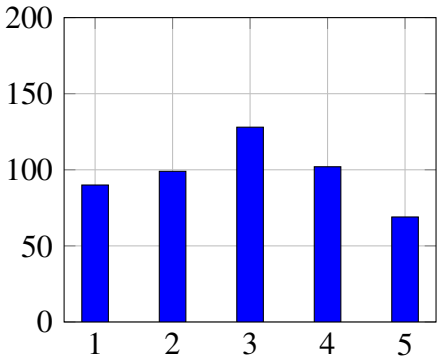

Figure 6.4: NYSIIS

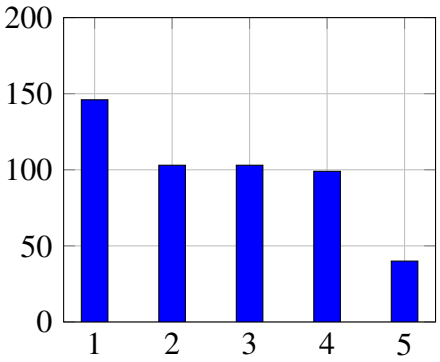

Figure 6.5: Metaphone

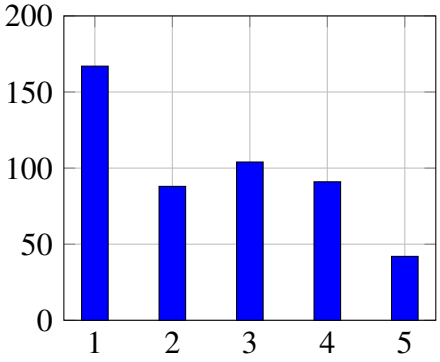

Figure 6.6: Phonetic vector

Figure 6.7: Individual breakdown of results for each metric

7 Conclusion

Bibliography

[1] 'Whatsapp encryption overview - technical white paper', Dec. 2017.

[2] G. A. Miller, 'The magical number seven, plus or minus two: Some limits on our capacity for processing information.', *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[3] S. Dechand, D. Schürmann, K. Busse, Y. Acar, S. Fahl and M. Smith, 'An empirical study of textual key-fingerprint representations', in *25th {USENIX} Security Symposium ({USENIX} Security 16)*, 2016, pp. 193–208.

[4] J. Tan, L. Bauer, J. Bonneau, L. F. Cranor, J. Thomas and B. Ur, 'Can unicorns help users compare crypto key fingerprints?', in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 3787–3798.

[5] R. Kainda, I. Flechais and A. Roscoe, 'Usability and security of out-of-band channels in secure device pairing protocols', in *Proceedings of the 5th Symposium on Usable Privacy and Security*, ACM, 2009, p. 11.

[6] H.-C. Hsiao, Y.-H. Lin, A. Studer, C. Studer, K.-H. Wang, H. Kikuchi, A. Perrig, H.-M. Sun and B.-Y. Yang, 'A study of user-friendly hash comparison schemes', in *2009 Annual Computer Security Applications Conference*, IEEE, 2009, pp. 105–114.

[7] A. Perrig and D. Song, 'Hash visualization: A new technique to improve real-world security', in *International Workshop on Cryptographic Techniques and E-Commerce*, 1999, pp. 131–138.

[8] C. Ellison and S. Dohrmann, 'Public-key support for group collaboration', *ACM Transactions on Information and System Security (TISSEC)*, vol. 6, no. 4, pp. 547–565, 2003.

[9] Y.-H. Lin, A. Studer, Y.-H. Chen, H.-C. Hsiao, L.-H. Kuo, J. M. McCune, K.-H. Wang, M. Krohn, A. Perrig, B.-Y. Yang *et al.*, 'Spate: Small-group pki-less authenticated trust establishment', *IEEE Transactions on Mobile Computing*, vol. 9, no. 12, pp. 1666–1681, 2010.

[10] M. Shirvanian, N. Saxena and J. J. George, 'On the pitfalls of end-to-end encrypted communications: A study of remote key-fingerprint verification', in *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACM, 2017, pp. 499–511.

[11] E. Uzun, K. Karvonen and N. Asokan, 'Usability analysis of secure pairing methods', in *International Conference on Financial Cryptography and Data Security*, Springer, 2007, pp. 307–324.

[12] M. M. Olembo, T. Kilian, S. Stockhardt, A. Hülsing and M. Volkamer, 'Developing and testing a visual hash scheme.', in *EISMC*, 2013, pp. 91–100.

[13] D. Loss, T. Limmer and A. von Gernler, *The drunken bishop: An analysis of the openssh fingerprint visualization algorithm*, 2009.

[14] A. Karole and N. Saxena, 'Improving the robustness of wireless device pairing using hyphen-delimited numeric comparison', in *2009 International Conference on Network-Based Information Systems*, IEEE, 2009, pp. 273–278.

[15] P. Juola, 'Whole-word phonetic distances and the pgpfone alphabet', in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, IEEE, vol. 1, 1996, pp. 98–101.

[16] M. T. Goodrich, M. Sirivianos, J. Solis, G. Tsudik and E. Uzun, 'Loud and clear: Human-verifiable authentication based on audio', in *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*, IEEE, 2006, pp. 10–10.

[17] N. Haller, 'The s/key one-time password system', 1995.

[18] K. Rieck, 'Fuzzy fingerprints attacking vulnerabilities in the human brain', *Online publication at http://freeworld. thc. org/papers/ffp. pdf*, 2002.

[19] M. Shirvanian and N. Saxena, 'Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones', in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2014, pp. 868–879.

[20] D. Mukhopadhyay, M. Shirvanian and N. Saxena, 'All your voices are belong to us: Stealing voices to fool humans and machines', in *European Symposium on Research in Computer Security*, Springer, 2015, pp. 599–621.

[21] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su and A. Mohaisen, 'You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones', in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2017, pp. 183–195.

[22] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, 'Spoofing and countermeasures for speaker verification: A survey', *speech communication*, vol. 66, pp. 130–153, 2015.

[23] R. L. Peterson and B. F. Pennington, 'Developmental dyslexia', *The Lancet*, vol. 379, no. 9830, pp. 1997–2007, 2012.

[24] *Mysql 5.5 reference manual :: 12.5 string functions and operators*. [Online]. Available: https://dev.mysql.com/doc/refman/5.5/en/string-functions.html#function_soundex.

[25] *Oracle - database sql reference*, Jul. 2005. [Online]. Available: https://docs.oracle.com/cd/B19306_01/server.102/b14200/functions148.htm.

[26] *F.15. fuzzystrmatch*. [Online]. Available: https://www.postgresql.org/docs/9.1/fuzzystrmatch.html.

[27] F. Patman and L. Shaefer, 'Is soundex good enough for you? on the hidden risks of soundex-based name searching', *Language Analysis Systems, Inc., Herndon*, 2001.

[28] L. Philips, 'Hanging on the metaphone', *Computer Language*, vol. 7, no. 12, pp. 39–43, 1990.

[29] *Metaphone*. [Online]. Available: https://www.php.net/manual/en/function.metaphone.php.

[30] G. P. Hettiarachchi and D. Attygalle, 'Sparcl: An improved approach for matching sinhalese words and names in record clustering and linkage', in *2012 IEEE Global Humanitarian Technology Conference*, IEEE, 2012, pp. 423–428.

[31] A. Parrish, 'Poetic sound similarity vectors using phonetic features', in *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2017.

[32] N. Chomsky and M. Halle, 'The sound pattern of english.', 1968.

[33] P. Ladefoged, 'The measurement of phonetic similarity', in *INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS COLING 1969: Preprint No. 57*, 1969.

[34] A. Bradlow, C. Clopper, R. Smiljanic and M. A. Walter, 'A perceptual phonetic similarity space for languages: Evidence from five native language listener groups', *Speech Communication*, vol. 52, no. 11-12, pp. 930–942, 2010.

[35] D. Difallah, E. Filatova and P. Ipeirotis, 'Demographics and dynamics of mechanical turk workers', in *Proceedings of the eleventh acm international conference on web search and data mining*, ACM, 2018, pp. 135–143.

[36] M. Cherubini, A. Meylan, B. Chapuis, M. Humbert, I. Bilogrevic and K. Huguenin, 'Towards usable checksums: Automating the integrity verification of web downloads for the masses', in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2018, pp. 1256–1271.