The Effect of Length on Key Fingerprint Verification Security and Usability

Dan Turner
dan@turnerhallow.co.uk

Siamak F. Shahandashti
University of York, UK
siamak.shahandashti@york.ac.uk

Helen Petrie
University of York, UK
helen.petrie@york.ac.uk

ABSTRACT

In applications such as end-to-end encrypted instant messaging, secure email, and device pairing, users need to compare key fingerprints to detect impersonation and adversary-in-the-middle attacks. Key fingerprints are usually computed as truncated hashes of each party's view of the channel keys, encoded as an alphanumeric or numeric string, and compared out-of-band, e.g. manually, to detect any inconsistencies. Previous work has extensively studied the usability of various verification strategies and encoding formats, however, the exact effect of key fingerprint length on the security and usability of key fingerprint verification has not been rigorously investigated. We present a 162-participant study on the effect of numeric key fingerprint length on comparison time and error rate. While the results confirm some widely-held intuitions such as general comparison times and errors increasing significantly with length, a closer look reveals interesting nuances. The significant rise in comparison time only occurs when highly similar fingerprints are compared, and comparison time remains relatively constant otherwise. On errors, our results clearly distinguish between security non-critical errors that remain low irrespective of length and security critical errors that significantly rise, especially at higher fingerprint lengths. A noteworthy implication of this latter result is that Signal / WhatsApp key fingerprints provide a considerably lower level of security than usually assumed.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; *Software security engineering*; • **Human-centered computing →** *Empirical studies in HCI*.

KEYWORDS

Key fingerprint verification, Device pairing, Out-of-band channel, Authentication, End-to-end encryption, Secure messaging, Signal safety number, WhatsApp security code, Usability, Security

ACM Reference Format:
Dan Turner, Siamak F. Shahandashti, and Helen Petrie. 2023. The Effect of Length on Key Fingerprint Verification Security and Usability. In *The 18th International Conference on Availability, Reliability and Security (ARES 2023), August 29-September 1, 2023, Benevento, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3600160.3600187

ARES 2023, August 29 – September 1, 2023, Benevento, Italy

1 INTRODUCTION

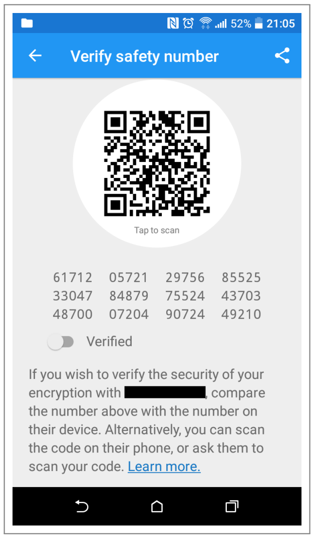
Authentic keys are required for secure communication. Devices negotiate these keys using a key exchange protocol, or use a public key that purportedly belongs to the other party. These keys may be authenticated using authenticated key exchange protocols such as password-based authenticated key exchange, or by verifying public key certificates. Such authentication is only possible when there is an existing shared security context between parties such as shared passwords or public key infrastructure (PKI). In the absence of authentication, adversaries may carry out adversary-in-the-middle (AitM, traditionally known as man-in-the-middle) or impersonation attacks to compromise security.

Digital devices have become ubiquitous, and hence there is a growing need for establishing ad hoc secure communication channels between devices, i.e. securely *pairing* devices, without a shared security context. Although impersonation and AitM attacks cannot be prevented, system designers can build in measures to *restrict* or *detect* such attacks. As an example of the restriction approach, distance bounding protocols in contactless payment systems limit the distance between the payment card or device and the point of sale terminal to minimise the possibility of AitM attacks [3], such as the so-called Mafia Fraud.

One of the most common methods to detect impersonation or AitM attacks is through an *out-of-band channel*. System designers assume that users have access to a separate secure communication channel with low bandwidth. The key observation is that the keys held by the communicating parties will differ when there is an impersonation or AitM attack, and will be identical in the absence of such attacks. The out-of-band channel is used to detect differences between the keys the two sides hold after the key exchange. Since the out-of-band channel is low bandwidth, devices usually apply a hash function to the keys and truncate the result to derive a short digest, which we call a *key fingerprint*. Comparing the short fingerprints through an out-of-band channel would provide the confidence in keys being identical bar any hash collisions.

Various formats for key fingerprints have been considered. OpenPGP, designed for email encryption, encodes public-key fingerprints as hexadecimal strings. The user then manually compares these against a trusted copy of the key fingerprint, e.g., on a business card. The ZRTP protocol for secure VoIP uses a Short Authentication String (SAS), which is a fingerprint of the key negotiated using Diffie–Hellman key exchange. The Silent Phone app shows the SAS as two words for users to verbally check. Loud and Clear, a device pairing method, creates a short sentence from the key fingerprint and speaks it aloud using a text-to-speech engine. The user checks it against a sentence shown on the other device [8].

Alphanumeric fingerprints are one of the most widely used as they are generally considered comparatively more usable. The most



61712 05721 29756 85525
33047 84879 75524 43703
48700 07204 90724 49210

If you wish to verify the security of your encryption with [redacted], compare the number above with the number on their device. Alternatively, you can scan the code on their phone, or ask them to scan your code. Learn more.

Figure 1: Safety number display in Signal for Android

widely deployed text-based format is likely to be numeric, thanks to WhatsApp's 2016 rollout of the Signal protocol for end-to-end encryption. Signal and WhatsApp use a string of 60 digits arranged in 12 chunks of 5 as shown in Figure 1. This is called a 'safety number' in Signal and a 'security code' in WhatsApp.

The key fingerprint length is usually set based on the security level required for the application. Signal and WhatsApp use 60-digit fingerprints since they need to provide long-term security against adversaries without any location restriction. However, a 4-digit fingerprint may be sufficient to safely pair two smart home devices if keys are freshly generated for one-time use and the communication protocol is short range.

There have been multiple studies on the usability of various key fingerprint formats and their susceptibility to error in the literature. However, apparently no study has investigated the effect of key fingerprint length. Intuitively, one expects that users can compare shorter key fingerprints more quickly and with fewer errors, but the veracity of this intuition does not seem to have been empirically tested yet. Such a rigorous study is also needed to clarify the parameters of the apparent trade-off between security and usability for a range of fingerprint lengths and provide crucial empirical evidence for designers when deciding on the specifications for key fingerprint verification methods.

In this work, we contribute to the understanding of the effect of key fingerprint length on the usability and security of manual key fingerprint verification. We focus on numeric key fingerprints because of their comparative usability, and specifically consider the Signal / WhatsApp format, as it is widely deployed. We present the result of a study in which participants were asked to compare Signal / WhatsApp-like key fingerprints of three different lengths. We measured how the key fingerprint length affects comparison time and accuracy. Analyses of our results provide evidence in support of a number of points that so far have been poorly understood in the literature. Namely, the results show that comparison time only changes significantly when fingerprint pairs of high similarity are being compared, but otherwise stays relatively constant. Furthermore, we present strong evidence showing that the security non-critical error rate remains fairly low even for long fingerprints, whereas the security critical error rate grows significantly at higher lengths. One of the main implications of these results is that Signal / WhatsApp key fingerprints provide considerably lower levels of security than intended.

Paper outline: Section 2 summarises the related work, Section 3 outlines our research questions and study design, Section 4 discusses the results, and Section 5 draws conclusions from the study.

2 RELATED WORK

There has been no previous study considering length as an independent variable. Therefore, in this section we provide a brief overview of the main results on manual fingerprint verification to set out the context in which our work is conducted.

Various formats for key fingerprints have been proposed in the literature or deployed in practice for manual comparison. Examples include *hexadecimal* e.g. GnuPG [5], *numeric* e.g. Signal, WhatsApp, and SafeSlinger [7], *words* (and pseudo-words) e.g. Bubble Babble encoding [11], *sentences* e.g. pseudo-random poems [2]. *graphical* e.g. abstract art [17], ASCII art [1], snowflakes [14], and unicorns [21], and *auditory* e.g. Loud and Clear [8].

Several teams investigated the comparative usability of various key representations including Kainda et al. [12], Dechand et al. [6], and Tan et al. [21]. These studies broadly found that alphanumeric and numeric representations offer better perceived usability, comparison speed, and accuracy. The considered numeric fingerprint lengths in these studies were 6, 34, and 48 digits, respectively.

The usability and security of key fingerprint verification for end-to-end encrypted instant messaging apps have been the subject of studies by Herzberg et al. [10], Schröder et al. [19], and Shirvanian et al. [20]. Evidence presented in these works unanimously points towards high error rates and low perceived usability of manual verification. More recently, Livsey et al. studied word-based manual fingerprint verification when compared visually or verbally and found that visual comparisons are more effective against security non-critical errors [15]. Considering the entire authentication ceremony in these apps, Vaziripour et al. found low usability, including low completion rates [24]. Follow-up studies showed rephrasing the task and redesigning the user interface is effective in helping users understand and perform the ceremony correctly [23, 26].

Evidence of low prevalence of manual verification has been reported in the literature. For instance, in an attempt to study whether users verify SSH key fingerprints, Gutmann approached two large organisations with 'several thousand computer-literate users', and found that staff were 'unable to recall a single case, or locate any records, or any user ever verifying any SSH server key out-of-band' [9].

Device pairing methods are related to manual fingerprint verification and have been studied for their comparative usability and security, notably by Kobsa et al. [13] and Uzun et al. [22]. Comparing numeric fingerprints has been consistently found to be perceived more usable, provide better speed, and lead to less security critical errors compared to other methods in these studies.

A pertinent research question here concerns the most effective adversarial strategy in crafting a similar fingerprint that would pass less attentive human verification. Cherubini et al. provide eye-tracking evidence that attention to compared strings is highest at the beginning of the string and decreases as progress is made towards the end [4]. Furthermore, several works have hypothesised that human attention is heavily biased towards the beginning and end of the compared sequences [6, 9, 18].

3 STUDY DESIGN

We consider the Signal / WhatsApp numeric key fingerprint format because of its comparatively higher usability and its wide deployment. As shown in Figure 1, these fingerprints are represented in three lines, each containing four 5-digit chunks, in their full format. To study the effect of length, we consider three length *conditions*:

- **1 Line** (1L): a fingerprint includes *four* 5-digit chunks in 1 line, corresponding to 1 line out of 3 of the full format,
- **2 Lines** (2L): a fingerprint includes *eight* 5-digit chunks in 2 lines, corresponding to 2 lines out of 3 of the full format, and
- **3 Lines** (3L): a fingerprint includes *twelve* 5-digit chunks in 3 lines, corresponding the full Signal / WhatsApp format.

To minimise the effect of inconsistent formats, we opted for a *between-participants* design with respect to length conditions, i.e. each participant will be randomly assigned to one condition and all the fingerprints they compare will be of the same length according to the condition they are assigned.

Compared key fingerprint pairs can be either matching or non-matching. An adversary may trade off attack success probability with computation and be happy with a nearly matching fingerprint that may fool a proportion of users. To be able to investigate the interplay of the effect of each of these possibilities with that of fingerprint length, we consider three comparison *types*:

- **Safe**: a comparison between a pair of *fully matching* (i.e. identical) fingerprints,
- **Adversarial** (Adv.): a comparison between a pair of *nearly matching* fingerprints with only 1 chunk being different, and
- **Random** (Rand.): a comparison between a pair of *randomly selected* (and hence highly dissimilar) fingerprints.

The above types represent scenarios where a user encounters an authentic key, an adversarially crafted one in case of an attack, or an erroneous key, respectively.

To closely follow what would happen in practice where the same user may compare safe, adversarial, or random fingerprints, we opted for a *within-participants* design with respect to comparison types, i.e. each participant will carry out comparisons of all types.

It is expected that in practice users will be comparing safe fingerprints most of the time and the occurrence of attack scenarios will be limited to rare occasions. Hence, a realistic study should contain as few adversarial pairs as possible. At the same time, gathering sufficient data to compute reliable security-critical error rates requires as many adversarial pairs as possible. We decided to strike a balance between these two competing goals by designing the study to show *12 safe, 4 adversarial, and 4 random* key fingerprint pairs to each participant. Dechand et al. follow a similar principle [6]. The 20 key pairs are shown to the participant in a random order different for each participant to counterbalance the possible effects of habituation and fatigue.

We emphasise that the scenario we consider is *manual* fingerprint verification carried out *individually*. This is also the approach taken by Kainda et al. [12], Dechand et al. [6], and Tan et al. [21]. Automated verification, such as scanning the QR code provided by Signal / WhatsApp using a smartphone camera, and collaborative verification, i.e. two users carrying out the comparison together, are both outside the scope of our study.

3.1 Adversarial Model

We consider adversaries that are able to intercept initial key exchange messages between user devices and replace them with adversarially chosen ones. However, the adversary does not have the ability to modify messages on the out-of-band channel, i.e. the channel through which key fingerprints are compared and verified. The goal of the adversary is to impersonate one or both of the entities, corresponding to impersonation or AitM attacks, respectively.

These capabilities allow the adversary to replace a user's authentic keys with their own which would result in key fingerprints being computed on different keys. Specifically for the Signal / WhatsApp key fingerprint format, we allow adversaries to create key fingerprints that matched all but one of the key fingerprint chunks. This is to keep the level of similarity high between adversarial pairs.

The Signal / WhatsApp fingerprint is made of two halves, each a 30-digit fingerprint of the so-called 'identity key' of one of the two parties [16, 25]. From each party's viewpoint, the adversary may only compromise one of these two halves since each party 'knows' the authentic version of their own key. Hence, we did not allow adversarial digits to cross the midpoint boundary and restricted the adversary to manipulating digits only in the second half of the fingerprint. The chunk not targeted for collision by the adversary was designated to be the one just after the key fingerprint midpoint. This is to maximise the likelihood that it would be overlooked since previous works suggest that users pay less attention to the middle sections of the compared fingerprints [4, 6, 9, 18].

Requiring all but one of the chunks to be identical in adversarial fingerprints corresponds to 'adversarial powers' outlined in Table 1 under 'no iteration' for each condition. For instance, for our 2 Lines condition, there are eight 5-digit chunks, four of which are computed from the key provided by the adversary. The adversary needs three out of these four chunks to be identical to those of the fingerprint half being impersonated, i.e. it needs a 3-chunk, i.e. 15-digit, collision. This is equivalent to finding a second preimage for a hash function with an output length of approximately 49.8 bits, since $10^{15} \approx 2^{49.8}$. Testing every preimage can be seen as a Bernoulli trial and hence the success probability of such an attack with respect to number of computed hashes follows the cumulative distribution function of a *geometric distribution*. It follows that the expected number of hashes that need to be computed in the attack is approximately $0.69 \times 2^{49.8}$. Despite this, the attack is said to require $2^{49.8}$ adversarial power by convention.

Modern applications use *iterated hashing* for fingerprint calculation to increase the computational cost for adversaries while keeping the cost of hashing for legitimate users within affordable bounds. For instance, WhatsApp and Signal iterate the hash 5200

Table 1: Adversarial power required to compute attack keys in each condition assuming either no iteration or 5200 iterations

	No. of Chunks		Adversarial Power	
Condition	All	Collision	no iteration	with iteration
1 Line	4	1	$2^{16.6}$	$2^{29.0}$
2 Lines	8	3	$2^{49.8}$	$2^{62.2}$
3 Lines	12	5	$2^{83.0}$	$2^{95.4}$

times to compute each fingerprint half. If such a design is used, the incurred computational cost of attacks will be about 5200 $\approx 2^{12.3}$ times higher than the base case where no iteration is used. The required adversarial powers, if 5200 iterations are used, are listed in Table 1 under 'with iteration'.

We have opted for variable adversarial power to mirror the fact that shorter fingerprints are only appropriate for safer environments, for instance use cases where adversaries are restricted in time or location. An adversary with high power would be able to easily compute keys that lead to full fingerprint collisions for shorter fingerprint lengths which would not allow us to see the effect of similar but not identical fingerprints on user performance.

3.2 Research Questions

The overall aim of our study is to investigate whether the length and similarity of key fingerprint have significant effects on a person's performance when comparing key fingerprints. We focus on user performance in the comparison task, as measured by effectiveness and efficiency. Perceived usability would be more appropriate for the overall confirmation ceremony and we do not consider it here. Accordingly, we developed three sets of hypotheses as follows.

Considering the speed with which participants can compare pairs of key fingerprints as a measure of efficiency, we tested the following high-level hypothesis $H_1^{t\sim\ell}$ on comparison time t with respect to fingerprint length ℓ, with the alternative hypothesis $H_0^{t\sim\ell}$ defined as the opposite:

$H_1^{t\sim\ell}$: Participants take longer time to compare longer numeric key fingerprints than shorter ones.

Since we are studying different comparison types, $H_1^{t\sim\ell}$ gives rise to three type-specific hypotheses for safe, adversarial, and random comparisons.

Considering safe, adversarial, and random fingerprint pairs as pairs with maximum, high, and low similarity, we tested the following high-level hypothesis $H_1^{t\sim s}$ on comparison time t with respect to fingerprint similarity s, or equivalently comparison type, with the alternative hypothesis $H_0^{t\sim s}$ defined as the opposite:

$H_1^{t\sim s}$: Participants take longer time to compare numeric key fingerprint pairs with higher similarity.

Similarly, $H_1^{t\sim s}$ is tested at three different fingerprint lengths, giving rise to three length-specific hypotheses.

Considering the accuracy with which participants can compare pairs of key fingerprints as a measure of effectiveness, we tested the following high-level hypothesis $H_1^{e\sim\ell}$ on error rate e with respect to fingerprint length ℓ, with the alternative hypothesis $H_0^{e\sim\ell}$ defined as the opposite:

$H_1^{e\sim\ell}$: Participants make more mistakes when comparing longer numeric key fingerprints than shorter ones.

Here, depending on the comparison type we consider, we have two types of errors:

- **False Acceptance Errors** occur when non-matching fingerprints are incorrectly accepted as matching, and
- **False Rejection Errors** occur when matching fingerprints are incorrectly rejected as non-matching.

Consequently, we test two error-type-specific hypotheses, i.e. $H_1^{e\sim\ell}$ for false acceptance and false rejection errors.

Note that the security implications of the two types of error can be considerably different. False acceptance errors, especially on adversarial fingerprints, would be security-critical as they would allow an AitM attack to go unnoticed. However, false rejection errors would only cause inconvenience.

It is clear that fingerprint length and comparison type are the independent variables, and comparison time, false acceptance and false rejection error rates are the dependent variables in this study.

3.3 Ethical Considerations

The ethical principles of avoidance of harm, informed consent, and data protection were followed throughout the design, data collection, and analysis phases of our study. No actual communication channels were attacked. Participants were asked for their consent after providing an information sheet at the start of the study. The participants could withdraw at any time for any reason. The information sheet explained the study and that participation was voluntary, and provided the contact details of the investigators. No personally identifiable information were collected from participants. Only general demographic data were collected to give contextual information. These were age range, gender, highest education level, and presence of a disability.

A pilot study was used to estimate the time taken to complete the study, based on which we calculated the amount to pay participants in the main study. We used the living wage for London and New York to ensure that all participants got fair pay for their time. Participants who withdrew were still paid for their time. The University of York's Physical Sciences Ethics Committee approved this work before we collected any data.

3.4 Pilot Study

First, we ran a pilot study to find any issues in the study design. We recruited participants locally by offering entry into a raffle for a £25 (GBP) Amazon gift card. We advertised the pilot study to friends and family on Facebook.

Participants publicly discussed the pilot study on Facebook. We did not intend this, but it gave us useful insights into how the participants were approaching the pilot study. Although we did not aim for many piloting participants, we recruited 60 participants, from which we excluded 17 for being inattentive as they indicated that at least one of the random fingerprint pairs matched. We asked each participant to compare 20 pairs of fingerprints, some identical and some different. We made several modifications to our study based on the pilot study feedback as explained below.

In each individual task, we asked each participant to compare a pair of fingerprints. Our question caused confusion for some of our participants, so we reworded the question from ‘Are Alice and Bob’s messages safe?’ to ‘Do the numbers match?’ to make it more clear. While the original question works well for those familiar with the purpose of key fingerprint verification, it requires a level of knowledge not generally expected of non-experts.

Some participants were unsure how to proceed, so we added more guidance. This was especially important as at least one pilot study participant commented ‘it took me waaay [sic] too long to work out it was essentially a “compare these numbers” exercise.’ We showed participants an extra screen before they started which explained the task and showed them where to find the key fingerprint on the screen. Besides, we added a counter to each page, so that progress through the study was clear to the participants.

3.5 Main Study

In the main study, each participant was randomly assigned one of the three fingerprint lengths, i.e. 1, 2, or 3 lines, and asked to compare 20 different key fingerprint pairs of the same length, comprising of 12 safe, 4 adversarial, and 4 random pairs in a randomised order. The browser window for each fingerprint pair comparison included two simulated phone screens side-by-side and asked participants ‘Do the numbers match?’ with response options ‘Yes, they match’ and ‘No, they don’t match’ as in Figure 2. Random key fingerprint pairs were used as *attention checkers*. Participants who got any of the attention checkers wrong were excluded from our analysis, but were still compensated for their time.

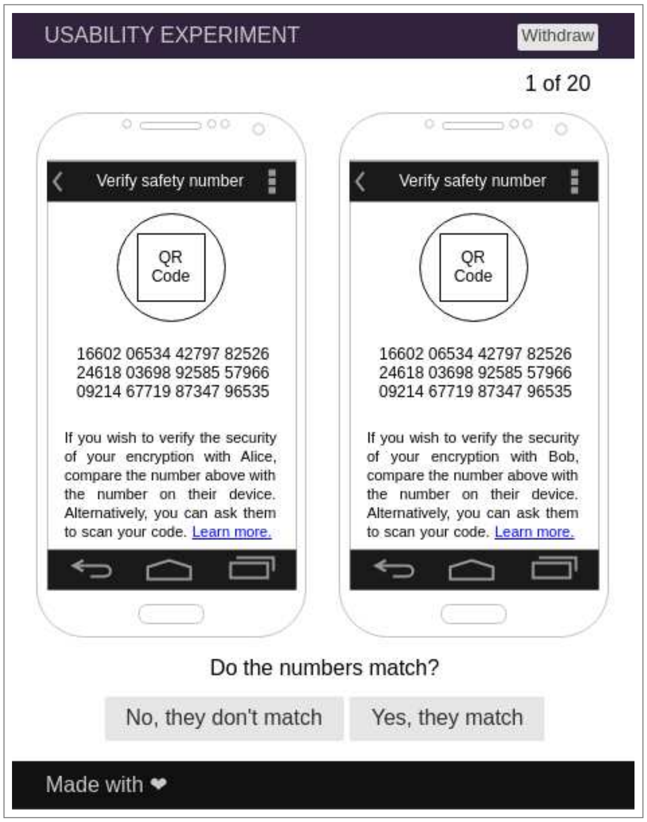
Participants were recruited through MTurk. We did not restrict which MTurk users could accept the task, other than stopping those who had already done the study. Each participant was paid $2 (USD) for their time. All of the guidance was written in English, so all participants needed a sufficient level of English reading comprehension to understand the tasks. Since the included participants all passed the attention checkers we assume this to be the case. Before starting the tasks, the participants read the information sheet and consented to take part in the study.

3.6 Technical Implementation

We built the experiment on Amazon Web Services (AWS) using Python and TypeScript. We used AWS Lambda to host the backend, stored the data encrypted in AWS DynamoDB, and fronted the site with a static site stored in AWS S3 and distributed through AWS CloudFront. We exposed the Lambda API using AWS API Gateway, which offers TLS by default, so all the participants’ data was encrypted in transit.

3.7 Study Participants

A total of 186 participants were recruited. 2 were excluded from our analysis for failing to complete the study and another 22 for failing the attention checkers. In all the following analyses, we report the results for the remaining 162 participants. Table 2 shows self-reported participant demographics. As the table shows, large proportions of our participants declared being male, young, educated, and not disabled. We had an even split however between conditions: 53, 55, and 54 participants were assigned to the 1L, 2L, and 3L conditions, respectively.



16602 06534 42797 82526
24618 03698 92585 57966
09214 67719 87347 96535

If you wish to verify the security of your encryption with Alice, compare the number above with the number on their device. Alternatively, you can ask them to scan your code. Learn more.

16602 06534 42797 82526
24618 03698 92585 57966
09214 67719 87347 96535

If you wish to verify the security of your encryption with Bob, compare the number above with the number on their device. Alternatively, you can ask them to scan your code. Learn more.

Do the numbers match?

No, they don't match | Yes, they match

Figure 2: A screenshot of the study interface for each comparison task as shown to participants.

4 RESULTS

In this section we give an overview of the collected data and the results of testing the hypotheses stated in Section 3.2, using the common $\alpha = 0.05$ significance level throughout.

We first tested for any significant demographic difference between groups of users in the three conditions. Fisher’s exact test found no significant difference in the reported gender, educational level, disability, or age between the three groups. The p-values were 0.56, 0.75, 0.91, and 0.28 respectively.

4.1 Comparison Time

We calculated each participant’s median comparison times for each three comparison types: safe, adversarial, and random comparisons. The distribution parameters of participant median comparison times by comparison type and condition are detailed in Table 3 and depicted in Figure 3. As expected, median comparison times for all nine combinations (3 conditions × 3 types) have skewed distributions with long tails. Shapiro-Wilk tests of normality were significant in all cases except for 1-line adversarial comparisons (1L

Table 2: Participant demographics

Demographic	Group	Count	Proportion
Gender	Female	40	≈25%
	Male	102	≈63%
	Other	1	<1%
	Preferred not to say	19	≈12%
Age	18–20	3	≈2%
	21–35	111	≈69%
	36–50	29	≈18%
	51–60	4	≈2%
	61 and above	0	0%
	Preferred not to say	15	≈9%
Education	High school diploma	35	≈22%
	Bachelor's degree	94	≈58%
	Master's degree	17	≈10%
	Professional degree	1	<1%
	Preferred not to say	15	≈9%
Disability	Declared a disability	12	≈7%
	Declared no disability	132	≈81%
	Preferred not to say	18	≈11%

Table 3: Distribution parameters (lower quartile, median, upper quartile) of median comparison times (in seconds) by comparison type (Safe, Adversarial, and Random) and condition (1, 2, 3 Lines)

Type	1 Line	2 Lines	3 Lines
Safe	(4.9, **5.7**, 7.7)	(4.9, **7.7**, 10.8)	(5.0, **10.7**, 14.8)
Adversarial	(4.1, **4.9**, 6.0)	(3.4, **5.9**, 7.1)	(4.7, **8.9**, 10.6)
Random	(2.7, **3.3**, 3.8)	(2.7, **3.2**, 3.8)	(2.9, **3.4**, 4.4)

safe $p < 0.001$, adv. $p = 0.071$, rand. $p = 0.004$, 2L safe $p = 0.001$, adv. $p = 0.007$, rand. $p < 0.001$, 3L safe $p = 0.016$, adv. $p = 0.028$, rand. $p < 0.001$) indicating that 8 out of 9 of the median time distributions are significantly non-normal. Hence, non-parametric tests were used for analysis. For analysing change with fingerprint length, we have independent samples and hence Kruskal–Wallis test was used, whereas for analysing change with comparison type, we have related measures and hence Friedman test was appropriate.

4.1.1 Change with Fingerprint Length. For safe fingerprint comparisons, Kruskal–Wallis test found statistically significant differences between median comparison times for fingerprints of various lengths ($\chi^2(2) = 13.3$, $p = 0.001$). The effect size was moderate ($\eta^2[H] = 0.071$). Pairwise Wilcoxon test between groups with Holm correction found significant differences between all conditions (1L–2L: $W = 1112$, $p = 0.049$, 2L–3L: $W = 1114$, $p = 0.049$, 1L–3L: $W = 905$, $p = 0.003$).

For adversarial comparisons, Kruskal–Wallis test found statistically significant differences between median comparison times for fingerprints of various lengths ($\chi^2(2) = 22.3$, $p < 0.001$). The effect size was moderate ($\eta^2[H] = 0.128$). Pairwise Wilcoxon test between groups with Holm correction showed that only the differences between 3-line comparisons and the other two groups were significant (1L–2L: $W = 1277$, $p = 0.269$, 2L–3L: $W = 889$, $p < 0.001$, 1L–3L: $W = 728$, $p < 0.001$).

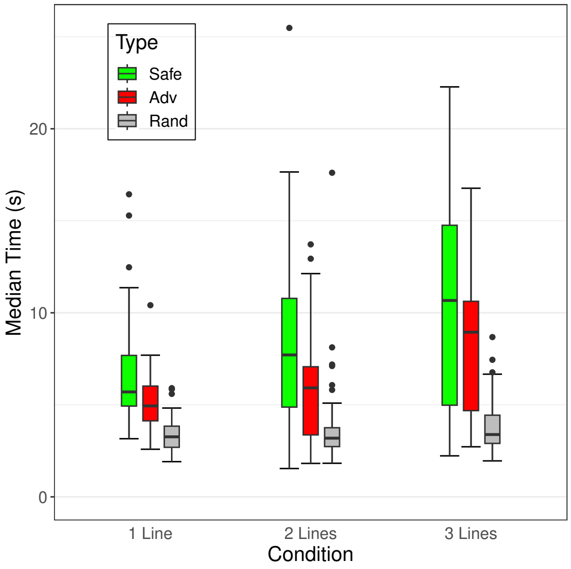


Figure 3: Distributions of participant median times to compare fingerprints by condition (1, 2, 3 Lines) and comparison type (Safe, Adversarial (Adv), Random (Rand))

For random comparisons, Kruskal–Wallis test did not find statistically significant differences between median comparison times for fingerprints of various lengths ($\chi^2(2) = 3.68$, $p = 0.159$).

The analysis above shows that although we can reject $H_0^{t\sim\ell}$ for safe comparisons and for adversarial comparisons at higher fingerprint lengths, namely for 2L–3L and 1L–3L comparisons, the same cannot be done for random comparisons. This means that our a priori expectation of median comparison time increasing with fingerprint length only holds when similarity between compared fingerprints is high (e.g. in the case of safe pairs that are identical), but as the differences between compared fingerprints grow larger (e.g. in random pairs) the differences between comparison times for various lengths become insignificant to the point that median comparison times stays approximately constant for 1-line, 2-line, and 3-line random fingerprints.

4.1.2 Change with Comparison Type. For 1-line comparisons, Friedman test found statistically significant differences between the distributions of median times for safe, adversarial, and random comparisons ($\chi^2(2) = 77.43$, $p < 0.001$). The effect size was large (Kendall $W = 0.73$). Nemenyi post hoc test indicated significant differences between median time distributions for all three pairs of comparison types (safe–adv.: $p = 0.001$, adv.–rand.: $p < 0.001$, safe–rand.: $p < 0.001$).

For 2-line comparisons, Friedman test found statistically significant differences between the distributions of median times for safe, adversarial, and random comparisons ($\chi^2(2) = 52.51$, $p < 0.001$). The effect size was moderate (Kendall $W = 0.48$). Nemenyi post hoc test indicated significant differences between median time distributions for all three pairs of comparison types (safe–adv.: $p < 0.001$, adv.–rand.: $p < 0.001$, safe–rand.: $p < 0.001$).

For 3-line comparisons, Friedman test found statistically significant differences between the distributions of median times for safe, adversarial, and random comparisons ($\chi^2(2) = 62.11, p < 0.001$). The effect size was large (Kendall $W = 0.58$). Nemenyi post hoc test indicated significant differences between median time distributions for all three pairs of comparison types (safe–adv.: $p = 0.011$, adv.–rand.: $p < 0.001$, safe–rand.: $p < 0.001$).

The analysis above shows that for all three different lengths of fingerprints we considered, our participants compare random pairs of fingerprints significantly more quickly than adversarial pairs, and adversarial pairs significantly more quickly than safe pairs. Therefore, we emphatically reject $H_0^{t\sim s}$ for all fingerprint lengths. In other words, the more the differences between the compared fingerprints, the less amount of time it takes on average to compare them and decide whether they are identical or not. This observation, coupled with the similar observations in Section 4.1.1, provide considerable evidence supporting the fact that users employ a 'short-circuit evaluation' like strategy for comparing fingerprints, i.e. as soon as a difference is observed a decision is made and the rest of the comparison is abandoned.

4.2 Error Rates

In this section we bring the results and analyses of the effect of length on false acceptance and rejection errors. Note that participants who made any errors in comparing random fingerprints were excluded from our study as inattentive participants and hence all attentive participants we consider have correctly identified such fingerprints as non-matching. Consequently, we do not consider random fingerprints in our analysis in this section. We are testing for change with fingerprint length for both error types, hence Kruskal–Wallis was deemed appropriate.

4.2.1 False Acceptance Errors. Each participant in our study carried out 4 adversarial comparisons. Table 4 lists the number and proportion of participants by number of false acceptance errors they made for different lengths of fingerprints.

The proportion of participants making no false acceptance error decreases from 72% for 1-line key fingerprints to 55% for 2-line fingerprints and eventually to the very low figure of 39% for 3-line fingerprints which are used by Signal / WhatsApp. On the other hand, while only 6% of the participants did not manage to spot any of the adversarial comparisons for 1-line fingerprints, this figure rose to 22% for 2-line fingerprints, and eventually to 31% for 3-line fingerprints.

Kruskal–Wallis test indicated significant differences between the number of false acceptance errors made by participants for different key fingerprint lengths ($\chi^2(2) = 15.03, p < 0.001$). The effect size was moderate ($\eta^2[H] = 0.082$). Pairwise comparisons using Wilcoxon rank sum test with Holm correction indicated significant differences only between 1-line and 3-line conditions (1L–2L: $p = 0.051$, 2L–3L: $p = 0.102$, 1L–3L: $p < 0.001$). Therefore we can reject $H_0^{e\sim\ell}$ for false acceptance errors for larger differences between fingerprint lengths. In other words, we find evidence that indicates false acceptance errors significantly increase when the length of the key fingerprint significantly increases.

Table 4: Number of participants (number/total, top row in each section) and proportion of participants (in bold) including 95% confidence interval lower and upper limits (bottom row in each section) by number of false acceptance errors out of 4 (denoted by #) and condition (1, 2, 3 Lines)

#	1 Line	2 Lines	3 Lines
0	38/53	30/55	21/54
	(62%, **72%**, 84%)	(44%, **55%**, 69%)	(26%, **39%**, 53%)
1	8/53	7/55	8/54
	(6%, **15%**, 28%)	(2%, **13%**, 27%)	(2%, **15%**, 29%)
2	4/53	5/55	6/54
	(0%, **8%**, 20%)	(0%, **9%**, 24%)	(0%, **11%**, 25%)
3	0/53	1/55	2/54
	(0%, **0%**, 13%)	(0%, **2%**, 16%)	(0%, **4%**, 18%)
4	3/53	12/55	17/54
	(0%, **6%**, 18%)	(11%, **22%**, 36%)	(19%, **31%**, 46%)

Table 5: Number of false acceptance errors (error/total, top row) and the mean rate (in bold) including 95% confidence interval lower and upper limits (bottom row) over all participants by condition (1, 2, 3 Lines)

1 Line	2 Lines	3 Lines
28/212	68/220	94/216
(9.0%, **13.2%**, 18.5%)	(24.9%, **30.9%**, 37.5%)	(36.8%, **43.5%**, 50.4%)

To distill these figures, we can compute overall average false acceptance error rates by looking at the number of such errors made over all comparisons across all participants. Since all participants make the same number of adversarial comparisons, this would be equivalent to first computing an average error rate for each participant and then averaging over all participants. Number of false acceptance errors for all participants and their respective rates, including 95% confidence intervals, are listed in Table 5. As the figures suggest, an adversary mounting an attack against random users is expected to have an estimated 13.2% success rate for 1-line, 30.9% for 2-line, and 43.5% for 3-line fingerprints.

4.2.2 False Rejection Errors. In our study, each participant compared 12 safe (i.e. identical) key fingerprints. The number and proportion of participants by number of false rejection errors they made for different fingerprint lengths are shown in Table 6. No participant made 7 or above errors and for all categories of 2 to 6 errors, there was at most 1 participant who made that number of errors. We therefore compressed the table for those categories.

As the table shows, the proportion of participants making no false rejection errors steadily decreases from 92% for 1-line fingerprints to 85% for 2-line fingerprints and eventually to 80% for 3-line fingerprints. However, the overwhelming majority of participants make no more than 1 error for all fingerprint lengths.

Table 6: Number of participants (number/total, top row in each section) and proportion of participants (in bold) including 95% confidence interval lower and upper limits (bottom row in each section) by number of false rejection errors out of 12 (denoted by #) and condition (1, 2, 3 Lines)

#	1 Line	2 Lines	3 Lines
0	49/53	47/55	43/54
	(87%, **92%**, 98%)	(78%, **85%**, 94%)	(70%, **80%**, 90%)
1	3/53	5/55	10/54
	(0%, **6%**, 12%)	(2%, **9%**, 18%)	(9%, **19%**, 29%)
2–6	0–1/53	0–1/55	0–1/54
	(0%, **0–2%**, 6–8%)	(0%, **0–2%**, 9–11%)	(0%, **0–2%**, 10–12%)
7–12	0/53	0/55	0/54
	(0%, **0%**, 6%)	(0%, **0%**, 9%)	(0%, **0%**, 10%)

Table 7: Number of false rejection errors (error/total, top row) and the mean rate (in bold) including 95% confidence interval lower and upper limits (bottom row) over all participants by condition (1, 2, 3 Lines)

1 Line	2 Lines	3 Lines
6/636	18/660	13/648
(0.3%, **0.9%**, 2.0%)	(1.6%, **2.7%**, 4.3%)	(1.1%, **2.0%**, 3.4%)

Kruskal–Wallis test did not find significant differences between the number of false rejection errors made by participants for different fingerprint lengths ($\chi^2(2) = 3.39, p = 0.184$). This shows that although the number of false rejection errors increase with fingerprint length, this increase is not statistically significant for the range of fingerprint lengths we considered and hence we cannot reject $H_0^{e\sim\ell}$ for false rejection rates.

We can again look at the global false rejection error rates over all participants as indicators of the rates with which safe comparisons might be erroneously rejected in general for different fingerprint lengths. These rates are listed in Table 7 and show that false rejection errors are rare, with upper confidence limits of less than 5% for all fingerprint lengths. Besides, there does not seem to be a considerable change in error rates as fingerprints get longer, especially at higher lengths.

4.3 Comparison with Previous Work

To put our results in context, in this section we list the comparison times and error rates reported in previous studies on numeric fingerprint verification alongside our results. These measurements are not directly comparable per se, since they are collected under different conditions. Nevertheless, we believe this comparison helps situate our results in the wider context.

Results show a gradual increase of comparison time with fingerprint length as expected. Kainda et al. reported a median of 5 and a mean of 6 seconds, respectively, for comparing 6-digit numeric fingerprints [12]. Other notable results are a median of 9.5 seconds for 34-digit fingerprints reported by Dechand et al. [6] and a median of 8.1 seconds for 48-digit fingerprints by Tan et al. [21]. These works all only report overall results and do not give a breakdown of the results by type, i.e. safe, random, and adversarial comparisons. The overall medians in our study can be computed as 5.1, 6.3, and 8.0 seconds for 20, 40, and 60-digit fingerprints (i.e. for 1, 2, and 3-line fingerprints), and the respective means as 6.1, 7.4, and 10.0 seconds. Overall median comparison times for our study and the previous studies are all shown in Figure 4. We have also included median comparison times for the three comparison types in our study, but excluded Uzun et al.'s reported mean of 12.5 seconds as it was for a pair of users carrying out the comparison collaboratively [22].

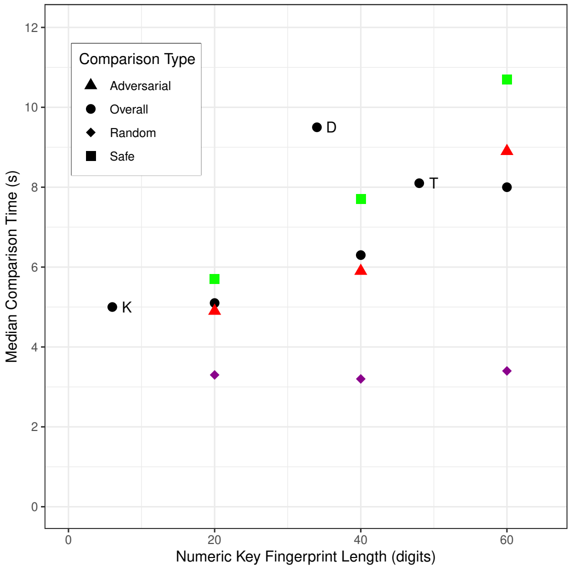


Figure 4: Median comparison times measured in our study (at lengths 20, 40, 60 digits) compared to those reported in the literature (annotated K: Kainda et al. [12], D: Dechand et al. [6], T: Tan et al. [21])

As the figure shows, our overall results and those of Kainda et al. and Tan et al. are more or less in line with each other, with Dechand et al.'s result seemingly being an outlier to some extent. Another important point depicted by our results is that overall medians only give reliable estimates in environments where occasional attacks and random comparisons are expected. In safer environments, where the overwhelming majority of the comparisons are expected to be safe ones, timing estimates should be considered to be considerably higher, e.g., by about a third for 60-digit fingerprints.

Kainda et al. did not observe any false acceptance errors (called 'security failure' there) in their 30-participant study for 6-digit fingerprints [12]. Dechand et al. reported a 6.3% rate (called 'fail rate') for 34-digit fingerprints [6] and Tan et al. a 35% rate (called 'fraction [of attacks] missed') for 48-digit fingerprints [21]. Our results of 13.2%, 30.9%, and 43.5% false acceptance error rates for 20, 40, and 60-digit fingerprints are broadly in line with the results above, except for that of Dechand et al.'s, as shown in Figure 5. A possible explanation for the discrepancy between Dechand et al.'s result and the rest, both in terms of comparison time and error rates,

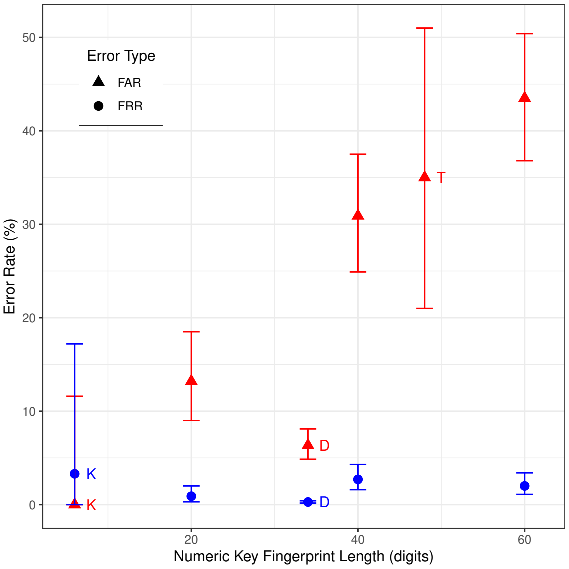


Figure 5: Average false acceptance and rejection rates (FAR, FRR) and 95% confidence intervals measured in our study (at lengths 20, 40, 60 digits) compared to those reported in the literature (annotated K: Kainda et al. [12], D: Dechand et al. [6], T: Tan et al. [21])

is that Dechand et al.'s participants were particularly attentive and hence took longer time for carrying out the comparisons, ending up with much lower error rates.

As for false rejection rates, Kainda et al. report a rate of 3.3% (called 'non-security failure') for 6-digit fingerprints [12] and Dechand et al. 0.28% (called 'false positive') for 34-digit fingerprints [6]. Tan et al. do not report the rate. Our rates of 0.9%, 2.7%, and 2.0% for 20, 40, and 60-digit fingerprints are largely consistent with the results above. As Figure 5 shows, mean false rejection rate remains below 5% irrespective of the length of compared fingerprints.

4.4 Limitations

It is not immediately clear what the best method is to control the similarity between pairs of fingerprints, ensuring adversarial pairs of different lengths have comparable similarity. For numeric fingerprints represented without chunking and in one line, one may keep the proportion of different digits constant for various fingerprint lengths. However when chunking and multiple lines come into play, factors such as where in each line and between chunks the differences appear and how many chunks are affected need to be taken into account. We aimed for a simple method of allowing one chunk of difference for all lengths, but this would mean that the proportion of different digits will not stay the same.

In our adversarial comparisons, we considered near-collision fingerprints differing only in one chunk immediately after the midpoint. This means that the non-identical chunk appeared in different positions in different conditions: in the middle of the line for the 1 Line condition, in the beginning of the second line for the 2 Line condition, and in the middle of the middle line for the 3 Line condition. This may have introduced a confound in comparing the 2 Line condition results with the other two conditions, but the comparisons between 1 Line and 3 Line conditions are not expected to have been affected.

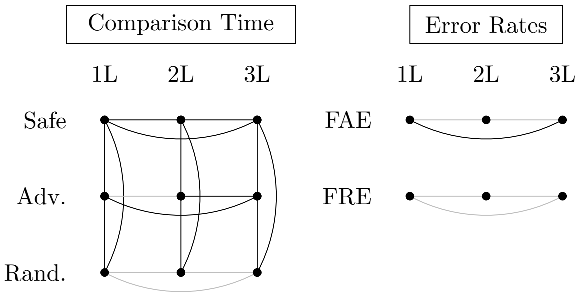


Figure 6: Summary of statistical significance results for comparison time by condition (1L, 2L, 3L) and type (Safe, Adv., Rand.) on the left, and for error rates (FAE, FRE) by condition (1L, 2L, 3L) on the right. Black lines indicate statistically significant differences and grey lines indicate non-significance.

We have simulated smartphone user interfaces within browsers. In practice, comparisons are made on two real smartphones that are likely to be different makes or models. However, we don't expect this issue to have had a considerable effect on our results in general.

Our participants were largely young (around 69% 21–35), male (around 63% male), and educated (around 69% with tertiary education). This needs to be kept in mind when considering the results.

5 DISCUSSIONS AND CONCLUSIONS

We discuss the implications of our results and some possible directions for future work in this section.

5.1 Implications of the Results

Figure 6 shows the summary of our results in terms of statistical significance for comparison time on the left and error rates on the right. We list the main takeaways from our study based on the results in the following. Although these are not mutually exclusive, it is instructive to look at the results from various perspectives.

Fingerprint length is a major determinant of efficiency. As the analysis in Section 4.1.1 shows, for safe comparisons, changes in comparison time are significant with respect to fingerprint length for all length differences. In the most common use cases of numeric key fingerprint verification, the overwhelming majority of comparisons are expected to be safe comparisons. Hence, our results provide strong evidence for the intuition that fingerprint length should be considered as a significant determinant of efficiency when designing numeric key fingerprint verification systems.

Overall time estimates can be misleading. Analysis in Section 4.1.2 demonstrates that time differences between comparison types are significant at all lengths, with timing estimates for safe comparisons being significantly higher than other types. Given that in most common use cases we expect safe comparisons to dominate, median comparisons times in practice are going to be closer to median safe comparison times. However, overall comparison times usually reported in the literature assume arbitrary and unrealistic proportions of safe, adversarial, and random comparisons. Hence, when considering efficiency, design decisions for common use cases should be made based on safe comparison times, when available, rather than overall comparison times usually reported in the literature. If safe comparison times are not available, our results show they can be estimated to be between a tenth to a third above overall times depending on fingerprint length.

Users are neither efficient nor effective in comparing highly similar long fingerprints. Focusing on adversarial fingerprints with high similarity, the results in Sections 4.1.1 and 4.2.1 show that although users take significantly longer time to perform the comparison, they make significantly higher false acceptance errors which can be security critical. This underlines the crucial role of providing alternative or complimentary means of key fingerprint verification for contexts where higher levels of security is required, as manual verification of long fingerprints suffers from low usability.

Manual key fingerprint verification provides a lower security level than usually assumed. Fingerprint lengths are usually chosen to provide desired levels of security. This level of security indicates the adversarial power required to achieve a (full) fingerprint *collision* (i.e. an adversarial fingerprint identical to an authentic one) and hence fool the user with a success probability of 1. For instance, the Signal / WhatsApp fingerprint is designed to provide 112-bit security since the adversarial power required for finding a second preimage for 30-digit key fingerprints computed with 5200 hash iterations is $10^{30} \times 5200 \approx 2^{99.7} \times 2^{12.3} \approx 2^{112}$. This means that with approximately 0.69×2^{112} hash computations, an adversary is expected to achieve a 50% success rate. Looking at another point of interest on the attack success probability curve (specified in Section 3.1), to achieve a 40% attack success rate, the adversary would be expected to perform approximately $0.51 \times 2^{112} \approx 2^{111}$ computations. However, as our results in Section 4.2.1 show, a *near collision* (i.e. an adversarial fingerprint sufficiently similar to an authentic one) is enough to achieve a considerable false acceptance error rate as high as 40%. As Table 1 shows, such a near collision would only require $2^{95.4}$ adversarial power, i.e. approximately $0.69 \times 2^{95.4} \approx 2^{94.9}$ hash computations. False acceptance error rate is strongly indicative of the success rate for attack campaigns targeting multiple victims repeatedly, which can be possible in many use cases of such fingerprints. Hence, it is more realistic to think of the Signal / WhatsApp fingerprint length providing approximately 96-bit security rather than 112-bit security, and in general, longer fingerprint lengths for which high false acceptance rates are possible should be considered to provide considerably less security than usually assumed.

Users are quite efficient and effective in recognising dissimilar fingerprints. Our results for random fingerprint comparisons in Sections 4.1.2 and 4.2.2 clearly show that not only users are pretty quick and accurate in recognising highly dissimilar fingerprint pairs, but also both comparison time and false rejection error rate stay low and roughly constant even with considerable changes in fingerprint length. As discussed before, this points toward a ‘short-circuit evaluation’ like behaviour exhibited by users in performing fingerprint comparison. Consequently, in an environment where users may be expected to perform higher proportion of such comparisons, designers can be confident that users can handle a wide range of fingerprint lengths with similar effectiveness and efficiency.

False rejection errors are rare. False rejection errors and rates stay quite low across a relatively wide range of fingerprint lengths as the results in Section 4.2.2 show. Indeed, even the 95% confidence interval upper limits stay below 5% in all the measurements we carried out. Therefore, when designing such mechanisms, decisions for fingerprint length can be made mainly based on efficiency and security (including false acceptance errors).

Similarity is a significant determinant of efficiency. The most emphatic results were given by the analysis of comparison time with respect to comparison type in Section 4.1.2: the differences between comparison time for safe and adversarial, as well as between adversarial and random, and hence between safe and random fingerprint pairs are found to be significant at all lengths. This shows that the effect of similarity between fingerprints is markedly significant on the efficiency of manual key fingerprint comparison.

5.2 Future Work

As with any other study, the scope of the parameters had to be limited in our investigation and further work is required to explore the parameter space more broadly. Of particular interest would be investigating a higher granularity of lengths and a wider range of similarity between fingerprints.

To test whether our results can be generalised to wider contexts, it would be crucial to replicate the investigation for other verification modes, including verbal and collaborative comparisons, and other fingerprint representations, including word-based ones.

Our results can be seen as part of a series of related works collectively demonstrating the poor usability of currently recommended methods for manual verification of long key fingerprints, e.g. those used by Signal / WhatsApp, and underlining the importance of developing better manual and automated verification methods.

Acknowledgement

We sincerely thank the reviewers of ARES’23 for their valuable and constructive comments.

REFERENCES

[1] OpenSSH 8.2. 2020. OpenSSH Release Notes. www.openssh.com.

[2] akwizgran. 2014. Basic English: Encode random bitstrings as pseudo-random poems. GitHub repository at https://github.com/akwizgran/basic-english.

[3] Stefan Brands and David Chaum. 1993. Distance-bounding protocols. In *Workshop on the Theory and Application of of Cryptographic Techniques at EUROCRYPT '93*. Springer, 344–359.

[4] Mauro Cherubini, Alexandre Meylan, Bertil Chapuis, Mathias Humbert, Igor Bilogrevic, and Kévin Huguenin. 2018. Towards Usable Checksums: Automating the Integrity Verification of Web Downloads for the Masses. In *CCS*. ACM, 1256–1271.

[5] Matthew Copeland, Joergen Grahn, and David A Wheeler. 1999. The GNU Privacy Handbook. https://www.gnupg.org/gph/en/manual.html.

[6] Sergej Dechand, Dominik Schürmann, Karoline Busse, Yasemin Acar, Sascha Fahl, and Matthew Smith. 2016. An empirical study of textual key-fingerprint representations. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX, Austin, TX, 193–208.

[7] Michael Farb, Yue-Hsun Lin, Tiffany Hyun-Jin Kim, Jonathan McCune, and Adrian Perrig. 2013. Safeslinger: easy-to-use and secure public-key exchange. In *Proceedings of the 19th annual international conference on Mobile computing & networking*. 417–428.

[8] Michael T Goodrich, Michael Sirivianos, John Solis, Gene Tsudik, and Ersin Uzun. 2006. Loud and clear: Human-verifiable authentication based on audio. In *26th IEEE International Conference on Distributed Computing Systems (ICDCS'06)*. IEEE, IEEE Computer Society, 10–10.

[9] Peter Gutmann. 2011. Do users verify SSH keys? *Login* 36 (2011), 35–36.

[10] Amir Herzberg and Hemi Leibowitz. 2016. Can Johnny finally encrypt?: evaluating E2E-encryption in popular IM applications. In *ACM Workshop on Socio-Technical Aspects in Security and Trust (STAST)*. ACM, New York, NY, USA.

[11] Antti Huima. 2000. The Bubble Babble Binary Data Encoding. Network Working Group Internet Draft, available at http://web.mit.edu/kenta/www/one/bubblebabble/spec/jrtrjwzi/draft-huima-01.txt.

[12] Ronald Kainda, Ivan Flechais, and A. W. Roscoe. 2009. Usability and Security of Out-of-Band Channels in Secure Device Pairing Protocols. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 11, 12 pages. https://doi.org/10.1145/1572532.1572547

[13] Alfred Kobsa, Rahim Sonawalla, Gene Tsudik, Ersin Uzun, and Yang Wang. 2009. Serial Hook-Ups: A Comparative Usability Study of Secure Device Pairing Methods. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 10, 12 pages. https://doi.org/10.1145/1572532.1572546

[14] Raph Levien and Donald Johnson. 1998. Snowflake. http://dlakwi.net/snowflake/snowflake.html.

[15] Lee Livsey, Helen Petrie, Siamak F Shahandashti, and Aidan Fray. 2021. Performance and Usability of Visual and Verbal Verification of Word-based Key Fingerprints. In *Human Aspects of Information Security and Assurance: 15th IFIP International Symposium, HAISA 2021, Virtual Event, July 7–9*. Springer, 199–210.

[16] Moxie Marlinspike. 2016. Safety number updates. Signal Blog. Availabe at https://signal.org/blog/safety-number-updates.

[17] Adrian Perrig and Dawn Song. 1999. Hash visualization: A new technique to improve real-world security. In *International Workshop on Cryptographic Techniques and E-Commerce*, Vol. 25.

[18] Konrad Rieck. 2002. Fuzzy Fingerprints Attacking Vulnerabilities in the Human Brain. *Online publication, available at http://ouah.org/ffp.pdf* (2002).

[19] Svenja Schröder, Markus Huber, David Wind, and Christoph Rottermanner. 2016. When SIGNAL hits the Fan: On the Usability and Security of State-of-the-Art Secure Mobile Messaging. In *Proceedings 1st European Workshop on Usable Security* (Darmstadt, Germany). Internet Society, Reston, VA.

[20] Maliheh Shirvanian, Nitesh Saxena, and Jesvin James George. 2017. 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* (Orlando, FL, USA) *(ACSAC 2017)*. ACM, New York, NY, USA, 499–511. https://doi.org/10.1145/3134600.3134610

[21] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. 2017. Can Unicorns Help Users Compare Crypto Key Fingerprints?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 3787–3798.

[22] Ersin Uzun, Nitesh Saxena, and Arun Kumar. 2011. Pairing Devices for Social Interactions: A Comparative Usability Evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2315–2324. https://doi.org/10.1145/1978942.1979282

[23] Elham Vaziripour, Justin Wu, Mark O'Neill, Daniel Metro, Josh Cockrell, Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent E Seamons, and Daniel Zappala. 2018. Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal.. In *SOUPS@ USENIX Security Symposium*. 47–62.

[24] Elham Vaziripour, Justin Wu, Mark O'Neill, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala. 2017. Is that you, Alice? A usability study of the authentication ceremony of secure messaging applications. In *13th Symposium on Usable Privacy and Security (SOUPS'17)*. 29–47.

[25] WhatsApp. 2017. WhatsApp Encryption Overview. Technical white paper, WhatsApp, Available from whatsapp.com.

[26] Justin Wu, Cyrus Gattrell, Devon Howard, Jake Tyler, Elham Vaziripour, Kent Seamons, and Daniel Zappala. 2019. "Something isn't secure, but I'm not sure how that translates into a problem": Promoting autonomy by designing for understanding in Signal. In *15th Symposium on Usable Privacy and Security (SOUPS'19)*.