Paper Notes

Aidan Fray

May 23, 2019

TODO: Need to find papers on study of words are a comparative measure.

Lots of research has been done into the secure pairing of devices with minimal human involvement:

- Using Camera Phones for Human-Verifiable Authentication
- Loud and Clear: Human-Verifiable Authentication Based on Audio
- A human-verifiable authentication protocol using visible laser light
- Secure Device Pairing based on a Visual Channel
- NFC
- Physical contact

OVERVIEW

- Lots of research performed on E2E applications (Telegram, WhatsApp etc)
- Lots of study into overall representations of fingerprints

1 Can Unicorns Help Users Compare Crypto Key Fingerprints

This paper tested 8 different fingerprint representations with 661 different participants. All of these representations are testing using compare-and-confirm (basic comparison process for fingerprints) and compare-and-select (select from a list)

- Hexadecimal
- Alternating vowel/consonant
- Words
- Numbers
- Sentences
- OpenSSH visual host key
- Vash
- Unicorns

The paper also emphasises the attention that has been taken to realism in the experimentation stage.

Talks about long term usability issues with public-key crypto and states fingerprints are a main source of the problem. ([28])

There seems to be mixed findings with compare-and-select and compare-and-confirm ([14, 17])

There has been research into OpenSSH fingerprints and how users interact ([19, 24])

Paper chose to target a security level of 160-bits

Papers attack strengths were 2^{40}, 2^{60} and 2^{80}

Paper states that MTurk users have been shown to be younger and better educated than the general US population ([15])

1.1 Findings

- Graphical representations have mixed success rates (with quick comparison times)
- Recommendations not to use **compare-and-select**
- **No method** really provided enough security for high risk situations, they recommend removing users from the loop by using smart phone cameras etc

2 An Empirical Study of Textual Key-Fingerprint Representations

Study involved 1047 participants evaluating 6 different textual representations on MTurk. The study also includes an evaluation into usability.

The study has an attacker power of around 80 of 112 bits.

- Alphanumerical
- Numerical
- Words
- Sentences

More mention to decentralised method finding it hard to find adoption such as Web of Trust and Namecoin ([7, 13, 30])

References as system called CONIKS and others ([24, 39, 27]) that aim to provided a directory of keys

States that many systems still rely on fingerprint comparison ([17])

States that the **hexadecimal** encoding is used in most systems.

States again that fingerprint comparison is seldom done in practice ([17, 37])

Studies have shown that users find it difficult to compare long and "meaningless" strings ([19])

References to other word lists used. PGP Word list [22] and english word list compiled by K.C. Ogden [31]. The Peerio word list is generated from the most common words in english subtitles.

Interesting recommendation into the use of *scrypt* [34] and *Argon 2* [3] can be used to shorten the fingerprint length. These work slowly and a prevent easy computation of pre-images.

When generating partial-collision fixing the bits at the start and end has been shown a best practice ([17, 37])

2.1 Findings

- Overall recommendation is to replace **hexadecimal** with **sentence** based encoding

- Findings show that **hexadecimal** and **Base32** perform worse than the other alternatives in realistic threat models. With over **10%** of users failing to detect attacks with these schemes.

- According to their study, **words** and **sentences** we some of the best methods for avoiding attack

- Words has the second highest usability ratings second to only sentences

3 Usability and Security of Out-Of-Band Channels in Secure Device Pairing Protocols

Paper looks into the positives of secondary channels used to authenticate device pairing. Such as device fingerprint comparison. It also aims to look into the fact that many studies performed do not take into consideration the aspect of human interaction.

Total of 30 paid participants were recruited where the experiment simulated a P2P payment system.

They collected:

- Time to completed the associated process
- Number of security or non-security related errors

And quantified using questionnaires:

- Ease of use with the representation
- Satisfaction with time spent with a method/representation
- Participant's confidence with the scheme

The paper also tested methods:

- Compare and Confirm
- Compare and Select
- Compare and Enter
- Barcode scanning

Representations:

- Numerical
- Alphanumerical
- Words
  - Dictionary of 1024 words each mapped to 10-bits
- Sentences
- Images

- Melodies
- Sound (Numerical/Alphanumerical)

"People are the weakest link in the security chain" [26]

Mentions a normal channel as a "Dolev-Yao" channel [5]. This is where an attacker can overhead, delete and modify messages.

Talks about systems that were secure on paper that ended up in practice being miss-handled to reduce security. One example is with the Russian army in WWI [1]

Bruce Schneier [27] also argues that systems and involves humans therefore often systems are broken due to improper use.

3.1 Findings

- Showed for **usability** methods ranked:
    - Comparing-confirm
    - Typing strings
    - Comparing-select
    - Barcode
- Showed for **combined security** and **usability** methods ranked:
    - Typing strings
    - Barcodes
    - Comparing-confirm
    - Comparing-select

  Showed that the best methods in terms of the SUM score [24] ranked the top 3 as:

    - Numeric(C&C) [73.7]
    - Alphanumerical(C&C) [72.5]
    - Words (C&C) [70.6]

  Numerical had 0% security failures, Word 3.3% and Alphanumerical had a high 13.3%

  NOTE: This is a very small sample size.

4 SafeSlinger: An Easy-to-use and Secure Approach for Human Trust Establishment

The aim of SafeSlinger is to solve the issue of key exchange. They aim to leverage the initial personal contact of people to facilitate trust. To solve the issue of people that would never meet they implement a system of "shared acquaintance"

Talks about issues with decentralised (PGP [42]) and centralised (SSL CA [27, 34]) key exchange.

"Tim Berners-Lee has called upon security researchers and professionals to design a public key encryption system for the people [11]"

Only uses **24-bits** of a SHA-1 hash of all the exchanged information, **could this be exploited?**

5 A Study of User-Friendly Hash Comparison Schemes

The aim of this study is to provide a study on what hash comparison scheme provides the best accuracy and shortest comparison time.

The paper also proposes new schemes: An extension to Flag and three asian character representations.

An interesting aspect is the effort put into the study to link characteristics about the participants, i.e. gender or age to how well they perform or rate the usability of the scheme. In total they had 436 participants.

Overall they compare:

- Base32
- English words
- Random Art
- Flag
- T-Flag
- Flag ext
- Chinese, Korean, Japanese characters

Key phrase: "We did not use hexadecimal digits because they're similar to Base32 and known to be error prone" [NO REFERENCE]. This is shown in *"An Empirical Study of Textual Key-Fingerprint Representations"* that this isn't true - "However, our work shows that numeric representations actually perform significantly better than Base32 and is less error prone."

They study also considers the comparison of "easy" and "hard" pairs of hashes. Hard is where the hashes are designed to be very similar. This is designed to provide a base-line (easy) and a worst-case (hard).

The paper deals with ranges of entropy of 22-28bits.

Another system that uses word is the Unmanaged Internet Architecture [4]

States that users make errors when comparing long strings [11]

Dicussions are also made into the pre-requirements needed for some of these schemes. Random-Art needs a colour display and requires a relativly high computational cost to generate a representation. Asian characters require unicode or codec support.

Paper also touches on "Additional Benefits" - the paper talks about the ability to verbally describe the scheme, this is so the encoding scheme can be verified using a channel that does not involve a visual feed such as a telephone conversation

Overall the paper provides a table of requirements for the schemes with the graphical representations requiring a large number of pre-requirements.

5.1 Findings

- Source, age and gender have no significant impact on the accuracy across the schemes. But younger participants were significantly faster
- In general, Base32, Random Art, T-Flag and Flag Ext provide fast and accurate comparisons
- Comprehension of a language when used as a encoding scheme affects the ability to distinguish hard pairs by a significant amount

6 Usability Analysis of Secure Pairing Methods

Paper is aiming to perform a comparative usability evaluation of **selection methods** of hash comparisons

This paper is doing this study in the context of secure device pairing in a diffie-hellman type key exchange. However, the results of the study could directly apply to that of hash comparison, however, it must be considered that the comparisons here are for **very** short strings. Around 4-digits

Comparison methods compared:

- Compare-and-Confirm
- Select-and-Confirm
- Copy-and-Confirm

The paper also compared to methods of passphrase comparison that weren't relevant.

Paper also included demographical information in the study, with two round of 40 participants in the usability studies.

Findings:

Round 1:

- Copy-and-confirm was percieved as hard to use
- Copy-and-confirm has the lowest fatal error rate but the longest comparison time and lowest percieved usability
- The schemes reduced fatal error rate as their comparison times rose
- Copy-and-confirm had issues with user confusion in terms of use

In the second round they only decided to continue with

- Compare-and-Confirm
- Select-and-Confirm

They dropped Copy-and-confirm due to its high similarity to the other scheme "Copy" that was being used for passphrase sharing

The second round included some UI changes that were backed up by external research.

This round the changes performed on the schemes resulted in much lower fatal error rates. However, Select-and-confirm had an unacceptable 5% error rate. Compare-and-confirm had 0% fatal error rate. The was an improvment from 20% to 0%.

Both schemes were again percieved as easy to use

Overall it was shown that compare-and-confirm was the superior choice.