Literature Review

Aidan Fray

May 26, 2019

1 Research questions

Overall the research will aim to investigate the strength of pEp's Trustword fingerprint mappings, and the ease in which partial collisions can be obtained for keys and how this will ultimately affect the end user(s).

Key areas that will be looked into:

- Is the recommended minimum number of Trustwords enough to provide a basic level of security.
- What attributes make a strong general wordlist for fingerprint mapping and does the Trustword implementation exhibit these features.
- How easy is it to generate similar keys that attack a targeted key pair?
- How can similarity be quantified in terms of words? Does this include pronunciation or visual aspects?
- As usability is the main justification for the use of Trustwords are there alternatives that provide the same usability but with more security?

2 Review

2.1 Fingerprint representation and comparison

Current research in ways to represent and validate fingerprints has almost exclusively focused on the usability of such schemes. The following section will individually discuss available research findings alongside an overall comparison of the research recommendations.

The first work in this area was performed by Hsu-Chun Hsiao, *et. al.* in their paper: *A Study of User-Friendly Hash Comparison Schemes*[1] in around 2009. The aim of the study was to provide the first insight into the best encoding scheme used to represent a hash fingerprint. The schemes used were Base32, English words, Random Art[1], Flag[2], T-Flag[3], Flag Ext[4] and finally Chinese, Korean and Japanese symbol encoding. 436 participants were assessed in the study.

Alongside the main aim of the paper, the authors wanted to provide empirical links between participant attributes such as age or gender to performance with the encoding scheme. This is something rarely seen in the related research. Hsu-Chun Hsiao *et al.* also provided two categories of similarities; "hard" and "easy". These were designed to be the worst and best case respectively. As a final aim and the reason for the inclusion of Chinese, Korean and Japanese characters was to research if being a native speaker assists the user in distinguishing subtle differences between encodings in the respective language.

The entropy of encodings ranged from 22-bits to 28-bits. This is extremely low in comparison to similar research. However, this can be attributed to the age of the paper due to its 10 year age at the time of writing.

The authors' decision to exclude hexadecimal and numerical encodings (some of the most widespread encoding schemes) was due to their similarity to Base32 and their well-known deficiencies. This is not consistent with available research. For example, in *"Empirical Study of Textual Key-Fingerprint Representations"* it was shown numerical representations performed better than that of Base32.

Overall, findings from the paper showed that age and gender do not affect the accuracy of the scheme, however, younger participants were significantly faster. The paper recommends Base32, Random Art, T-Flag or their own improvement; Flag Ext. These encoding recommendations were also supplemented with a review of the requirements required by these schemes. The recommendation of graphical encoding schemes is inconsistent with alternative literature, where most other papers have found graphical encodings easy to use but insecure. In terms of language comprehension; being able to speak the language assists in the ability to discern the differences between hard pairs. Knowledge of the language subsequently did not help with easy pairs.

In conclusion, the paper provided a very strong foundation for further research due to the wide variety of topics touched upon. However, its age means that results cannot be applicable to modern day scenarios, and oversights in the development of the study such as the exclusion of hexadecimal encoding and the lack of information gained on the perceived usability of the scheme from users results in an ultimately incomplete study.

[1] A scheme proposed by A. Perrig *et al.*[2]
[2] Encoding scheme proposed by C Ellison *et al.*[3]
[3] Yue-Hsun Lin *et al.*[4] improvement on Flag
[4] The author's own improvment on Flag

Further research bu Kainda *et al.*[5] in 2009 investigated fingerprint representations and comparison methods in the context of using humans as the Out-Of-Band (OOB) channel for secure device pairing. This study was comprehensive and assessed a wide variety of concepts. This is one of the few papers in the literature that assessed the way to compare representations alongside the representations themselves.

For comparison methods that paper looked into Compare-and-Confirm, Compare-and-Select, Compare-and-Enter and Barcode scanning. This set includes the most common ways of comparison according to available research. Alongside this, the paper reviews Numerical, Alphanumerical, Word, Sentences, Images, Melodies and Sound (Numerical/Alphanumerical). Again, this in comparison to similar literature is highly comprehensive.

In terms of empirical data gathered by the paper, they reviewed the time completed to asses the representation and the number of security and non-security related errors. Furthermore, the authors also collected user opinion on ease-of-use, satisfaction and confidence of the reviewed schemes. The overall schemes were then quantified using a single metric known as the Single Usability Metric (SUM)[6]. This provides a quick way to rank and compare the schemes, thus improving the usability of the results. However, the one apparent downside to the experimental side of the study is the 40 participants enrolled. This in comparison to other studies is one of the smallest sizes of enrolment and thus, puts a possible limit on the accuracy and applicability of the obtained results.

Overall, if usability is the only consideration the paper ranked the comparison methods like so: Compare-and-Confirm, Compare-and-Enter, Compare-and-Select and Barcode. If, however, the consideration of security alongside usability is required the best is Compare-and-Enter, Barcode, Compare-and-Confirm and Compare-and-Select. Research has been limited in this area of modes of comparison, however, the results from this paper are consistent with the alternative literature.
In terms of encoding schemes, the paper ranked Numerical, Alphanumerical and Words as the top three encoding schemes (These were all with Compare-and-Confirm). This took into consideration the usability and security of the scheme. This is relatively consistent with other literature where numerical and written encoding schemes seem to be the most effective overall.

In conclusion, this paper provided a highly comprehensive review of all the aspects of human fingerprint verification, where it reviewed all the elements involved in the process. The only limitation of the paper is the size of the participants enrolled for the study.

A paper by Ersin Uzun *et al.*[7] exclusively reviewed comparison methods. This is performed in the context of "Secure device pairing" and thus some comparison schemes are tested that are not relevant in a fingerprint comparison context, for example "Choose-and-Enter" (A participant chooses a passphrase from an avail-

able list and enters it into the other device). Thus, the only relevant comparison scheme and their results have been extracted and compared.

The relevant comparison methods tested were, Compare-and-Confirm, Compare-and-Select and Compare-and-Enter. These are almost exactly the same encoding schemes assessed in the relevant research literature. The paper also had two rounds of experiments including 40 participants each where changes were made to UX and format between these studies.

Due to the "Secure device pairing" context, the paper was written under, consideration needs to be made into the strings encoded. Due to secure device pairing commonly utilising Short Authentication Strings (SAS) the paper only reviewed encodings of 4 digit long strings. This, therefore, limits the applicability of the results to fingerprint encoding. This will be considered later when comparing the results to alternative literature.

Between the two rounds, the authors dropped Confirm-and-Enter due to its low usability and over similarity to a similar scheme "Copy" used for passphrase verification. The low rating for Confirm-and-Enter is in contrast with research performed by Kainda *et al.*[5] They ranked Confirm-and-Enter as the 2nd most usable system.

With changes to the UX of Confirm-and-Confirm backed up by external research [8][9] the second round of research was performed with a new set of participants. The alterations took Compare-and-Confirm from 20% to 0% fatal error rate (FER). The same trend for Confirm-and-Select with it reducing its FER from 12.5% to 5%. With the changes, the participants perceived the systems as easy to use. The authors, therefore, recommended Compare-and-Confirm as the overall best choice for comparison methods. Again, comparing this word to the most similar work done by Kainda *et al.*[5] this is consistent with their findings where they ranked Compare-and-Confirm as the most secure comparison method.

Overall, the paper provides a sizeable insight into suitable methods of comparison, this is due to the cross over between fingerprint comparison and secure device pairing. This paper is limited in terms of the short strings compared, the different overall research aim and the size of the number of participants used in the studies. In terms of security of the schemes, the results were consistent with that of alternative research but clashed when usability was considered. This, however, may be due to the differing use-cases of the comparison methods and the sizes of strings being compared.

Work by Dechand Sergej, *et al.*[10] empirically investigated the usability of 4 distinct textual representations evaluated with an experiment involving a total of 1047 participants. The textual representations were: Alphanumeric, Numeric, Words and Sentences. They assessed the number of attacks missed with each scheme alongside results from a questionnaire on the participants preferred scheme and perceived usability.

The paper touches upon issues with decentralised methods of identification such a PGP's Web of Trust and the problems these solutions have with user adoption. These points are made in an attempt to validate the requirement for manual comparison of key based fingerprints. This is a common theme that appears in the majority of the reviewed papers.

Other references made within the paper touch upon the vulnerabilities and usability issues present with the way humans interact with the security systems. Example of these are studies showing humans find it difficult to comprehend long and "meaningless" strings and the lack of actual comparison performed by users in live scenarios. These are, however, acknowledged as limitations in the later stages of the paper.

The paper has defined the upper and lower bound costs of the attacker's resources and strength as $610K to $16B, with an ability to control 80-bits of the fingerprint. This in comparison to other papers is high and is almost encroaching into the realm of a highly sophisticated attacker. Therefore, the lack of consideration for the lower-end of the attack resource spectrum can be considered a limitation of this study.

Overall, findings from the paper state that conventional encodings such as Hexadecimal and Base32 perform worse than all other alternatives in a realistic threat model with over 10% of users failing to detect an attack on these encoding schemes. The recommendations of the authors are to replace these encoding schemes with Words or Sentences due to their very high success rate and high usability scores. The performance of the Alphanumerical encoding schemes is relatively consistent with the literature, however, it is hard to say conclusively due to the small number of total schemes assessed and lack of visual encoding schemes making it hard to place alphanumerical encodings in an overall rating.

J Tan *et al.*[11] work is very similar to that of the research already discussed. 8 distinct fingerprint representations were tested with over 661 participants. Each representation was tested using Compare-and-Confirm and Compare-and-Select. The small difference with this paper is the inclusion of two novel graphical encodings Vash[5] and Unicorns[6]. The compared schemes were: Hexadecimal, Alternating vowel/consonants, Words, Numbers, Sentences, OpenSSH visual host key, Vash and Unicorns.

The target security level chosen was an entropy of 160-bits. This is a very high level of security in comparison to alternative literature. Alongside this, the paper defined attack strengths of an average 2^{60} with other specific use-cases for lower and higher attacker strength. Users were recruited via MTurk and assigned an experiment attempting to emulate a real-life condition (Comparison of a fingerprint on a business card). The real-world accuracy was one unique element that this paper concentrated on, this was portrayed through the author's

[5]https://github.com/thevash/vash
[6]https://unicornify.pictures/

attention to detail when designing the experiment.

Overall, results from the experiments emulated previously discussed literature. Textual representations faired effectively well with Words and Sentences again being some of the best options overall. Performance of graphical representation was mixed with OpenSSH having performance matching that of a strong textual mapping while Unicorns having the worst performance overall. This performance with graphical representations is consistent with the available literature apart from the positive performance of OpenSSH. The results of different modes of comparison were highly consistent with other studies where again the recommendation was to not use Compare-and-Select as its performance across the board was below even basic levels of security.

2.2 Topic conclusion

In conclusion for this section, there has been extensive work in the investigation of the usability of textual and visual encoding schemes alongside methods of comparison. There are mixed opinions about visual representations, however, the majority defines them as easy to use but severely lacking in security. Research on textual representations, however, is in agreement; natural language based encodings such as words or sentences provide some of the highest usability and security over the baseline performance of alphanumerical encodings. The literature also concurs on the use of Compare-and-Confirm as the best method of comparing fingerprints overall, with Compare-and-Select being highlighted for its poor security and usability.
The starkest limitations of the literature are the range of participant and encoding entropy, where at its worst only 22-bits of entropy were tested. Another limitation is the scant amount of research solely investigating modes of comparison, where the literature only contains two papers with this research exclusive research aim. Even these limitations, as a whole the literature on this topic has provided a relatively comprehensive and coherent view on the usability of textual and visual hash encoding schemes.

3 Fingerprint representation 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*[2]. 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 (as discussed in the previous section).

Further research into the creation of unique visual hash schemes have been performed by C Ellison *et al.* [3] (Flag), Yue-Hsun Lin *et al.*[4] (T-Flag) and work by MM 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[7] and Robots[8]. 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 the 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 is 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 short 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 Patrick Juola and Philip Zimmermann [15] and their work in generating a word list where phonetic distinctiveness was prioritised. Each word is mapped to one byte where the unique aspect of the word list is the separation of "even" and "odd" words where even byte positions sample from the even list and odd from the odd list. The effectively creates two sub-word lists. The pairing of these word lists were determined through the use of a Genetic algorithm. The paper also includes a study on effective measures of linguistic distances of words

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

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 effectiveness of the word list. However, this was later evaluated in work by Dechand *et al.* [10] 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[9] 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 implementations. The first being by Michael Rogers[10]. Rogers' implementation is a program designed to map fingerprints to pseudo-random poems. This implementation was again, empirically evaluated by Dechand *et al.*[10]. 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.
A very recent implementation of a word list can be found in Pretty Easy Privacy (pEp) implementation of TrustWords[11]. The unique aspect of TrustWords is its mapping of words 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. This is arguably higher than most users vocabulary. This deviation from the norm has not been currently backed up by research. This, therefore, highlights a gap in the current literature.

4 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

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

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 security and usability.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[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] J. Sauro and E. Kindlund, "A method to standardize usability metrics into a single score," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 401–409.

[7] 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.

[8] J. Palmer, "Attentional limits on the perception and memory of visual information." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 16, no. 2, p. 332, 1990.

[9] R. Hammer, T. Hertz, S. Hochstein, and D. Weinshall, "Category learning from equivalence constraints," *Cognitive Processing*, vol. 10, no. 3, pp. 211–232, 2009.

[10] 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.

[11] 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.

[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*, vol. 1. IEEE, 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.