Summary of Neuroscience Meets Cryptography: Designing Crypto Primitives Secure Against Rubber Hose Attacks

Saad Usman Khan
saad@ut.ee

Abstract

Majority of currently deployed authentication systems rely on a secret string or a key. For the system to be secure it is required that secret remains a secret. However, using rubber hose attacks it is possible to persuade or force someone to reveal the secret leading to a security breach. The authors in this paper present a security system in which the secret is learnt by users without any conscious knowledge. Since the secret is stored in the subconscious of the users, there is no way to forcefully extract it from them. This is achieved by asking users to play a game consisting of character sequence anticipation. Some of the sequences are random while the other sequences are repeated. It is expected that the users will learn the repeating sequences and will be able to "identify" them subconsciously during authentication. The authors also included few user studies using Amazon's Mechanical Turk (a framework for experiments) to verify that users indeed learn the patterns over time. It was also demonstrated that users can re-authenticate using the learnt secret and occasionally can also recognize short parts of the learnt secret.

1. Introduction

This article summarizes the research paper Neuroscience Meets Cryptography [4]. Sections 1 through 6 provide the summary of corresponding sections in the original paper. Section 7 and 8 have been added by the author of this summary as his contribution.

Consider a scenario where users have to authenticate themselves into a high value facility which requires users to use all three methods of access control, namely:

1. Something they have,
2. Something they know,
3. Something they are [1].

However all of these mechanisms can be broken. Something that users have can be stolen. They can be coerced to reveal what they know and what they are can be faked. It is not very likely that all of this can be broken at once but it is possible.

Users have proven to be the weakest link in security [2]. Hence the author has focused on creating an authentication scheme which is not affected by mistakes of human conscious. Social engineering and coercion are two of common ways to exploit the humans.

In Hristo et al [4] the authors are exploiting the implicit learning capabilities of human brain. Human beings are capable of performing certain tasks which have to take place so fast that the conscious part of the brain is not involved in that decision making, rather those skills can be learnt by performing them repeatedly. The examples include sports such as ping pong, bicycle riding and a number of other similar activities.

This concept can be used to create a coercion resistant authentication system. The system requires a training phase followed by an authentication phase. During the training phase which lasts around 30 to 60 minutes the users will learn to play a Serial Interception Sequence Learning (SISL) game. Game requires users to intercept sequences consisting of some random data and some specific repeating sequences. It is not possible from game to find out which of the data is random. It is also not trivial to find out when a sequence starts or ends. The sequences which are repeated are implicitly learnt by the user without his conscious knowledge. The end of this game marks the end of the training session. Now whenever a trained user wants to authenticate himself he is presented with the same game again which lasts around 15 minutes. This time around, the user is presented with some random sequences and some of the sequences that he was trained on. If the user performs considerably better at trained sequences compared to the random sequences, this means that he was indeed trained and is the correct user and hence he gets authenticated.

Threat Model:

The scheme presented is valid only in the situation where a user can authenticate in presence of guards to

make sure that someone is not coercing him to authenticate and then enter the building along with him. The facility where this system is deployed has to be very high value and should have extremely low traffic. Low traffic and presence of guards ensures there is no tailgating. Tailgating refers to the situation where an unauthorized user enters the building along with an authorized user before the door closes [3]. Moreover guards are supposed to make sure that every entering person does not use any electronic devices to cheat the authentication. The guards are also supposed to ensure that the user is not recording the sequences for later analysis or some other user is not observing the valid user to learn the sequences.

The only factor which this scheme is resistant against is offline attacks. A valid user may be forced to reveal all the information he remembers but the scheme ensures that he doesn't consciously know enough to let an invalid user successfully pass authentication.

The authors also briefly mention that this scheme is better than biometrics because it cannot be stolen or faked and secondly it is possible to change the password (learned sequence in this case) by simply training user to a different sequence which is not possible in case of biometrics.

User Studies:

Three user studies were performed to check the feasibility of authentication via implicit learning. The experiments were performed using Amazon's Mechanical Turk which provides human experimentation services. The answer to following two questions was attempted to be found:

- Can the user authenticate after training?
- Can the sequence be reverse engineered from performance data?

The results from the experiments indicate that indeed authentication is possible and sequence cannot be reverse engineered with high probability.

2. An Overview of Human Memory System

A human brain consists of multiple memory systems. Two major memory systems are memory of verbally reportable facts or explicit memory and implicit memory which exists but cannot be recalled by conscious effort.

Memory of reportable facts "depends on the medial temporal lobe memory system (including the hippocampus)" [4] [5]. Medial temporal lobe consists of amygdala, brainstem and hippocampus. Amygdala is located right above the hippocampus.

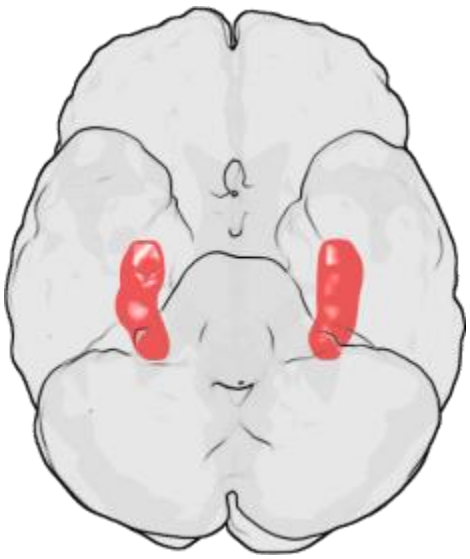

Figure 1: Location of Hippocampus in Brain [6]

Studies have shown that even if the explicit memory system is damaged due to a disease or injury the implicit memory system may potentially still keep working [7].

Several techniques exist which can be used to install information in implicit memory of users. Although the users are not consciously trying to learn something, still the results show signs of learning [8].

One such scheme involves that users are presented with a sequence of statistical data and they are required to somehow respond to it. If they have been presented with same data many times the speed of responses improve. However speed goes back to normal if the user is presented with similar but not exactly the same data in the same scheme. This behavior shows that due to repeatedly performing the same task, users learn those specific tasks. However, they have no conscious knowledge of what they learnt. This implicit memory depends on basal ganglia and connections to motor cortical areas of brain.

During initial phases of research on implicit memory it was believed that brain can only learn small sequences containing 10 to 12 items. However, recent research shows that brain can learn sequences of much more length and learning process is not affected by noise in the data [9]. Brain can recognize the repeating data and can keep learning it while ignoring the non-repeating data. Because of these reasons the brain can be used to embed information which can be recalled in the same implicit way and it can be used for cryptographic purposes.

2.1. The SISL Task and Applet

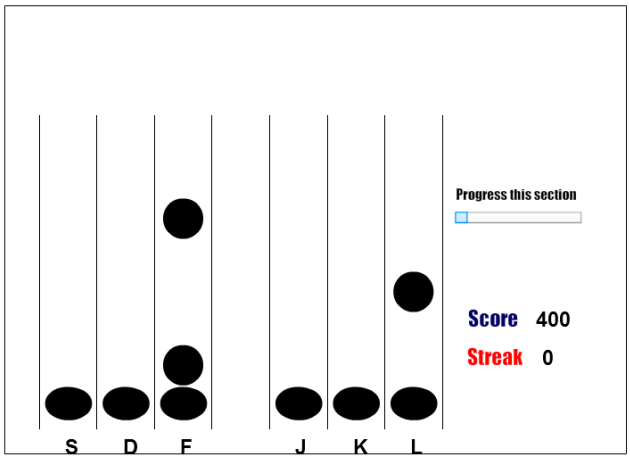

Figure 2: Screenshot of the SISL task in progress.

SISL (Serial Interception Sequence Learning) task is used in the designed scheme both in training and authentication phases.

The task looks like a game which consists of 6 columns. Each column contains a black oval at the bottom end of it labelled with one of following characters SDFJKL. When the game starts new circles appear in columns at top in different order and those circles are dropping with a certain speed. As soon as the circle reaches the oval at the end of column the user is supposed to press the corresponding key (SDFJKL) on keyboard. If the user manages to press the key in time some points gets scored.

A typical game session consists of 30 to 60 minutes duration and the user tries to intercept several thousand of those circles. 80% of the falling circles consist of repeating sequences in a non-predictable way while the remaining 20% are just random characters. Furthermore, the speed of the falling circles change so that the user can only intercept around 70% of characters correctly. If the user is intercepting more the speed is increased and otherwise speed is decreased. The speed of user's performance on trained sequence compared to his performance on untrained sequence is what determines that user is learning. The same thing is used later in authentication.

The sequences are designed so that no easy to remember sequences are present. A sequence contains one pair of characters exactly once and no same character twice to ensure that users can't consciously remember the sequences. During testing of this scheme, the users who were training were asked if they remember any of the sequences consciously.

Each training sequence is 30 items long which makes the possible keys to around 248 billion combinations.

3. Basic Authentication System using implicit learning

During the training phase a key is stored in the brain of the user. The key consists of a sequence of 30 characters over the alphabet $S = \{s, d, f, j, k, l\}$. Not all the possible combinations are used to prevent the user from consciously remembering parts of sequence. Rather, Euler cycle graph is used to create the sequences. Euler cycle graph has the property that every non repeating bigram appears only once. Furthermore, BEST theorem [10] provides a formula which can be used to compute the number of ways possible to traverse an Euler cycle graph which computes the total possible keys in this case to be $6^4.24^6$ which is approximately equal to $2^{37.8}$.

The resulting sequence gives an entropy of 38 bits which is quite reasonable for a secret.

Technical Details of Training:

During the training phase a user is presented with a 30 item sequence which will be the actual key three times. After appearance of sequence three times an 18 items random sequence is presented to mix things up and complicate the conscious memorizing of the sequence. This totals in 108 characters

These 108 characters are repeated 5 times to get a bigger sequence of 540 items. This big sequence is repeated 7 times to get a gigantic sequence of 3780 items. After the user goes through the 540 items chunk a pause takes place before the start of next 540 item sequence.

After completion of authentication the user tries to pass the authentication to check if the training was successful. Another important thing which is noted during the training is the final speed at which the user can play the game with 70% hit rate. This speed will play the crucial rule during authentication phase where it is checked that user performs better at learnt sequences compared to random sequences.

Technical Details of Authentication:

During the authentication phase user is presented with the sequence he was trained on which consists of 30 items and 2 additional randomly chosen 30 item sequences. Each of these 3 sequences are shown to the user exactly 6 times in no specific order. So, the user has to go through 18, 30 items sequences which means intercepting 540 characters.

The characters are dropped at the speed at which user finished playing during the training phase. If the accuracy of user on the trained sequence is greater than his average accuracy on random sequences plus some error margin the authentication is considered successful.

$$p_t > average(p_r + p_r) + \sigma$$

Where p_t denotes accuracy on trained sequence and p_r denotes accuracy on random sequence, σ represents small error margin.

4. Usability

In this section authors mostly discusses if the authentication is possible to be carried out reliably over the time.

To demonstrate that the user conducts a number of experiments using Amazon's Mechanical Turk platform which provides a number of users who can perform the specified tasks for small amount of money. The experiments were carried out in three stages

1. In the first set of experiments the focus was on the fact that learning is indeed possible.
2. Users can retain the learnt information for a period of time like one or two weeks.
3. Scheme are attack resistant so the secret cannot be stolen.

4.1. Experiment 1

During this experiment users are presented with the standard training test consisting of 3780 total trails as discussed in the training section. After completion of the training, users are presented with the authentication scheme as discussed previously.

The results showed clear evidence that users learn the sequences good enough to pass the authentication. Experiments were carried out by 35 individuals and during the authentication phase their accuracy was 79.2% on the sequence they were trained on and 70.6% on the random sequences. The difference in performance was 8.6% with a standard error of 2.4%.

As a second part of this experiment users were presented five sequences at the end of the test to check if they consciously remember them. Users were asked to rank their familiarity with a scale from 0 to 10. The average familiarity score of trained sequence was 6.5 with SE (Standard Error) 0.4 and the same score for untrained sequence was 5.15. With SE 0.3. The author claims that although the trained sequence is a bit more familiar still the familiarity is not big enough to reliably differentiate between random sequence and the key sequence, furthermore it is very difficult to reconstruct the sequence from memory when the recognition is this hard.

4.2. Experiment 2

During this experiment the users were trained on the sequence using the standard method and they were asked to complete the authentication after the time of one week or two weeks. 32 participants participated in the one week delay test while 80 participants participated in two weeks delay test.

For the one week delay test 15 out of 32 participants demonstrated reliable sequence knowledge while 49 out of 80 participants recognized the sequence reliably in the two week delay test. However the authors claim that as a group their information retention was reliable and as a group the information was retained. Furthermore authors claim that since the information loss between first and second week is not much, it is expected that information will also be retained for longer periods of time.

4.3. Mechanical Turk

In this section the author explains how the experiments were conducted using Mechanical Turk. He mentions different incentives that were given to users for conducting experiments and how successful different techniques were.

5. Security Analysis

The security feature that will be focused on in this section is not that the system is secure under the assumption that secret is safe, rather the main focus will be on the fact that the secret will remain secret even under coercion.

The user is first trained and a secret is embedded in the brain of the user. The user can then be asked to perform the experiment which will verify if he contains the secret or not. This is achieved by performing the authentication. Other than performing the authentication, the carrier of the secret or anyone else cannot extract the secret from the brain.

The basic coercion threat model:

The authentication system is designed such that some physical guards check whether a human being is taking the authentication test. The test taker is not allowed to take the test with any other help. It is also not allowed that an adversary will coerce a trained user to pass the authentication test and let the adversary inside the building. Furthermore the adversary is not allowed to take multiple authentication tests. If a user tries to authenticate and fails, the guards will arrest him. Another limitation of the system is that users are not allowed to cheat the training system. If they do not follow proper instructions during training they may be able to break the system. Apart from above mentioned limitations system is secure under the following attack.

An adversary can kidnap or coerce a trained user into revealing his secret or to reveal as much of the secret as he can reveal. After extracting the secret the adversary can go to the authentication site and tries to pass the test.

To show the security of this system let's suppose that the attacker captures *u* trained users which were all trained on

the same sequence and he makes them go through q tests each. After the tests the adversary will check whether the user performed better on any of his randomly generated sequences. The probability of finding a real sequence would be then be

$$qu/(key\ space)$$

If 100 users are intercepted and 10^5 queries are asked from each then the probability of finding a sequence is 2^{-16}. Another point worth noting is that if you make the user take many test randomly, he might actually forget the secret.

This scheme is not secure if the user can perform the same authentication multiple times because each time he tries authentication he is being trained on the sequence. There are a total of 3 sequences in authentication and two of them are random. If the user does it enough times he will be trained and will pass authentication. Another possible attack here is that user can try to memorize some sequence and train himself on it offline and try again next time. At a probability of 1/3 he had memorized the correct sequence and will succeed during next authentication attempt. Furthermore, the attacker can memorize all three sequences and will train on them and authenticate the next time. However, its less likely as memorizing these many symbols during playing are difficult. Recording the game using a video camera is not allowed in this threat model.

Since the authentication depends on the fact that user performs better at correct sequence compared to others, the attacker might just perform deliberately poor at one sequence compared to other two. This can be defeated by not training the user on one sequence but training them on multiple sequences such as four. Human brain is capable of storing multiple patterns without intermingling them. This will increase the training time but it will increase the security as well. During the authentication phase the user is presented with 4 correct sequences and 8 random sequences and the probability of cheating this system by randomly performing worse at some sequences compared to others becomes as low as 1/500.

The attacker can try to memorize all the symbols and create a game out of those and coerce a legitimately trained user to play the game to determine which of those sequences the correct ones are. However, the defense against that can be that during authentication users are not presented with new key sequences unless they show a reasonable familiarity with currently presented key sequences. If the user fails to demonstrate reasonable familiarity with a key sequence, remaining sequences shown to him will be random. Therefore the memorization will become less and less effective. The attacker can still repeat it many times to get correct sequences using coercion but if during authentication the user shows that he knows some sequences very well and some sequences not at all, the system can recognize this as an attack and alert the guards.

This system is also vulnerable to eavesdropping attacks such as shoulder surfing or traffic stealing.

5.1. Extracting Sequence Fragments

The training sequences are constructed such that each 30 item sequence consists of 30 unique trigrams. The total possible unique trigrams are 150. The attacker can try to create a sequence consisting of all the possible combination of trigrams and let a trained user play the game. It is suspected that the trained user might perform better on the trigrams belonging to the key compared to rest of trigrams. If this is possible then the key sequence might be reconstructed from learnt pieces.

To check if this attack can be carried out, the authors performed an experiment in Mechanical Turk where trained users were presented with 10 sequences, each containing of all 150 possible combinations. The performance of users was measured by checking if the users managed to intercept the learnt trigrams or not.

Results of this experiment showed that the performance on the trigrams present in sequence was same as the performance on random sequences and no considerable difference was observed which could be used to reconstruct the sequence. To be more specific the 34 participants who took part in experiment on average showed 73.9% correct for trigrams from trained sequence and 73.2% on random trigrams. The difference was not reliable or consistent.

If the same attack is tried with a bigger fragments e.g. 4 letter subsets of the sequence the total possible combinations raise to 750 and they increase exponentially with further increase in fragment size.

Future Work:

More research on the size of fragment which is needed to express the sequence knowledge will be conducted. Also, more focus will be put in the cases where users can be coerced during training phase.

6. Related work

The work presented in this paper is superior to biometrics, passwords and security tokens due to the fact that it provides coercion resistance and cannot be stolen. Furthermore it is also superior to the techniques which relies on permanent human behaviors such as walking pattern or typing style etc. The major difference is that the learnt key in the scheme of this paper can be changed by

revoking it and letting the user learn a new one, however this is not possible or convenient with walking style etc.

This scheme is also better than schemes which require humans to memorize image patterns or any other thing which requires conscious memory. Those schemes are good against brute force attacks but are not good against coercion attacks.

7. Criticism

The scheme presented in this paper can be subdivided in many components. One of the most important such component is the ability to learn implicit information and to be able to reproduce it at a later time. Authors proved this technique through a number of tests they conducted which shows that it is indeed possible to learn sequences implicitly without any explicit knowledge of the sequence. This technique can be useful in many applications e.g. authentication as discussed in this paper. It may be possible to extend this idea to create a secret sharing scheme where rather than giving a piece of information to each user, they are trained on a specific sequence out of many possible sequences. If multiple users manage to reproduce their sequences the task which has to take place can be conducted.

This scheme also has many shortcomings but it should be noted that most of shortcomings are related to application of implicit memory in authentication which is not the complete work of the paper. Some of these shortcomings will be mentioned below.

The training phase of this scheme is very time consuming and is very non-user friendly. Users have to repeat the same procedure over and over again without any reward. The author of this summary tried to play the game and found it not very engaging and difficult to play for longer.

The logging in procedure takes very long to complete. It may take up to 15 minutes before a user can actually log in to the system. The facilities which are extremely critical which would require such authentication system are also usually very time critical. Due to this limitation very few practical applications are possible. Furthermore this limitation makes this scheme not friendly for critical high traffic buildings e.g. the office of some billion dollars corporation or a sensitive government building where users want to go every day to work. Even with quick authentication methods as a key card, tailgating is common in busy buildings.

Some of the assumptions made under which the security was proven are very strict and they leave the scheme with much less to no practical applications. The assumption that guards will be present to check whether a user is performing the test fairly leaves us to the same attack which this paper is trying to prevent. A user can bribe or coerce the guards while he is cheating the test. Another attack can be that the adversary can coerce a trained user to pass the test while bribing the guards to let him inside the building after the authentication is already completed. If guards themselves can enter the building without authentication they can lead to security breaches, even if guards are required to pass this test themselves, who will make sure that they are not cheating.

Furthermore the system does not allow the coercion resistance during training which can lead to attacks.

The system also does not allow the user to perform the authentication test multiple times because it leads to the fact that user is getting trained on the sequence which is the key. If a user attempts authentication 10 times he may actually pass the test 11^{th} time. This further limits the usability of the system and puts pressure on user that he has to get the authentication correct the first time.

Another thing which author neglected to mention is how this scheme will perform under pressure. Such as if a user is very sad or is extremely nervous because his authentication will lead to annihilation of another country by a nuclear bomb.

It is also possible to measure a person's EEG signals and replay them during authentication phase.

Some of the claims made in the paper do not have solid reasoning behind it, one of them was that since information was retained for up to 2 weeks and the loss of information from first week to second week was lesser than between day zero to first week it is highly likely that knowledge will not decay over longer periods of time. More research is needed to prove this claim. Furthermore, the percentage of users who successfully authenticated after one week or two weeks was not 100%. Majority of the users remembered the key and managed to authenticate but there was quite a few users who did not remember the key enough to authenticate it. This information is good enough for statistical results, but in practice if an important trained user forgets (sub-consciously) his key it can be very critical.

8. Acknowledgements:

This text is a summary of work done by Hristo Bojinov, Daniel Sanchez, Paul Reber, Dan Boneh and Patrick Lincoln in their paper "Neuroscience Meets Cryptography: Designing crypto primitives secure against rubber hose attacks". Special thanks to Tambet Mattisen for his thoughts on this paper which were incorporated in the criticism section.

9. References

[1] Professor Fred B. Schneider, Tom Roeder, “Something You Know, Have, or Are." 2011. Accessed on 11 Oct. 2013
<http://www.cs.cornell.edu/courses/cs513/2005fa/nnlauthpeople.html>

[2] Schneier B: ‘Secrets and Lies’, John Wiley and Sons (2000).

[3] "Piggybacking (security) - Wikipedia" 2007. Accessed on 11 Oct. 2013
<http://en.wikipedia.org/wiki/Piggybacking_(security)>

[4] Hristo Bojinov, Daniel Sanchez, Paul Reber, Dan Boneh and Patrick Lincoln. Neuroscience Meets Cryptography: Designing Crypto Primitives Secure Against Rubber Hose Attacks, 2012.

[5] Paul Reber. Cognitive neuroscience of declarative and non-declarative memory. Parallels in Learning and Memory, Eds. M.Guadagnoli, M.S. deBelle, B. Etnyre, T. Polk, A. Benjamin, pages 113–123, 2008.

[6] User:Washington irving, 2004, Accessed on 13 Nov 2013.
<http://commons.wikimedia.org/wiki/Hippocampus>,

[7] Brooks, D.N. & Baddeley, A.D. (1976). What can amnesic patients learn? Neuropsychologia, 14, 111-129.

[8] Mary J. Nissen and Peter Bullemer. Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19(1):1–32, January 1987.

[9] D.J. Sanchez and P.J. Reber. Operating characteristics of the implicit learning system during serial interception sequence learning. Journal of Experimental Psychology: Human Perception and Performance, in press.

[10] T. van Aardenne-Ehrenfest and N. G. de Bruijn. Circuits and trees in oriented linear graphs. Simon Stevin, 28:203–217, 1951.