Systematic Overview of Password Security Problems

Viktor Taneski, Marjan Heričko, Boštjan Brumen

Faculty of Electrical Engineering and Computer science, University of Maribor, Koroška cesta 46, 2000 Maribor, Slovenia

viktor.taneski@um.si, marjan.hericko@um.si, bostjan.brumen@um.si

Abstract: Alphanumeric passwords are the first line of defense in security for most information systems. Morris and Thompson identified passwords as a weak point in an Information System's security, 35 years ago. Their findings showed that 86% of the passwords were too short, contained lowercase letters only, digits only, were easily found in dictionaries and/or easily compromised. The objective of this paper is to perform a systematic literature review in the area of passwords and passwords security, in order to determine whether alphanumeric passwords are still weak, short and simple. The results show that only 42 out of 63 relevant studies propose a solid solution to deal with the identified problems with alphanumeric passwords, but only 17 have statistically verified it. We find that only 3 studies have a representative sample, which may indicate that the results of the majority of the studies cannot be generalized. We conclude that users and their alphanumeric passwords are still the "weakest link" in the "security chain". Careless security behavior, involving password reuse, writing down and sharing passwords, along with an erroneous knowledge concerning what constitutes a secure password, are the main problems related to the issue of password security.

Keywords: authentication; password security; password security problems; systematic literature review

1 Introduction

The rapid growth of the Internet technology and the widespread use of the World Wide Web (WWW) has changed the way people operate nowadays. The increased number of online services, online social networks (e.g. Facebook, Twitter, etc.) and other websites that have content that is tailored to the users' interests, has increased the need for authentication mechanisms. Authentication is the core of today's Web experience [1]. Online services, social networks and websites require an authentication so that users can create a profile, post messages and comments, and tailor the website's content so it can match their interests.

In an information security sense, authentication is the process of verifying someone's identity. Typically, authentication can be classified into three main categories: *knowledge-based authentication* - *"what you know"* (e.g., textual or graphical passwords), *biometrics authentication* - *"what you are"* (e.g., retina, iris, voice, and fingerprint scans), and *token-based authentication* - *"what you have"* (e.g., smart cards, mobile phones or other tokens). Lately, another alternative authentication method is becoming more available - the two-step verification. The problems with these alternative authentication methods are not related to the security itself, in fact, these methods also provide excellent security for the system. Instead, the weaknesses of these authentication methods are that they can be expensive (biometrics, smart cards), they must be carried around at all times when access to the system is required (smart cards, two-step verification), they are difficult to implement on a large scale, and they are not widely accepted by the users. Single Sign-On (SSO) is another method for authentication that is recently becoming more available, that provides access to many resources once the user is initially authenticated. However, a recent study [2] found that SSO solutions impose a cognitive burden on web users, and users have significant trust, security, and privacy concerns, which hinders the wide acceptance and usage of SSOs.

We focus on the textual passwords and their security simply because the username-password combination used to be [3] [4] and still is the most widely used method for authentication [5]. Even though passwords suffer from a number of problems, they continue to be one of the most common control mechanisms to authenticate users in information systems, due to their simplicity and cost effectiveness. The problems related to textual passwords and password security are not new. Morris and Thompson [6] were first to identify textual passwords as a weak point in information system's security. More than three decades ago, they conducted experiments about typical users' habits about how they choose their passwords. They reported that many UNIX-users have chosen passwords that were very weak: short, contained only lower-case letters or digits, or appeared in various dictionaries. Zviran and Haga [7] had similar findings in their study conducted 20 years later. They came to the conclusion that users are one of the biggest threat to information system's security. In their study, almost half of the users created passwords composed of five or fewer characters, 80% had only alphanumeric characters, and 80% never changed their password.

The objective of this paper is to perform a systematic literature review of studies related to textual passwords and textual passwords security. There are three reasons for conducting the review in this specific field. The first reason is to identify any problems that may arise in creating or managing textual passwords. The second reason is to assess the current situation of passwords with respect to password strength, password management and password memorability. Finally, the third reason is to find out whether the users are still considered the "the weakest link?" in information security.

The paper is based on our previous work [8] where we presented the preliminary results of our systematic literature review. We extend our preliminary work by including additional papers that were published in the period from 2014 to 2018. We improve our systematic literature review in a more detailed and strict manner. We also perform quality assessment for the relevant studies and categorize the data extracted from the studies in order to answers our research questions more effectively.

2 The Review Methodology

A systematic literature review (SLR) is "a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest." [9]. Most research studies begin with the process of literature review. The extent and the properties of the review are not necessarily fully consistent with the research methodology of systematic literature review. If a literature review is not conducted in a thorough and proper way, its scientific value is low. To this end, we need a systematic literature review carried out in accordance with a pre-defined search strategy. When performing our systematic literature review, we took in consideration the guidelines by Kitchenham and Charters [9] for performing SLR in software engineering. These guidelines propose carrying out the systematic literature review in the form of three major phases: *planning the systematic literature review*, *conducting the review* and *reporting the review (presenting the results)*. The tasks performed in each phase are described in more detail in the following subsections.

2.1 Planning The Systematic Literature Review

A review protocol should be defined prior to conducting the systematic literature review in order to reduce the researcher's bias [9]. The protocol prescribes pre-review activities and, in our case, includes *defining the research questions*, *defining the search strategy*, *defining the study selection procedure*, *defining the quality assessment checklist* and *data extraction strategies*.

2.1.1 Research Questions

We used a modified version of the Population, Intervention, Comparison, Outcomes, Context (PICOC) [9] [10] structure, in order to construct well-formulated research questions. This structure contains the attributes that can help us define the research questions. The *population* is represented by textual passwords, while the *intervention* is represented by different approaches (methods, strategies or techniques) used for creating and managing textual passwords. The

attribute *comparison* is not applicable, because we do not compare textual passwords with other types of passwords, since our subject of interest are strictly textual passwords. The *outcome* refers to the problems that may arise when creating and using (or managing) textual passwords. The *context* are the user-selected textual passwords and the relationship between the users and their textual passwords.

Considering the above structure, we formulate the following research questions:

- **RQ1**: What are the major problems with creating and managing textual passwords?
- **RQ2**: What is the current situation of textual passwords with respect to the password strength, password management and password memorability?
- **RQ3**: What is the relationship between users and textual passwords?
  - **RQ3.1**: Are the users still "the weakest link"?

We defined the RQ1 to get a better view if the past, and already known problems related to creating and managing textual passwords, still exist today. Please note that the term "managing" is referred to the way users use and store their passwords, and manage the aspects surrounding it (e.g. how often do they change it, do they reuse the password on several other services and accounts etc.).

In addition, we are interested in the current situation of textual passwords with respect to the password length, password management and password memorability. This issue is covered by RQ2 and helps us assess the current situation of textual passwords so we can make a comparison with earlier findings in this area.

Furthermore, RQ3 helps us assess the current relationship between the users and their textual passwords. Users were already identified earlier as "the weakest link" because they used very week passwords (short, contained only lower-case letters or digits, or appeared in a dictionary). Confirming that this statement still holds, combined with the answers to the RQ2, can help us outline directions for future research that can be used in aiding the users when selecting and managing their textual passwords.

2.1.2 Search Strategy

The important phases of our literature search strategy are: *initial search* and *reference search*. The initial search was performed over digital libraries. When selecting the digital libraries, we followed the recommendations in [9]. We also took into account our knowledge and practical experience, and the fact that we do not have access to *all* digital databases. We carried out the search through the following databases: *IEEEXplore*, *ScienceDirect*, *SpringerLink* and *ACM Digital*

Library. The Google Scholar database was also considered, but was not included. Later in the paper we will explain the reasons for not including this database. We used these databases since they provide for many of the leading publications in the Computer Science field. Furthermore, these databases allow searching by keywords. We restricted the initial search to articles in journals, conference papers and books/book chapters written in English that were peer-reviewed and published since 1979, i.e., the year when the first article in the area of password security was published. We conducted this search in 2018. Therefore, this systematic literature review includes studies that were published before and including 2018.

We composed a search string for searching through the digital databases. The search string contains major search terms from our research questions connected by using Boolean OR:

("password security" ***OR*** *"password strength"* ***OR*** *"password memorability"* ***OR*** *"password cracking"* ***OR*** *"password management").*

During the search of the digital databases it was necessary to slightly modify the search string and to modify it in such a way so it could fit the syntax requirements and capability of the search engine of each digital database used.

After the completion of the initial search, we performed a reference search by reviewing the reference lists of studies found in the previous step in order to identify additional studies that are relevant to our review.

2.1.3 Study Selection Procedure

We performed the search by using the search string and the search result was a set of documents in which the search string appears partially or entirely. We excluded irrelevant studies and publications, and select those that are relevant to our study and may very likely provide answers to our research questions. We systematically selected the relevant studies by applying the following steps:

1. We examined the paper titles and excluded the papers and publications that were clearly irrelevant to our search focus.
2. We examined the abstracts and keywords in the remaining studies to select relevant studies.
3. For filtering the remaining studies, we used inclusion and exclusion criteria given in Table 1. To carry out the selection in an objective manner and to reduce the likelihood of bias, we defined the inclusion and exclusion criteria during the definition phase of the review protocol (with a possibility of later adjustments during the search).

The titles and abstracts of the documents do not always provide clear information whether the document meets the specified criteria. If this was the case, we took a further step and read the whole document to determine whether the it meets the inclusion and exclusion criteria, which is presented in Table 1.

Table 1
Inclusion and Exclusion Criteria

Inclusion criteria	**Exclusion criteria**
1. Studies that focus mainly on textual passwords	1. Studies that focus on graphical passwords or any other type of user authentication (biometrics, tokens or smart cards, etc.)
2. Studies that focus on password security	2. Studies that deal with computer security or cryptography in general
3. Studies that present method(s) for password creation	3. Studies that are not peer-reviewed
4. Studies that deal with issues or problems with password use, password management or password memorability	

2.1.4 Quality Assessment Checklist

The quality assessment is a means of weighting the importance of the relevant studies and relates to the extent to which the study minimizes bias [9]. We evaluated the quality of the selected relevant studies and we based our quality assessment on a *quality instrument* which is a checklist that needs to be evaluated for each relevant study. Our quality assessment checklist comprises of three main questions, each of which directly corresponds to one of our main research questions. The questions are answered with 'Yes', 'No' or 'Partially' to which values '1', '0' or '0.5' are assigned, respectively. Each of the quality assessment questions also contains additional sub-questions. The scores for the sub-questions are divided so that the overall score of each question would range between '1' (very good) and '0' (very poor). For example, the first question has four sub-questions, which can be answered with 'Yes', 'No' or 'Partially' for which values '0.25', '0' or '0.125' are associated, respectively. The three quality assessment questions are:

1. Does the study address any problem related to creating or managing textual passwords?
    a. Is it clear what problem is identified?
    b. Is the problem clearly defined?
    c. Is there a solution proposed for solving the identified problem?
    d. Is the proposed solution experimentally or statistically verified?
2. Is the approach towards acquired data for password strength, password management or password memorability sound?
    a. Is the data retrieved through a questionnaire?
    b. Is the sample size known?

c. Is the sample representative?

d. Is the data retrieved through an experiment?

e. Is the experiment set in a realistic setting using real data?

3. Does the study address the issue of users being "the weakest link"?

a. Is the approach that addresses the issue well-defined?

b. Is there a proposed solution for improving the relationship between users and their textual passwords?

The academic studies about password security and usability can be divided into two major categories: studies of real world passwords (e.g. leaked/cracked password lists like the RockYou or MySpace password databases) and user studies [11]. Furthermore, the most common choices for a user study are *online study (in a form of an online survey)* and *laboratory studies* [11]. When it comes to such password studies an important issue is the *ecological validity*. Ecological validity refers to whether or not the findings of a research study are able to be generalized from observed behavior in the laboratory to real-life settings [12] i.e. do the participants of the study behave the way users would in real life. Ecological validity is very important in user studies, since it is believed that the description of the study can influence user behavior from the beginning of the study. The authors in [11] explored the impact user study setups actually have on the ecological validity of these studies. They came to a conclusion that participants are biased and their behavior changes due to the fact that they are participating in a password study.

The terms "experiment", "realistic setting", and "real data" are closely related to the context of ecological validity. In our case, the term "experiment" can be defined as a laboratory study where users are not in their natural environment or an analysis done over leaked password lists. The term "realistic setting" can be defined as an environment where users are not aware that they are being studied (e.g. at home, at work etc.). The last term "real data" represents real world passwords that users are using in their everyday life.

An experiment about users' passwords, conducted over real world passwords (real data) or in an environment where users are not aware that they are being studied (realistic setting), can significantly reduce the potential bias. Furthermore, a combination of an experiment conducted in a realistic setting, using real password data, with a survey can additionally increase the value and the quality score of the study.

2.1.5 Data Extraction Strategy

In addition to the quality assessment check list, we need to extract relevant information from the selected studies for answering the research questions. To that end, and also to make sure that the task is performed in an accurate and consistent

manner, we used a data extraction form based on the research questions. Again, in order to prevent bias, it is important that this form is defined during the definition phase of the review protocol. Table 2 shows the data extraction form used for retrieving relevant data from the selected studies.

Table 2
The Data Extraction Form

Data Item	Description
Basic information about the study	
Title	The title of the study
Author(s)	The author(s) of the study
Venue	Venue where the study is published
Type	Type of the article (journal/conference/book section)
Year	Year of publication
Analysis of the abstract	
Problem	The problem statement in the abstract
Idea	The idea of the paper described in the abstract
Research data	
Domain	The domain of textual passwords
Research methodology	The research methodology used for retrieving the results (experiment, questionnaire or both)
Sample	The type and the size of the sample (if there is one)
Realistic setting	Is the survey or the experiment conducted in a realistic setting using real data?
Identified problems	Identified problems related to password use, password management or password memorability
Proposed solutions	Proposed solution for the identified problems
Interpretation of results	
Conclusions	Conclusions and findings from the research
Main results	Main results of the research
Future work	Future work stated in the study

2.2 Conducting The Review

2.2.1 Initial and Reference Search Results

Table 3 lists the results from the initial (keyword based) search. The first column represents the electronic databases. The second column shows the number of studies found in each database, while the number of studies that have already occurred in another digital database is presented in the third column. The last column shows the total number of relevant studies (excluding the duplicates). The search resulted in relevant studies published in journals, conference proceedings

or book titles. When selecting the relevant studies, we used the predetermined and detailed inclusion/exclusion criteria, presented in subsection 2.1.3 Study Selection Procedure. The study selection in the initial search, in our case, was performed by one of the co-authors. After selecting the relevant studies from the initial search, the co-author consulted and discussed included and excluded papers with the other co-authors. We acknowledge that there is still a possibility of researcher bias in the process of study selection.

Table 3
Summary of Found and Selected Studies

Electronic database	**Found studies**	**Duplicates**	**Relevant studies**
IEEEXplore	192	20	27
ScienceDirect	88		9
SpringerLink	1676		17
ACM Digital Library	144		29
Total	**2100**	**20**	**82**

The initial search found a total of 2,100 candidate studies. We first examined all 2,100 studies by reading the paper titles and removed studies that were unrelated to our research focus. The next step included reviewing the abstract of each remaining paper to exclude additional studies that are not relevant to our research. In some situations, the abstract did not provide enough information to determine whether a study is relevant to our research. In this case, we reviewed the introduction and the conclusions of the article, as well. Next, we examined the content of the remaining studies by reading the whole documents, and filtered them by applying the inclusion and exclusion criteria. After applying these three steps, 101 studies remained. As we were searching through different search engines, we encountered some duplicates, i.e., studies that already appeared in more than one digital database. In this sense, 20 studies were excluded because they were duplicates. In the end, we have **82** relevant studies from the search into 4 electronic databases. Table 3 shows that the ACM Digital Library contributed the largest number of relevant studies (35.36%), while ScienceDirect contributed the smallest number of relevant studies (11.11%). Figure 1 shows the distribution of published relevant studies per year.

We identified the relevant studies that provide the needed information for answering our research questions through a keyword based search in four digital databases. When using keyword based search, there is a potential risk of incomplete identification of relevant studies. There is a possibility that there are some relevant studies that do not explicitly mention the keywords that we use. Hence, there is always a risk that we might have missed some relevant studies during the initial search. When performing the initial search for relevant studies, we did not include the Google Scholar digital library because of a few reasons. The first reason is that Google Scholar only supports keyword search by title and full text and does not support keyword search within paper keywords and abstract.

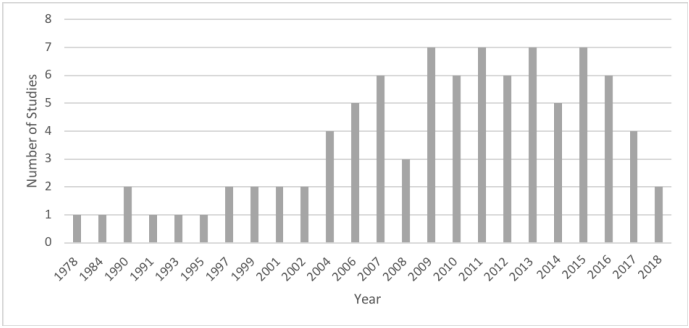

Figure 1
Distribution of relevant studies published per year

The second reason is that this increases the number of hits and complicates the selection of relevant papers. The third reason for not including this digital library is that we performed a test search by using our search string, and while reviewing the first 10 pages (according to a study made in 2013 the first 10 pages in Google search results are the most significant and the most visited [13]), we came to a conclusion that they were containing articles that we have already found in other digital libraries, that we have included in our study. We found no additional relevant papers with this search.

In addition, the search of the references of the primary studies found 8 additional articles that were not found by the initial search of the electronic databases. Thus, we used a total of **90** articles as relevant studies for our systematic literature review.

2.2.2 Quality Assessment

We performed quality assessment on the relevant studies based on three quality assessment questions presented in subsection 2.1.4. Due to the insufficient or unclear data in some of the papers (e.g., unknown sample size, unclear procedure for performing the study, unclear whether real data are used when performing the experiment or whether the experiment is set in a realistic setting, etc.), there is a slight possibility of bias when answering these quality assessment questions. This bias can later affect the data extracted from the relevant studies. In order to reduce the likelihood of bias, when fulfilling the quality assessment checklist and assessing the relevant studies, and to provide better quality assessment results, the quality assessment task was performed by two independent evaluators. The results were statistically compared in order to find out if there is an inter-ratter agreement between the two evaluators. We can achieve this if we compare the scores of the two evaluators using the Kappa coefficient, so we can find out if there is any inter-ratter agreement. Kappa coefficient measures inter-ratter agreement for qualitative (categorical) items and takes into account the agreement occurring by chance.

Since there were only two evaluators/ratters per subject we used the Cohen's kappa instead of Fleiss' kappa, which is used in case there are more than two evaluators/ratters [14] [15].

Each evaluator assessed each of the 90 relevant papers by answering the quality assessment questions. This resulted in two sets of scores, which come from the same pool of relevant studies. The two evaluators/ratters were independent (i.e., one ratter's judgement did not affect the other ratter's judgement) and physically apart from each other. With these precautions we removed the potential for bias from the quality assessment evaluation as much as possible.

The statistical comparison of the quality assessment scores of the two reviewers gives us the following Kappa Coefficients: $k = 0.806$ ($p < 0.0005$), $k = 0.844$ ($p < 0.0005$), and $k = 0.796$ ($p < 0.0005$) for each quality assessment question pair respectively. These coefficients show that there is excellent agreement beyond chance [15]. Furthermore, since $p <= 0.000$ (or $p < 0.0005$), our kappa (k) coefficient is statistically significantly different from zero.

The further classification of the relevant studies will follow the organization of the quality assessment. It is important to note that we only used the average quality assessment scores to organize the relevant studies, which are suitable for answering each research question, into tabular form so as to provide concrete and concise answers to our research questions. Our intention is not to objectively assess the quality of the studies, or in any other way to criticize any of the studies, since that is not the purpose of this research.

3 Results

In this section, we provide answers to our research questions, defined in Section 2.1.1. We took a comprehensive analysis of the relevant papers to extract the necessary data from the selected relevant studies. The data extracted and used to answer each research question is organized in a tabular form. Only the studies that are associated with a score higher than '0' (quality score '0' means that the study is not relevant for answering the corresponding research question) for the corresponding quality assessment question are taken into account. The studies are sorted by multiple criteria: a descending order with respect to their average quality score and an ascending order with respect to the year that the study was published. For every research question, we present a summary of the results and a discussion.

3.1 RQ1: What are the major problems with creating and managing textual passwords?

In order to answer this research question, we analyzed the relevant studies regarding the identified problems and proposed solutions for those problems. 27 relevant studies have a quality score of ’0.5’ or lower. Such studies are further not taken into account, since provide an incomplete answer to this research question because the addressed problems are either not clearly defined in the study or there is no proposed solution.

We analyzed the relevant studies to identify the most common problems and most common proposed solutions related to creating and managing textual passwords. We identified 11 different categories of problems related to creating and managing textual passwords: *Human limitations, Multiple passwords, Weak passwords, Password reuse, Information overload, Password writing down, Users lack security knowledge, Strong password policies, Password sharing, Poor password management, Outdated password strength metrics*. The studies that 1.) do not belong to any of these 11 categories, or 2.) neither identify a common problem nor specify a solution, are classified under the category *Other*. Due to the nature and the interrelationship of the problems related to creating and managing textual passwords, some problems were address by multiple studies.

Figure **3** shows the number of studies for each identified category of the most common problems.

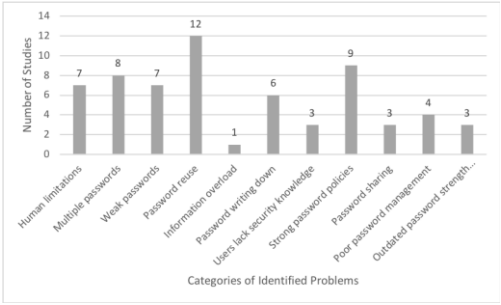

Figure 2
Identified categories of most common problems

We found that almost all (86 out of 90) relevant studies address a problem related to creating or managing textual passwords. The most common problems are related to *password reuse* and are addressed by 12 out of 63. Reusing the same password for more than one account can cause serious damage and can compromise other accounts in the system. Enforcement of *strong password policies* is the second most common problem and is addressed by 9 out of 63 studies. The problems related to users having *multiple passwords* to maintain are

addressed by 8 out of 63 studies. Further down the list are the problems related to users choosing *weak passwords* (7 out of 63 studies), *human limitations* (7 out of 63 studies), users *writing their passwords* down (6 out of 63 studies), *poor password management* (4 out of 63), *password sharing* (3 out of 63 studies) and *outdated password strength metrics* (3 out of 63), the problems related to *users lacking security knowledge* (3 out of 63 studies) and the *information overload* (1 out of 63 studies) as a reason for users having problems to remember and manage all their passwords.

After identifying the most common problems related to creating and managing textual passwords, we went further analyzing whether those studies have proposed some solution for coping with the identified problems. The evidence from these relevant studies helped us identify the most common solutions proposed by the reviewed studies, for creating and managing textual passwords. By the studies we identified 13 different categories of proposed solutions related to creating and managing textual passwords: *Mnemonic passwords, Password meter / password rule presentation, Cognitive passwords, Proactive password checker, A "user-centered" approach, Password policy, Persuasive technology, Information security training, Associative passwords, Password manager, New password strength metrics, New password security scheme, Recommendations*. The studies that do not belong to any of these 13 categories are classified under the category *Other*. Figure 3 shows the number of studies for each category of most commonly proposed solutions.

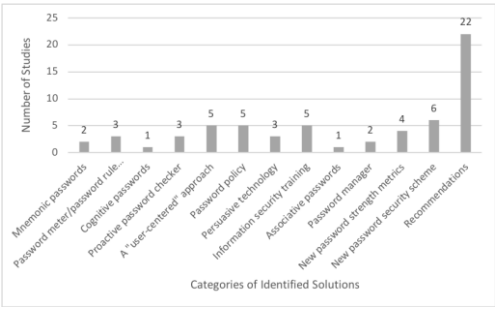

Figure 3
Identified categories of most common solutions

We found that 42 out of 63 relevant studies propose a solid solution for an identified problem related to creating or managing textual passwords. The most common solutions are *a new password security scheme* (addressed by 6 out of 63 studies). A *"user-centered" approach* (addressed by 5 out of 63 studies) and *password policies* (5 out of 63 studies) are the second most common proposed solutions. Following are *information security training* (5 out of 63), and *new password strength metrics* (4 out of 63) as an approach for encouraging users to choose stronger passwords. Further down the list are *password meter or other*

presentation styles (addressed by 3 out of 63 studies), *proactive password checker* (3 out of 63), and *persuasive technology* (3 out of 63) also proposed as solutions for encouraging users to choose stronger passwords, *mnemonic passwords* (addressed by 2 out of 63 studies), *password manager* (2 out of 63), and finally *cognitive passwords* (1 out of 63), and *associative passwords* (1 out of 63). Only 17 of these studies have statistically or experimentally verified their proposed solutions. 22 studies propose a set of (unverified) recommendations, while 26 out of 90 propose neither a solution nor a recommendation.

3.1.1 Discussion on Research Question 1

Morris and Thompson [6] are the first authors that addressed the issue of password security in 1979. The goal of their experiments was to determine typical users' habits related to choosing passwords. They found that users very often choose short and simple passwords that are constructed from a restricted character set (e.g., alphanumeric password with all lower-case letters) and can be found in dictionaries. To increase the difficulty of password cracking and to prevent the fast, simple attacks it is important that systems implement certain password policies that will require the passwords to contain a certain amount of entropy [16].

Unfortunately, users do not always comply with password policies. Basically, human limitations are one of the most common problems related to information security, that is still addressed today [17]. Research has proven that despite the recommendations by information system professionals and their efforts to educate users about secure password policies, users still tend to choose weak passwords that are easy to remember. Very often these passwords are based on user's personal data, or a combination of meaningful details [7]. These problems can be related to users' lack of security motivation and understanding of password policies. In our previous work [18] we performed a questionnaire where we tried to motivate and educate users about frequent password change. We analyzed the effect of password security training on user's practices regarding password creation, frequent password changes and their consciousness about security and the importance of creating strong and hard-to-guess passwords. We found that educating users about password security and assisting them with creating secure passwords can raise the security consciousness of system users and can help achieve greater security. Unfortunately, in order to provide better password security, some security systems incorporate stricter password policies that are forcing users to create stronger passwords with higher amount of entropy. Entropy is a two-edged sword, since higher entropy increases the difficulty of the user to memorize a password [19]. One should be careful with creating a password policy. One "side-effect" of strong password policies is that users tend to circumvent password restrictions for the sake of convenience [20]. Another "side-effect" of strong password policies noted in the literature is users writing down their passwords [21] [22]. Furthermore, due to the increased number of online

services requiring password based authentication, the number of different passwords for different accounts that one user has to maintain is increasing [23]. We expect users to follow common recommendations that say different passwords should be used for every account in order to prevent their other accounts and the accounts of other users in the system to be compromised [24]. Such recommendations are not in line with the issues related to human limitations that we mentioned earlier. This results in the users having many different accounts to maintain and many different passwords to remember, and tend to reuse their passwords.

One of the most common vulnerability related to password security is the password reuse (using the same password or a very similar one for multiple accounts or secure items) [21] [25]. In a system where one password is used for authentication for more than one different accounts, password reuse can cause serious damage, if the password is successfully cracked for a single (not-so-important) account. Information may be revealed that can aid the hackers in infiltrating into other accounts (including the ones that are far more secure that the first one) [24].

Since there is 35 years of research in this area, we were expecting that a proper and useful solution for solving the trade-off between memorability and security would have already been found. The above results give evidence that this is not the case. Contrary to our expectations, we find that only about half (42 out of 90) of the relevant studies propose a solid solution for better coping with the identified problems with textual passwords. Most of the relevant studies propose recommendations for better password creation and password management. We summarize the most common recommendations to the following:

- A secure password should not appear in dictionaries, should not be too short and should not contain personal data
- The use of special characters is strongly advised in order to increase the password security
- Some of the studies recommend a strategy which consists of creating different passwords suitable for different accounts regarding their level of security (e.g. simpler passwords should be used for accounts that contain less important information)
- Associative passwords (i.e. passwords based on associations) combined with guidelines for categorized passwords can ease the construction of strong and easy-to-remember passwords
- The use of enterprise single sign-on is advised as a coping mechanism with password overload and eliminates the need for users to remember multiple passwords

Overall, this is a very low number of proposed solutions, given the fact that these problems have been known for some time now. Furthermore, in the next section we present that only a small number of these studies have verified their solutions. Despite all these solutions and recommendations, there is no common solution proposed or verified by the academic world, and businesses are far from a standardized solution.

3.2 RQ2: What is the current situation of textual passwords with respect to password strength, password management and password memorability?

We searched for attributes in our relevant studies that can help us assess the current situation of textual passwords with respect to their strength, management and memorability. We were interested in what type of research methodology is used (e.g., questionnaire, experiment), whether a realistic setting and real data are used and whether the sample is representative or not.

We found that 70 out of 90 studies are relevant for answering our second research question. For this analysis we took into account all of the studies that had a research methodology. We only excluded the studies with score of '0'. By examining the relevant studies found that half of the studies (35 out of 70, or 50%) are neither conducted in a realistic setting nor use real data (real textual passwords that users use in their daily life).

By analyzing the relevant studies, we identified 2 research methodologies that are used. All 70 studies have retrieved the data either through a survey (questionnaire), an experiment or a combination of a questionnaire and an experiment. A questionnaire as a research methodology was a choice for 23 out of 70 (32.86%) studies, while an experiment was used in 30 out of 70 (42.86%) studies. Both research methodologies were used in 17 studies. A surprising fact is that only 3 out of 70 studies have a representative sample (as claimed by authors).

3.2.1 Discussion on Research Question 2

The findings presented in the previous subsection may indicate that the data related to textual passwords, collected by these studies, may not be accurate or may not reflect the reality. Conducting laboratory experiments and surveys in which participants are aware that they are being monitored, may lead to biased, less accurate or fake data. For example, participants may create fake passwords in order to protect their real ones or to quickly conclude the survey; or create stronger passwords if they are expecting additional effort. The fact that all 70 studies have retrieved the data either through a questionnaire, an experiment or a combination of both, but only 3 out of 70 studies have a representative sample, may indicate that the retrieved results are neither statistically significant, nor

represent the population, or both. This is very important, since this increases the standard error and we cannot conclude whether the results are reliable or can be generalized. By analyzing the relevant studies for this research question we noticed that users are becoming more aware about the threats related to password security and the importance of creating passwords that are strong and hard-to-guess [26]. Also, passwords became more secure over time regarding password length. The average password length has increased to slightly less than 7-8 characters [25] [27]. Despite that, almost nothing has changed regarding password composition and password management. User-selected passwords are still weak (composed only of lower-case letters, upper-case letters or numbers), users still tend to write their passwords down, so they can easily remember them, still tend to share their passwords with their friends, and they also rarely change their passwords [28] [29].

Because of the ease with which random passwords are recoverable offline, we can expect that, in the future, the security of any information system that is based on passwords will be related to the availability of the material for passwords disclosure and not on how random and strong passwords are [30]. Therefore, stronger passwords may not be always the right solution, as long as the security mechanisms and protocols are well designed (e.g., freezing the account for a time if the wrong password is entered for a certain number of times) [31]. This can be more useful for smaller institutions with hundreds of users where more complex security protocols can be easily applied. Users can also be encouraged to design strong passwords using elements associated with a given service together with a personal factor [32]. We discussed in subsection 3.1.1 that the growth of Web-based services will bring additional challenges for the users, since they will have to memorize even more passwords in the future. This can develop the need for some other usable alternatives to textual passwords in the future [27]. On the other hand, as discussed in subsection 3.1.1, it is very important to prevent users from entering weak passwords into the system, since this can lead to compromising other users' accounts. In order to reach that goal a certain number of new password strength metrics and password meters have been developed [33]–[36] [37]. Nevertheless, due to the widespread use of the World Wide Web and the increased number of Web accounts that a user has to maintain, it is debatable whether these solutions can help users to cope with the large number of accounts and passwords.

3.3 RQ3: What is the relationship between users and their textual passwords? (Are the users still "the weakest link"?)

The issue of users being the weakest link in password security is addressed by 13 out of 90 studies. By analyzing the 13 relevant studies we came to a conclusion that the user behavior is a common issue in the security of information systems. User are often treated by the security departments as a security risk that needs to

be controlled, consequently creating security mechanisms whose usability is rarely investigated [38] [39]. From what has been presented and discussed so far we can argue that users usually are not aware about the security threats and their importance in the security of any information system. The lack of communication between users and organizations (or their security departments) is still present and often leads to the development of useless security mechanisms because they are badly matched to users' capabilities and their tasks [38] [39]. Therefore, if we want more usable security mechanisms for the users, then maybe we should use a "user-centric" approach for designing "usable security" (i.e., human factors should be given priority over technological factors) [40] [41]. Some preliminary studies even imply that "nudges" using multiple psychological effects could serve as important design cues towards making users to perform the intended behavior more easily [17]. On the other hand, Vidyaraman et al. [42] claim that users are nonetheless the enemies of the system and different security policies should be tailored for different types of users. They divide the users to *ignorant* and *non-compliant* users. They argue that the solution to cope with ignorant users is to educate them about security mechanisms, and the solution to cope with non-compliant users is to persuade them to follow the security best practices. Users and their textual passwords will continue to be "the weakest link" in any password system. Security departments should consider implementing a "user-centered" design in order to motivate the users to behave in a secure manner [20] [38]. Users have to be treated as partners in the endeavor to secure organization's systems, not as the enemy within.

Conclusions

This paper presents the results of a systematic review of 90 relevant studies in the area of password use and password security. To the best to our knowledge, this is the first systematic literature review about password security problems. We identified the most common problems related to creating and managing textual passwords. We also outlined the various solutions proposed and used over the years. Because passwords continue to be one of the most common authentication mechanisms, we expected to find a considerably high number of relevant studies in the area of password use and password security. Contrary to what we expected, we found only 90 relevant studies, out of 2201 potential search results. Almost all of them (86), address a problem regarding to creating or managing textual passwords, but only 42 propose a solution for coping with the identified problems, which is a very low number of proposed solutions, given the fact that these problems have been known for almost 35 years. Furthermore, only 17 studies have statistically verified their solutions and used real data in their surveys or experiments, which may raise a suspicion that the retrieved data in the remaining studies is biased or may not reflect the reality. Finally, the most important finding is that only 3 studies have a representative sample, which may indicate that the results of the majority of the studies are not statistically significant and cannot be

generalized to the population. In other words, only the results of 3 studies can be regarded as scientifically acceptable.

Overall, our results demonstrate that not much has changed in password management in almost 35 years. For example, an average user has 6.5 passwords (each of which is shared across 3.9 different websites), resulting in users, to very often, write them down, so they can easily remember them [43]. Lax security behavior involving password reuse, writing down and sharing passwords still exists, along with a lack of, or erroneous knowledge, about what constitutes a secure password. The main weakness in any password system is the end user, because they often choose weak and easily guessed passwords: dictionary words, names, birthdates, etc., only because they are easy to remember. Users' awareness about the consequences of their password choice is not at a high level and a common solution regarding to password problems has not been proposed.

In order to solve many password-related problems, much more research into the matter should be conducted. One way to increase password strength and decrease password "guessability" is devising future security policies, guidelines and education, in such a way, that will take into account human capabilities and strategies for dealing with password overload. A password manager could be used as a way of dealing with password overload. It could greatly reduce the need to remember or write down a password. The problem with common password managers is that they have a number of critical vulnerabilities (e.g. authorization vulnerabilities, user interface vulnerabilities, Web vulnerabilities etc.) [44]. They are also a single point of failure of the system, which is not quite recommended for achieving better security. Another way is to restrict the passwords that are entered in the system, by using a password checker, that filters out weak and easily guessed passwords. Most of the password checkers that we encountered during our systematic literature review, basically check for password length, perform a brute force or a dictionary check of the password, or entropy based checking for presence of non-alphabetic and upper-case characters [45] [46]. Lately, new ways of checking password strength are incorporated into password meters or password checkers [34] [35], that also check the probability of a given password to be chosen by the user. This means that meaningless but pronounceable passwords (which are easier to remember) should take precedence, thus, sacrificing some strength for usability. Understandably, such password checkers should be supported by an appropriate password policy.

We have noted that stricter password policies can pose an additional burden to the users. There is a possibility that this kind of thorough and prudent proactive password checker that forces users to choose complex passwords, can add some additional difficulty for users, when selecting their passwords. Hence, our future research will focus on creating flexible password policies tailored specifically for certain types of users, following the recommendations from [22]. Furthermore, we want to combine flexible password policies with a proactive password checker, based on Markov models. Such a password checker could check the probability of

a given password to be chosen by the user. This approach could help users create strong and easy-to-remember passwords.

Acknowledgement

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0057).

References

[1] S. M. Taiabul Haque, M. Wright, and S. Scielzo, "Hierarchy of users bew ′ passwords: Perceptions, practices and susceptibilities," *Int. J. Hum. Comput. Stud.*, Vol. 72, No. 12, pp. 860-874, Dec. 2014

[2] S.-T. Sun, E. Pospisil, I. Muslukhov, N. Dindar, K. Hawkey, and K. Beznosov, "What Makes Users Refuse Web Single Sign-on?: An Empirical Investigation of OpenID," in *Proceedings of the Seventh Symposium on Usable Privacy and Security*, 2011, pp. 4:1-4:20

[3] K. D. Loch, H. H. Carr, and M. E. Warkentin, "Threats to Information Systems: Today's Reality, Yesterday's Understanding," *MIS Q.*, Vol. 16, No. 2, pp. 173-186, 1992

[4] W. Tzong-Chen and S. Hung-Sung, "Authenticating passwords over an insecure channel," *Comput. Secur.*, Vol. 15, No. 5, pp. 431-439, Jan. 1996

[5] S. Creese, D. Hodges, S. Jamison-Powell, and M. Whitty, "Relationships between Password Choices, Perceptions of Risk and Security Expertise," in *Human Aspects of Information Security, Privacy, and Trust*, Vol. 8030, L. Marinos and I. Askoxylakis, Eds. Springer Berlin Heidelberg, 2013, pp. 80-89

[6] R. Morris and K. Thompson, "Password security: a case history," *Commun. ACM*, Vol. 22, No. 11, pp. 594-597, 1979

[7] M. Zviran and W. J. Haga, "Password security: an empirical study," *J. Manag. Inf. Syst.*, Vol. 15, No. 4, pp. 161-185, 1999

[8] V. Taneski, M. Heričko, and B. Brumen, "Password security — No change in 35 years?," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1360-1365

[9] B. Kitchenham and S. Charters, "Guidelines fro performing Systematic Literature Reviews in Software Engineering," 2007

[10] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences - A Practical Guide*, 1 edition. 2006

[11] S. Fahl, M. Harbach, Y. Acar, and M. Smith, "On the Ecological Validity of a Password Study," in *Proceedings of the Ninth Symposium on Usable*

Privacy and Security, 2013, pp. 13:1-13:13

[12] M. A. Schmuckler, "What Is Ecological Validity? A Dimensional Analysis," *Infancy*, Vol. 2, No. 4, pp. 419-436, 2001

[13] Chitika Insights, "The value of Google result positioning." p. 10, 2013

[14] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, Vol. 20, No. 1, pp. 37-46, 1960

[15] J. L. Fleiss, B. Levin, and M. C. Paik, *The Measurement of Interrater Agreement*. John Wiley & Sons, Inc., 2004

[16] D. Feldmeier and P. Karn, "UNIX Password Security - Ten Years Later," *Adv. Cryptol. — CRYPTO' 89 Proc. SE - 6*, Vol. 435, No. November 1988, pp. 44-63, 1990

[17] S. Kankane, C. DiRusso, and C. Buckley, "Can We Nudge Users Toward Better Password Management?: An Initial Study," *Ext. Abstr. 2018 CHI Conf. Hum. Factors Comput. Syst.*, p. LBW593, 2018

[18] V. Taneski, M. Heričko, and B. Brumen, "Impact of security education on password change," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 2015, pp. 1350-1355

[19] P. Cisar and S. M. Cisar, "Password - A form of authentication," *5th Int. Symp. Intell. Syst. Informatics, SISY 2007*, pp. 29-32, 2007

[20] A. Adams, M. A. Sasse, and P. Lunt, "Making Passwords Secure and Usable," in *People and Computers XII*, 1997, pp. 1-19

[21] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas, "Generating and remembering passwords," *Appl. Cogn. Psychol.*, Vol. 18, No. 6, pp. 641-651, 2004

[22] P. G. Inglesant and M. A. Sasse, "The true cost of unusable password policies," *Proc. 28th Int. Conf. Hum. factors Comput. Syst. - CHI '10*, p. 383, 2010

[23] G. Notoatmodjo and C. Thomborson, "Passwords and perceptions," *Conf. Res. Pract. Inf. Technol. Ser.*, Vol. 98, No. Aisc, pp. 71-78, 2009

[24] B. Ives, K. R. Walsh, and H. Schneider, "The domino effect of password reuse," *Commun. ACM*, Vol. 47, No. 4, pp. 75-78, 2004

[25] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, "Does my password go up to eleven? The impact of password meters on password selection.," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 2379-2388, 2013

[26] A. Moallem, "Did You Forget Your Password?," in *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, 2011, pp.

29-39

[27] E. Von Zezschwitz, A. De Luca, and H. Hussmann, "Survival of the shortest: A retrospective analysis of influencing factors on password composition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, Vol. 8119 LNCS, No. PART 3, pp. 460-467, 2013

[28] S. Riley, "Password security: what users know and what they actually do," *Usability News*, Vol. 8, No. 1, pp. 2833-2836, 2006

[29] L. Tam, M. Glassman, and M. Vandenwauver, "The psychology of password management: a tradeoff between security and convenience," *Behav. {&} Inf. Technol.*, Vol. 29, No. 3, pp. 233-244, 2010

[30] L. St. Clair *et al.*, "Password Exhaustion: Predicting the End of Password Usefulness," *Inf. Syst. Secur. Lect. Notes Comput. Sci.*, pp. 37-55, 2006

[31] J. Yan, B. Alan, R. Anderson, and A. Grant, "Password memorability and security: Empirical results," *IEEE Secur. Priv.*, Vol. 2, No. 5, pp. 25-31, 2004

[32] K. Helkala and N. K. Svendsen, "The security and memorability of passwords generated by using an association element and a personal factor," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, Vol. 7161 LNCS, pp. 114-130, 2012

[33] M. Aliasgari, N. Sabol, and A. Sharma, "Sesame: A secure and convenient mobile solution for passwords," in *2015 1st Conference on Mobile and Secure Services, MOBISECSERV 2015*, 2015, pp. 1-5

[34] Y. Guo and Z. Zhang, "LPSE: Lightweight password-strength estimation for password meters," *Comput. Secur.*, Vol. 73, pp. 507-518, 2018

[35] D. Wang, D. He, H. Cheng, and P. Wang, "FuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars," *Proc. - 46th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Networks, DSN 2016*, pp. 595-606, 2016

[36] M. Dell'Amico and M. Filippone, "Monte Carlo strength evaluation: Fast and reliable password checking," *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, pp. 158-169, 2015

[37] P. Y. Lee and Y.-Y. Choong, "Human Generated Passwords -- The Impacts of Password Requirements and Presentation Styles," in *Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings*, T. Tryfonas and I. Askoxylakis, Eds. Cham: Springer International Publishing, 2015, pp. 83-94

[38] A. Adams and M. A. Sasse, "Users Are Not the Enemy," *Commun. ACM*, Vol. 42, No. 12, pp. 40-46, Dec. 1999

[39] M. A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'weakest link' - A human/computer interaction approach to usable and effective security," *BT Technol. J.*, Vol. 19, No. 3, pp. 122-131, 2001

[40] C. A. Fidas, A. G. Voyiatzis, and N. M. Avouris, "When security meets usability: A user-centric approach on a crossroads priority problem," *Proc. - 14th Panhellenic Conf. Informatics, PCI 2010*, pp. 112-117, 2010

[41] M. Adeka, S. Shepherd, and R. Abd-Alhameed, "Resolving the password security purgatory in the contexts of technology, security and human factors," *Int. Conf. Comput. Appl. Technol. ICCAT 2013*, Vol. 2013, 2013

[42] S. Vidyaraman, M. Chandrasekaran, and S. Upadhyaya, "Position: The User is the Enemy," in *Proc. of the 2007 Workshop on New Security Paradigms*, 2008, pp. 75-80

[43] D. Florencio and C. Herley, "A large-scale study of web password habits," *Proc. 16th Int. Conf. World Wide Web - WWW '07*, p. 657, 2007

[44] Z. Li, W. He, D. Akhawe, and D. Song, "The Emperor's New Password Manager: Security Analysis of Web-based Password Managers.," in *USENIX Security Symposium*, 2014, pp. 465-479

[45] M. Bishop and D. V. Klein, "Improving system security via proactive password checking," *Comput. Secur.*, Vol. 14, No. 3, pp. 233-249, 1995

[46] J. J. Yan, "A Note on Proactive Password Checking," *Proc. 2001 New Secur. Paradig. Work.*, pp. 127-135, 2001