Leveraging Semantic Transformation to Investigate Password Habits and Their Causes

Ameya Hanamsagar
anahamsa@usc.edu
USC
Los Angeles, USA

Simon S. Woo
simon.woo@sunykorea.ac.kr
SUNY Korea
Incheon, Republic of Korea

Chris Kanich
ckanich@uic.edu
UIC
Chicago, USA

Jelena Mirkovic
sunshine@isi.edu
USC/ISI
Marina del Rey, USA

ABSTRACT

It is no secret that users have difficulty choosing and remembering strong passwords, especially when asked to choose different passwords across different accounts. While research has shed light on password weaknesses and reuse, less is known about user motivations for following bad password practices. Understanding these motivations can help us design better interventions that work with the habits of users and not against them.

We present a comprehensive user study in which we both collect and analyze users' real passwords and the reasoning behind their password habits. This enables us to contrast the users' actual behaviors with their intentions. We find that user intent often mismatches practice, and that this, coupled with some misconceptions and convenience, fosters bad password habits. Our work is the first to show the discrepancy between user intent and practice when creating passwords, and to investigate how users trade off security for memorability.

ACM Classification Keywords

K.6.5. Security and Protection: Authentication

Author Keywords

passwords; bounded rationality; risk perception

INTRODUCTION

We know that current advice for password creation—to create a strong, unique password for every online account—is unreasonable. Users have many online accounts, and cannot remember many different, unrelated, complex passwords. But we do not know how users reason about this trade-off between memorability and security, or if they engage in their bad password habits knowingly or unknowingly. These are the questions that we seek to address in this paper.

There is ample prior research on how people choose and reuse passwords. Some works analyze real passwords through leaked datasets [4, 27, 14, 23, 1] or browser plugins [5, 26] to learn about trends related to weak passwords and password reuse. These methodologies, however, do not allow for user input about *why* they engage in bad password habits. On the other hand, researchers have used lab studies with fake servers [22, 18] to investigate user perceptions about passwords and risk. This approach may paint an incorrect picture if users are unaware of their actual password habits, or if users interact differently with fake servers than with real ones.

To the best of our knowledge, simultaneously studying both user actions regarding real password usage and user intent has not been previously attempted. Performing an integrated investigation of these components of password choice and usage has the potential to create a better understanding of the motivations behind bad password habits, and better interventions that are aligned with those motivations.

Open questions that we seek to address are:

(1) How prevalent are bad password habits—such as weak or reused passwords—in our participant population of 50 college students? While others have studied similar questions on larger populations, their work covered only frequently used accounts [5, 26], while we also study rarely used ones. Understanding how users manage their entire portfolio of accounts can help us design better interventions to minimize password reuse.

(2) How do a user's intent, risk perception, a site's importance to a user and a site's password policy influence password strength? Answers to these questions can help improve password policies and inform user education.

(3) How well do users understand their password practices? Does their practice align with their intent? Answers to these questions can help create tools that improve user understanding and help them make more informed choices.

Studying these open questions is hard, as it requires access to real passwords, along with interviews with their owners, but it must protect users' privacy.

Our **first contribution** lies in our novel study methodology that enables us to: (1) collect information about many accounts belonging to a user, including rarely-used ones, (2) collect information about *semantic structure* of a user's real passwords in a privacy-safe manner that allows us to detect similar passwords and understand password composition. In our study we first scan each participant's Gmail account, looking for account creation or password reset emails. From this initial pool of accounts, we ask the participant to select 12 to log

CHI 2018 CHI Conference on Human Factors in Computing Systems, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: 10.1145/3173574.3174144

into during the study. As they do, we extract each password's semantic structure, length and strength automatically (e.g., name+number, 12 chars, 10,000 guesses needed for cracking), and then we transform the original password using a consistent but irreversible mapping of semantic segments (e.g., Joe123 transforms into Maya422). We then store the password's features and the transformed password, and discard the original. These actions uniquely enable us to study password structure and password reuse, while keeping study participants safe. We supplement these objective findings with user surveys, giving us an insight into user reasoning about passwords, and enabling us to compare observed behavior with user-narrated intent. We describe our study design in detail in Section 4. The study was reviewed and approved by our IRB.

Our **second contribution** lies in our findings, which we obtain by applying our methodology to a cohort of 50 participants. Although our participant pool is small, we make several statistically significant findings. First, as expected, we find that bad password habits abound. 12% of accounts were vulnerable to online password-guessing attacks and 90% were vulnerable to offline attacks. Further, all users in our study reused passwords, 98% verbatim and 2% with slight versioning. Second, we find that bounded rationality, misconceptions about risk and user desire for memorability are the main causes of bad password habits. Bounded rationality becomes evident when we compare a user's narration of her password habits with her behavior observed in our study: all users underestimate the number of accounts they have, they narrate more rational reuse strategies than they exhibit, and they narrate different password-composition strategies than they employ. In addition to bounded rationality, misconceptions about risk and preference for memorability over security contribute to bad password habits. 18% of users have misconceptions about password-reuse attacks, and 76% assume password-guessing attacks would be of an online nature ($< 10^6$ guesses per second). Further, 28% of our participants knowingly created weak passwords, and 44% knowingly engaged in reuse because they valued memorability over security.

Some of our findings update prior results, which is our **third** contribution: (1) We find a median of 80 accounts per user, updating the Florencio et al. [5] estimate of 25 from 2006; (2) We find no significant correlation between password strength and a user's risk perception, as observed by Creese et al. in [3]; (3) We find that weak and strong passwords are equally reused, while Wash et al.[26] found that strong passwords are reused more often.

BACKGROUND AND RELATED WORK

In this section we present prior research on password creation, reuse patterns, user behaviors, and risk perceptions.

Password Reuse

Florencio et al. [5] conducted a large-scale password reuse study in 2006 by instrumenting Microsoft Windows Live Toolbar. The study included half a million users monitored over a three-month period. They found that each user had about 25 accounts and 6.5 passwords, each shared across 3.9 sites. Our study provides an updated estimate of 80 online accounts per user and potentially 25 accounts per password, but it is conducted over a much smaller and less diverse user sample.

Wash et al. [26] examined the types of passwords that are more frequently reused. They developed a Web browser plugin to collect user passwords, and conducted a user study with 134 participants. They found that strong and more frequently used passwords were reused more often. We could not confirm this on our dataset. Instead, in our study weak and strong passwords were reused comparably often. One possible reason for this discrepancy may lie in the types of accounts we study. Wash et al. study analyzes passwords only for those accounts that a user accesses frequently, while we also analyze passwords for rarely-accessed accounts.

Similarly, Pearman et al. [13] recently studied 154 participants' password habits by instrumenting their browsers to record both password inputs and other computer behaviors that may reveal a participant's security habits. This population is larger and more diverse than ours. The main differences between our work and Pearman et al.'s are: (1) In addition to the participants' passwords, we also collect their subjective responses that reveal attitudes about risk and security and their reasoning about passwords, while Pearman et al. attempt to infer these attitudes from security habits they record. (2) Our password collection strategy differs; We use semantic transformation, while Pearman et al. use hashing of substrings. We thus can study password structure (e.g., dictionary word vs. a random set of characters), while Pearman et al cannot. (3) We study both frequently and rarely used accounts, while Pearman et al. only study accounts accessed during the study's duration. Interestingly, even though our methodology differs, several of our findings match: low influence of password managers on password strength, high partial reuse (versioning) of passwords and higher reuse frequency of weaker passwords.

In a lab study, Ur et al. [22] examined password behaviors of 49 users, creating accounts at three fictitious servers. They found that password reuse is common, and that users are not good at making value decisions about their online accounts. Due to the fictitious nature of accounts, Ur et al. were able to collect user passwords and examine them for versioning, whereas our study can do this. Ur et al. found that users had serious misconceptions about how to compose strong passwords. We find that users generally understand how to *compose* strong passwords but have misconceptions about password length. The size and composition of our participant population is comparable to that used in the Ur et al. study.

Das et al. [4] examined how people reuse passwords using leaked password datasets. This study is limited, because very few users appeared in more than two datasets, while our users accessed many more online accounts. Das et al. estimated that 43-51% of users reuse passwords, while we find that 98% of our participants do so. Shay et al. [16] show that, when asked for a new password, more than half of participants modify an old password or reuse it verbatim. Similarly, E. von Zezschwitz et al. [24] found through user interviews that 45% of users reuse passwords verbatim, while 70% version them. We find that 98% reuse passwords verbatim and the remaining 2% version them.

Users' Perceptions About Passwords

Creese et al. [3] examined the relationship between perceptions of risk and password choice. They found that users whose risk assessment differs from the experts' assessment on six chosen questions (out of 20) tended to use passwords with a smaller keyspace. In our work, we repeat their approach, but find no such correlation. Our and their population sizes and diversity match (both studies use 50 college students and/or staff). Ur et al. [21] investigated the relationship between users' perceptions of the strength of specific passwords and their actual strength, and found that users had serious misconceptions. Also, they showed that users do not really understand how password attacks work. We confirm this second finding.

Causes of Weak Passwords

Redmiles et al. [15] investigated reasons for selective adoption of broad digital-security advice by users; among this was inconvenience. We also find that inconvenience, or desire for memorability, plays an important role in password reuse. Coventry et al. [2] argued that the general public does not generally follow best practice because there is no clarity about required actions. Our results support this argument, especially with regard to password length.

DESIGN

In this section we describe our research goals and privacy protection goals, and how they shaped our user study design.

Research Goals. We wanted to study how users design passwords for different sites, and how and why they reuse their passwords. For this we needed: (1) information about real passwords on real sites, (2) ability to detect the passwords of a given user that are similar but not the same, and (3) ability to discuss specific password practices and choices with the user to understand causes of bad password habits. One way to collect necessary data would have been to ask each user to list all their accounts and passwords. However, users may forget where they have created accounts, and they may be reluctant to share their real passwords. For these reasons we designed an automated way to extract necessary information with minimal user or researcher involvement.

Privacy Protection Goals. Asking users to give us their passwords directly would be risky to their privacy. Thus we formulated the following privacy protection goals: (1) no storing of any identifying information, (2) no human access to users' real passwords, (3) no intentional (by us) or accidental (by our browsers) storing of real passwords.

To satisfy both our research and our privacy protection goals, we designed our study as shown in Figure 1. This study was reviewed and approved by our Institutional Review Board. The study was performed in our lab on our laptop, in a Chrome incognito window. We opened the window for each study participant and closed it after the participant completed the study. This ensured that no login credentials or sessions/cookies remained stored in the browser.

In our study we asked participants, in addition to other actions, to answer questions on six surveys shown in Table 1. Questions for the "risk perception" survey were taken from Creese et al. [3]. Others were designed by us and refined on a five-person volunteer group (not included in participant pool) for clarity.

Pre-study surveys (Step 1). First, we asked a participant to fill "statistics" and "password strategy" surveys, designed to collect their subjective assessment of their password behavior, and the "risk perception" survey from Creese et al. [3], which measures their general attitude towards risk.

Compiling a list of websites (Step 2). Next, we scanned the participant's GMail account, using the CloudSweeper tool [17] to compile a list of sites where they may have an account.

Collecting participant login information (Steps 3-7). Next, we showed the list of sites to the participant. The participant could delete sites where they did not have an account, or sites that were sensitive. The participant could also add to the list other sites where they had an account. We next asked the participant to mark each remaining site as important to them or not, and as the one they frequently visited or not. We provided no guidance to participants on how to assign these tags, but we find that participants generally marked sites as important if they cared about security of the content at these sites (see Results section). Finally, we asked the participant to choose at least four important and frequently visited sites, four important but not frequently visited sites, and four non-important sites (regardless of the visit frequency) to log onto. The participant could choose to visit more sites. We chose this blend of sites because prior research [26, 22, 7] found that a site's importance and frequency of use may affect password strength.

We developed a Google Chrome extension to capture the password from each login attempt, and collected character-length information for the original password. We also noted whether there was a capitalization or mangling in the original password, and if it was there at the beginning, in the middle, or at the end. We did not store more detailed data about positions of such changes because we believed that this would unduly increase privacy risk, while not bringing much research benefit. Finally, we fed the original password into our local installation of the `zxcvbn` [28] strength meter and retrieved the resulting strength. We then transformed the original password into its semantic equivalent. Such transformed passwords do not expose any information about the original password beyond its semantic structure, e.g., noun+verb+number. We then stored the transformed password and deleted the original.

Post-study surveys and discussion (Steps 8 and 9). After the logins, we asked each participant to respond to questions from the "mental model of attackers" survey (Table 1), which measures if a participant knows how password-guessing attacks occur. After that, we utilized the "impact reasoning" survey. For each site where the participant attempted to log on, we asked them to rate on a Likert 1–5 scale how affected they would be if a stranger or a friend impersonated them on that site, or if their data from that site were made public. This provided the second source of information about the site's importance to the participant (the first one being the important/non-important

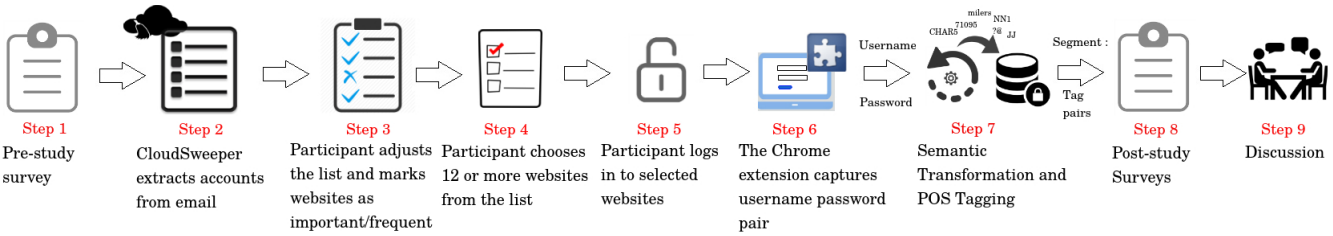


Figure 1. User study flow.

Type	Acronym	Question	Type	Acronym	Question
Statistics, pre-study					
Num	NUMACC	How many online accounts do you have?	Num	NUMPAS	How many different passwords do you have?
Num	NUMEML	How many email accounts do you have?	Y/N	PRIMAR	Is this Gmail account your primary email account?
Password strategy, pre-study					
Free	STRATG	How do you choose your passwords? Is this good?	Lik	NOCHNG	I do not change my passwords, unless I have to.
Free	MENAGR	Do you save passwords in browser/manager?	Lik	SPECCH	I always include special characters in my passwords.
Lik	COMPLX	I create passwords that go beyond min requirements	Lik	NORUSE	I use different passwords for different accounts.
Risk perception, pre-study					
Lik	ONBANK	Online banking is risky.	Lik	GEOTAG	Geotagging content is risky.
Lik	CCAMAZ	Using credit card on Amazon is risky.	Lik	UNKEML	Opening an email from an unknown sender is risky.
Lik	CCEML	Sending credit card over email is risky.	Lik	CCBAR	Leaving a credit card at a bar for the tab is risky.
Lik	PPLEBY	Using eBay and Paypal is risky.	Lik	CLICK	Clicking on a link in an email from a stranger is risky.
Lik	PUBWIF	Using WiFi in a coffee shop is risky.	Lik	DATE	Using online dating services is risky.
Lik	PIRTSW	Using pirated software is risky.	Lik	FLY	Flying from the UK to the US is risky.
Lik	NLCCAR	Leaving your car unlocked on parking is risky.	Lik	CBRCAF	Using a cybercafe is risky.
Lik	OPENPR	Using OSN with open privacy settings is risky.	Lik	OSUPD	Not updating your OS is risky.
Lik	CLSEPR	Using OSN with closed privacy settings is risky.	Lik	BRUPD	Not updating your web-browser is risky.
Lik	PHOTSH	Using photo-sharing sites is risky.	Lik	APPUPD	Not updating other applications is risky.
Impact reasoning, post-study					
Type	**Acronym**	**Question**			
Lik	STRANG	If a stranger could impersonate me on this site this would bring me personal or financial harm.			
Lik	FRIEND	If a friend/family could impersonate me on this site this would bring me personal or financial harm.			
Lik	PUBDAT	If the data from my account became public this would bring me personal or financial harm.			
Password strategy reasoning, post-study					
Lik	WHYSAM	What is the reason behind using the same password?			
Lik	REUSE	Are you concerned that an attacker may obtain your password on site A and use it to access site B?			
Lik	WHYSIM	If the passwords are not same, but similar, ask why did the user change the password?			
Mental model of attackers, post-study					
Lik	NUMGSS	How many guesses could an attacker make in 1 minute?			
Lik	HOWGSS	How would an attacker come up with guesses?			
Lik	OFFLIN	How might an attacker guess a password with an unlimited number of trials?			

Table 1. Survey questions we used (types are *Num:* numeric, *Free:* narrative, *Lik:* Likert or *Y/N:* yes/no). Due to space, some questions are paraphrased.

tag). Finally, we used the "password strategy reasoning" to collect the participant's reasons for password reuse.

METHODOLOGY

We now provide more details about our detection of user accounts from emails, recording of the login attempts, and semantic transformation of passwords.

Detecting Sites With Participant Accounts

The Cloudsweeper tool [17] extracts the cleartext passwords in emails by connecting to Gmail's IMAP using OAuth tokens. We use it to identify emails that match certain common patterns in new account registration and password reset e-mails. These patterns include strings: "welcome to," " reset password," "thank you/thanks for registering/creating/signing," and "your *site name* account has been created." The identified e-mails are further processed to reduce false positives by filtering out e-mails that have more than one recipient, or those where URL and welcome text in the body do not match the domain name of the sender.

We measured the accuracy of our automated account extraction tool on a pool of account creation e-mails collected from a catch-all mail archive of the `geeemail.com` domain. We acquired this domain in 2010 and set it up with a mail server. It catches cases when users mistype their `gmail.com` address as our domain. Most emails arriving at this server are for account registration purposes, which made it easy to establish ground truth. We manually identified account subscriptions to 176 unique domains out of 2,968 e-mails on the server. Out of these, our tool successfully identified 91, i.e., recall was 52%. We also made 5 false identifications, where an email was not for the account registration but for promotion purposes; thus our precision was 94%.

Password Extraction

We developed a Google Chrome extension to extract a participant's password from login attempts, while preserving privacy. The extension is enabled manually during each study instance. During each login attempt, on each key press in the password field, we capture the participant's input. The extension detects login events using the JavaScript window object's `onbeforeunload` event [25], and this triggers sending of the last captured input for semantic transformation. All used passwords were successfully captured by our extension.

The extension is also responsible for recording successful logins if the page following the login attempt does not have a password input field. This approach was reliable for most websites, but a few necessitated a manual recording.

Semantic Transformation

The extracted password is sent for semantic transformation to an application running on the same laptop as the Chrome browser. We illustrate this process in Figure 2. The application was our modified version of a tool developed by Veras et al. [23] for semantic segmentation and part-of-speech (POS) tagging of strings. We first undo any mangling before feeding the password into the semantic segmentation and tagging tool. This is done by reversing the KoreLogic's L33t password cracking rules [11].

The semantic segmentation and tagging tool transforms an input string into the list of segments and their (POS) tags. The tool uses POS tags from the CLAWS7 tagset [20]. For example, for a string "applerun" the string would return segments (apple)(run) and tags (NN1)(VV0) indicating a singular noun and a base form of a verb. Some segments may be returned untagged, such as random sequences of letters, numbers and special characters.

Next, we transform each segment into a different segment in the same semantic category to preserve privacy for the participants. The goal is to achieve consistent but irreversible transformation of segments. For example, if a participant had two passwords "john352@" and "john222," the semantic segmentation and tagging would result in POS tags indicating (proper-name)(3-digit-number)(special-char) and (proper-name)(3-digit-number). We would transform the proper name "john" into another proper name consistently, so that the resulting two passwords continue to have one common segment. We also would transform the 3-digit numbers 352 and 222 into different 3-digit numbers and the special character "@" into a different special character. For example, "john352@" and "john222" could be transformed into "bob475!" and "bob687."

We treat POS-tagged and untagged segments differently for the transformation. We achieve the consistent and irreversible mapping for POS-tagged segments by employing a keyed one-way hash function with a random per-participant key, and a dictionary of words for each POS tag. We used the same dictionary and Python pickle files as those used in [23]. For each participant, we generate a random key from the range $[2, 2^{32}]$. This key is appended to each segment and the resultant string is hashed using a one-way hash algorithm—SHA512. We append the key to enlarge the space of possible inputs to the hash function. This is especially important for segments that may have few unique inputs, such as locative nouns. As the next step, we extract only the digits from the hash function's output and calculate the modulo of the resulting number and the size of the dictionary for the given POS tag. We use this result as an index into the dictionary to find the word which will replace the original segment. Consistency is achieved because the same segments result in the same input to the hash. Privacy is achieved because of our use of the per-participant random key. This key remains in memory during the participant's engagement in the study, and is deleted when we close the application. Because we cannot reproduce the input to a one-way hash function without this key, neither we nor anyone else can reverse the mapping.

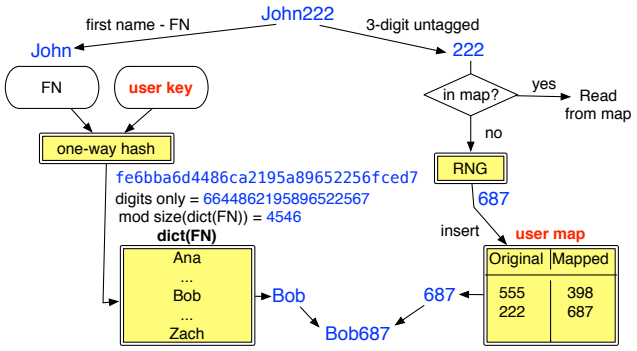


Figure 2. Semantic transformation example: John222 is first semantically tagged as a first name and 3-digit number combination. John is then transformed into another name (Bob) using per-participant random key and one-way hash. The 3-digit number 222 is transformed into another random 3-digit number 687, which is stored in the per-participant map.

For some POS tags, our dictionary has fewer than 100 words. We add these words into their parent's category. For example, words belonging to NNL1 and NNL2 (locative nouns) were added to the category NN (common noun).

Untagged segments mainly include random alphanumeric or special characters, but may also include words in a foreign language or misspelled words. For such segments, we generate random sequences of characters in the same category as the original ones (alphabetic, numeric or special), and achieve consistency by storing this mapping in memory. Irreversibility is guaranteed by the randomness of the mapping, and because we delete this mapping when the participant exits the study.

RESULTS

We first discuss limitations of our study, and how we process results. We then present our findings on how users choose and reuse passwords, and their reasoning about password habits. We preface those results that came from a participant's narration with "subjective" and those that came from measuring actual passwords with "objective."

Limitations and External Validity

Our study required a significant amount of interaction. The research staff had to be present during the study to explain the process to each participant and to interview participants for

some survey questions. Each participant also spent 30–45 minutes in the study. Hence, we ended up with a small participant pool of only 50. While small, this pool size is comparable to other recent lab-based password studies (25 participants in [15], 49 in [22]). We present several statistically significant findings here, which build a new understanding of how our participant population approaches the problem of maintaining passwords across different classes of online services. Further research is needed to test whether these observations generalize to larger populations.

Because our research was self-funded, we could not compensate participants with a large payment. We thus recruited participants from our own university, which minimized their cost, and paid them $10. Forty-eight of our 50 participants were local students (54% males and 46% females), and 23 majored in a technical field.

Another limitation of our study is that we asked participants to log into at least 12 accounts, and not all of the accounts they had because we had to limit the burden on participants, which was already very high (30-45 minutes) for the compensation we offered. Unfortunately, this prevented us from studying passwords for all the accounts of any given participant. Allowing participants to select accounts to log on to may also bias our study towards weaker passwords that they feel comfortable revealing, and may bias it towards those passwords that they can remember. While we cannot measure this impact directly, we note that sites, which are likely to have strong passwords, such as financial and school/work were well represented (33% of all accounts) in our study. Further, login success at all sites was around 50%, which means that our participants have selected many sites whose passwords they could not recall.

Statistics

In this section we present general statistics on our study population, their accounts and login attempts.

To analyze statistical significance of our findings across different participant groups, and factors, we used multiple regression on continuous variables, with participant ID being one of the independent variables. We performed χ^2 (Pearson's Chi-squared test with Yates' continuity correction), on categorical data with $p = 0.05$. To measure correlation between a participant's survey response, and their password reuse, we use the Spearman's rank correlation.

Account composition and login success. Participants attempted to log into 621 accounts in our study. We show account breakdown across important/not important and frequently/infrequently used categories in Table 2 along with the login success rate in parentheses. 392 of accounts were marked as important by participants, 241 were marked as frequently used, and 212 were in both of these categories. Additionally 29 accounts were marked as frequently used but not important, and 180 were marked as important but not frequently used.

Subjective. We provided no guidance to participants about what "important" means; therefore they may have flagged an account as important based on their preference for content, rather than security considerations. We investigate this by comparing the participant responses to STRANG question

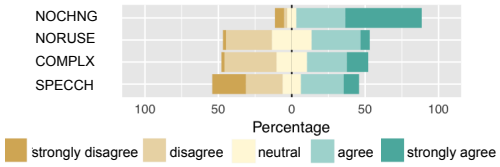


Figure 3. Participant ratings for password strategy questions.

in our impact reasoning survey, which asks how affected the participant would be if a stranger could access their given account. There is a significant difference ($p = 9.6213 \times 10^{-14}$, R^2=0.104, Cohen's effect size= 0.1161 – small) between participants' ratings of important versus non-important accounts in replies to STRANG, with higher ratings assigned to important ones. We thus conclude that participants tag an account as "important" if they would be negatively affected by a breach of that account.

Objective. Participants successfully logged into 470 accounts. The success rate for important accounts (79%) was higher than for non-important accounts (59%), with χ^2_1= 26.734, $p = 2.334 \times 10^{-7}$. The success rate for frequent accounts (85%) was higher than that for non-frequent accounts (63%), with χ^2_1= 30.993, $p = 2.589 \times 10^{-8}$.

Account	Frequent	Not frequent	Total
Important	212 (86%)	180 (71%)	392 (79%)
Not important	29 (76%)	200 (57%)	229 (59%)
Total	241 (85%)	380 (63%)	621 (72%)

Table 2. Account types in our study and login success rates.

Many accounts: subjective+objective. Out of 50 participants, 29 reported (question PRIMAR, statistics survey) that the GMail account they used in the study was their primary e-mail account. We now compare the subjective measure of the number of online accounts (question NUMACC, statistics survey, blue line in Figure 5(a)) with our objective estimate, based on GMail account scans (black and red lines in Figure 5(a)), updated to correct for our tool's underestimate. The median subjective estimate is 15 accounts, but median objective estimate is 80 (primary GMail account) and 30 (non-primary GMail account). The estimate of 80 is much higher than 25 accounts found by Florencio et al. [5] in 2006. This is expected, as the number of online services has increased considerably since then.

A few passwords: subjective+objective. When asked how many distinct passwords they had (question NUMPAS, statistics survey), participants estimated between 3 and 30 passwords, with an average of 6.6 and the median 5. Because we only asked subjects to log in to 12 different sites, we cannot calculate how many passwords they actually have. But we do not expect that human memory limitations have changed since 2006, when Florencio et al. [5] found 6.5 passwords per user. This matches our findings.

Reuse

We now investigate how often users reuse their passwords. There were 160 unique passwords in 446 successful logins.

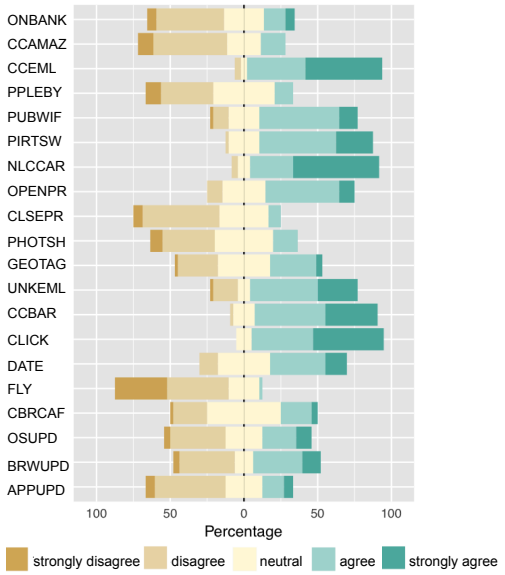


Figure 4. Participant ratings for risk perception questions.

Out of these, 72 or 45% were used only at one site, while the rest were reused.

User estimate of reuse is large: subjective. In responses to our statistics survey (NUMACC and NUMPAS), 98% of participants stated that they had fewer passwords than accounts. Based on these subjective measures, we believed that participants shared a password among 4.7 accounts on average. However, we found that participants underestimate the number of accounts they have, by a factor of 5.3. If the subjective estimate of the number of passwords were correct, this puts the actual password reuse close to 25 accounts per password. This is higher than findings in [5, 26], likely because our study also examines rarely used accounts, which are not investigated in other studies.

Actual reuse is large and indiscriminate: objective. Table 3 summarizes our findings about reuse, counting a password as versioned if it shared at least one segment, three or more characters long, with another password by the same participant. We find that reuse is rampant! 98% of participants reuse their passwords among accounts, and the remaining 2% have similar passwords between accounts. Further, 84% of participants reuse a password from an important site at a non-important site, and an additional 6% have similar passwords between important and non-important accounts. Also, 98% (100% including similar passwords) of participants reuse their important-site password at another important site, but only 64% (72% including similar passwords) reuse their non-important-site passwords at another non-important site. This data indicates that many participants create a limited number of passwords and reuse them across different categories of accounts indiscriminately.

Users are not aware of their reuse patterns: subjective+objective. We do not compare subjective versus objective reuse of passwords for each participant because our study design limits our ability to observe passwords for all participant accounts. Instead we seek to understand how a participant's intended reuse strategy matches the actual one. We examine responses to NORUSE "I use different passwords for different accounts that I have." Figure 3 shows that participants were quite divided on this question. We use the Spearman's rank correlation to measure the correlation between a participant's response, and the subjective and objective estimate of reuse. There is a statistically significant negative correlation between the response to NORUSE and a subjective estimate of reuse ($r = -0.4091$ and $p = 0.0032$). Thus, participants who self-report stronger intentions to use different passwords also estimate a lower incidence of password reuse.

Type of reuse	Verbat.	Verbat. or similar
All accounts	98%	100%
Important/Non-imp	84%	90%
Important/Important	98%	100%
Non-imp/Non-imp	64%	72%

Table 3. Percentage of participants that reuse in a given way.

On the other hand, we did not find a significant correlation between a participant's response to NORUSE and an objective estimate of reuse ($r = -0.0046, p = 0.9749$), whereas Wash et al. [26], reported significant correlation in a similar setting. We attribute this difference to differences in accounts accessed by our two studies—Wash et al. observe only frequently used accounts, while we also observe rarely used accounts.

Users reuse both strong and weak passwords: objective. Similar to Wash et al. in [26] we measure correlation of password strength versus number of accounts where this password is reused verbatim. We find significant *negative* correlation between these measures (Spearman's rank correlation, $r = -0.2$, $p = 0.01089$). This disagrees with findings in [26] where they found that stronger passwords are reused more often ($r = 0.063$, $p = 0.007$), but agrees with findings of Pearman et al [13]. The difference in findings between us and Wash et al. may result from different strength measures – we use statistical guessing measure from `zxcvbn` strength meter while Wash et al. use a weaker measure of password entropy. However, when we repeat our test using entropy we still find no significant correlation ($r = -0.16$, $p = 0.08$). Another possible reason for the difference may lie in the ability of different studies to uncover password habits for rarely used accounts. Wash et al. follow 134 participants for 42 days, and Pearman et al. follow 154 participants for 147 days. Longer duration makes it more likely to observe rarely used accounts, and thus observe reuse patterns at more samples. Our study also observes both rarely and frequently used accounts.

Wash et al. [26] also found that participants heavily reused their university password. In our study only 13 out of 50 participants chose to log into their university account, and successfully completed that login. Nine out of these 13 reused their university password on the average at 3.5 other accounts. This result is consistent with that in [26], which finds reuse at 3.2 additional accounts.

Unintentional reuse patterns: subjective+objective. When we detected reuse of the same password verbatim, we asked the participant about its cause (WHYSAM, password strategy reasoning survey). 5 out of 49 participants (10%) said they share passwords only among accounts they do not care about. We investigated this claim by examining important and non-important site passwords for these participants. In all five cases these participants shared a password between at least two important sites, and they also shared a password between an important and a non-important site. This is contrary to their narrated intent.

Simple password versioning: objective. We compare pairs of passwords by the same participant to detect password versioning—slight changes in the password structure that may be easily guessed by an attacker. We say that two passwords are *similar* if they have at least one common segment (3 or more characters long) and at least one different segment. 34 out of our 50 participants have at least one pair of similar passwords. Overall, we found 61 similar pairs. We then examined the changes between passwords and detected eight change patterns. 62% of passwords are versioned in a very simple manner, by changing or adding a number, a special character, one word, or by introducing capitalization or mangling. 38% of passwords experience more complex transformations that combine two or three simple techniques.

Users do not take advantage of password managers: subjective+objective. We asked participants if they allow browsers or password managers to save their passwords (MANAGR, password strategy survey). 60% of participants said they allow this always, 24% said they do not allow it, and 16% allow it sometimes. To examine whether a password manager (browser-based or stand-alone) helps participants make better password choices, we compare the password strength and reuse between always-use, and never-use groups, using multiple regression, with password manager use and participant ID being the independent variables. We did not find a significant difference in strength (p=0.986 $\gg$ 0.05) or reuse (p=0.949 $\gg$ 0.05). We further did not find a significant difference in login success between these three participant groups (p=0.796 $\gg$ 0.05), but we would expect to see such differences if participants relied on password managers more than on their memory. Investigating how much of this failure to take advantage of managers is intentional versus simply a force of habit is an interesting direction for future research.

Password Strength

In this section we present our findings about password strength in our participant population. We obtain the strength estimate from `zxcvbn` as the expected number of guesses before success. Thus our strength estimates the number of guesses in a specific statistical guessing attack.

Weak passwords: objective. Figure 5(b) shows the distribution of password strength for successful logins to important and non-important sites, and across all sites. Florêncio et al. [8, 6] suggested that 10^6 and 10^{14} guesses were the reasonable estimate of a password strength, necessary to withstand online and offline attacks, respectively. Unfortunately, 14% of important-site and 27% of non-important site passwords were vulnerable to online attacks, and 92% of important-site and 95% of non-important site passwords were vulnerable to offline attacks. While one could argue that most servers should be secure and users should thus only worry about online attacks, rampant password reuse and many online accounts make any vulnerable server a serious threat to the user. For example, a user may have an account at some vulnerable server that gets compromised through an offline attack. If the user uses the same credentials for their work or bank server, the attacker now can impersonate the user at these secure servers.

Important sites have longer and stronger passwords: objective. We find that important sites have passwords that are on the average 1–2 characters longer and 20 times stronger than passwords at non-important sites (multiple regression, length: $p = 0.000281$, R^2=0.0315, and Cohen' effective size= 0.315 – medium; multiple regression, strength: $p = 0.000849$, R^2=0.032, Cohen's effective size= 0.033 – small).

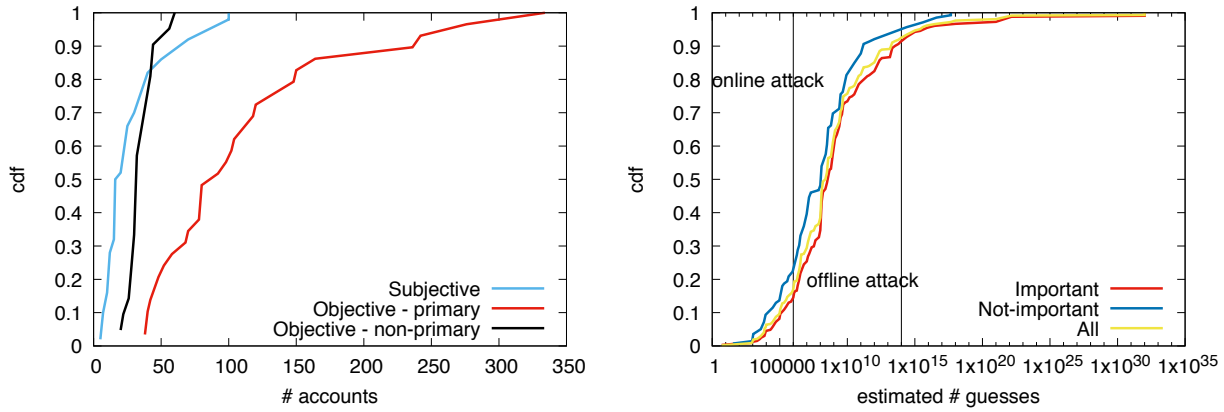
User Perceptions

In this section we summarize our findings about user perceptions about risk, attacks and password strength that may influence their password choices.

Risk-averse users do not have stronger passwords: subjective+objective. A user may believe that password attacks are rare or that they are not likely to be attack target, because the user has an open attitude towards risk in general. Creese et al. in [3] investigated this and we repeat their approach, using the same instrument (risk perception survey). Participants' ratings are shown in Figure 4.

Creese et al. in [3] found that answers to questions PPLEBY, CLSEPR, FLY, OSUPD, BRUPD and APPUPD were correlated with password strength. Specifically, when user's ratings and the experts' ratings differed for these questions, users were found to have weaker passwords. We repeat their approach on our data, and use their experts' ratings, but find no significant correlation between participant's misconceptions about risk and their password strength (Pearson's correlation $r = 0.06$, $p = 0.67$ for maximum strength and $r = 0.086$, $p = 0.89$ for average strength). For completeness, we also attempted to find correlations between each of the risk perception survey responses and password strength, but found none.

Users underestimate password attacks, but better understanding does not lead to stronger passwords: subjective+objective. A user may create weak passwords because they do not understand how powerful password-guessing attacks are. We code responses to narrative questions in the mental model of attacker survey. These questions ask how many password guesses an offline attacker could make per minute (NUMGSS), and how he would craft these guesses (HOWGSS). Question OFFLIN asks if a participant has heard of offline guessing attacks. We round the answers to question NUMGSS to the nearest power of 10. Similar to Wash et al. [26] we find that users severely underestimate the speed of the password cracking: 76% assume the attack to be of an online nature, and the rest assume an offline attack. Around half of the users do not know how attackers formulate guesses or else believe they use personal information about a user.



(a) Number of accounts per participant as estimated by the participant (subjective), and as measured in our study (objective).

(b) Distribution of password strength for important, non-important and all sites.

Figure 5. Number of accounts per participant and distribution of password strength.

But better understanding does not lead to stronger passwords. We found no significant correlation between participant responses to these survey questions and their password strength (Spearman's rank correlation $-0.1 \leq r \leq 0.11$, $p \geq 0.49$).

Users narrate good password composition strategies: subjective. A user may understand the need for strong passwords, but may not know how to create one. In our password strategy survey we ask participants to narrate how they create passwords and use these answers to infer whether they know how to create strong passwords. For space reasons, we summarize our findings. The majority of participants (93%) used names and words of personal significance in passwords and increased the strength by adding numbers and symbols, and capitalizing parts of passwords. 80% of participants said they use two or more character classes, and 30% use three or more. Thus most participants understand that they cannot rely on content that is personally significant to them, and must add random content to make passwords stronger.

Some users choose to create weak passwords: subjective+objective. A user may have all the right knowledge, but choose to disregard it in favor of memorability, or because they do not care if their accounts get hacked. We asked the participants in our password strategy survey (question STRATG) if they thought their strategy was good. 28% of participants said they knew their strategy was bad but continued to follow it, 10% thought it was OK, and 62% thought it was good. There is significant statistical difference (multiple regression, p=0.003779, R^2=0.030607, Cohen's effective size= 0.0316 – small) in password strength between participants with good and bad strategies. Passwords of bad-strategy respondents were indeed weaker than those of good-strategy respondents.

We further analyzed narrative responses by bad-strategy participants to question STRATG. One of the participants said: "Its probably not good, but I am not terribly worried about my passwords being found out." Another participant said: "I choose whatever is easy to remember. I think its bad. But, I don't want to use password resets frequently." Also, two participants remarked that their strategy is not good but it is easy to use. Thus memorability and convenience seem to motivate these participants to continue bad password practices.

Users' passwords are well-composed but short: subjective+objective. We infer a participant's actual password strategy by using the segmentation of each successful login password. We regard all POS-tagged segments as "meaningful-word segments" and those that were untagged as "random segments." We then classify passwords where length of random segments exceeds that of meaningful-word segments as "random," and the rest as "dictionary." Thus passwords classified as random may not be fully random, but they have more randomness than meaningful content. We also infer the character mix of a password using information about capitalization, mangling and presence of character/digit segments. 60% of passwords used mostly dictionary words, and the remaining 40% used more random content than dictionary words. Further, 27% of passwords used a single-character class, 54% used two-character classes, and 19% used three-character classes. This shows that participants understand how to compose strong passwords—40% create passwords where half or more characters seems random, and 73% use more than one character class. But password length plays a considerable role in determining strength, along with composition. Half of the passwords were shorter than 10 characters (and 90% shorter than 15).

Interestingly, participants' intended and actual password strategies rarely match. We mine a participant's intended strategy from answers to question STRATG, and infer the actual strategy from transformed passwords for that participant. We categorize both strategies in the same manner. Those that use more random than dictionary characters are labeled as "random." Otherwise, they are labeled as "dictionary." We also note the number of character classes that a strategy uses. We then say that a strategy A is stronger than B if A is random and B is dictionary, or if both have the same label but A has higher number of character classes. We find that 28% of passwords use a weaker strategy than narrated by a participant, 48% use a stronger strategy and only 24% match.

DISCUSSION AND RECOMMENDATIONS

We now discuss main reasons for bad password habits that we observed, and provide recommendations for improvement.

Bounded Rationality

Bounded rationality is the idea individuals do not make decisions in a rational manner, but are limited by the tractability of the problem, human cognitive limitations, and the time available to make the decision [9]. Bounded rationality drives people to make unhealthy decisions in spite of being well informed about their health risks [10], or to under-save for retirement [12]. We believe that bounded rationality plays a significant role in password choice and reuse. Our study found numerous discrepancies between password strategies narrated by a user and the actual strategies that same user engaged in: (1) Users underestimate the number of accounts they have by almost six times. (2) Users self-report strong intentions to have diverse passwords but do not follow through. (3) Users narrate intentions to reuse passwords only within certain accounts classes, yet reuse indiscriminately in practice. (4) Users rely on memory for password recall even when they use password managers. (5) User-intended password composition matches the actual one only 24% of the time.

It is no wonder that users struggle to keep track of their accounts and passwords. Not only do users have many accounts (median 80), but they create them over a long time, and thus cannot keep track of their behavior. Automated assistants (browsers and password managers) could help users by long-term tracking of their accounts and passwords, and by producing periodical summaries and analyses of this data. For example, once a month a user may see a report that states "You have 100 online accounts but have only used 15 in the past year. You have used only 3 different passwords on these 100 accounts." The assistant could further suggest random passwords for rarely used accounts.

Misconceptions

Similar to prior approaches, we found that users underestimate the risk of attacks and attacker abilities: (1) When we asked users about password-reuse attacks, 18% were ill-informed, 10% did not know about password-reuse attacks, and 8% knew but thought that strong passwords were immune, which is incorrect. Further 15% believed that such an attack is unlikely. (2) When asked about password-guessing attacks, 76% of users assumed that the attack would be of an online nature. Around half of the users did not know how attackers formulate guesses or else thought that they use personal information.

We also found that users knew how to compose strong passwords—40% created passwords where half of the content seems random, and 73% used more than one character class. But users did not create sufficiently long passwords. Half of the passwords were shorter than 10 characters (and 90% shorter than 15) and thus crackable by brute force.

This misconception about the importance of password length may come from password policies. We surveyed the policies of all 210 websites where participants successfully logged in during our user study, and found that 27% did not have a minimum length requirement, 33% required a minimum of 6 characters, and 28% required a minimum of 8 characters. Sites that required 8 characters or less had significantly weaker passwords than those with a stronger requirement. Users may assume that the minimum requirement is sufficient for a strong password, but it can be brute-forced if the site does not use slow hashing [19]. Sites should thus require longer passwords.

Willingly Trading Security for Memorability

Some users make a conscious decision to reuse passwords or create weak passwords for memorability reasons. User willingness to trade security for memorability is not surprising, as the frequency of needing to recall passwords is far greater than the frequency of attacks on any specific user. 100% of 49 participants who reused a password verbatim said they did it for memorability. Further, 44% of users said they were familiar with password-reuse attacks, but continued to reuse because memorability was more important to them than security.

Similarly, some users consciously chose to create weak passwords. 28% of participants said they knew their strategy was bad but continued to follow it. These participants had weaker passwords than those that thought their strategy were good, and their comments indicated that memorability and convenience were the main reasons for bad password habits.

We should make good choices convenient. We could achieve this by making password managers more proactive, such as suggesting random, long passwords for each new account or suggesting to the user, when they access a rarely-visited site to replace their password with a long, random, unique string.

CONCLUSIONS

In theory, good password hygiene and risk management are straightforward: strong, unique passwords for all accounts, but especially for more important ones. However, the proliferation of accounts, weak password policies, and difficulty remembering all of these passwords make good password behaviors hard to implement in practice.

Throughout this research, we have observed that users' security perceptions and intent rarely match their security realities. Some reasons for this lie in misconceptions about risk and a desire for convenience, identified by other researchers. But another large reason, uncovered by our research, lies in the sheer complexity of managing many accounts over a large time span – a task that is cognitively hard for humans. We have recommended development of tools that reduce this cognitive load and identify cases where password sharing increases user risk. Our future research will investigate whether this kind of intervention can measurably improve user password strategies.

ACKNOWLEDGEMENT

We thank anonymous reviewers, and Xiaojun Bi for providing helpful comments to improve our paper. This work was supported in part by the National Science Foundation under grant no. 1351058.

REFERENCES

1. Joseph Bonneau. 2012. The science of guessing: analyzing an anonymized corpus of 70 million passwords.

In *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, Oakland, USA, 538–552.

2. Lynne Coventry, Pamela Briggs, John Blythe, and Minh Tran. 2014. *Using behavioural insights to improve the public's use of cyber security best practices*. Technical Report. Northumbria University.
3. Sadie Creese, Duncan Hodges, Sue Jamison-Powell, and Monica Whitty. 2013. *Relationships between password choices, perceptions of risk and security expertise*. Springer, Berlin, Heidelberg. 80–89 pages.
4. Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. 2014. The Tangled Web of Password Reuse. In *NDSS*, Vol. 14. Internet Society, San Diego, USA, 23–26.
5. Dinei Florencio and Cormac Herley. 2007. A large-scale study of web password habits. In *Proceedings of the WWW*. ACM, Banff, Alberta, Canada, 657–666.
6. Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. 2014a. An Administrator's Guide to Internet Password Research.. In *LISA*. USENIX, Seattle, USA, 35–52.
7. Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. 2014b. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *23rd USENIX Security Symposium*. USENIX, San Diego, USA, 575–590.
8. Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. 2016. Pushing on string: The "don't care" region of password strength. *Commun. ACM* 59, 11 (2016), 66–74.
9. Gerd Gigerenzer and Reinhard Selten. 2002. *Bounded Rationality: The Adaptive Toolbox*. The MIT Press.
10. Xiaoyan Huang. 2017. KevinMD, Why are health care consumers not making smart decisions? `http://www.economicsonline.co.uk/Behavioural_economics/Bounded_rationality_and_self_control.html`. (2017).
11. KoreLogic Security. 2010. KoreLogic's Custom rules - DEFCON 2010. `http://contest-2010.korelogic.com/rules.html`. (2010).
12. Economics Online. 2017. Bounded Rationality and Self-Control. `http://www.economicsonline.co.uk/Behavioural_economics/Bounded_rationality_and_self_control.htm`. (2017). Accessed: 2017-7-14.
13. Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. 2017. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *Proc. of the ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 295–310.
14. Ashwini Rao, Birendra Jha, and Gananand Kini. 2013. Effect of grammar on security of long passwords. In *Proceedings of the third ACM conference on Data and application security and privacy*. ACM, ACM, San Antonio, Texas, USA, 317–324.
15. Elissa M Redmiles, Amelia Malone, and Michelle L Mazurek. 2016. I Think They're Trying To Tell Me Something: Advice Sources and Selection for Digital Security. In *2016 IEEE Symposium on Security and Privacy*. IEEE, San Francisco, CA, USA, 272–288.
16. Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2010. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM, Redmond, Washington, USA, 2.
17. Peter Snyder and Chris Kanich. 2013. Cloudsweeper: enabling data-centric document management for secure cloud archives. In *Proceedings of the 2013 ACM workshop on Cloud computing security workshop*. ACM, Berlin, Germany, 47–54.
18. Elizabeth Stobert and Robert Biddle. 2014. The password life cycle: user behaviour in managing passwords. In *Symposium On Usable Privacy and Security (SOUPS 2014)*. ACM, Menlo Park, USA, 243–255.
19. Wikipedia John the Ripper. 2012. Differences Between Fast Hashes and Slow Hashes. (2012). `http://openwall.info/wiki/john/essays/fast-and-slow-hashes`
20. UCREL CLAWS7 Tagset. 2016. `http://ucrel.lancs.ac.uk/claws7tags.html`. (2016).
21. Blase Ur, Jonathan Bees, Sean Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Do Users' Perceptions of Password Security Match Reality?. In *CHI'16: 34th Annual ACM Conference on Human Factors in Computing Systems*. ACM, San Jose, California, USA, 3748–3760. `DOI: http://dx.doi.org/10.1145/2858036.2858546`
22. Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2015. "I added '!' at the end to make it secure": Observing password creation in the lab. In *SOUPS '15: Proceedings of the 11th Symposium on Usable Privacy and Security*. USENIX, Ottawa, Canada, 123–140. `http://www.ece.cmu.edu/~lbauer/papers/2015/soups2015-password-creation.pdf`
23. Rafael Veras, Christopher Collins, and Julie Thorpe. 2014. On the semantic patterns of passwords and their security impact. In *Network and Distributed System Security Symposium (NDSS'14)*. Internet Society, San Diego, USA.
24. Emanuel Von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2013. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In *IFIP Conference on Human-Computer Interaction*. Springer, Berlin, Heidelberg, 460–467.

25. W3Schools. 2017. `onbeforeunload Event`. `http://www.w3schools.com/jsref/event_onbeforeunload.asp`. (2017).

26. Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding Password Choices: How Frequently Entered Passwords are Re-used Across Websites. In *Symposium on Usable Privacy and Security (SOUPS)*. ACM, Denver, CO, USA, 175–188.

27. Matt Weir, Sudhir Aggarwal, Breno De Medeiros, and Bill Glodek. 2009. Password cracking using probabilistic context-free grammars. In *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, Oakland, USA, 391–405.

28. Dan Lowe Wheeler. 2016. zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security*. USENIX, Austin, TX, 157–173.