Forgetting of Passwords: Ecological Theory and Data

Xianyi Gao, Yulong Yang, Can Liu, Christos Mitropoulos, and Janne Lindqvist,
Rutgers University; **Antti Oulasvirta,** ***Aalto University***

https://www.usenix.org/conference/usenixsecurity18/presentation/gao-xianyi

This paper is included in the Proceedings of the 27th USENIX Security Symposium.

August 15–17, 2018 • Baltimore, MD, USA

ISBN 978-1-939133-04-5

Open access to the Proceedings of the 27th USENIX Security Symposium is sponsored by USENIX.

Forgetting of Passwords: Ecological Theory and Data

Xianyi Gao[†], Yulong Yang[†], Can Liu[†], Christos Mitropoulos[†], Janne Lindqvist[†], Antti Oulasvirta[*]
[†]*Rutgers University,* [*]*Aalto University*

Abstract

It is well known that text-based passwords are hard to remember and that users prefer simple (and non-secure) passwords. However, despite extensive research on the topic, no principled account exists for explaining when a password will be forgotten. This paper contributes new data and a set of analyses building on the ecological theory of memory and forgetting. We propose that human memory naturally adapts according to an estimate of how often a password will be needed, such that often used, important passwords are less likely to be forgotten. We derive models for login duration and odds of recall as a function of rate of use and number of uses thus far. The models achieved a root-mean-square error (RMSE) of 1.8 seconds for login duration and 0.09 for recall odds for data collected in a month-long field experiment where frequency of password use was controlled. The theory and data shed new light on password management, account usage, password security and memorability.

1 Introduction

This paper contributes to understanding the security of text-based passwords, the most prevalent method of authentication [43]. This paper builds on an ecological theory [10] of human memory to address the well-known tension between the security of a password and its usability. For example, common password creation guidelines predominantly focus on security objectives, yet users are reluctant to invest adequate effort in creating passwords that meet these criteria [47]. A large proportion of real-world passwords are weak and easy for attackers to guess [14]. Further, when a password is hard to remember, users may resort to practices that compromise security, such as reusing passwords [26]. Password managers have not solved this issue [2]. For example, one study suggested that the prevalence of password managers for text-based passwords is only at one percent [44]. The reasons users reported not adopting password managers include concerns about security, trust issues in vendors, uncertainty on software functions, limited support for web applications, and the fear of losing control of passwords [2]. Improving password memorability and usability is a worthy endeavor because password forgetting can even be associated with significant financial losses with password resets [76, 55].

At the core of the memorability–security issue is the psychological question *why* people remember some passwords and forget others. The key issue is forgetting: we need to understand why users are at times unable to remember passwords and unwilling to invest in creating complex passwords. Although one may understand system security as a technical subject, memorability is a fundamental factor in practical system security. Although previous studies have measured the memorability of passwords in the context of different authentication systems or strategies [21, 23, 33, 48, 60, 83], it is not known what makes a password memorable.

Several known principles of long-term memory functioning are relevant in this context. Based on the *depth of processing theory* [20], the way we attend to a password affects how well it is remembered. A password generated quickly will be not as well remembered as a password generated when one pays attention to it. The *encoding–retrieval match* suggests that similarity between cues (e.g. visual design of the login screen or presence of company logo) during encoding (when creating a password) and retrieval affects the probability of retrieval [61]. These two theories, however, do not predict password recall over time, because they do not include any time-related predictor. *Decay theory* suggests that memory traces decay over time when not activated, and several models have been proposed to capture this effect [56]. This could mean that longer time ago one used a password, the less likely one can remember it. *Interference theory* suggests that forgetting can be due to interference between similar memory traces, such as when the pass-

words have same words or are used in similar-looking applications [17]. *Activation theory* suggests that temporal effects and interference also depend on the level of activation [67]: the higher the activation to begin with, the more robust it is for memory recall.

Our paper investigates and empirically evaluates an *ecological theory of human long-term memory* [10] in the context of password recall. The ecological theory of memory suggests that long-term memory evolved to help survival by anticipating organismically important events [18]. The most important predictor of recall is the organismic importance of recalling it: in other words, the predicted value of remembering it in the future. Since most memory usage is not directly related to survival, Anderson and Schooler proposed an adaptation for daily stimuli such as emails and newspaper headings. Their model is a statistical mapping between occurrence probability and the probability of recall [9]. In the context of passwords, we propose that the more likely it is that one will need it in the future, the more likely it is recalled. Following Anderson and Schooler, this principle can be used to derive mathematical models of password retrieval probability. In this paper, we present and study these models, comparing their predictions on memorability against empirical data in the context of online authentication. The design of our field experiment tries to minimize the confounding effects of password security, log-in frequency, account type, and password managers.

This paper makes two main contributions: (1) we present models of password memorability based on the ecological theory of memory; (2) we present model fit and qualitative observations from a field experiment of online authentication. The results largely support the ecological hypothesis and the suggestion that forgetting is a major limiting factor leading to poor password practices and compromising of systems security. Our model enables system designers and security engineers to predict the probability of password forgetting given a level of system usage and potentially impose appropriate memory practice for users to mitigate forgetting. We discuss the implications of these findings on the design of authentication systems, policies, and guidelines.

2 Related Work

Previous studies have identified a number of factors affecting password memorability.

Repetitions improve memory of passwords. By asking participants to memorize secrets gradually and repeatedly, a study found that 88% of its participants were able to recall a 56-bit secret code after three days [16]. Another study also utilized spaced repetitions to help participants memorize Person-Action-Object (PAO) stories as a password management approach to generate strong passwords [13]. 77% of their participants recalled all their stories four months later, with at most 12 tests over the period. The study also found that the majority of forgetting occurred within the first 12 hours. Another study suggested that recalling after a short delay is an effective way to help retention [80]. In addition to the number of repetitions, the frequency of such repetitions also affects password memorability. A diary study reported that people seldom forget their passwords if they are used frequently [47].

Memorability also depends on the number of accounts and passwords [80, 23, 34]. A study found that password strength and use of symbols and digits in passwords can predict the likelihood of password reuse [65]. Another study reported that undergraduates had an average of 7.8 accounts per person [37]. They also found that majority of participants had only three or fewer unique passwords. A three-month study, collecting password-usage data with a browser plug-in, showed that people managed seven unique passwords in average, each of which were used for about 5.67 different sites [36]. A two-week diary study estimated that participants had an average of 11.4 accounts per person [42].

Password generation and mnemonic strategies have been found to affect memorability. For example, appending additional characters and digits noticeably reduces memorability [80]. A study showed that number chunking, a memorization technique, improved the memorability of system-assigned PINs [45]. Similarly, passwords generated by associating selected cognitive items could yield acceptable guessing rates while being less susceptible to forgetting than conventional passwords [19]. A study testing password creation guidelines explained that the password phrase strategy was secure against cracking while being easy to remember [83]. Similarly, another study suggested mnemonic phrase-based passwords are still secure and appropriate for some uses today although they could become more vulnerable in the future [53].

The human limits of memorability have also been linked to security issues in password management strategies. One study suggested that a maximum of four or five passwords per person reaches the limit of most users' memory capabilities [1]. Another study showed that people categorize their passwords into a limited set of categories, with varied security, with some accounts (e.g. financial accounts) being more important [40]. They also found that it is possible to crack passwords across categories if passwords from lower-value categories are known, as the passwords are similar across categories.

Other research has focused on studying whether different password-strength meters and password policies can affect user's password selection [51, 33, 72]. Although these studies also measured password memorability, the purpose was to examine the usability of corresponding

password meters and policies. Our study focuses on investigating password memorability and understanding how different factors affect password memorability quantitatively. We are the first to apply major memory theories and build mathematical models of text passwords for online authentication.

3 Modeling Password Memorability

This section presents mathematical models for password retrieval. The models are based on the ecological memory hypothesis. To derive quantitative predictions for recall odds and retrieval time, we used an established cognitive model called ACT-R (Adaptive Control of Thought – Rational [4]). The ACT-R model includes two key parts: (1) a model of memory activation and (2) a model of retrieval as a function of memory activation. These can be mapped to events in password use such as frequency.

The model assumes that the higher the activation, the more accessible a memory representation of a password is. Activation is related to the historical use of this memory element and contextual associations related to the memory recall [3, 6]. Based on this, the equation of activation for a memory element i (or chunk i) is

$$A_i = B_i + \sum_{j=1}^{n} W_j S_{ij} \tag{1}$$

where A_i is the activation level for element i, W_j is the source activation of element j, S_{ji} is the strength of association from element j to i, and B_i is the base-level activation. The second term with W_j and S_{ij} is related to the contextual setting in the current memory recall, the former affected by available cues and the latter by the level of attention to a cue. The base-level activation B_i is based on the history of use (e.g. previous retrievals). These two terms are independent of each other and can be added when estimating the memory activation. B_i can be obtained through equation [8]:

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) \tag{2}$$

where t_j is the time for the jth use of this memory element, and d is the memory decaying parameter. This equation aligns with how human memory works with spaced repetitions. It includes effects from both practice (summation of n times memory usage) and memory decay over time (power function with negative factor).

The memory recall time is exponentially related to the memory activation [4, 7, 9]:

$$Time_i = Fe^{-M_i} \tag{3}$$

where F is a time scale constant, and $M_i = A_i - P$. A_i denotes the activation of element i, and P is the mismatch penalty referring to the similarity of component i to conditions. In case of online account logins, users are presented with the same login page each time, so the conditions and context information are similar each time. P can be seen as a constant. As the value of P only changes the scale factor of Equation 3, we can simply set it to zero and combine the effect of P to term F.

The recall odds (R_o, ratio of the probability of successful recalls and the probability of failed recalls) can be calculated using the following equation [4]:

$$R_o = e^{(M_i - \tau)/s} \tag{4}$$

where τ is a memory threshold parameter, and s is a parameter related to the variance of activation.

Login Duration: To predict login duration, we assume that most variability in recall comes from retrievability of the associated memory. In the study reported in this paper, we assigned password logins for online accounts with different login frequencies (e.g. once per day or once per five days). We can consider the successful login duration as the summation of memory recall time ($Time_i$) and action time ($Time_{act}$, including the time for users to navigate the login page, to type, and to enter:

$$Time_{login} = Time_i + Time_{act} \tag{5}$$

We can calculate the expected value of successful login duration:

$$E[Time_{login}] = E[Time_i] + E[Time_{act}] \tag{6}$$

where $Time_{act}$ is a random variable with a mean $E[Time_{act}]$. After substituting (Equations 1, 2, and 3 to Equation 6), we can obtain

$$E[Time_{login}] = \frac{E[Fe^{-C}]}{\sum_{j=1}^{n} t_j^{-d}} + E[Time_{act}] \tag{7}$$

where C is the contextual term ($\sum_{j=1}^{n} W_j S_{ij}$), which can be considered as a random variable with a constant mean. Therefore, we can simply use a constant parameter K to represent the value of $E[Fe^{-C}]$. The time variable t_j is equal to $f \cdot j$ where f is the login frequency (e.g. login in every f days). n is the amount of practice with the same password.

$$\sum_{j=1}^{n} t_j^{-d} = \sum_{j=1}^{n} (f \cdot j)^{-d} = f^{-d} \sum_{j=1}^{n} j^{-d} \tag{8}$$

By applying an integral approximation [5] for the summation term,

$$\sum_{j=1}^{n} j^{-d} \approx \int_{j=0}^{n} j^{-d} dj = \frac{n^{1-d}}{1-d}, (d < 1) \tag{9}$$

which has bounded error for a fixed value of d, we obtain an equation for the average successful login duration:

$$E[Time_{login}] \approx \frac{Kf^d(1-d)}{n^{1-d}} + E[Time_{act}] \quad (10)$$

Recall Odds: Using a similar approach, we can derive the equation to predict recall odds, defined by the probability of successful logins divided by the probability of failed logins. We can substitute Equations 1 and 2 into Equation 4 and then apply the approximation in Equation 8 to obtain:

$$R_o \approx e^{-\tau/s+C/s}(1-d)^{-1/s} f^{-d/s} n^{(1-d)/s} \quad (11)$$

where C is the contextual term ($\sum_{j=1}^{n} W_j S_{ij}$) as above after Equation 7, τ is the threshold parameter, s is a parameter related to the variance of activation, and d is the memory decay parameter.

The experimental value of recall odds can be a good estimation of the expected value of the theoretical recall odds ($R_{o\_Measured} = E[R_o]$). Therefore, we can further obtain that

$$R_{o\_Measured} \approx A f^{-d/s} n^{(1-d)/s} \quad (12)$$

where $A = e^{-\tau/s}(1-d)^{-1/s} E[e^{C/s}]$.

4 Method

This study focuses on the effects of account login frequency, account types, and password strength on the password recall success rate and time. Participants generated passwords for several accounts and were asked to recall the passwords multiple times at different points of time afterwards. Asking participants to generate passwords for a study is a common approach for password studies [40, 51, 57, 72, 78, 80, 83] and this approach has received empirical support when compared against real passwords [35].

Each participant was required to participate in our study for about one month. We stored all collected data (e.g. passwords generated for our study, time, and account information) in our secure server for later analysis. We recruited participants over approximately four months from July 2017 to October 2017.

Our study was approved by the Institutional Review Board of Rutgers University.

Many password studies have used crowdworking sites, such as Amazon Mechanical Turk [77, 46, 78, 73, 72, 71, 51, 45, 33]. We decided that crowdsourced recruitment is not ideal for our purposes, because participation was needed for a sustained period of one month and participation required a face-to-face meeting for instructions and the survey. During the study period, we kept in touch with participants through emails for any questions and concerns. We also sent out reminders to make sure that most tasks were completed. Based on our experience from a pilot study, meeting in person to give instructions, explain tasks, and show task examples results in less misunderstanding and lower drop-out rate compared to crowdsourcing approach (e.g. watch tutorial videos and read instructions).

4.1 Participants

109 participants were recruited by posting flyers around the university campus, web sites (e.g. reddit and craigslist), and university mailing lists. During the study, four participants decided to quit (due to change in summer vacation schedule). Five participants took too long to complete more than one third of the tasks, and were excluded from analyses. Having too many expired tasks would have affected the independent variables. Therefore, the results in this paper are based on the remaining 100 participants. Based on our pilot study, we estimated that the sample size is sufficient for modeling.

Our participants' ages ranged from 18 to 62 with a mean of 24. 52% of them were women and 48% were men. Most of our participants were college students who were pursuing a variety of majors (e.g. engineering, computer science, business, psychology, and biology): 57% were undergraduate students and 29% were graduate students. The remaining 14% included employed engineers, IT professionals, administrative support workers, and others.

4.2 Experiment Design

Our experiment asked participants to create passwords for eight online accounts and log in to these accounts with certain frequencies. Participants performed tasks using a web application. This type of design allowed participants to perform tasks anytime and anywhere, which fitted the real usage of passwords better compared to lab studies.

4.2.1 Password Memorability Metrics

In our study, we used login success rate and login duration as password memorability metrics. Login success rate, defined as the ratio of successful logins over the number of total logins satisfying a certain condition, is a commonly-used metric to measure the memorability [21, 22, 23, 24, 33, 60, 45, 16]. Login duration has been used in previous studies to measure memorability as well [73, 22, 23, 45]. In our study, the login duration is the time period from when the login page appears to when the participant logs into the home page or sees the login failure message.

4.2.2 Study Variables

In our study, we focus on investigating major prediction variables including account type, login frequency, and password strength.

Account Type: Account type variable is a within-subject variable because participant is required to generate passwords for different accounts. Similar web based studies have shown that people purposefully generate passwords with different levels of security and behave differently for different accounts [11, 63, 40]. The purpose of this variable is to study whether such difference exists in the memorability of passwords as well.

We used the account categories proposed in literature [40, 15]: identity accounts, financial accounts, content accounts, and sketchy accounts (we refer to them as advertisement accounts in this paper). This categorization provides a reasonable separation of different accounts, and has been shown to match the subjective perception of importance people have regarding their accounts [40].

We designed eight different accounts in our study: one email account and one social networking account as identity accounts, one banking account and one shopping account as financial accounts, one news reading account and one music streaming account as content accounts, and one daily deal posting account and one coupon posting account as advertisement accounts. We selected these eight accounts for their common appearance on the Internet. Our account categories also match the accounts people typically use online [42].

Login Frequency: Login frequency variable is a within-subject variable indicating how frequently a participant needed to log in to an account. It has been shown that people access their passwords at various frequencies [42], and login frequency plays an important role in password memorability [47].

There were eight different login frequencies: once a day, once every two days, once every three days, ..., and once every eight days. Previous studies utilized different log-in frequencies from once per hour to once per two weeks [31, 23, 21, 16, 84]. Our frequencies were also within this range adjusted for the case of online accounts.

Eight different frequencies were randomly assigned to eight different accounts with 8! possible assignments in total. Each participant was randomly given one of these assignments. A diary study on password usage found that most users accessed their accounts 40 to 110 times in two weeks and users had a mean of 8.6 accounts [42]. In our study, the number of logins for each participant per two weeks was about 50 which was within the normal range.

Password Strength: Password strength is a variable related to password security. It is common for users to self-generate passwords instead of being assigned them for online accounts. Therefore, password security varies based on our participants chosen passwords to ensure the ecological validity of our study. We do post-hoc analysis for the effect of password strength.

We included a password strength meter in the login page to provide participants feedback on passwords. Password strength meters are well-studied and shown to have an observable impact on password security as well as user behavior [33, 51, 78, 71]. Also, they have been wildly deployed in industry to help users generate passwords. Therefore, participants are familiar with them and they are effective at influencing password generation. We used the *zxcvbn* password strength meter from Dropbox [29]. It is open-source and has been deployed in many practical applications such as WordPress [38], Dropbox [29], Stripe [75], and Coinbase [25]. Prior work has shown that compared with meters that primarily focus on character sets and length requirement, zxcvbn meter measures the password strength based on the structure of passwords, and found to be consistent with most publicly-available password datasets [27]. It was shown to be accurate and suitable for mitigating online attacks [81].

Recently, researchers have also used neural networks with password meters to provide real-time text feedback on why the password is weak and how to make it strong [77]. Although this data-driven password meter is effective, it generates a lot of password guidelines leading users to only generate passwords that are considered to be secure (e.g. they contain more than 8 characters, include several symbols, do not use date and year, include a number in the middle, and do not to use common phrases or words) but could be hard to remember. Since we are interested in both security and memorability, we did not want to provide participants with too many guidelines to restrict the natural variation of our password data.

In addition to the online password meter zxcvbn, we applied off-line methods to evaluate password guessability. Off-line password crackers and estimators allow intensive computations compared to online password meters. We used Hashcat 3.00 [41] to perform the rule-based dictionary attacks on our collected password set. Hashcat is a popular password cracker that has been applied to many password studies [66, 79, 57, 30]. The password dictionary that was used is a shuffled combination of different wordlists including Google 1-gram English dataset [39], UNIX dictionary [54], RockYou leaked password dataset, and phpbb leaked password dataset. The dictionary contained 38M unique words. We used the rules (i.e. functions that modify, cut or extend the dictionary words) from KoreLogic [52] for our password cracking. KoreLogic contains 42M rules and it has been used to imitate the real-world attacker behavior in the latest text password cracking study [79]. To obtain a good estimation of password strength, we also applied an existing password estimating model trained with neural networks [58].

We asked our participants not to reuse passwords for different accounts because the number of different passwords a person needs to manage can largely affect memorability. In our study, password reuse needed to be controlled in order to examine other interesting factors ef-

fectively. We focused on the quantitative modeling of password memorability instead of exploring the factors related to memory load such as number of passwords or accounts that others have studied [80, 23, 34]. To examine password reuse and similarity, we used *edit proportion*, which is a normalized version of the Damerau-Levenshtein string-edit distance [12]. For two passwords, we calculated first the edit distance, and then normalized it by dividing it to the length of the longer password. The edit proportion ranges from 0 (exactly the same) to 1 (completely different). Passwords from different accounts need to have edit proportion larger than 0.25. Previous studies have used similar approach with Damerau-Levenshtein distance to measure password similarity [26, 62].

4.2.3 Task Scheduling

It is unlikely anybody creates eight accounts in a day during their normal daily lives. Therefore, we designed our study so that participants created one new account a day, regardless of the login frequency the account had. For each account created, the corresponding login tasks were scheduled based on the login frequency starting from the creation day. The order of accounts was randomly shuffled to avoid bias.

The time of the day for sending a registration or login task was randomly chosen between 6:30 AM to 10:00 PM. A previous study showed people primarily use their passwords within this time range [36]. Using this randomization ensured we had creation and login tasks distributed throughout the day.

For login tasks, we prefilled the corresponding username for participants because we only focus on studying the memorability of passwords in this paper. Forgetting usernames is different from forgetting passwords as usernames and passwords can be managed very differently by users. Prefilling the username can rule out the cases of forgetting usernames which should be studied differently.

For each login task, participants had five attempts. If they failed to login to an account with five attempts, they received a link to reset the password. We decided to allow five attempts after referring to real-world applications. Given that some prominent services still limit attempts to three nowadays (e.g. Facebook), we would like to set the number of attempts for our study low as well. However, if the number of attempts is too low, participants may keep resetting passwords which would generate less data on the memorability of a password over time. Our choice of a maximum of five attempts was based on a pilot study where these factors were considered.

Each task was generated with a unique id. Each link participants used to access their tasks was based on its unique id. By making each link unique and attaching a status flag to it, we could control when participants could

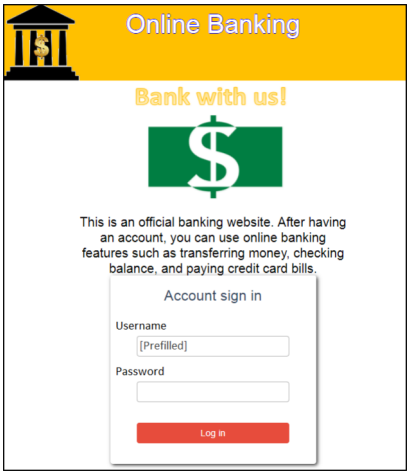

Figure 1: An example user interface of the login page for an online banking account.

access each task. Each link expired 24 hours after it was sent to participants. The email recovery link also followed similarly. Each recovery email participants received contained a unique recovery link for the password reset. The recovery link was set to expire in one hour. In this way, we ensured only participants themselves could proceed in their recovery process. We also ensured that each account had a set of unique email templates to distinguish from each other.

4.3 Apparatus

We designed and built a web application for this study. The application was written in Javascript, using Meteor framework [59]. The application generated different emails depending on the type of the task. In addition, for each task, the application generated and sent a reminder email automatically if the participant had not finished it after three hours.

We disabled auto-fill password function of web browsers and password managers. For example, we customized the password input field to read-only, as web browsers would only autofill the field if the fields were writeable. Our application also checked if the password field was already filled with texts.

To prevent participants to simply put their username or account header text (e.g. Online Banking) as the password, our web application examined the similarity of username and account header text to the generated password. The editing proportion among them should be larger than 0.25 when measured with normalized Damerau-Levenshtein distance. Many existing online accounts have similar restrictions [82, 32, 49, 69, 64, 28]. We also asked par-

ticipants in the exit survey whether the participants had written down their passwords or used password managers.

We used representative icons and headings in the web page for each account (e.g. see the online banking login page in Figure 1) to make sure participants were aware of these accounts being different and of their real-world usage and importance. In the login page, we also included brief text to explain online service features for the corresponding account.

4.4 Procedure

First, participants were introduced to the study and asked for consent to participate in the experiment. After consenting, we explained how to use our web application with an example demo and encouraged participants to ask questions if any.

Next, we asked participants to complete an entry survey (see Appendices for questions). The entry survey asked participants to report their email, demographic information, and answers to questions about password management. For password management questions, we asked participants how many passwords and accounts they were using, how many passwords they could remember without checking notes, and how often they forgot passwords. These password management questions were inspired by a prior password managing study [74]. We also asked our participants' opinions towards using password saving features in browsers or other password managers, and whether they had any strategy to help them memorize passwords in their daily lives.

The study lasted approximately one month. Participants needed to monitor their email account daily for new tasks. Each email contained a link to access the web application and to complete a task which was either registration or login.

After one month, we asked our participants to come back and complete an exit survey (see Appendices section at the end for questions) in our lab. In the exit survey, we had questions confirming whether they wrote down any passwords and whether they used any password managers or other password saving options during our study. We also asked participants the importance of different types of online accounts.

4.5 Survey Response Coding Approach

For coding open-ended responses, we followed a coding guideline for qualitative analysis [68]. First, open coding was used to generate labels from participants' responses. Then, several themes emerged from responses on each topic. We applied axial coding for further categorization to find the overall concepts and themes. Ambiguous cases were discussed among our group. At the end, we proofread coding and re-coded several times to ensure the reliability of our results. Some of the representative samples of participants' quotes will be shown as we present our findings.

zxcvbn Score	Score 0	Score 1	Score 2	Score 3	Score 4
Password Strength	Too Weak	Very Weak	Medium	Strong	Very Strong
Password Distribution	2%	22%	24%	31%	21%

Table 1: Distribution of our collected passwords for different zxcvbn scores [29]. There are 1443 different passwords in total. Most passwords (31%) are in score 3 (strong) and very few passwords (2%) are in score 0 (too weak).

Frequency (days per login)	1	2	3	4	5	6	7	8
Success Rate	0.967	0.942	0.912	0.871	0.871	0.833	0.813	0.802

Table 2: Login success rates for different login frequencies. The login success rate of a frequency is the ratio of the number of successful logins in this frequency to the total number of logins of this frequency.

5 Results

We start with an overview of our collected data. We then analyze how each variable affects password memorability and discuss model fitting. At the end, we present findings from survey responses.

Overall, 10680 login tasks were sent. Of these, 10041 tasks were completed and 639 tasks expired. Participants completed 800 account creation tasks and 9241 account login tasks.

Our participants generated 1443 passwords, which had minimum length of 3 and maximum length of 31. In the account creation page, we used the zxcvbn password strength meter [29] to estimate password security: score 0 (too weak) – passwords with this strength are considered as risky and can be guessed with fewer than 10^3 guesses, score 1 (very weak) – passwords are very guessable with fewer than 10^6 guesses, score 2 (medium) – passwords are somewhat guessable with fewer than 10^8 guesses, score 3 (strong) – passwords are safely unguessable with fewer than 10^{10} guesses, and score 4 (very strong) – passwords are very unguessable with more than 10^{10} guesses. Table 1 shows the distribution of our collected passwords across different zxcvbn scores.

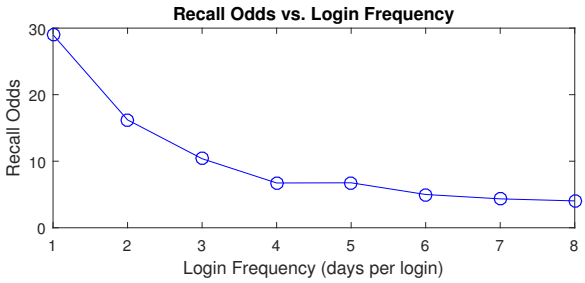

Figure 2: Recall odds vs. password login frequency. Recall odds drops fast initially and slows down as login frequency continues to change.

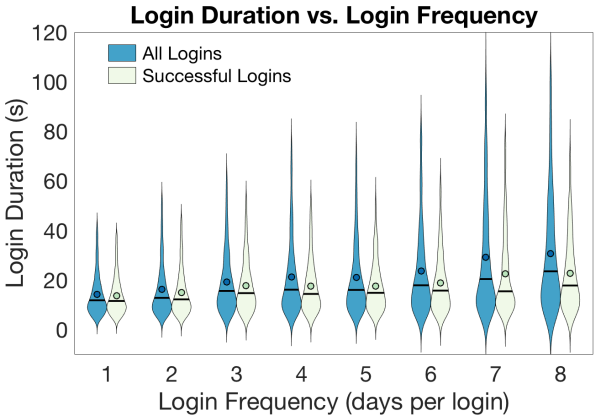

Figure 3: Violin plots of login duration vs. login frequency for successful logins and all logins. Violin plots show the probability density of the data at different login duration. On the violin plots, we marked the means (small circles) and medians (horizontal line) of the login duration when grouped based on login frequency.

5.1 Frequent Logins Help Memorability

In the dataset, participants' overall login success rate drops when the login frequency becomes less frequent (Table 2 with login frequency changing from 1 day per login to 8 days per login). Because recall odds has been shown to have functional relationship with time and practice [8, 9], we can simply convert success rate, p, to recall odds ($Ro = p/(1-p)$).

Figure 2 shows more logins help people to memorize their passwords since the recall odds decreases when the login frequency changes from 1 day per login to 8 days per login. Curve fitting with common functions (e.g. linear, polynomial, logarithmic, exponential, and power) shows power function as the best match ($R^2 = 0.9901$) with the fitting function: $Ro = 29.97 f^{-0.98}$. Exponential function shows the second best match ($R^2 = 0.9060$) with the fitting function: $Ro = 27.45 e^{-0.27 f}$. This finding is in agreement with a study that showed that the power function had a better match for memory decay compared to the exponential function proposed by very early psychology studies [9].

Figure 3 shows that the mean and variance of login duration both increase as frequency changes from 1 day per login to 8 days per login. It means people need more time to input their passwords when the passwords are less frequently used. This pattern exists in both the overall login data and the successful login data. When compared with successful logins, overall logins have higher means and variances of login durations. This makes sense as overall login data include failed logins which usually have longer login durations than successful logins. Login durations for successful logins are plotted separately because they are highly related to memory recall time (i.e. successful login duration is the recall time plus action time). On the other hand, a login duration for a failed login is the time for a participant to try all five attempts and give up because we limited number of attempts to five.

Figure 4 shows that the average login duration increases as the login frequency changes from 1 day per login to 8 days per login and the difference is statistically significant ($p < 10^{-15}$ for both successful logins and overall logins). It means the password login frequencies affect the time that people needed to input their passwords. This result and Figure 2 suggests that people need more time to input their passwords when the passwords are less frequently used. We used Scheffé's test for the pairwise comparison between different frequency groups. We chose Scheffé's test instead of other common ones (e.g. Tukey's test, Bonferroni method, Dunn and Sidák's approach, and Fisher's test) because Scheffé's test allows unbalanced sample sizes for different groups and it provides a simultaneous confidence level for comparisons [70]. As tradeoff, Scheffé's test is very conservative compared to other tests. Figure 4 also shows the confidence intervals for pairwise comparison with Scheffé's test.

5.2 More Logins Help Memorability

Figure 5 shows that logging in more helps people to memorize their new passwords. We call this login practice. We plotted the reciprocal of the recall odds instead of recall odds because recall odds could reach infinity when the recall success rate reached to 1 after enough practice. The recall odds can be calculated by the success rates, and the success rate for Nth login with a password is the number of successful logins divided by the total number of logins when grouped by the practice variable. As the number of logins or practice increases, the reciprocal of recall odds decreases and quickly reaches near zero, meaning that the recall odds increases and reaches near infinity quickly. Curve fitting with common functions (e.g. linear, polynomial, logarithmic, exponential, and power) shows the

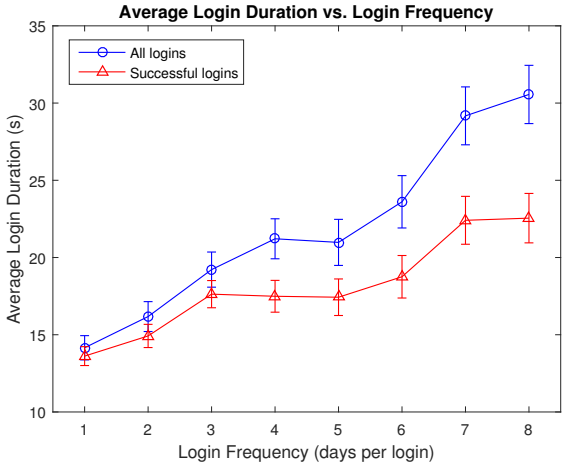

Figure 4: Average login duration vs. login frequency for overall logins and successful logins. The confidence interval (CI) for each mean is shown as a vertical bar. This figure also shows pairwise comparison for different frequency groups. If the CIs of two frequency groups do not overlap, the means of their login durations are statistically significantly different. For example, with successful logins, the mean login duration for 4 days per login is statistically significantly different from the means of 1, 2, 7, and 8 days per login but not statistically significantly different from the means of 3, 5, and 6 days per login.

power function as the best match ($R^2 = 0.9978$) with the fitting function: $1/Ro = 0.60n^{-2.22}$ where Ro is the recall rate (note that this recall odds grouped by practice is different from the recall odds grouped by login frequency in previous section) and n is the Nth login with a password.

Figure 6 shows that logging in more helps to decrease the needed time for inputting the passwords. Again, all login data indicates the overall login duration, while successful login data is highly related to the recall time. Both successful and all logins show the decreasing of average login duration when practice increases. They also both show the increase of variance for login duration when practice increases. Overall, the mean of login duration across different practice groups are statistically significantly different ($p < 10^{-15}$ for both overall logins and successful logins).

We applied Scheffé's test [70] for the pairwise comparison between different groups with different practice (see Figure 6 for CIs and comparisons). For example, from the upper figure with all login data, the mean of login duration for 1st login is statistically significantly different from all other groups. The mean for 2nd login is statistically significantly different from 1st, 6th, 12th, and 13th login groups. From the lower figure with only successful logins, the mean for 1st login is statistically significantly different from all other groups except the 2nd login group. The mean for 2nd login is statistically significantly different from 1st, 12th, and 13th login groups.

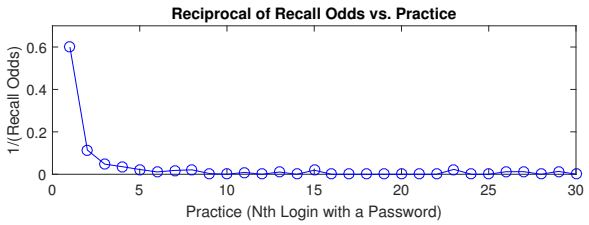

Figure 5: Reciprocal of recall odds vs. practice. The Nth number of login with the same password is the practice variable (horizontal axis). Note that Nth login with the same password is not the same as the Nth login to an account, as a password can be reset and the participant can restart the practice with the new password for the same account. We only concern about the practice on the same password in this case.

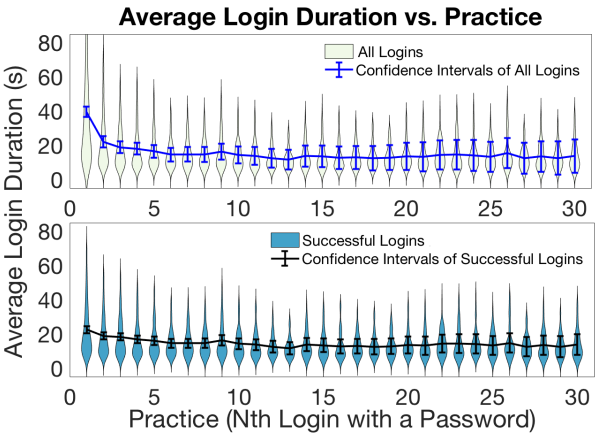

Figure 6: Average login duration vs. practice for all logins (upper figure) and only successful logins (lower figure). The confidence interval (CI) for each mean is shown as a vertical bar in the data distribution. Two figures also show pairwise comparison for different practice values. If the CIs of two groups do not overlap, they are statistically significantly different.

5.3 Secure Passwords Are Less Memorable

Table 3 shows that the average login duration (for both successful and all logins) increases when the password strength increases from 1 to 4 and the differences are statistically significant (see confidence intervals in the table). The group with password strength equal to 0 has very small sample size to draw meaningful conclusion (recall Table 1 that only 2% of passwords have score 0 compared to other groups that all have above 20% of passwords). We did not find any interesting relationship between recall odds and password strength estimated by zxcvbn (see Table 3 for recall odds).

After performing rule-based dictionary attacks to our collected password set, we found that with four password cracking rules we applied (best64 with $3x10^9$ guesses, generated2 with $2x10^{12}$ guesses, rockyou-3000

Password Strength (zxcvbn)		0	1	2	3	4
Login Success Rate		0.865	0.934	0.911	0.921	0.896
Recall Odds		6.40	14.24	10.17	11.73	8.57
Login Duration (Mean, 95% CI)	All Logins	20.25 (16.49, 24.00)	15.65 (14.95, 16.35)	17.92 (17.20, 18.63)	19.12 (18.50, 19.75)	20.84 (20.05, 21.64)
	Successful Logins	15.15 (12.18, 18.12)	13.78 (13.26, 14.31)	15.58 (15.04, 16.12)	16.81 (16.33, 17.28)	18.03 (17.43, 18.63)

Table 3: Results of login success rates, recall odds, and login duration (means and confidence intervals) when the logins were grouped based on password strength. Each password strength was estimated by zxcvbn password meter: 0 (too weak), 1 (very weak), 2 (medium), 3 (strong), and 4 (very strong).

with $1x10^{12}$ guesses, and incisive-leetspeak with $6x10^{11}$ guesses), the recall odds for cracked passwords are higher than the recall odds for uncracked passwords (see the upper figure in Figure 7). In addition, Figure 7 shows that the passwords that are easier to recall are also less secure under rule-based dictionary attacks.

We applied a neural network model [58] to further estimate our passwords. We found that recall odds decrease as the number of guesses increases (see the upper figure in Figure 8) and the average successful login duration increases as the number of guesses increases (see the lower figure in Figure 8). This means that the more secure passwords are with neural network model are also less memorable and need significantly more time for entry. We only analyzed data with password length equal to or greater than 8 because the pre-trained neural network model does not provide estimation for passwords with length shorter than 8 [58]. The figure plots the logarithm of the number of guesses with base 10. The grouping was done after splitting the number of guesses into five intervals (10^0-10^6, 10^6-10^{12}, $10^{12}-10^{18}$, $10^{18}-10^{24}$, and $10^{24}-10^{30}$). We split the range into five intervals because it is the largest number of intervals to guarantee that each interval has at least ten different passwords (e.g. 0-6, 6-12, 12-18, and 18-24 in Figure 8). Although the average login duration for the last group (24-30) is smaller than the previous group (e.g. 6-12), the difference is not statistically significant (confidence interval is very large for 24-30 and it overlaps with 6-12, 12-18, and 18-24).

5.4 Account Types Do Not Affect Memorability

We found that the average successful login duration for financial accounts is statistically significantly longer than the ones for content accounts and advertisement accounts (see comparison in Table 4). Financial accounts and identity accounts have lower recall odds than content accounts and advertisement accounts. Overall, the differences of recall odds and average login durations for different account types are small.

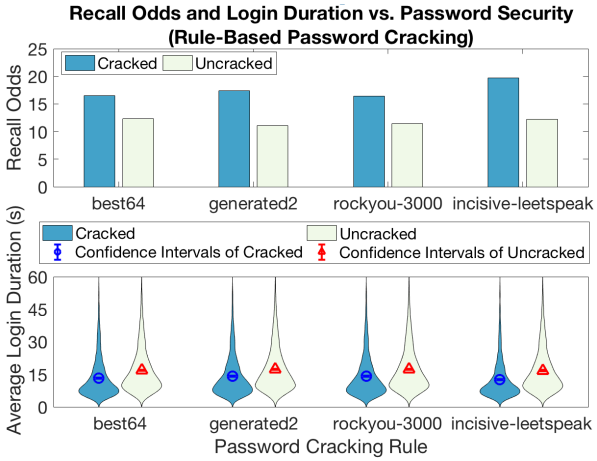

Figure 7: Recall odds (upper figure), and data distribution of successful login duration with their averages (lower figure) vs. password crackability using different rules: best64 cracked 22.6% of passwords, generated2 cracked 37.0%, rockyou-3000 cracked 37.4%, and incisive-leetspeak cracked 15.7%. Recall odds were calculated using login success rates. A login success rate was the total number of successful logins from cracked (or uncracked) passwords divided by the total number of logins from these cracked (or uncracked) passwords. The lower figure shows distributions of login durations along with their means (circles and triangles) and 95% confidence intervals (vertical bars) for cracked and uncracked password groups.

Account Types	Financial Accounts	Identity Accounts	Content Accounts	Ad. Accounts
Recall Odds	10.4	9.1	13.9	12.2
Login Duration (Mean, 95% CI) Successful Logins	16.82 (16.33, 17.31)	15.97 (15.50, 16.43)	15.76 (15.28, 16.24)	15.36 (14.90, 15.83)

Table 4: Recall odds and successful login duration for different types of accounts. For successful login duration, means and 95% confidence intervals are shown in the table.

5.5 Model Fitting

Login frequency and practice have similar mathematical functions that fit well with our data (see Figure 2, Figure 5 and their fitting results). In addition, we show that password security has a very interesting effect on the password memorability (see Figure 7 and Figure 8). However, given that there is no existing work proposing any functional relationship between memorability and password security and the current quantitative measurement of password security is highly dependent on the password attacking al-

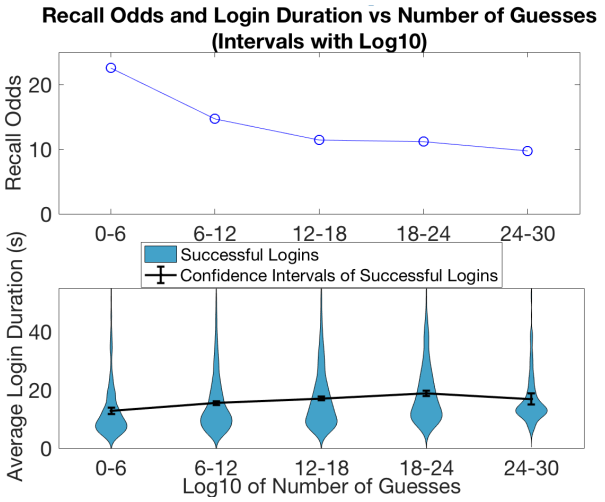

Figure 8: Recall odds (upper figure) and distributions of their successful login duration with their average (lower figure) vs. number of guesses. The number of guesses, based on the neural network estimator [58], is in logarithmic scale with base 10. The lower figure shows both means and 95% confidence intervals for different groups of passwords with different numbers of guesses. The difference of means from any pair of groups among 0-6, 6-12, 12-18, and 18-24 is statistically significant (i.e. their confidence intervals do not overlap). The mean from the last group 24-30 is only statistically significantly higher than the first group 0-6.

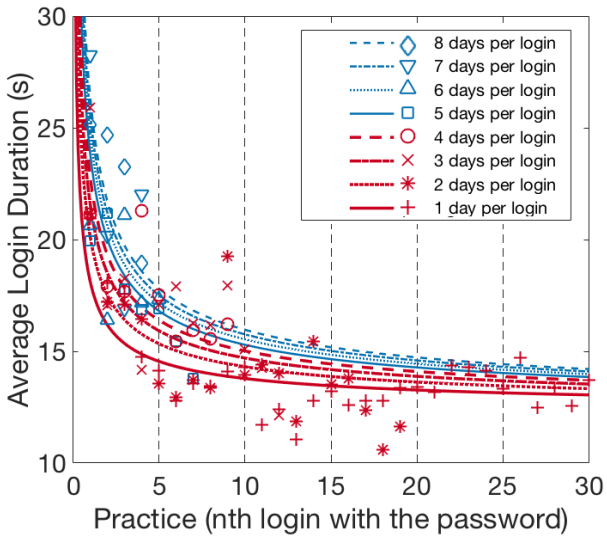

Figure 9: Average successful login duration for different values of login frequency and practice. The fitting curves based on the derived equation are shown as lines in the figure. We can see the curves shifting upward as frequency changes from 1 to 8 days per login.

gorithm (e.g. neural network estimator does not estimate passwords with length shorter than 8, different rule-based cracking gives different number of guesses for the same password), our mathematical model only combines the effect of password login frequency and practice.

5.5.1 Average Successful Login Duration

Figure 9 shows the fitting of our data to the derived equation (Equation 9):

$$E[Time_{login}] \approx \frac{Kf^d(1-d)}{n^{1-d}} + E[Time_{act}]$$

The fitted parameter values are $d = 0.4213$, $K = 10.21$, and $E[Time_{act}] = 12.23$. The memory decay parameter d is dependent on the specific application. Previous research has suggested that the value of d is near 0.5 for many applications [4], which matches our result. The fitted value of $E[Time_{act}]$ also appears to be reasonable because we can see average login duration stabilize near 12 seconds at the end (see Figure 6). Figure 9 shows that our data generally follows the fitted function curves of different login frequencies and the fitting curves shift upwards as the login frequencies changes from 1 to 8 days per login. It makes sense since the people should spend more time on recalling their passwords if the passwords are less frequently used. We found that the data points for some specific login frequencies do not fit into the fitting curve with optimal parameters. The main reason is that the parameters in Equation 9 are optimized based on the error between the data points of all login frequencies and their corresponding fitting curves. Most of the observation points lie on the early part of the x-axis. With our memorability model, we obtained very small root mean square error of 1.8 seconds for a successful login duration (see Figure 6 for the data range of login duration).

5.5.2 Recall Odds

Our model for recall odds yields Equation 11:

$$R_{o\_Measured} \approx Af^{-d/s}n^{(1-d)/s}$$

Note that we have already obtained the value of d through fitting the login duration function ($d = 0.4213$). It is the same d in this equation as is derived from the same activation function. Therefore, A and s parameters need to be fitted from our data. Obtaining value d from previous fitting makes the fitting of this complicated function feasible. It is challenging to fit both s and d unknown since s and d have ratio relationship within the power term. In addition, A is related to d ($A = e^{-\tau/s}(1-d)^{-1/s}E[e^{C/s}]$), making fitting even more challenging if d is unknown.

Equation 11 can produce an infinite value when a recall is perfect at certain combination of login frequency and practice values. As computation and fitting do not work well with infinite values, we need to take the reciprocal of the measured recall odds for function fitting and plotting. Therefore, we transform to following equation:

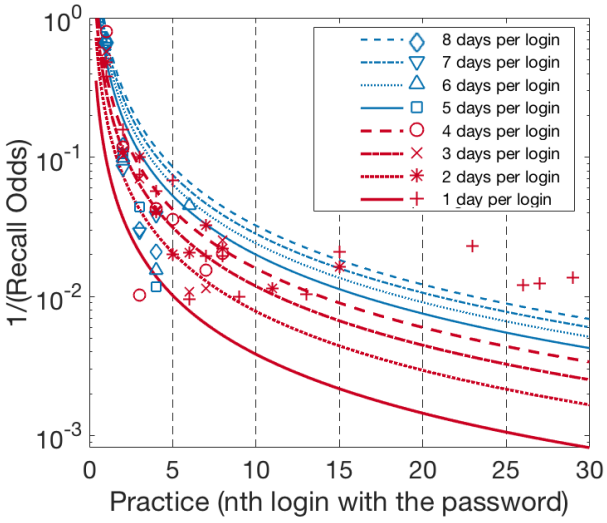

Figure 10: Reciprocal of measured recall odds for different values of login frequency and practice. The best fitted curves based on the derived equation are shown as lines in the figure. We can see the curves shifting upward as frequency changes from 1 to 8 days per login.

$$\frac{1}{R_{o\_Measured}} \approx \frac{1}{A} f^{d/s} n^{(d-1)/s} \tag{13}$$

Figure 10 shows the fitting of the reciprocal of recall odds to our study data. The best fitted parameter values based on our data are $1/A = 0.0980$ and $s = 0.4113$. d is 0.4213 as the memory decay parameter. Our data follows nicely with the fitted curves in Figure 10 and the curves shifting upward with the frequencies changes from 1 to 8 days per login for the same reason of Figure 9. Observe that the data of 1 day and 2 days per login do not fit well in the latter logins. This is mainly because we used Log scale to be able to visualize the fitting. We obtained a relatively small root mean square error of 0.0868 for the reciprocal of recall odds (the data range is about 0 to 1).

5.6 Survey Responses

In this section, we present the major findings from entry and exit surveys.

5.6.1 Online Account Usage

We found that the total number of online accounts from our participants ranged from 2 to 50 with mean of 13 accounts. For the most frequently used account, 73% of our participants logged in several times per day, 18% logged in once per day, and 9% logged in once per week or less. For the least frequently used account, 16% of participants logged in only once per years or less, 31% logged in several times per year, 28% logged in once per month, and 25% logged in once per week or more.

We analyzed participants' survey responses against their task performance during our study and found that participants having more accounts in their daily life performed better in our study tasks (successful recall rate 0.93 vs. 0.89 with $p < 0.0001$). The comparison was done by grouping our participants based on the total number of online accounts they had in their daily life (i.e. one group had fewer than 13 accounts, and one group had at least 13 accounts, given that the average was 13).

We asked survey questions about the importance of different online accounts in their daily life and found the order of importance to be banking accounts, email accounts, social networking accounts, shopping accounts, music streaming accounts, daily deals accounts, news accounts, and coupon accounts. In these questions, we asked participants to rate the importance of accounts with five levels: very important, important, neutral, not important and not important at all. The ranking was based on participants rating. For example, for banking accounts, 77% of our participants considered them very important and 20% considered them important. For email accounts, 51% considered them very important. For social networking accounts, 18% considered them very important. Other accounts were ranked in the similar way and they had less than 10% considering them as very important. We asked these questions in the exit survey to avoid introducing bias to our study data.

5.6.2 Password Usage and Management

From survey responses, the total number of different passwords ranged from 1 to 20 with an average of 5.8 different passwords. 91% of our participants reused at least one of their passwords for different accounts. We asked a question about the total number of different passwords that they memorized without the need to check notes or use password managers. We participants' responses ranged from 1 to 12 with an average of 4.6 (or 5) different memorized passwords.

Due to forgetting, 30% of participants had to reset passwords a few times in past years, 59% of them reset passwords about once or several times per year, 9% of them reset once per month, and 2% of them reset more than once per month.

Table 5 shows our participants' responses on password management. More than half of our participants wrote down some of their passwords in their daily life. While only 10% of our participants used dedicated password managing software, 73% of our participants used password saving feature in the browser.

Based on responses, the major reason for writing down passwords and using password saving features was to

Password management	Yes	No
Write down any password in daily life?	57%	43%
Use any dedicated password manager in daily life?	10%	90%
Use browser password saving feature in daily life?	73%	27%

Table 5: Table shows distribution of participants' response on password management survey questions.

prevent forgetting. When asked whether there was any concerns or disadvantages of using the password saving feature or the password manager, participants mostly mentioned: (1) risks of getting hacked (e.g. "I think it is subject to hacking", "security and privacy issues", "if the database of the password manager is leaked, then hackers have the access to all of the passwords I use."), (2) concern about device or software sharing (e.g. "people who have access to my browser will also be able to login into the websites.", "someone using my device can log into my accounts"), (3) lost of practice (e.g. "using it does not force me to commit the password to memory"), and (4) concern about accidental password loss and software failure (e.g. "if the password history gets cleared it might be hard to recall the password", "you will lose all of them if it fails or they get erased for some reason").

After the study, we asked participants to share whether they had used any browser password saving feature or written down any password during our study and mentioned that their response would not affect their compensation. As our study focuses on memorization, using these questions, we could have removed participants if they largely relied on writing down passwords for our tasks. None of the participants shared that they were writing down passwords or using password saving features.

5.6.3 Password Memorization Strategies

We asked our participants how they memorized their passwords and their strategies. Based on their responses, the major strategies included: (1) creating passwords with certain pattern or meaning such as inclusion of phases, names, familiar items, school names, and dates (e.g. "[use] family, school, personal information", "passwords have a certain pattern or a year that corresponds to the current year", "words or numbers that have meaning"), (2) memorizing based on keyboard layout (e.g. "memorizing keyboard layout (the way I press the key in a certain order)"), (3) recalling the password frequently (e.g."use it again and again and I'll remember them naturally"), (4) associating it with the corresponding website (e.g. "I associate each website/platform name with a certain password stored in my memory"), and (5) generating simple passwords (e.g. "make it simple, think of last names of myself and family members", "keep them simple").

After the study, 89% of our participants found that more frequently used passwords were easier to memorize based on the exit survey responses.

We also investigated how memorization strategies can help on task performance. As memory recall is related to contextual associations of the memory element [3, 6], we grouped our data based on whether a participant encoded the contextual information (e.g. account information) while generating the password. We found that there were 297 passwords (out of 1443 passwords – 21%) that contained the account information (i.e. include some parts of the account name or have slight variations, for example, "shoptillyoudrop!" for an online shopping account). The remaining 1146 passwords did not include any information about the account. We found that passwords that were generated with encoding of account information were easier to recall than those without considering the account information (successful login rate 0.94 vs. 0.91 with $p < 0.0001$). This result supports the ecological memory theory that having strong connection between the memory element and the contextual setting helps on memory recalling [3, 6].

6 Discussion

This paper is the first to apply the ecological theory of long-term memory to model the forgetting of passwords. The model is rooted in decades of memory research which were previously applied to memory of emails and newspaper articles in psychology [18, 9, 8, 4]. It predicts recall odds and login duration from login frequency and number of logins in the past. In our work, online authentication with text passwords, the model predicted successful login duration with RMSE of 1.8 seconds and recall odds with RMSE of 0.0868 (for the reciprocal of recall odds). We consider this a very promising first result and supportive of the tenet of studying password use from the perspective of ecological view of memory.

At a theoretical level, the finding points to a new understanding of passwords. What makes passwords hard to remember is not their complexity per se, but the fact that the human memory is opportunistic in what it attempts to remember or to forget. Instead of looking at the password itself, we need to look at the environment in which it is used. The more important a password is to the user, and the more it is likely to be used in the future, the higher the chances of recalling it.

The finding and the model have direct practical use. The model can be used to obtain a reasonable estimation on the probability of password forgetting given its use. To mitigate password forgetting, system designers and security engineers can provide guidelines emphasizing

the importance of memory practice for a new password. In some cases, high-value account services could use our model to control when to ask for user logins. Increased frequency of password usage improves probability of remembering the password and reduces the need for users to generate weak passwords for important accounts.

While login frequency may be straightforward to identify empirically, how about organismic importance? Our participants' survey responses tentatively suggest that financial (e.g. banking and shopping) and identity (e.g. email and social networking) accounts are more important than content and advertisement accounts. Interestingly, recall odds for financial and identity accounts were slightly lower than content and advertisement accounts. Participants also appeared to take more time to recall passwords for financial and identity accounts than content and advertisement accounts even if the only difference in our study was the decoration of the login screen. This indicates that password memorability is better for less important accounts than for more important accounts.

However, the difference of recall odds and mean login durations for different account types was small (see Section 5.4). This means that the effect of account types on memorability is much smaller than the other controlled variables such as login frequency, practice, and password security. There are indeed two possible ways that the account importance can affect memorability: (1) users create very secure passwords for important accounts and these passwords are harder to remember than the ones created for less important accounts; (2) users spend more effort generating passwords for important accounts, resulting passwords for important accounts better memorized (*depth of processing theory* [20]).

We also learned that most participants were capable of memorizing their passwords in their daily lives but still chose to write down passwords or use password saving features to prevent forgetting. Note that the participants shared that they did not write down passwords during our study (see Section 5.6.2) Based on our results, the average number of total passwords (5.8) is only slightly larger than the average number of memorized passwords (4.6). This indicates that most participants were able to memorize most of their passwords. Outside of our study, the participants reported that 57% of them still chose to write down their passwords and 73% of participants chose to use browsers to save passwords during their daily password management. When asked about it, the major reason was to prevent forgetting. Therefore, the major cause of writing down passwords could be participants' false belief that they were not able to remember passwords or their overestimation of the password resetting effort.

In addition to login frequency and practice, we found that password security has an independent effect on password memorability. For example, passwords with higher zxcvbn score have somewhat longer average successful login duration. Although we did not find recall odds to follow an interesting pattern with zxcvbn scores, results from more dedicated password cracking and neural network password estimators both showed that recall odds drop when passwords are more secure (see Section 5.3).

Although past studies have mentioned that very secure passwords can be hard to remember [50, 83, 74], the results reported here show it with a dedicated experimental study using state-of-the-art password crackers and estimators. However, this does not mean that all secure passwords are hard to remember. There are existing studies providing good strategies on creating both memorable and secure passwords [80, 83].

Limitations: Similar to other password studies, a few limitations must be considered in interpreting our findings. Our participants were mostly young adults with a mean age of 24. Second, we cannot directly collect participant's actual passwords for their actual online accounts. Therefore, similar to other password studies about online accounts, our study is based on researcher-designed online accounts which may not align with the real-world importance of these accounts to participants. However, with our careful study design and special consideration for ecological validity in each step, we have ensured our design to match as closely as possible to the daily online account usage.

7 Conclusions

In this paper, we explored and analyzed how account type, login frequency, amount of practice, and password security can affect password memorability. We combined login frequency and amount of practice to construct a model that can predict successful login duration and recall odds in an understandable mathematical form derived from major memory theories. Our data largely shows that human memory of passwords follows the ecological theory of memory. Importantly, our finding points to a new understanding of password forgetting: instead of looking at the password itself (e.g. password complexity), we need to consider the environment in which it is used and how memory functions over time. Compared to solely statistical group comparisons, our modeling approach provides quantitative predictions that can be directly applied by designers and can transform the knowledge in the field to an actionable form.

In addition, the study shows that when participants were allowed to self-generate passwords (which is how current online authentication systems work), password security can affect password memorability: stronger passwords were harder to remember. This shows that our participants have not mastered password generating strategies to generate both secure and memorable passwords.

In addition, based on our results from survey data, we found that most participants were capable of memorizing their passwords during their daily lives but still chose to write down or save passwords to prevent forgetting.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Numbers 1228777 and 1750987. Xianyi Gao was supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number 1433187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional material available at `http://scienceofsecurity.science`.

References

[1] ADAMS, A., SASSE, M. A., AND LUNT, P. Making passwords secure and usable. In *Proceedings of HCI on People and Computers XII* (London, UK, UK, 1997), HCI 97, Springer-Verlag, pp. 1–19.

[2] ALKALDI, N., AND RENAUD, K. Why do people adopt, or reject, smartphone password managers? In *1st European Workshop on Usable Security* (2016), vol. 18, pp. 1–14.

[3] ANDERSON, J. R. *Rules of the Mind*. L. Erlbaum Associates, 1993.

[4] ANDERSON, J. R., BOTHELL, D., LEBIERE, C., AND MATESSA, M. An integrated theory of list memory. *Journal of Memory and Language 38*, 4 (1998), 341 – 380.

[5] ANDERSON, J. R., FINCHAM, J. M., AND DOUGLASS, S. Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition 25*, 5 (1999), 1120.

[6] ANDERSON, J. R., AND LEBIERE, C. J. *The atomic components of thought*. Psychology Press, 2014.

[7] ANDERSON, J. R., REDER, L. M., AND LEBIERE, C. Working memory: Activation limitations on retrieval. *Cognitive psychology 30*, 3 (1996), 221–256.

[8] ANDERSON, J. R., AND SCHOOLER, L. J. Reflections of the environment in memory. *Psychological Science 2*, 6 (1991), 396–408.

[9] ANDERSON, J. R., AND SCHOOLER, L. J. The adaptive nature of memory. *The Oxford handbook of memory* (2000), 557–570.

[10] ATKINSON, R., AND SHIFFRIN, R. Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation 2* (1968), 89 – 195.

[11] BAILEY, D. V., DÜRMUTH, M., AND PAAR, C. *Statistics on Password Re-use and Adaptive Strength for Financial Accounts*. Springer International Publishing, Cham, 2014, pp. 218–235.

[12] BARD, G. V. Spelling-error tolerant, order-independent passphrases via the damerau-levenshtein string-edit distance metric. In *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers - Volume 68* (Darlinghurst, Australia, Australia, 2007), ACSW '07, Australian Computer Society, Inc., pp. 117–124.

[13] BLOCKI, J., KOMANDURI, S., CRANOR, L., AND DATTA, A. Spaced repetition and mnemonics enable recall of multiple strong passwords. *arXiv preprint arXiv:1410.1490* (2014).

[14] BONNEAU, J. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy* (May 2012), pp. 538–552.

[15] BONNEAU, J., AND PREIBUSCH, S. The password thicket: Technical and market failures in human authentication on the web. In *WEIS* (2010).

[16] BONNEAU, J., AND SCHECHTER, S. Towards reliable storage of 56-bit secrets in human memory. In *23rd USENIX Security Symposium (USENIX Security 14)* (2014), pp. 607–623.

[17] BROWN, G. D., NEATH, I., AND CHATER, N. A temporal ratio model of memory. *Psychological review 114*, 3 (2007), 539.

[18] BRUCE, D. The how and why of ecological memory. *Journal of Experimental Psychology: General 114*, 1 (1985), 78.

[19] BUNNELL, J., PODD, J., HENDERSON, R., NAPIER, R., AND KENNEDY-MOFFAT, J. Cognitive, associative and conventional passwords: Recall and guessing rates. *Computers & Security 16*, 7 (1997), 629–641.

[20] CERMAK, L. S., AND CRAIK, F. I. *Levels of processing in human memory*. Lawrence Erlbaum, 1979.

[21] CHIANG, H.-Y., AND CHIASSON, S. Improving user authentication on mobile devices: A touchscreen graphical password. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services* (New York, NY, USA, 2013), MobileHCI '13, ACM, pp. 251–260.

[22] CHIASSON, S., FORGET, A., BIDDLE, R., AND VAN OORSCHOT, P. C. Influencing users towards better passwords: Persuasive cued click-points. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1* (Swinton, UK, UK, 2008), BCS-HCI '08, British Computer Society, pp. 121–130.

[23] CHIASSON, S., FORGET, A., STOBERT, E., VAN OORSCHOT, P. C., AND BIDDLE, R. Multiple password interference in text passwords and click-based graphical passwords. In *Proceedings of the 16th ACM Conference on Computer and Communications Security* (New York, NY, USA, 2009), CCS '09, ACM, pp. 500–511.

[24] CHOWDHURY, S., POET, R., AND MACKENZIE, L. Passhint: Memorable and secure authentication. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 2917–2926.

[25] COINBASE. Coinbase zxcvbn, 2016. Retrieved Sep 07 2016 from `https://libraries.io/github/coinbase/zxcvbn`.

[26] DAS, A., BONNEAU, J., CAESAR, M., BORISOV, N., AND WANG, X. The tangled web of password reuse. In *NDSS* (2014), vol. 14, pp. 23–26.

[27] DE CARNÉ DE CARNAVALET, X., AND MANNAN, M. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security (NDSS) Symposium 2014* (February 2014), Internet Society.

[28] DISCOVER. Change my user id or password, 2017. Retrieved April 14 2017 from `https://www.discover.com/credit-cards/help-center/faqs/user-password.html`.

[29] DROPBOX. dropbox/zxcvbn: a realistic password strength estimatior, 2016. Retrieved April 4 2016 from `https://github.com/dropbox/zxcvbn`.

[30] DÜRMUTH, M., AND KRANZ, T. *On Password Guessing with GPUs and FPGAs*. Springer International Publishing, Cham, 2015, pp. 19–38.

[31] EBBINGHAUS, H. *Memory: A contribution to experimental psychology*. No. 3. University Microfilms, 1913.

[32] EDITORIALMANAGER. Password security, 2017. Retrieved April 14 2017 from http://www.editorialmanager.com/robohelp/current/Editorial_Manager_Help/Password_Security.htm.

[33] EGELMAN, S., SOTIRAKOPOULOS, A., MUSLUKHOV, I., BEZNOSOV, K., AND HERLEY, C. Does my password go up to eleven?: The impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), CHI '13, ACM, pp. 2379–2388.

[34] EVERITT, K. M., BRAGIN, T., FOGARTY, J., AND KOHNO, T. A comprehensive study of frequency, interference, and training of multiple graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2009), CHI '09, ACM, pp. 889–898.

[35] FAHL, S., HARBACH, M., ACAR, Y., AND SMITH, M. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (New York, NY, USA, 2013), SOUPS '13, ACM, pp. 13:1–13:13.

[36] FLORENCIO, D., AND HERLEY, C. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web* (New York, NY, USA, 2007), WWW '07, ACM, pp. 657–666.

[37] GAW, S., AND FELTEN, E. W. Password management strategies for online accounts. In *Proceedings of the Second Symposium on Usable Privacy and Security* (New York, NY, USA, 2006), SOUPS '06, ACM, pp. 44–55.

[38] GOODING, S. Ridiculously smart password meter coming to WordPress 3.7, 2016. Retrieved April 1st 2016 from http://wptavern.com/ridiculously-smart-password-meter-coming-to-wordpress-3-7.

[39] GOOGLE. Google ngram viewer, 2013. Retrieved May 07 2017 from http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

[40] HAQUE, S. T., WRIGHT, M., AND SCIELZO, S. Hierarchy of users' web passwords: Perceptions, practices and susceptibilities. *International Journal of Human-Computer Studies 72*, 12 (2014), 860 – 874.

[41] HASHCAT. Hashcat: advanced password recovery, 2017. Retrieved May 07 2017 from https://hashcat.net/hashcat/.

[42] HAYASHI, E., AND HONG, J. A diary study of password usage in daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 2627–2630.

[43] HERLEY, C., AND OORSCHOT, P. V. A research agenda acknowledging the persistence of passwords. *IEEE Security Privacy 10*, 1 (Jan 2012), 28–36.

[44] HOONAKKER, P., BORNOE, N., AND CARAYON, P. Password authentication from a human factors perspective: Results of a survey among end-users. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2009), vol. 53, SAGE Publications, pp. 459–463.

[45] HUH, J. H., KIM, H., BOBBA, R. B., BASHIR, M. N., AND BEZNOSOV, K. On the memorability of system-generated pins: Can chunking help? In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)* (Ottawa, July 2015), USENIX Association, pp. 197–209.

[46] HUH, J. H., KIM, H., RAYALA, S. S., BOBBA, R. B., AND BEZNOSOV, K. I'm too busy to reset my LinkedIn password: On the effectiveness of password reset emails. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 387–391.

[47] INGLESANT, P. G., AND SASSE, M. A. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 383–392.

[48] JAKOBSSON, M., AND AKAVIPAT, R. Rethinking passwords to adapt to constrained keyboards. *Proc. IEEE MoST* (2012), 1–11.

[49] JOTFORM. The easiest online form builder: Jotform support forum, 2017. Retrieved April 14 2017 from https://www.jotform.com/answers/1094950-cannot-log-on-using-the-same-user-and-password-please-advise-is-there.

[50] KLEIN, D. V. Foiling the cracker: A survey of, and improvements to, password security. In *Proceedings of the 2nd USENIX Security Workshop* (1990), pp. 5–14.

[51] KOMANDURI, S., SHAY, R., KELLEY, P. G., MAZUREK, M. L., BAUER, L., CHRISTIN, N., CRANOR, L. F., AND EGELMAN, S. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 2595–2604.

[52] KORELOGIC. Crack me if you can, 2010. Retrieved May 07 2017 from http://contest-2010.korelogic.com/rules.html.

[53] KUO, C., ROMANOSKY, S., AND CRANOR, L. F. Human selection of mnemonic phrase-based passwords. In *Proceedings of the Second Symposium on Usable Privacy and Security* (New York, NY, USA, 2006), SOUPS '06, ACM, pp. 67–78.

[54] LAWLER, J. The web2 file of english words, 1999. Retrieved May 07 2017 from http://wwwpersonal.umich.edu/œjlawler/wordlist.

[55] LEE, J. Forgot a password? try way2many; better online security has meant more passwords, and more frustrated users, new york times, 1999. Retrieved Sep 07 2017 from http://www.nytimes.com/1999/08/05/technology/forgot-password-try-way2many-better-line-security-has-meant-more-passwords-more.html?pagewanted=all.

[56] MCGEOCH, J. A. Forgetting and the law of disuse. *Psychological review 39*, 4 (1932), 352.

[57] MELICHER, W., KURILOVA, D., SEGRETI, S. M., KALVANI, P., SHAY, R., UR, B., BAUER, L., CHRISTIN, N., CRANOR, L. F., AND MAZUREK, M. L. Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2016), CHI '16, ACM, pp. 527–539.

[58] MELICHER, W., UR, B., SEGRETI, S. M., KOMANDURI, S., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. Fast, lean, and accurate: Modeling password guessability using neural networks. In *25th USENIX Security Symposium (USENIX Security 16)* (Austin, TX, 2016), USENIX Association, pp. 175–191.

[59] METEOR. The fastest way to build javascript apps, 2016. Retrieved August 14 2016 from https://www.meteor.com/.

[60] MONCUR, W., AND LEPLÂTRE, G. Pictures at the atm: Exploring the usability of multiple graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 887–894.

[61] NAIRNE, J. S. The myth of the encoding-retrieval match. *Memory 10*, 5-6 (2002), 389–395.

[62] NIELSEN, G., VEDEL, M., AND JENSEN, C. D. Improving usability of passphrase authentication. In *Privacy, Security and Trust (PST), 2014 Twelfth Annual International Conference on* (2014), IEEE, pp. 189–198.

[63] NOTOATMODJO, G., AND THOMBORSON, C. Passwords and perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security - Volume 98* (Darlinghurst, Australia, Australia, 2009), AISC '09, Australian Computer Society, Inc., pp. 71–78.

[64] OPENATHENSMD. About usernames, passwords and expiry dates, 2017. Retrieved April 14 2017 from `https://docs.openathens.net/display/public/MD/About+usernames\%2C+passwords+and+expiry+dates`.

[65] PEARMAN, S., THOMAS, J., NAEINI, P. E., HABIB, H., BAUER, L., CHRISTIN, N., CRANOR, L. F., EGELMAN, S., AND FORGET, A. Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2017), CCS '17, ACM, pp. 295–310.

[66] RAO, A., JHA, B., AND KINI, G. Effect of grammar on security of long passwords. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy* (New York, NY, USA, 2013), CODASPY '13, ACM, pp. 317–324.

[67] REDER, L. M. *Implicit memory and metacognition*. Psychology Press, 2014. 53.

[68] SALDAÑA, J. *The coding manual for qualitative researchers*. Sage, 2015.

[69] SALEFORCE. Passwords, 2017. Retrieved April 14 2017 from `https://help.salesforce.com/articleView?id=security_overview_passwords.htm&language=en&type=0`.

[70] SCHEFF, H. The relation of control charts to analysis of variance and chi-square tests. *Journal of the American Statistical Association 42*, 239 (1947), 425–431.

[71] SHAY, R., BAUER, L., CHRISTIN, N., CRANOR, L. F., FORGET, A., KOMANDURI, S., MAZUREK, M. L., MELICHER, W., SEGRETI, S. M., AND UR, B. A spoonful of sugar?: The impact of guidance and feedback on password-creation behavior. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 2903–2912.

[72] SHAY, R., KOMANDURI, S., DURITY, A. L., HUH, P. S., MAZUREK, M. L., SEGRETI, S. M., UR, B., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 2927–2936.

[73] STOBERT, E., AND BIDDLE, R. Memory retrieval and graphical passwords. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (New York, NY, USA, 2013), SOUPS '13, ACM, pp. 15:1–15:14.

[74] STOBERT, E., AND BIDDLE, R. The password life cycle: user behaviour in managing passwords. In *Proc. SOUPS* (2014).

[75] STRIPE. Stripe account registration, 2016. Retrieved Sep 07 2016 from `https://dashboard.stripe.com/register`.

[76] THELOCAL.DE. 'Forgot password' resets cost VW 1 million a year, 2016. Retrieved Sep 07 2016 from http://www.thelocal.de/20151211/forgot-password-resets-cost-vw-1-million-per-year.

[77] UR, B., ALFIERI, F., AUNG, M., BAUER, L., CHRISTIN, N., COLNAGO, J., CRANOR, L. F., DIXON, H., EMAMI NAEINI, P., HABIB, H., JOHNSON, N., AND MELICHER, W. Design and evaluation of a data-driven password meter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), CHI '17, ACM, pp. 3775–3786.

[78] UR, B., KELLEY, P. G., KOMANDURI, S., LEE, J., MAASS, M., MAZUREK, M. L., PASSARO, T., SHAY, R., VIDAS, T., BAUER, L., CHRISTIN, N., AND CRANOR, L. F. How does your password measure up? the effect of strength meters on password creation. In *Proceedings of the 21st USENIX Conference on Security Symposium* (Berkeley, CA, USA, 2012), Security'12, USENIX Association, pp. 5–5.

[79] UR, B., SEGRETI, S. M., BAUER, L., CHRISTIN, N., CRANOR, L. F., KOMANDURI, S., KURILOVA, D., MAZUREK, M. L., MELICHER, W., AND SHAY, R. Measuring real-world accuracies and biases in modeling password guessability. In *24th USENIX Security Symposium (USENIX Security 15)* (Washington, D.C., 2015), USENIX Association, pp. 463–481.

[80] VU, K.-P. L., PROCTOR, R. W., BHARGAV-SPANTZEL, A., TAI, B.-L. B., COOK, J., AND EUGENE SCHULTZ, E. Improving password security and memorability to protect personal and organizational information. *Int. J. Hum.-Comput. Stud. 65*, 8 (Aug. 2007), 744–757.

[81] WHEELER, D. L. zxcvbn: Low-budget password strength estimation. In *25th USENIX Security Symposium (USENIX Security 16)* (Austin, TX, 2016), USENIX Association, pp. 157–173.

[82] YAHOO. Get message your password cannot include your name or username, 2017. Retrieved April 14 2017 from https://forums.yahoo.net/t5/Password-and-sign-in/Get-message-quot-your-password-cannot-include-your-name-or/td-p/4988.

[83] YAN, J. J., BLACKWELL, A. F., ANDERSON, R. J., AND GRANT, A. Password memorability and security: Empirical results. *IEEE Security & privacy 2*, 5 (2004), 25–31.

[84] YANG, Y., LINDQVIST, J., AND OULASVIRTA, A. Text entry method affects password security. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2014)* (2014).

A Pre-study (or Entry) Survey Questions

A.1 Demographic Information

1. What is your email address?
2. What is your gender?

   [Options: • Male, • Female]
3. What is your age?
4. Which of the following best describes your primary occupation?

   [Options: • Administrative support, • Art, writing, or journalism, • Business, management, or financial, • Legal e.g. lawyer, • Medical, • Engineering or IT professional, • Service, • Skilled labor, • Unemployed, • Retired, • College (undergraduate) student, • College (graduate) student, • Other, • Prefer not to share]

A.2 Online Accounts and Password Management

1. How many personal online accounts do you have in total? (You may count and add up the number of accounts in each category to get the total.)

2. In your daily life, do you reuse your passwords across different accounts? (Password reuse means using the same password for different accounts.)

   [Options: • Yes, • No]

3. In your daily life, for accounts you have, how many DIFFERENT passwords do you use? (You may write down the password for each account by yourself to help counting. Do not write your passwords in this answer. Please only indicate the number of different passwords.)

4. How often do you reset your passwords because of forgetting?

   [Options: • Several times per day, • About once per day, • About once per week, • About once per month, • Several times per year, • About once per year, • About a few times in past years, • Never]

5. In your daily life, how frequent do you log into your MOST-frequently-used account?

   [Options: • Several times per day, • About once per day, • About once per week, • About once per month, • Several times per year, • About once per year, • About a few times in past years, • Never]

6. In your daily life, how frequent do you log into your LEAST-frequently-used account?

   [Options: • Several times per day, • About once per day, • About once per week, • About once per month, • Several times per year, • About once per year, • About a few times in past years, • Never]

7. Do you use password saving feature in the browser to help you remember passwords?

   [Options: • Yes, • No]

8. Do you use any dedicated password manager software to help you remember passwords?

   [Options: • Yes, • No]

9. If you use any password manager or password saving feature, what are the advantages of using it?

10. If you use any password manager or password saving feature, what are the disadvantages of using it?

11. Do you write down (or type down) your passwords in a certain place?

    [Options: • Yes, • No]

12. How many passwords do you memorize? (without the need to check notes or using password manager)

13. If you memorize passwords, what is your strategy to help memorizing?

B Post-study (or Exit) Survey Questions

B.1 Importance of Online Accounts in Daily Life

Following are 5-point Likert scale questions with options: • 1: Not important at all, • 2: Not important, • 3: Neutral, • 4: Important, • 5: Very important

1. How do you rate the importance of online banking accounts?

2. How do you rate the importance of email accounts?

3. How do you rate the importance of shopping accounts?

4. How do you rate the importance of social networking accounts?

5. How do you rate the importance of news accounts?

6. How do you rate the importance of music accounts?

7. How do you rate the importance of coupon recommendation accounts?

8. How do you rate the importance of deal recommendation accounts?

B.2 Our Study and Passwords

1. During our study, did you write down any of the passwords so you could remember them better? (There are no consequences for you if you did this)

   [Options: • Yes, • No]

2. During our study, did you use a password manager to save the passwords for you? (There are no consequences for you if you did this)

   [Options: • Yes, • No]

3. During our study, did you allow web browsers to save the passwords for you? (There are no consequences for you if you did this)

   [Options: • Yes, • No]

4. In the study, did you find more frequently used passwords were easier to memorize?

   [Options: • Yes, • No]