How Users Choose and Reuse Passwords

Ameya Hanamsagar, Simon S. Woo, Christopher Kanich and Jelena Mirkovic

***Abstract*—Weak or reused passwords are guilty for many contemporary security breaches. It is critical to study both how users choose and reuse passwords, and the causes that lead users to adopt unsafe practices. Existing literature on these topics is limited as it either studies patterns but not the causes (using leaked or contributed datasets), or it studies artificial patterns and causes that may not align with the real ones (lab interviews and/or fictional servers). Our research complements the existing works by studying the semantic structure, strength and reuse of real passwords, as well as conscious and unconscious causes of unsafe practices, in a population of 50 participants. The participants took part in a carefully designed, ethical and IRB-approved lab study, where we harvested their existing online credentials, and interviewed them about their password strategies and their risk perceptions. We found that: (1) an average password is weak and used at more than four sites, (2) important-site passwords are only 1-2 characters longer and 10 times stronger than those for non-important sites, (3) main causes of weak passwords are security fatigue and short password length, (4) 98% of users reuse their passwords with no changes and the rest make slight changes, which can be easily brute-forced, (5) 84% of users reuse passwords between important and non-important sites, and (6) main causes for password reuse are misconceptions about risk, and preference for memorability over security.**

I. Introduction

We know that the current advice—to create a strong, unique password for every online account—is unreasonable. Users have many online accounts, and human memory is limited and ill-suited to remember many different, unrelated, complex passwords. When this policy is inevitably broken by users, what do they do instead? How do users balance security, memorability, and convenience? A perfect storm of vulnerability exists at the intersection of extensive password reuse, the hundreds of millions of passwords breached at sites like Yahoo [14], and the attacks which exploit password reuse to gain access to sensitive data or capabilities [1], [3], [11].

Investigating these factors in a lab setting is difficult: one must both deeply analyze the real, everyday passwords of multiple users, as well as the attitudes, strategies, and concerns of these users with respect to their security. Creating a better understanding of the security behaviors that people currently follow, will help researchers design password policies or alternative authentication methods that are better aligned with user capabilities.

We start our paper by reviewing extensive literature on password practices in Section II. Many researchers have studied password choices and reuse in leaked datasets [6], [31], [16], [27], [2]. While this data is valuable to show wide-spread user trends, it cannot be used to understand why users make unsafe choices, and thus cannot inform interventions. Other researchers have studied password choices in a lab setting [25], [22], where participants created and narrated their choice of passwords for fictional servers. But users may be biased towards security in these are artificial conditions, and may exhibit and describe behaviors that do not align with their actual practices. Yet other researchers [8], [30] studied users' login behaviors by monitoring authentication attempts via browser plugins. While these approaches enable study of real user login behaviors, they do not shed light on causes of these behaviors.

A. Open Questions

While the existing literature on password practices shed light on frequency of unsafe practices, there are several important yet challenging questions that remain open:

1) What are the causes for weak and reused passwords? Are users uninformed about what they should do? Do they underestimate risk from attacks? Or do they simply not care? Answers to these questions can help focus interventions and user education, as well as research into more usable authentication methods.
2) How prevalent is *password versioning*, where a password is reused in a slightly changed form? How do users version passwords? Answers to these questions can help focus user education and inform more realistic attack models.
3) How well do users understand their password practices? Does a user perception of a site's value to them influence their password practice at the site, or their sharing of passwords? Answers to these questions can help create password management tools and strategies that build on users' existing reasoning, thus increasing chances of adoption.

Studying these open questions is hard, as it requires access to real passwords, along with interviews with their owners. It is challenging to design a study that produces such data without jeopardizing user safety.

B. Our Contributions

Our paper seeks to answer the remaining open questions we have listed above. Toward this goal we carefully designed a user study where participants were asked to log into sites where they already have an account, and then explain their password choices, as well as answer a number of questions about the site's value to them, and their risk perception. We also captured and analyzed user passwords to evaluate how well user perceptions match the reality. Because this study deals with real passwords, it was very important to design it in a way that fully protects user identity and password data, while preserving necessary information to answer our research questions.

Our first contribution lies in our novel study methodology that achieves these goals. We extract relevant password features, such as semantic structure, length and strength, and then transform the original password using a consistent but irreversible mapping of semantic segments. We then store the password features and the transformed password, and forget the original password. These actions, together with no user-identifying information, enable us to study password structure, strength and reuse patterns while keeping study participants safe. We describe our study design in detail in Section III. The study was reviewed and approved by our Institutional Review Board (IRB).

Our second contribution lies in our findings, which we present in Section IV. Although significant technical and policy efforts have been made to improve user awareness and enforce strong passwords, we find that: (1) a median user has 40 online accounts, (2) most passwords are very weak and could be brute-forced in a day or less, (3) passwords for sites that are important to a user are only 1-2 characters longer and up to 10 times stronger than those for non-important sites, (4) main causes of weak passwords are security fatigue and weak password-length policies, (5) password reuse is rampant and indiscriminate; 98% of users reuse their passwords verbatim and 84% reuse an important password at a non-important, and likely less secure site, (6) main causes for password reuse are poor understanding of risk and preference for memorability over security.

Our final contribution lies in our suggestions for user interventions and authentication strategies, which are motivated by our findings and presented in Section V. Users need help from password assistants (browsers and password managers) to better understand their password habits. They need suggestions for slight improvements that align with user cognitive abilities. Sites need better password-length policies and better password meters.

II. Background and Related Work

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

A. Password Reuse

Florencio et al. [8] conducted a large scale password reuse study 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 of which is shared across 3.9 sites. Their results showed that the large number of weak passwords were heavily re-used. This study was conducted in 2006, thus the number of online accounts and user password creation and reuse behaviors may have changed. Our study provides updated estimates of 40 online accounts per user.

Recently, Walsh et al. [30] 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 passwords were more frequently reused. However, they do not provide details on the reused password structures and their study cannot detect password versioning, nor collect information about causes of password reuse. Our study contributes these findings. They further found that a median user has 16.5 online accounts, while we find he has 40. Walsh's study discovers only those accounts that a user accesses frequently, while we also discover rarely and never-accessed accounts. These accounts pose a significant risk to the user as she is not aware of them, but they still can be used to harvest her passwords.

In a lab study, Ur et al. [25] examined password behaviors of 49 users, creating accounts at three fictitious servers – a bank, an e-mail and a news server. They found that password reuse is common, and that users are not good at making value decisions about their online accounts. Due to fictitious nature of accounts, Ur et al. were able to collect verbatim passwords and examine them for versioning. They found that users had serious misconceptions that making minor or incremental additions to dictionary words would result in secure passwords. We validate their findings on our real-passwords dataset. We further investigate other causes of weak and reused passwords in addition to users' misconceptions about password composition. We find that misconceptions about password length, and about the interplay between random-content and length are the main cause of weak passwords among our participants.

In addition, when users are required to adopt a new password composition policy for a university account, Shay et al. [18] showed that more than half of participants reported that they either modified an old password or reused it verbatim. Our findings put the number of users that reuse passwords at 98%, 100% if we count similar passwords as reused.

B. Password Structures

Several articles [16], [26], [12], have pointed out that the people use predictable password patterns when creating multiple passwords. Weir et al. [31] formally modeled and analyzed passwords using probabilistic context-free grammars (PCFG). Veras et al. [27] investigated the semantic patterns in the millions of cracked passwords using WordNet, Part-of-Speech (POS) tagger, and other natural language processing (NLP) tools. We use their tool to semantically transform collected passwords to detect and analyze password versioning by each participant.

Bonneau [2] demonstrated that many passwords are vulnerable to *statistical guessing* due to use of dictionary words or popular password strings. Additionally, self-reported user behaviors from [18] showed that nearly 80% of users based their password on a word or a name, with special characters added at the beginning or at the end.

While we similarly observed that many users started with a dictionary word and make simple changes to it, we focus on understanding per-participant habits, and not on common passwords across participants.

C. Password Reuse Attack

Das et al. [6] examined how people reuse and mangle passwords based on leaked password datasets. The limitation of their study is that 97.75% of reused password pairs are only reused across two sites. Thus they cannot analyze reuse across many sites, by a single participant, while we can. Their study estimated that 43-51% of users reuse the same password across multiple sites, while we find that 98% of our participants do so. They further developed the first cross-site password-guessing algorithm, which can guess 30% of transformed passwords within 100 attempts. Wang et al. [29] developed an effective approach for targeted online guessing, which exploits a leaked victim's password along with other personal information. Such attack achieves success rates as high as 73% within 100 guesses. These findings quantify the great dangers of password reuse.

D. User's Perception on Passwords

Creese et al. [5] examined the relationship between perceptions of risk associated with online tasks and password choice. They explored how individual's perceptions of risk vary depending on whether the password user is a security expert or not, and whether they have experienced some form of attack in the past. In our work, we repeat their approach to investigate link between user risk perception and password strength, but find no such correlation.

Furthermore, Ur et al. [24] investigates the relationship between users' perceptions of the strength of specific passwords and their actual strength. They found out that participants had serious misconceptions of using common phrases as base form for their passwords Also, they showed that there was large variance in participants' understanding of how passwords may be attacked. While ideally users would use more random content in their passwords, realistically this conflicts with user need for memorable passwords. Thus we must find strategies that better align with user cognitive abilities.

E. On Password Re-use Strategy

Florencio et al. [9] explored a game-theoretical approach to manage a portfolio of passwords, given a limited user memory. They outlined an optimal password-sharing strategy as follows: (1) group strong passwords within accounts with high value and low probability of compromise; and (2) group weak passwords within accounts of low value and high probability of compromise. In this work, we adopted "important" and "non-important" tags for these two categories of accounts, and analyzed user sharing strategies. We found that users share both within and across important/non-important categories. Based on our findings, we propose a different password sharing strategy, which may be better aligned with users' current practice.

Stobert and Biddle [22] investigated how users keep track of many accounts and passwords, by conducting a series of interviews with 27 participants. They showed that most of users reused passwords by writing them down and almost all of them reported using password managers. They reported that participants have their own personal model of account groups (semantic or thematic grouping) and apply different password strategies to different groups. We complement this work by both interviewing users about their password habits, and correlating these subjective narratives with observation of their actual behavior.

Id	Question
ST1	How many online accounts do you think you have?
ST2	Roughly, how many different passwords do you think have for all online accounts?
ST3	How many email accounts do you have?
ST4	Is the Gmail account you are using for this study your primary email account?

TABLE I
STATISTICS, PRE-STUDY SURVEY.

F. Security Fatigue

Security fatigue [10] is defined as a saturation point, where it simply gets too difficult or burdensome for users to maintain security. Stanton et al. [21] examined the difficulty of remembering multiple passwords through interviews with 40 participants. They showed that users had many misconceptions about the security risks. Security policies were not able to eliminate users' irrational or sub optimal behaviors, because users reached security fatigue, resulting in a "don't care" attitude about security. Stanton et al. suggested policies that take security fatigue into account, and emphasized the need for consistency across different cybersecurity approaches so that users would feel less resigned or lacking control. We also find that security fatigue and misconceptions play an important role in creation of weak passwords and password reuse. But we further find that guidelines about password length are a more prominent reason for weak passwords.

Redmiles et al. [17] investigated reasons for selective adoption of broad digital-security advice by users. They identified four main reasons for user rejection of security advice: 1) too much marketing information, 2) lack of risk, 3) over-saturation, and 4) inconvenience. In this study, we found that inconvenience, or desire for memorability, plays the most important role in password reuse.

Coventry et al. [4] argued that "best practices are broadly identified but there is less clarity around the specific recommended actions one can follow." They assert that general public does not generally follow best practice. And it is not hard to expect that the lack of clarity around actions would be a contributing factor to user noncompliance, leading to confusion. Our results support this argument. We find that participants generally understand how to compose strong passwords and implement this in practice, but do not understand how long passwords should be nor how to trade off between randomness and length. This leads to many low-strength passwords.

III. METHODOLOGY

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

Research Goals. Our research goal was to study how users design passwords for different sites, and how and why

Id	Question
PS1	How do you choose your passwords? Do you think this is a good practice?
PS2	Do you allow your browser to save passwords for you, or use a password manager?
PS3	I change my passwords even when I don't have to.
PS4	I use different passwords for different accounts that I have
PS5	When I create a new online account, I try to use a password that goes beyond the site's minimum requirements
PS6	I include special characters in my password even if it is not required.

TABLE II
PASSWORD STRATEGY, PRE-STUDY SURVEY.

they reuse their login credentials (username/password). This necessitated:

- Access to information about real usernames, real passwords and real sites where they are used,
- Ability to detect passwords by the same user that are very similar to each other,
- Ability to show to the users their password practices and ask them to explain specific cases of their weak or reused passwords.

The easiest way to collect this information would have been to ask each user to list all their accounts and passwords. However, this would have been a great risk to the users' privacy, since it would provide us with real login credentials. These credentials could potentially be misused by someone on our team, or they could be stolen by outside attackers who could have compromised our study server. These considerations led us to formulate our privacy protection goals.

Privacy Protection Goals. We aimed to protect privacy of the users whose password habits we study in the following manner:

- No storing of any identifying information,
- No intentional (by us) or accidental (by our browsers) storing of real usernames and 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 consists of the following steps, which are performed in our lab on our laptop, in a Chrome incognito window. We open the window for each study participant and close it after the participant completes the study. This ensures that no login credentials or sessions/cookies remain stored in the browser.

Pre-study surveys. First, we ask a participant to fill "Statistics" (Table I), "Password Strategy" (Table II) and "Risk Perception" (Table II) surveys. The questions for these surveys are given in the appendix.

Compiling a list of websites. Next, we scan a user's GMail account, using CloudSweeper tool [20], to compile a list of sites where a user may have an account. The CloudSweeper [20], uses OAuth2 protocol to access GMail, thus a participant's GMail credentials are not seen by our software. We use regular expressions, as described in Section III-A to identify account creation and password reset e-mails sent by online sites to a

Id	Question
R1	Online banking is risky
R2	Using Amazon to purchase items using a credit-card is risky
R3	Sending credit card details over email is risky
R4	Using eBay to purchase items using Paypal is risky
R5	Using unsecured WiFi in a coffee shop is risky
R6	Downloading and using pirated or cracked versions of software is risky
R7	Leaving your car unlocked in city centre multi-story car park is risky
R8	Using social networking sites (e.g. Facebook, LinkedIn) with open privacy settings is risky
R9	Using social networking sites (e.g. Facebook, LinkedIn) with closed privacy settings is risky
R10	Using photo sharing sites (e.g. Flickr, Instagram) is risky
R11	Geotagging content in Twitter or "Checking-in" to a location on Facebook / Foursquare is risky
R12	Opening an email from an unknown sender is risky
R13	Leaving a credit card behind a bar to guarantee a tab is risky
R14	Clicking on a link in an email from an unknown sender is risky
R15	Using online dating services is risky
R16	Flying from the UK to the US is risky
R17	Using a cybercafe is risky
R18	Not updating your operating system (e.g. Windows, Mac OS X) is risky
R19	Not updating your web-browser (e.g. Internet Explorer, Firefox, Google Chrome) is risky
R20	Not updating other applications (e.g. Adobe PDF reader, Microsoft Office / Word, iTunes) is risky
R21	How many guesses could an attacker make in 1 minute?
R22	How would an attacker come up with guesses?
R23	How might an attacker guess a password with an unlimited number of trials?

TABLE III
RISK PERCEPTION STUDY. QUESTIONS R1–R20 ARE SERVED PRE-STUDY AND THE REST POST-STUDY.

Id	Question
IR1	If a stranger could impersonate me on this site this would bring me personal or financial harm
IR2	If a friend or family member could impersonate me on this site this would bring me personal or financial harm
IR3	If the data from my account became public this would bring me personal or financial harm

TABLE IV
IMPACT REASONING, POST-STUDY SURVEY.

Id	Question
PR1	What is the reason behind using the same password?
PR2	Are you concerned that an attacker may crack your password on site A and then use it to access site B?
PR3	If the passwords are not same, but similar, ask why did the user change the password?
PR4	Would you follow the following password strategy. Important & Frequent - unique password, strong password. Non-important - one password, reasonably strong. Important & Infrequent - resettable password.

TABLE V
PASSWORD STRATEGY REASONING, POST-STUDY SURVEY.

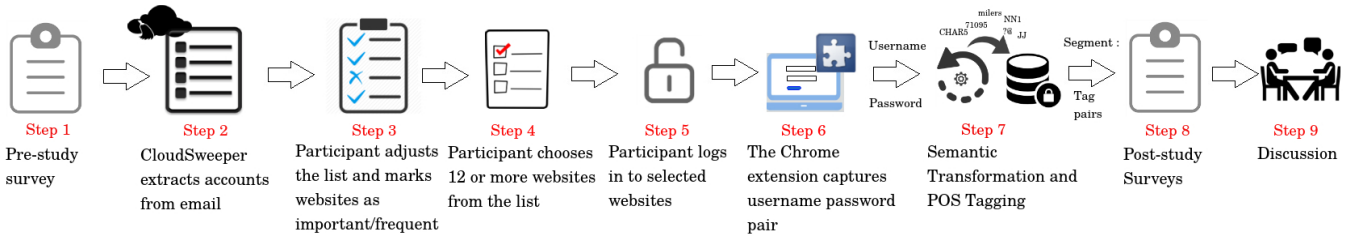


Fig. 1. User study flow

user-supplied e-mail address. We extract the site's URL from such messages. While we could have asked users to provide a list of sites where they have an account, we believed, and our results confirm this, that users may not be fully aware of all the accounts they have, since they are created over a long time.

Collecting participant login information. Next, we show the list of site URLs to the participant. The participant can delete sites where she does not have an account, or sites that are sensitive. The participant can also add to the list other sites where she has an account. We next ask the participant to mark each site in the remaining pool as important to her or not, and as the one she frequently visits or not. We provided no guidance to users how to assign these tags, but we find that users generally marked sites as important if they cared about security of their content at these sites (see Section IV-B). Finally we ask the participant to choose at least 8 important and 4 non-important sites to log on to. The participant can choose more if she wishes.

We have developed a Google Chrome Extension to capture the username and password pair from each login attempt and describe it in Section III-B. We collect character-length information for the original password. We also note if there is capitalization or mangling in the original password and if they are at the beginning, in the middle or at the end. We do 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 input the original password into our local installation of the `zxcvbn` [32] strength meter, and retrieve the resulting strength. We then transform the original password into its semantic equivalent, as described in Section III-C. Such transformed password does not expose any information about the original password beyond its semantic structure, e.g., noun+verb+number. We then store the transformed password and forget the original.

Post-study surveys. After the logins, we ask each participant to respond to questions R21–R23 from the "Risk Perception" survey (Table III). Then, for each site where the participant attempted to log on we ask the "Impact Reasoning" survey (Table IV), which asks her to rate on a Likert 1–5 scale, how affected she would be if a stranger or a friend impersonated her on that site, or if her data from that site were made public. Finally, we serve the 'Password Reasoning" survey (Table V), which asks the participant for each reused or versioned password to narrate the reasons for reuse.

Pattern
welcome to
reset password
thank you for registering
thanks for registering
thank you for creating
thanks for creating
thank you for signing
thanks for signing
your .* account has been created

TABLE VI
PATTERNS IN AN E-MAIL MESSAGE THAT INDICATE THAT A USER MAY HAVE AN ACCOUNT ON THE SENDER'S SITE.

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

A. Detecting Sites Where Participant Has an Account

Cloudsweeper tool extracts emails that contain a search pattern by connecting to Gmail's IMAP using OAuth tokens. We leverage this functionality to identify emails that match certain patterns relating to new account registration and password reset e-mails. We identified nine commonly used patterns by websites in "new registration" and "password reset" emails, and we show them in Table VI. We encode these as regular expressions and use Gmail's "X-GM-RAW" IMAP extension to search from them in e-mail body.

The identified e-mails are further processed to reduce the likelihood of falsely identifying a site as having the participant's account. First, we filter out e-mails that have more than one recipient in "To" and "CC" fields, because account-related e-mails are sent only to the account owner. Second, we use regular expressions on the email subject and body to: (1) filter out e-mails where the URL in the body of the e-mail does not match the domain name of the e-mail's sender, and (2) filter out e-mails where the string following the "Welcome to" does not match the domain name of the e-mail's sender. In each of the match checks, we use python's SequenceMatcher [15], which calculates similarity between two strings. If the SequenceMatcher returns similarity ratio of 0.5 or higher, we say that the given e-mail has passed the check.

B. Password Extraction

We developed a Google Chrome extension to extract the username and password pairs from login attempts, while preserving users' privacy. The extension is enabled manually during each study. During each login attempt, on each key press on username and password fields, we capture the user input. We detect login events using JavaScript window object's `onbeforeunload` event [28], and this triggers sending of the last captured input for semantic transformation.

The extension is also responsible for recording successful attempts. We record attempts as successful if, after the login event is triggered, the following page that loads does not have a password input field. We found this approach reliable for wide range of websites, though some sites have a password input field on their dashboard after successful log in (for e.g., GNU mailman dashboard). For such sites, we recorded the successful attempt manually in the database.

C. Semantic Transformation

The extracted username and password are sent for semantic transformation to an application running on the same laptop as the Chrome browser. The application was our modified version of a tool originally developed by Veras et al. [27] for semantic segmentation and part-of-speech (POS) tagging of strings. We first remove any mangling of the original strings (username and password), before feeding them into the tool. This is done by following the KoreLogic's L33t password cracking rules [13]. We then feed the resulting strings into the semantic segmentation and tagging tool. If the username is in the form of an email address, we only transform the part before the @ sign, and keep the latter part intact.

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 CLAWS7 tagset [23]. 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 user. The goal was to achieve consistent but irreversible transformation of segments. For example, if a user 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 then wanted to transform the proper name "john" into another proper name consistently, so that the resulting two passwords continue to have one common segment. We further wanted to transform the 3-digit numbers 352 and 222 into different 3-digit numbers and the special character "@" into a different special character. For example, the transformed passwords could be "bob475!" and "bob681". 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 [27]. 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 SHA512. Next, we extract only the digits from the hash and we modulo of this number with 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 same segments result in the same input to the hash. Privacy is achieved thanks to our use of a 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. This guarantees that neither we, nor anyone else, can reverse the mapping.

For some POS tags, our dictionary has fewer than 100 words, which could facilitate brute-force search for the original input. To prevent this, we aggregate such tags with the tags in their parent category. For example, words belonging to NNL1 and NNL2 (locative nouns) were added to the tag NN (common noun). If a tag has fewer than 100 words even after this combination it is grouped into the "OTHER" category. In our participant data, we did not have any segment tagged with "OTHER".

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 for the duration of the participant's engagement with the study. Irreversibility is guaranteed by the randomness of the mapping, and because we delete this mapping when the participant exits the study.

Finally, we store the semantically transformed segments, their POS tags, capitalization and mangling information for all segments, authentication status, and the length and the strength of the original password.

IV. Results

We now discuss limitations of our study, and present our findings about how users create passwords, how they reuse them and the causes of such behaviors.

A. Limitations and External Validity.

Our experimental methodology requires a significant amount of interaction: both the experimenter interviewing the user, as well as the user spending time to log in to websites using their real passwords. To balance our sample size with time and monetary constraints, we chose to pay subjects $10 to participate in this study. This constraint is why 48 of our 50 participants were local students (54% (27) were males and 46% (23) were females), as they had no travel expenses or large time overhead to participate in our study. While the statistically significant findings here build a new understanding of how this population approaches the problem of maintaining several passwords across different classes of online services, further

Account	Frequent	Not frequent	Total
Important	212	180	392
Not important	29	200	229
Total	241	380	621

TABLE VII
ACCOUNT TYPES IN OUR STUDY

Account	Frequent	Not frequent	Total
Important	86% (63%)	71% (51%)	79% (58%)
Not important	76% (49%)	57% (37%)	59% (39%)
Total	85% (61%)	63% (43%)	72% (50%)

TABLE VIII
LOGIN SUCCESS RATES PER ACCOUNT (PER LOGIN ATTEMPT)

research is needed to test whether these effects generalize to larger populations.

B. Statistics

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

Account composition. Participants attempted to log into 621 accounts in our study. We show their breakdown across important/not important and frequently/infrequently used categories in Table VII. 392 of accounts were marked as important by users, 241 were marked as frequently used and 212 were in both of these categories. Additionally 29 sites were marked as frequently used but not important, and 180 were marked as important but not frequently used.

We provided no guidance to participants about what "important" means, thus they may have flagged a site as important based on their preference for content, rather than security considerations. We investigate this by comparing the participant responses to IR1 question in our Impact Reasoning survey, which asks a participant how affected she would be if a stranger could access her account. We compare the responses to IR1, given on a Likert 1–5 scale, for sites that a user marked as important versus those that a user marked a non-important. There was a significant (p-value < 0.05) difference in means between these ratings. Users on the average agree (rating 4) that compromise of an important-site account would significantly impact them, while they are mostly neutral (average rating 3.2) about compromise of a non-important-site account.

Site categories. We categorized each site where a participant logged in successfully, into the following categories: (f) financial, (c) e-commerce, (e) e-mail, (s) social, (w) school/work and (o) other. We engaged Mechanical Turk workers to perform site classification. Each site was tagged by 2–4 workers and we adopted the tag that received the majority vote. Table IX shows the total number of sites per category, and the percentage of sites that were tagged as important by any participant. Financial sites were tagged as important the most often (89%), followed by e-mail (81%), social (73%) and school/work (71%) sites. Other and e-commerce sites were tagged as important 57% and 54% of time.

Login success. Participants successfully logged into 446 accounts. We show the login success rate across categories

Category	Sites	Important	Login success
Financial	37	89%	51%
E-commerce	90	54%	44%
E-mail	26	81%	64%
Social	158	73%	54%
School/work	126	71%	51%
Other	63	57%	45%

TABLE IX
SITE COUNT AND PERCENT TAGGED AS IMPORTANT BY PARTICIPANTS, PER SITE CATEGORY

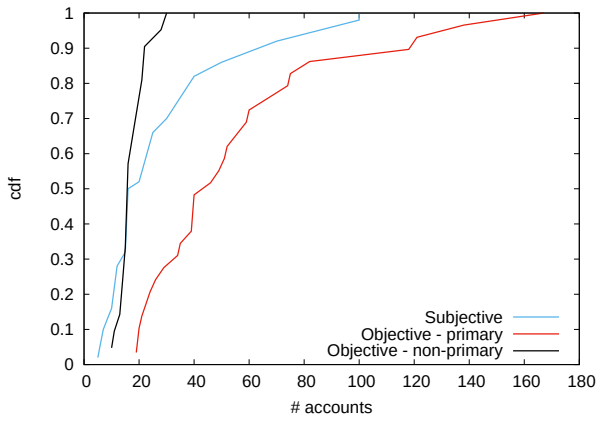


Fig. 2. Number of accounts per participant as estimated by the participant (subjective), and as measured in our study (objective).

in Table VIII, per account and per login attempt (shown in parentheses). Accounts that are both important and frequently used have the highest login success per attempt (63%) and per account (86%). Those that are only important or only frequently used have somewhat lower login success rates – 71–76% per account and 49–51% per login attempt. Finally, those that are neither important nor frequently used have the lowest login success – users successfully logged into 57% of such accounts and succeeded 37% of time. We show the login success per site category in the third column of Table IX. The E-mail accounts have somewhat higher login success (64%), while other sites have 44–54% success rate.

Number of accounts. Out of 50 participants, 29 reported (question ST4, Statistics survey) that the GMail account they used in the study was their primary e-mail account, where "primary" means that they use this account when creating new online accounts. We also asked participants how many E-mail accounts they had (question ST3, Statistics survey). All responded that they had 2 or more, and 76% had 3 or more.

We now compare the subjective measure of the number of online accounts (question ST1, Statistics survey) to our objective estimate, based on the GMail account scans. We show the distribution of these subjective and objective measures of the number of accounts per participant in Figure 2. When we asked participants how many online accounts they thought

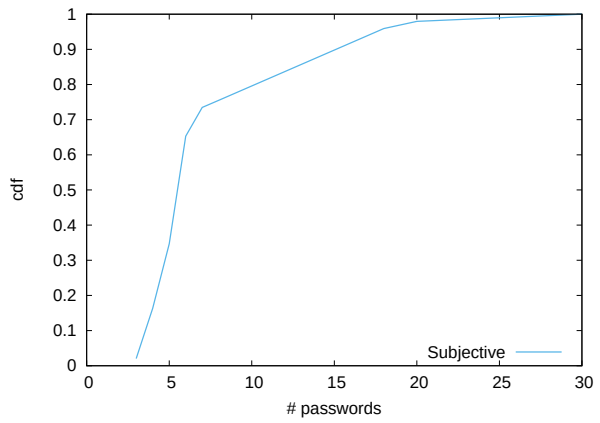


Fig. 3. Number of passwords per participant as estimated by the participant (subjective).

they had, the median estimate was 15. However, when we scanned their GMail inboxes, the median number of accounts for participants with the primary GMail account was 40, and for those with the non-primary GMail account was 15. We report the median and not the mean, as there are several users with a very high number of accounts. We expect that both of our objective measures (for primary and non-primary accounts) underestimate the actual number of online accounts because: (1) some sites do not send welcome messages nor use e-mail address for password resets, (2) we will miss accounts for which a user supplied another e-mail account than the one we scanned, (3) our account identification approach from e-mails may be imperfect, and (4) we allowed participants to remove sites that they consider as sensitive from the list generated by Cloudsweeper. We conclude that in our population, users severely underestimate the number of accounts they have. Users are likely unaware how often they create accounts, as creation happens over a long time period. Our finding of 40 median (52 average) accounts per user updates the finding from a password study by Florencio and Herley [8] done in 2006, which reported 25 accounts per user. This increase is expected as many more online services became available to users over the past nine years.

Number of passwords. When asked how many distinct passwords they had (question ST2, Statistics survey), participants estimated between 3 and 30 passwords. We show the distribution of these estimates in Figure 3; the average is 6.6 and the median in 5. Because we only asked subjects to log in to 12 different sites, we do not know the full distribution of distinct passwords for each user. However, we examine subjective and objective estimates of the level of password sharing in Section IV-E.

C. Password Strength

In this section we present our findings about password strength in our user population.

Long but weak. Figures 4 and 5 show the distribution of password length and strength for successful logins to important and non-important sites, as labeled by participants, and across all sites. We obtain the strength estimate from `zxcvbn`, Dropbox's password strength estimator. `Zxcvbn` runs entirely within the browser on the client side and is trained on password leaks and dictionaries as employed by attackers [32]. The strength metric is given as the expected number of attempts before a password is successfully guessed. Looking at length and strength across all sites, a median password had 11–12 characters – an improvement over the common 8-character minimum length password policy. This illustrates a clear effort on the part of the user to create strong passwords for important sites. Unfortunately, a median password from our dataset requires only 238 million guesses to be broken, making it easily broken by an offline attack. `zxcvbn` estimates that an offline attack on a slow-hashed password can perform 10,000 guesses per second. Thus, the median strength password from our experiment would be broken within 6 hours. The median password strength is fifteen orders of magnitude smaller than the number of guesses required to brute-force a random letter, number, and symbol password of 12 characters, indicating low presence of randomness in passwords. Overall, these findings indicate a disconnect between a user's desire to make a password strong (by making it longer), and the outcome.

Another way to evaluate strength of passwords is to measure the percentage of passwords whose length or strength exceeds some desired value. Let us assume that our security goal is to make passwords at least 8 characters long and resistant to offline attacks (assuming a slow hash) within one year ($10^{11.5}$ tries). Then 92% of passwords meet the length requirement but only 16% meet the strength requirement, as shown in Figures 4 and 5.

Longer is stronger. While long passwords in our dataset are still weak, making a password longer helps. We confirm that there is significant correlation between password length and strength (p-value is $3.8 \cdot 10^{-13}$), and the Pearson Product-Moment correlation is strong ($r = 0.82$).

Important sites have longer and stronger passwords. Figures 4 and 5 show that important sites have passwords that are on the average 1–2 characters longer and about 10 times stronger than those at non-important sites. We confirm that this difference in length and strength is significant by running the Welch's two-sample t-test (p-values less than 0.05).

Unfortunately, this effort by users to create longer and stronger passwords for important sites does not significantly improve their security. While, 91% of important-site passwords and 85% of non-important-site passwords meet our length requirement (8+ characters), only 19% meet our strength requirement.

Site category does not matter. We also investigated strength and length of passwords per site category (financial, social,

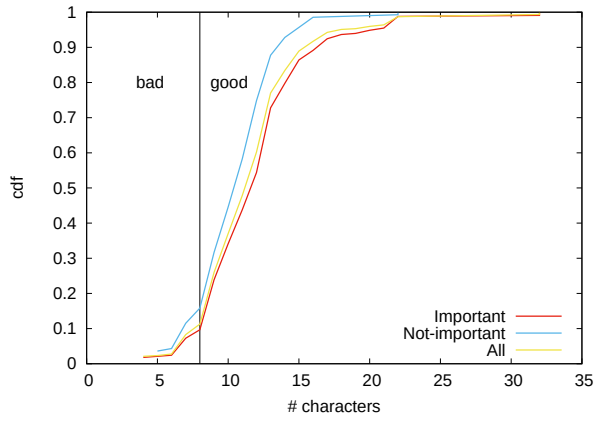


Fig. 4. Distribution of password length for important and non-important sites

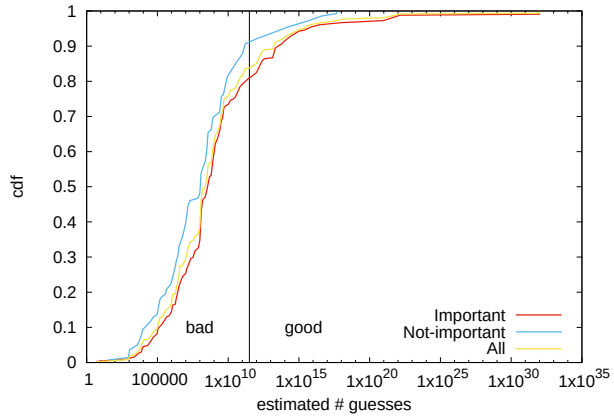


Fig. 5. Distribution of password strength for important and non-important sites

etc.). Distributions of these measures look very similar across categories, and the t-test shows no significant difference in means. Yet, financial and e-mail accounts pose much higher risk to users, if compromised. A user may lose money from her bank account or important secrets from her e-mail account. We conclude that users seem to have a simple mental model of a site's importance, which does not necessarily correlate with the site's category and does not lead to stronger passwords for categories that pose higher risk to user, if compromised.

D. Why Are Passwords Weak?

We now focus on understanding why users create weak passwords. We pose several possible hypotheses and then mine data from our study to confirm or refute them:

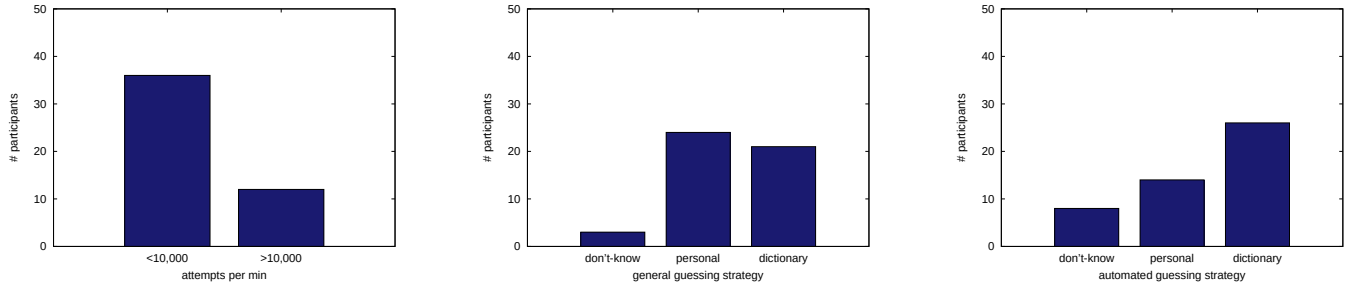
Hypothesis 1: Users do not understand risks of attacks. A user may believe that password cracking attacks are rare or that she is not likely to be their target. She may also have a poor mental model of how these attacks occur and how powerful they are. In our Risk Perception survey, we measure users' attitudes toward risk and understanding of password cracking. This provides both an aggregate understanding of how users perceive risk, and per-participant data that we test for correlation with password strength.

We first analyze the participants' ratings of questions R1–R20 on a 5 point Likert in the Risk Perception survey. These questions have been proposed by Creese et al. in [5], who found that answers to questions R4, R9, R16, R18, R19 and R20 were correlated with password strength. Specifically, they found that the magnitude of the difference between a participant's ratings and the experts' ratings was positively correlated with 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. We further show the summarized ratings in Figure 7. Participants found activities in questions R3, R5, R6, R7, R8, R12, R13, R14 and R15 very risky (most ratings fall in 4–5 range), while the remainder were considered less risky (most ratings fall in 1–3 range). These specific questions touch upon the dangers of mishandling one's credit card, car, social media account, and potentially harmful software.

Next, we code responses to narrative questions R21–R23 in the Risk Perception survey. These questions ask how many password guesses an offline attacker could make per minute (R21), and how he would craft these guesses (R22). Question R23 tests if a participant has heard of offline guessing attacks against passwords. We round the answers to question R21 to the nearest power of 10. We code the answers to questions R22 and R23 as: (d) don't know, (p) using personal info, (d) using dictionary. Figures 6(a)–6(c) show the distribution of user responses to these questions. 75% of users underestimate the speed of the password cracking, and around half do not know how attackers formulate guesses or believe they use personal information about a user. These misconceptions are dangerous as they may lead to passwords that are too short or that predominantly use dictionary words. However, we found no significant correlation between user responses to questions R21–R23 and their password strength or length. Thus better informed users did not create stronger or longer passwords. Based on the lack of correlation between user risk perception and password strength/length we *reject hypothesis 1*.

Hypothesis 2: Users do not know how to create strong passwords. A user may understand the need for strong passwords but may not know how to create one. In our Password Strategy survey we ask users to narrate how they create passwords. This provides both the summary data of how users create passwords, and per-participant data that we attempt to correlate with password strength.

The first question is narrative-based. We break it into three sub-questions asking about choices for password segments, the blend of character classes and if the user considers her



(a) How many guesses can attacker make per min

(b) How would attacker generate guesses

(c) How would attacker generate guesses in an offline attack

Fig. 6. User mental model of attacks.

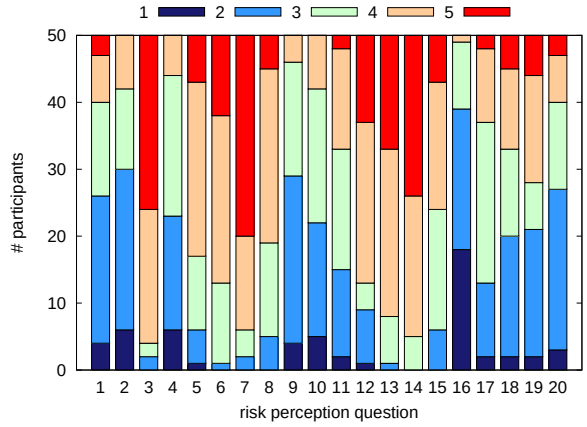


Fig. 7. The breakdown of participant's ratings on the Risk Perception survey.

password strategy good. We then code responses to the first two sub-questions. Together these provide information about a user's password composition.

- PS1-1: How are the password segments chosen: (d) dictionary words, personal names, significant locations, something in the environment, or (r) random character-s/digits
- PS1-2: How many classes of characters are used: (1) one class, (2) two classes, (3) three or more classes

Finally, we combine codes for PS1-1 and PS1-2 to arrive at a password strategy. For example, strategy d-2 would mean that a user starts with one or more common words (a name, a dictionary word) and adds to them one more character class. Figure 8 show the distribution of responses for question PS1. The counts for password strategies add up to 36 because some participants provided insufficient or vague information about their specific password strategy. The most popular password strategies were d-2 and d-3, favored by 72% of participants. They start with a dictionary word and add one or two more character classes. Majority of participants – 93% – used names and words of personal significance in passwords, some of which may be mined from the environment. They increase the strength of such passwords by adding numbers and symbols, and capitalizing parts of passwords. 80% of participants use two or more character classes, and 30% use three or more. These results indicate that *most users understand how to create strong passwords with regard to password composition*.

Do users with better password composition end up having stronger passwords? We answer this question by looking for significant differences in password strength between users that narrated different password strategies. Because there were only two participants with r-1 strategy, and only one participant with r-2 strategy we do not have enough samples to consider strategies at this granularity. Instead, we can ask if those three participants that intended to use random segments had stronger passwords than the 33 that intended to use dictionary words. The Welch two-sampled t-test on password strength between these two groups of users shows no significant difference in means. Similarly, t-test on password strength between users that use only one character class versus two or three classes shows no significant difference in means. This indicates that a user's intended password composition does not significantly influence their password strength. It is possible, however, that their *actual* password composition differs from their intended one. We will explore this in hypothesis 4.

A password strategy includes not just the choice of composition, but also the choice of length. Password length was significantly and strongly correlated with password strength, as we discussed in Section IV-C. We illustrate how password composition and length interplay in Fig. 9(a) that plots password strength vs length, and uses different markers for different password compositions. Many passwords with a user-intended robust strategy such as d-3 or r-1 or r-3 were simply too short to be strong (below the bar in the Figure). On the other hand, several passwords with seemingly bad d-1 strategy, and a little better d-2 strategy, were long enough to achieve considerable strength (above the bar in the Figure).

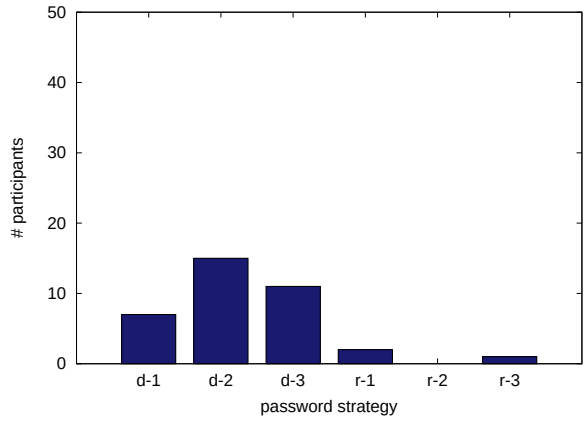


Fig. 8. Count of participants that report using different password strategies.

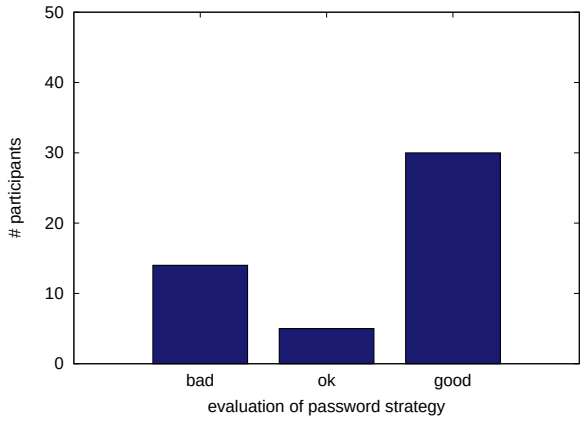


Fig. 10. Count of participants that believe their password strategy is bad, OK or good.

Finally, we also investigate if there is a correlation between a user's answers to questions PS3–PS6 on the Password Strategy survey and the strength or length of her password. These questions suggest four good password strategies, such as use of special characters, going beyond minimum requirements suggested by a password policy, etc, and ask the user to rate her compliance with these strategies on a Likert scale. We find no significant correlation.

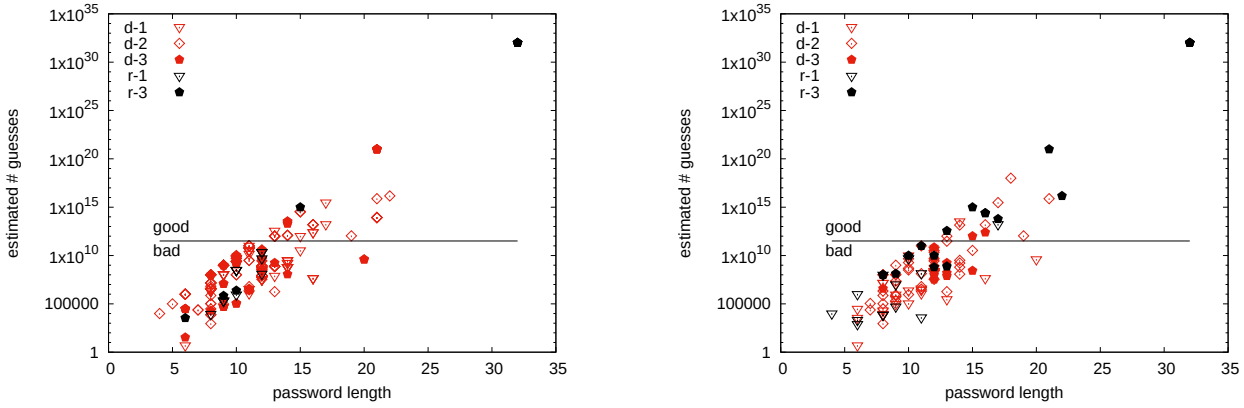
Based on all these results we *confirm hypothesis 2, with regard to password length but reject it with regard to password composition*. Thus educating users to create long passwords should significantly increase password strength. It also appears that educating users to create more random passwords or passwords with more character classes should not influence password strength; but this conclusion is based on the user's *intended* and not *actual* password strategy. We explore this matter further in hypothesis 4.

Hypothesis 3: Users know how to create strong passwords, but choose not to do it. 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 PS2) if they thought their strategy were 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. Welch two-sampled t-test on password strength, and on password length, shows significant difference in means between the participants that say their strategy were bad, versus those that believe their strategy were good. Good-strategy participants have passwords that are on the average 2.3 characters longer and 10 times stronger than bad-strategy participants. This *confirms hypothesis 3*. On the other hand, OK-strategy participants have a wide range of password strengths and lengths and there was no significant difference between this group and the other two groups.

We further analyzed narrative responses by bad-strategy participants to question PS2. 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 users to continue unsafe password practices. Funnel [10] and Redmiles [17] found that security fatigue plays a significant role in user choice of weak passwords and our findings confirm this.

Hypothesis 4: Users know how to create strong passwords in theory, but struggle to implement this in practice. A user may have all the right knowledge in theory, but fail to implement it in practice. This could happen for many reasons including convenience, misunderstanding of good password strategies, and bounded rationality [19]. We detect instances where theory mismatches practice by comparing a participant's intended, subjective password strategy (Password Strategy survey) and their actual, objective password strategy (extracted from transformed passwords).

We derive the objective 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 encode passwords where length of random segments exceeds that of meaningful-word segments as (r) random, and the rest as (d) dictionary. Thus passwords encoded as random may not be fully random, but they have more random than meaningful content. We also encode the character mix of a password using information about capitalization, mangling and presence of character/digit segments. Thus we arrive at the same tags we used for subjective password strategy. Before we compare



(a) Intended

(b) Actual

Fig. 9. Password strategies intended and actual, showed jointly with password length and strength mental model of attacks.

subjective and objective password strategies, we first evaluate if objective password strategy has significant influence on password strength.

We consider the prevalence of different password strategies among *unique* passwords from successful-logins in our study. The actual password strategies are very evenly distributed, with the most popular choice being d-2 (31% of passwords), and the least popular choice being d-3 (9% of passwords). Overall, 58% of passwords use dictionary words and the remaining 42% use only random characters and/or digits. Further, 23% of passwords use a single character class, 54% use two character classes and 23% use three character classes. We find significant difference in means (Welch t-test, p-value < 0.05) between the strength of passwords in d-1, d-2 and d-3 groups. d-2 passwords are around 363 times stronger than d-1 passwords, and d-3 passwords are 52,480 times stronger. We further find no significant difference in length between d-1 and d-2 passwords, but we find it between d-2 and d-3 passwords. d-3 passwords are on the average almost 5 characters longer than d-1 and d-2 passwords. When comparing password strength and length between r-1, r-2 and r-3 groups we find significant difference in means between r-1 and r-2 groups, but no difference between r-2 and r-3 groups. r-2 passwords are on the average 316 times stronger than r-1 passwords. There is further no significant difference in password length between r-1, r-2 and r-3 groups. When we compare between d and r groups, we find that random passwords are around 457 times stronger than those using dictionary words, and we find no significant difference in length.

We illustrate how the actual password composition and length interplay in Figure 9(b). The actual password composition influences somewhat the password strength, with many robust strategy passwords being above the bar, and many weak-strategy passwords being below. Yet, length still plays a considerable role. Around half of the random passwords, and all but two d-3 passwords, are too short and thus too weak.

We conclude that *users know how to compose strong passwords in practice, but may not understand long their passwords should be*. One successful strategy starts from words of personal significance and adds more character classes to them, making the password longer and also stronger. Another successful strategy generates random passwords of reasonable length.

We next compare subjective and objective password strategies and note how often an objective strategy is weaker, stronger or equal to subjective strategy. For example, if a user declared to use r-1 strategy, but ended up using d-1 we would say that her objective strategy is weaker than her subjective strategy. We find that 29% of passwords use a weaker strategy than narrated by a user, 46% use a stronger strategy and only 25% match. This confirms that users do not consistently implement their subjective password strategy, but also shows that they often use a stronger-than-intended strategy. Thus we *refute hypothesis 4*.

Hypothesis 5: Password policies lead users to create weak passwords. It is possible that users do not independently decide on a given password strategy, in a rational manner, but instead adopt a site-suggested password policy. To test for this we surveyed, with the help of Mechanical Turk workers, the password policies of 174 randomly selected websites where participants attempted to log in during our user study. 29% of these websites did not have a minimum length requirement, 33% required a minimum or 6 characters, and 27.6% required a minimum of 8 characters. Yet, in our dataset none of 8-character or shorter passwords were stronger than our goal of $10^{11.5}$ guesses to success. In fact, the shortest password that met this goal had 12 characters.

With regard to password composition, 13.9% sites required 8 characters and at least one number and symbol. Further, 18.4% employed some password meter, but de Carnavalet

and Mannan [7] found that commonly used meters are highly inconsistent, fail to supply coherent feedback to users, and sometimes provide strength measurements that are obviously misleading. Overall, we conclude sites in our study had weak, potentially confusing password policies, which undoubtedly led users to choose weak passwords, believing them to be strong. This *confirms hypothesis 5*.

Hypothesis 6: Bounded rationality. A password strategy combines a lot of information together. Users must choose passwords they can recall, that have strong composition and sufficient length. They further must reason about the site's importance to them and choose strong passwords for sites they believe are important. Further password policies communicated by sites are confusing and user perception of risk may be incorrect. All these factors may lead to bounded rationality [19], a phenomena that leads people to make sub-optimal decisions in situations where they must act quickly on lot of confusing information. While we cannot rigorously test this hypothesis, we offer some observational evidence that bounded rationality is a plausible explanation. We first identify those participants that created passwords of sufficient strength (those that require $> 10^{11.5}$ guesses). We then remove from this set the participants whose important-site passwords were weaker than this target strength at least 50% of time. This leaves us with 7 participants or 14% that had more than half of their important-site passwords of sufficient strength. These participants were exposed to the same password policies as the rest, yet they made better decisions. We call this group of participants *power users* and investigate what differentiates them from non-power users. These user groups have no significant difference in their risk perception, nor in their comprehension of attacks. They also do not narrate significantly different password strategies. Yet the outcome is different.

We now closely examine all the passwords of the power users against those of non-power users, using the objective tagging of the actual password strategy. Table X shows the composition of the passwords. Power-users have more random content in their passwords (57% vs 42%) and tend to mix 3 character classes more often than non-power users (33% vs 22%). Further, we find a significant difference (t-test has p-value < 0.05) in password strength and length for random-content passwords between these two groups. Power-users create random-content passwords that are 10^{11} times stronger and 11 characters longer than non-power users. We find no significant difference in length and strength of dictionary-content passwords between power and non-power users. We conclude that *power users have more random content in their passwords and their random-content passwords are longer. This is the main factor that makes their passwords strong.* While we cannot confirm or refute hypothesis 6, our observational data indicates that trade-offs that power users balance well are complex enough that they could confuse many users, and lead them to exhibit bounded rationality.

E. Reuse

We now examine how often users reuse their passwords.

Class/Comp	Power			Non-power		
	d	r	Tot	d	r	Tot
1	14%	10%	24%	19%	9%	28%
2	29%	14%	43%	34%	23%	57%
3	0%	33%	33%	12%	10%	22%
Tot	43%	57%		65%	42%	

TABLE X
PASSWORD COMPOSITION OF POWER VS NON-POWER USERS.

Subjective estimate of reuse is large. Looking at the users responses to our Statistics survey (ST1 and ST2) 98% of participants stated they have fewer passwords than accounts. Based on these subjective measures, participants believed to share a password among 4.7 accounts on the average. Further, participants that estimated to have a higher number of accounts did not report to have correspondingly more passwords. Except for three participants who reported to have 18–30 different passwords, the rest reported to have up to 10 different passwords. Florencio and Herley's study conducted in 2006 [8] show that the average user has 6.5 passwords, each of which is shared across 3.9 different sites. Wash et al. [30] in 2016 found that people re-use each password on 1.7–3.4 different websites. Hence, our findings are consistent with prior research. However, we found in Section IV-B that participants underestimate the number of accounts they have, by a factor of 2.6. If the subjective estimate of the number of passwords a participant has were correct, this puts the actual password reuse close to 10 accounts per password. Both prior studies [8], [30] tracked number of user accounts by instrumenting their browser to record successful logins. However, users have many accounts that they create once and then use rarely or never again. Our study has potentially revealed these accounts through the inbox scan, and thus produced a higher estimate for password reuse.

Objective estimate of reuse is large. We now calculate percentage of participants that reuse their passwords either verbatim or have similar passwords between two accounts, in the login attempts they make in our study. We say that two passwords are similar if they have at least one common segment. Table XI summarizes our findings about reuse. 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 additional 6% have similar passwords between accounts. Password sharing across sites of different importance is dangerous. A non-important site may belong to a smaller company than an important site, and may be poorly protected against server compromise. If an attacker compromises such site, he gains access to user accounts at important sites. Also, 98% (100% including similar passwords) of users 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 users create a limited number of strong passwords and reuse them without discrimination at both important and*

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

TABLE XI
PASSWORD REUSE: PERCENTAGE OF PARTICIPANTS THAT REUSE IN A GIVEN WAY.

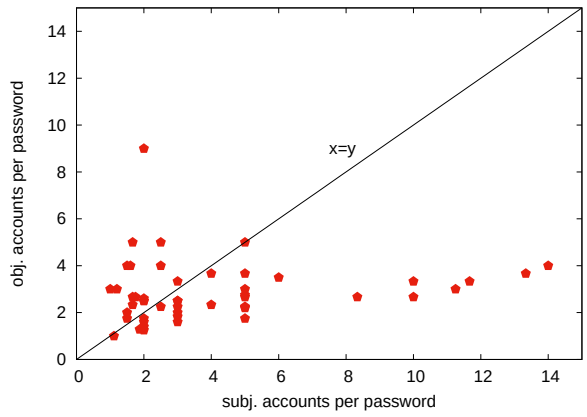


Fig. 11. Subjective vs objective reuse of passwords across accounts: the numbers above the x=y line show participants that underestimate their password reuse.

non-important sites. Average number of accounts per password in our study, among successful logins, was 2.9. This figure is low primarily because participants only logged into 12 accounts during the study.

Users under-estimate their reuse patterns. We now compare subjective vs objective reuse of passwords for each participant. We obtain subjective estimate of reuse by dividing user-reported number of accounts with a user-reported number of passwords. We obtain the objective reuse of passwords by dividing the number of the sites a participant successfully logged into, with the number of unique passwords for that participant. This comparison is not unbiased, as the accounts the participant chose to log into may not be an unbiased sample out of all her accounts. Figure 11 shows the subjective vs objective estimate of reuse for each participant. While a number of points lie under the x=y line, 17 or 34% are above it. Thus a number of users that believe their password reuse is low (subjective accounts-per-pass 1—3) actually have high password reuse (objective accounts-per-pass 2–9).

Reuse leads to password leakage. Reuse has a further detrimental impact on password security. Users tend to forget *where* they reused *which* password and thus attempt all their passwords on a failed login. We track a number of participants that met the following conditions in our study: (1) successfully logged on into at least one important and one non-important site, and (2) there is at least one password for an important-site I that the participant does not share with some non-important site N. All the participants in our study met both conditions. We then note for each I, N pair of sites, how many participants tried to input their password for site I into the site N. 44 participants or 88% did so. This practice poses a risk to users, since it makes the following scenario possible. An attacker can advertise a service, and attract users to create accounts on her site. The attacker then rejects but logs each user login attempt. This way the attacker "milks" a user for all her passwords.

Simple password versioning. A user may not reuse her password verbatim, but apply small changes to it, either due to policy requirements at the new site or due to user concern about password-reuse attacks. 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 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 patters, shown in Table XII. 62% of passwords are versioned in a very simple manner, by changing or adding a number, a special character, one dictionary word, or by introducing capitalization or mangling. If an attacker obtains one password from a pair he can guess the other one with a very small number of tries. 38% of passwords experience more complex transformations that combine 2–3 simple techniques and may change or add two dictionary words. The attacker will need more tries to explore the space of these transformations, yet much fewer tries than if he were to brute-force the more complex password in the pair.

F. Why Is Reuse Prevalent?

When we detected reuse of the same password verbatim, we asked the participant what was the cause for reuse. Total of 49 participants have reused some password verbatim. All participants said they do it for memorability reasons. In addition, 14% of participants said they share passwords only among accounts they do not care about.

We further asked participants if they were concerned about password-reuse attacks. 10% of participants did not know about password-reuse attacks and further 10% heard of them but thought that they are immune because their password is strong. This is a serious misconception as any password is susceptible, when reused, to password-reuse attack.

Finally, we asked users that had similar passwords why they changed their password. 62% did so due to policy requirement and 38% did it out of free will. Thus users are conscious that verbatim reuse is bad, and they attempt to find alternatives through password versioning. Yet their versioning is simple and does not make them immune to password-reuse attacks.

V. RECOMMENDATIONS

Single factor authentication for important accounts is particularly dangerous when paired with our finding that even technologically savvy users, like college students, engage in

	Type of change	Pass.	Example
simple	Change/add numbers	15	qklt → qklt18
	Change/add spec. chars	4	romine6719 → romine6719]
	Change/add alph. chars	2	jdofabergs36859 → cqkfabergs36859
	Capitalize first/last char	4	coumadinjamkaran → Coumadinjamkaran
	Switch digit/spec. positions	1	vwuondfk.3 → vwuondfk3.
	Change/add one word	12	2735770770 → benigno 2735770770
	Total	38	
cmplx	Combination of 2–3 techniques above	20	tranquillizersdenham → tranquillizersdenham9?
	Change/add two words and opt. num/spec. char	3	ddn1ddn → ddn1ddnauditionsboorstin372
	Total	23	

TABLE XII
PASSWORD REUSE: PERCENTAGE OF PARTICIPANTS THAT REUSE IN A GIVEN WAY.

unsafe password practices. We draw on our findings of how and why users engage in unsafe password practices to propose interventions that are aligned with user capabilities.

Help users in understanding their current choices. Users have many online accounts created over long time periods. They cannot keep track of them, nor rationalize their password behaviors. Automated assistants (browsers and password managers) could help users by long-term tracking of their accounts and password, and periodical summaries and analysis of this data. For example, once monthly 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."

Suggest and automate better strategies. When a user is given a summary report, like above, we should develop mechanisms that suggest meaningful actions, and implement them automatically. For example, the assistant may ask "Do you want me to delete these 85 unused accounts or reset your password there to a random string?" and proceed to do so if allowed. The user should also have an option of examining unused accounts, aided by the assistant and selecting only some for automated remediation.

Make better use of password assistants. We asked users if they allow browsers or password managers to save their passwords, in question PS2 in our Password Strategy survey. 59% of users said they allow this always, 25% said they do not allow it and 16% allow it sometimes. Yet we found no significant difference in strength or length of passwords between the group that never uses an assistant and the one that always uses it. Further, use of random-content passwords increased only slightly from 37% (never-use group) to 43% (always-use group). Finally, we found no significant difference in password sharing between these groups. This is a missed opportunity. Users that use password assistance could have very long, very random and unique passwords for each account. Password assistants could suggest this policy and aid users by automating its implementaton.

Improve password policies. Current password policies have very low guidance for password length. Sites should require longer passwords, and suggest strategies that help users create those. For example, a site could suggest "use two meaningful words and add two numbers" as a strategy. Sites should also implement more realistic password meters, and suggest specific improvements when a password fails the meter. For example, a site could provide the following feedback to the user "Your password is weak because it is based on one dictionary word. Your options are: (1) add a 4-digit number, (2) add another dictionary word and a special character, (3) add a 5-character random string or a 5+ character foreign word." Finally, password meters that tell a user "This password can be brute-forced within 1 second" deliver more meaningful information that meters that show red/orange/green color.

Suggest better password compartmentalization. Users tend to share passwords indiscriminately between important and non-important sites. Some sites may have 2-factor authentication and thus it may be OK to use a weaker or shared password on these sites. Other sites may host important content for the user (social, financial, email) and may not use 2-factor authentication. Passwords for these sites are crucial to be protected. Password assistants could aid users by detecting a small number of such unprotected sites (5–8) that users use frequently. They could then suggest that users create strong, unique passwords for these sites and a shared, strong password for all the other sites.

VI. CONCLUSIONS

Users are overburdened today with many online accounts. They are also confused with weak, inconsistent password policies, simplistic password meters and with many factors that influence a password's strength. Like much other research, we find that many passwords are weak. We find that users have a good will to create stronger passwords, but they do not understand how long their passwords should be and how randomness of the content interplays with length. We further find that users share passwords a lot, due to memorability issues, even when they use a password manager. While good strategies for password management have been proposed (e.g., [9]), users struggle to implement them in practice. We hope that our recommendations may be better aligned with user cognitive abilities and thus better adopted in practice.

REFERENCES

[1] T. G. Blog. GitHub Security Update: Reused password attack. https://github.com/blog/2190-github-security-update-reused-password-attack, 2016.

[2] J. Bonneau. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*, May 2012.

[3] K. Conger and M. Lynley. Dropbox employee's password reuse led to theft of 60M+ user credentials. https://techcrunch.com/2016/08/30/dropbox-employees-password-reuse-led-to-theft-of-60m-user-credentials/, 2016.

[4] D. Coventry, P. Briggs, J. Blythe, and M. Tran. Using behavioural insights to improve the public's use of cyber security best practice. In *Government Office for Science, London*, 2014.

[5] S. Creese, D. Hodges, S. Jamison-Powell, and M. Whitty. Relationships between password choices, perceptions of risk and security expertise. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 80–89. Springer, 2013.

[6] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The Tangled Web of Password Reuse. In *NDSS*, volume 14, pages 23–26, 2014.

[7] X. de Carné de Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. In *Network and Distributed System Security Symposium (NDSS 2014)*. Internet Society, 2014.

[8] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proceedings of the WWW*, 2007.

[9] D. Florêncio, C. Herley, and P. C. Van Oorschot. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *23rd USENIX Security Symposium*, pages 575–590, 2014.

[10] S. Furnell and K.-L. Thomson. Recognising and addressing 'security fatigue'. *Computer Fraud & Security*, 2009(11):7–11, 2009.

[11] D. Goodin. Then there were 117 million. LinkedIn password breach much bigger than thought. http://arstechnica.com/security/2016/05/then-there-were-117-million-linkedin-password-breach-much-bigger-than-thought/, 2016.

[12] D. Goodin. Why passwords have never been weaker, and crackers have never been stronger. http://arstechnica.com/security/2012/08/passwords-under-assault/, 2016.

[13] KoreLogic Security. KoreLogic's Custom rules - DEFCON 2010. http://contest-2010.korelogic.com/rules.html.

[14] S. Philion. An Important Message About Yahoo User Security. https://investor.yahoo.net/releasedetail.cfm?ReleaseID=990570, 2016.

[15] Python SequenceMatcher Objects. https://docs.python.org/2.4/lib/sequence-matcher.html, 2016.

[16] A. Rao, B. Jha, and G. Kini. Effect of grammar on security of long passwords. In *Proceedings of the third ACM conference on Data and application security and privacy*, pages 317–324. ACM, 2013.

[17] E. M. Redmiles, A. Malone, and M. L. Mazurek. I Think They're Trying To Tell Me Something: Advice Sources and Selection for Digital Security. In *2016 IEEE Symposium on Security and Privacy*, May 2016.

[18] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 2. ACM, 2010.

[19] H. A. Simon. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press, 1982.

[20] P. Snyder and C. Kanich. Cloudsweeper: enabling data-centric document management for secure cloud archives. In *Proceedings of the 2013 ACM workshop on Cloud computing security workshop*, pages 47–54. ACM, 2013.

[21] B. Stanton, M. F. Theofanos, S. S. Prettyman, and S. Furman. Security Fatigue. *IT Professional*, 18(5):26–32, 2016.

[22] E. Stobert and R. Biddle. The password life cycle: user behaviour in managing passwords. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 243–255, 2014.

[23] UCREL CLAWS7 Tagset. http://ucrel.lancs.ac.uk/claws7tags.html, 2016.

[24] B. Ur, J. Bees, S. Segreti, L. Bauer, N. Christin, and L. F. Cranor. Do users' perceptions of password security match reality? In *CHI'16: 34th Annual ACM Conference on Human Factors in Computing Systems*. ACM, May 2016.

[25] B. Ur, F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, and L. F. Cranor. "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, July 2015.

[26] A. Vance. If Your Password Is 123456, Just Make It HackMe. http://www.nytimes.com/2010/01/21/technology/21password.html.

[27] R. Veras, C. Collins, and J. Thorpe. On the semantic patterns of passwords and their security impact. In *Network and Distributed System Security Symposium (NDSS'14)*, 2014.

[28] W3Schools. `onbeforeunload Event`. http://www.w3schools.com/jsref/event_onbeforeunload.asp.

[29] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang. Targeted Online Password Guessing: An Underestimated Threat. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1242–1254. ACM, 2016.

[30] R. Wash, E. Rader, R. Berman, and Z. Wellmer. Understanding password choices: How frequently entered passwords are re-used across websites. In *Symposium on Usable Privacy and Security (SOUPS)*, 2016.

[31] M. Weir, S. Aggarwal, B. De Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 391–405. IEEE, 2009.

[32] D. L. Wheeler. zxcvbn: Low-budget password strength estimation. In *Proc. USENIX Security*, 2016.