Password Creation in the Presence of Blacklists

Hana Habib, Jessica Colnago, William Melicher, Blase Ur†, Sean Segreti,
Lujo Bauer, Nicolas Christin, and Lorrie Cranor
Carnegie Mellon University
{hana007, jcolnago, billy, ssegreti, lbauer, nicolasc, lorrie}@cmu.edu
†University of Chicago
blase@uchicago.edu

***Abstract*—Attackers often target common passwords in guessing attacks, leading some website administrators to make common passwords ineligible for use on their sites. While past research has shown that adding such blacklists to a password policy generally makes resulting passwords harder to guess, it is important to understand whether users go on to create significantly stronger passwords, or ones that are only marginally better. In this paper, we investigate how users change the composition and strength of their passwords after a blacklisted password attempt. Additionally, we analyze differences in sentiment toward password creation based on whether a user created a blacklisted password. Our examination utilizes data collected from a previous online study evaluating various design features of a password meter through a password creation task. We analyzed 2,280 password creation sessions and found that participants who reused even a modified version of a blacklisted attempt during the task ultimately created significantly weaker passwords than those who did not attempt to use a blacklisted password. Our results also indicate that text feedback provided by a password meter mitigated this effect.**

I. Introduction

Some Internet services, including those offered by Microsoft[1] and Google,[2] attempt to reduce the predictability of passwords on their systems by rejecting users' attempts to create passwords that are in a blacklist of common passwords. While past research has studied how people create and use passwords [27], [34], [35], [44] and has found that robust blacklists can reduce how easily user-chosen passwords can be guessed [22], [34], it is important to also understand how users respond to having their password attempts rejected for being on a blacklist: do users make only small (and perhaps predictable) alterations to a blacklisted password, do they create passwords that are substantially different but not much harder to guess, or do they create passwords that are significantly less guessable than the blacklisted password? How can we encourage users to create less guessable passwords after their blacklisted passwords are rejected? What effect does this have on user sentiment toward password creation? In this paper, we investigated these questions through the analysis of 2,280 password-creation interactions, including 350 in which participants typed in blacklisted passwords.

System administrators often look to the National Institute of Standards and Technology (NIST) for guidance on password policy [6]. NIST recently released a draft of its Special Publication 800-63B, in which it proposes new requirements for "memorized secrets" (i.e., passwords and PINs) [16]. The draft document recommends that memorized secrets be at least eight characters in length and advises against other composition policies, such as requiring a minimum number of different character classes. It also proposes that passwords not be in a list of "commonly-used, expected, and/or compromised values." This last requirement relates to the fact that large-scale password breaches have shown that many of the passwords leaked, such as "12345678," are commonly used across websites [41]. As such, listing these values on a *blacklist* and denying their use is a seemingly simple solution for improving users' password strength against modern brute-force attacks, without the added difficulty and frustration associated with composition policies [17], [19], [24], [33], [47].

Furthermore, the NIST draft proposal requires that a user who has selected a blacklisted password be "advised that they need to select a different secret because their previous choice was commonly used, and be required to choose a different value" [16]. However, there is no recommendation for any additional feedback to be provided to help them create better passwords, even though recent work has found that providing such feedback can lead to stronger passwords [42].

In this paper, we analyzed a subset of the data collected during a prior online study evaluating the design of a password meter. We evaluated 2,280 password-creation interactions, created under the same policy recommended in the NIST proposal, and explored the composition and strength of both the blacklisted password attempts and final passwords participants created. We then manually inspected 350 candidate passwords that were rejected because they matched passwords in our blacklist, as well as the final passwords that these participants subsequently created, to determine how participants changed their blacklisted password attempt into one that passed the blacklist check. We also evaluated how attempting to use a blacklisted password affected participants' sentiment toward the password-creation task. With these analyses, we introduce recommendations for feedback that can nudge users away from weak passwords after a blacklisted password attempt.

Our analyses found that the final passwords created by

[1] Microsoft Corporation. https://www.microsoft.com

[2] Google. https://www.google.com

USEC '17, 26 February 2017, San Diego, CA, USA
Copyright 2017 Internet Society, ISBN 1-891562-47-9
http://dx.doi.org/10.14722/usec.2017.23043

participants who previously had a password rejected because of blacklisting were less varied in their composition and weaker than those created by participants who did not have a blacklisted attempt. Providing text feedback to participants had a stronger effect on those with a blacklisted attempt, suggesting that even users inclined to create simple passwords can be nudged into creating stronger ones. Additionally, approximately 69.4% of participants who had a blacklisted attempt either used some sort of transformation (e.g., inserting digits, using a different keyboard pattern) of their blacklisted password for their final password, or exactly reused their blacklisted attempt as a part of their final password. Participants who changed their previously blacklisted password attempt more comprehensively created stronger passwords, but reported password creation to be more difficult and annoying than those who did not.

The primary contribution of this work is the analysis of how users respond to having a password attempt rejected for being on a blacklist of popular passwords. We provide data-driven recommendations for the best way to leverage blacklists, and build upon previous findings that website operators and system administrators should provide feedback to users on how to improve the strength of their password [42], highlighting its positive effect on those who attempted blacklisted passwords.

The remainder of this paper proceeds as follows. We first, in Section II, provide an overview of prior work studying various aspects of password creation. In Section III, we then describe the details of the online study, the methodology used in our analyses, and potential limitations of this work. We provide a description of the demographics of our participants in Section IV. We present our results in Section V, describing the differences in password composition and strength of blacklisted and final passwords, how blacklisted passwords are changed, and the effect of blacklisted attempts on password creation sentiment. In Section VI, we discuss our findings and recommendations for website operators and system administrators for helping their users create stronger passwords. We conclude in Section VII with a summary of our results and recommendations.

II. Background and Related Work

Passwords are widely used today even though people create easily guessed passwords, reuse them across multiple accounts, and write them down [44]. These shortcomings have led to a move toward multi-factor authentication, where factors belonging to different categories among "something you know" (e.g., passwords), "something you have" (e.g., tokens), and "something you are" (e.g., biometrics) are combined to provide a more secure authentication process [7]. The trend toward multi-factor authentication has been reinforced by technology companies claiming that "passwords are dead" [40] and by the U.S. government through the launch of a national campaign to "move beyond passwords" [29]. However, even if only as part of a more complex multi-factor system, it is clear that passwords will still be relevant to the technical ecosystem for at least the immediate future.

Password blacklists are a vital mechanism for protecting users from adversarial guessing attacks. These guessing attacks take two primary forms. In online guessing attacks, in which an attacker tries to authenticate to a live system by guessing users' passwords, attackers are generally limited in the number of guesses they can make, because systems that follow security best practices will rate-limit authentication attempts and may require secondary authentication following a number of incorrect attempts. In an offline guessing attack, an attacker takes a password file to another system where guessed passwords can be hashed and compared with hashed passwords in the file, so that there is no limit to the number of guesses that can be tried, but efficient guessing can allow passwords to be guessed more quickly [14].

Attackers rely on guessing two types of passwords that have a relatively high probability of success. First, commonly used passwords are a source of high-success password guesses. Passwords frequently contain words and phrases [4], [25], [44], as well as keyboard patterns (e.g., "1qaz2wsx") [45] and dates [46]. If a password contains uppercase letters, digits, or symbols, they are often in predictable locations [3]. Furthermore, most character substitutions (e.g., replacing "e" with "3") found in passwords are predictable [21]. The intuition behind blacklisting the N most common passwords is that users who otherwise would have chosen one of these common passwords will instead choose from a larger space of potential passwords, rather than one of the N next-most-common passwords. The empirical analysis we report in this paper is the most in-depth analysis to date of whether this intuition holds in practice.

Reused credentials are the second source of high-success guesses. If an attacker has compromised the password store on another system and discovered a user's password through an offline guessing attack, he or she will try the same credentials on other systems because users frequently reuse the same password across different accounts [9], [13], [20], [39]. Following best practices, system administrators will store passwords using hash functions like Argon2, bcrypt, or scrypt, which are specifically designed to substantially slow down password-guessing attacks [2], [30], [32]. While system administrators do not always follow these best practices [14], a well-implemented system will again limit the attacker to guessing the most probable passwords. As a result, a blacklist that leads users to choose less predictable passwords in practice defends against both online and offline guessing attacks.

Password blacklists can be created using a number of approaches, including making lists of commonly used passwords discovered in leaked password databases or blacklisting the initial guesses made by password-guessing algorithms. Blacklists can range very widely in size, from dozens of extremely common passwords [10], [43] to lists of potentially billions of blacklisted passwords that are stored server-side [22]. In typical usage, a user is prohibited from using a password that appears on a blacklist, although some systems may still allow the selection of a blacklisted password despite discouraging it. Furthermore, different systems take different views of what constitutes a blacklisted password. Among many possibilities [10], [22], [43], design choices include, for instance, whether blacklists are case-sensitive or not, and whether blacklists only apply to full passwords or to mere substrings.

Some of the prior work has superficially analyzed the aggregate effect of blacklists on password security and usability. In analyzing leaked sets of passwords alongside potential blacklists ranging in size from 100 to 50,000 passwords, Weir et al. observed that the password sets' resistance to guessing attacks would substantially improve if the blacklisted passwords were removed [48]. Because Weir et al. were retroactively studying sets of passwords, however, they were unable to examine what passwords the affected users would pick in place of the forbidden, blacklisted passwords.

Kelley et al. analyzed passwords created with various blacklists under different password composition policies—namely, requiring at least eight characters, requiring at least 16 characters, and requiring at least eight characters and all four character classes (lower letters, uppercase letters, digits, and symbols). Their blacklists varied based on their size, complexity (dictionary words only versus both dictionary words and common passwords), and modification detection (direct match, case insensitive, pre-processed to strip non-alphabetic characters). They found that bigger and more complex dictionaries led to stronger passwords being created. While they analyzed the overall impact on security and usability, they did not deeply investigate how the blacklist impacted user behavior [22].

In another study, Shay et al. analyzed passwords created under the requirement that they be at least 12 characters long and contain three character classes (lower or uppercase letters, numbers or digits). For their blacklist, they used common substrings of passwords that were cracked in a previous study, as well as substrings thought to be easily guessable (e.g. four sequential digits or letters, parts of the word password, years, character repetition, etc.). This led to a blacklist with 41,329 strings, and any password that contained one of these banned substrings was forbidden. Shay et al. found that having a blacklist increased security without making password recall significantly more difficult, yet decreased other aspects of usability in password creation [34].

In this work, we move beyond these prior studies by delving into how users behave after their prospective password is flagged as blacklisted, as well as how these different behaviors affect password strength and sentiment toward the task of password creation. Better understanding user behavior in response to blacklists is crucial both because many major service providers use password blacklists in the wild [10], [14], [43] and the use of blacklists features prominently in current NIST draft password guidance [16].

Blacklists are often used in concert with other interventions designed to guide users toward stronger passwords. Password composition policies are one such intervention. These policies specify characteristics a password must have, such as containing particular character classes. While these policies can improve the resultant passwords' resistance to a guessing attack, users often find complex password policies unusable [1], [17], [19], [24], [37], [47]. Proactive password checking, such as showing the user an estimate of password strength through a password meter, is another common intervention. Researchers have found password meters can help guide users toward stronger passwords [11], [43]. Different meters rely on client-side heuristics [10], [49], server-side Markov models, or artificial neural networks [28] to gauge password strength. Beyond displaying a strength score to users, some proactive password checkers give users detailed feedback about their password's characteristics [42], show users predictions of what they will type next to encourage them to pick something different (and thus harder to predict) [23], or compare the strength of that user's password to other users' passwords [36].

III. Methodology

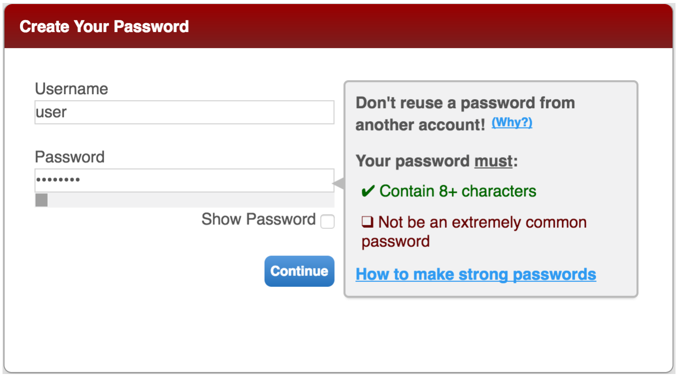
The data analyzed in this study was collected from our group's prior work evaluating the security and usability impact of a data-driven password meter [42]. We recruited participants on Amazon's Mechanical Turk,[3] limiting ourselves to participants aged 18 and older, and located within the United States. Participants were required to complete the task in Firefox, Chrome/Chromium, Safari, or Opera, as the password meter being evaluated had only been tested on those browsers. During the task, participants were shown a variation of the password meter that guided them through creating a password. To be in alignment with the NIST proposal, in this paper we focus only on those passwords that were created under a policy that required passwords to contain eight or more characters (referred to as "1class8") and had no other restrictions on their composition beyond prohibiting passwords that were on a blacklist.

The blacklist used to prohibit common passwords was built off the Xato corpus, a filtered list of 10 million passwords out of billions that were captured from several password leaks and made available to security researchers [5]. The Xato data set was chosen due to its use in prior passwords research [49], and because it allowed the detection of passwords that were common across websites and not specific to a single website. A blacklist of around 100,000 passwords was used in this work since it produced a blacklist file on the order of a few hundred kilobytes (or less using compression). This is small enough to feasibly transfer to a client for client-side blacklist checking, which would avoid a server performing the blacklist check on a plain-text candidate password. Specifically, using the threshold of a password appearing four or more times in the Xato corpus resulted in 96,480 passwords being included in the blacklist.

Each keystroke performed by the participant during password creation was captured, and the feedback displayed by the meter adapted to changes in the password as it was being typed. When a participant typed in a password string found in the blacklist, a message saying "Your password must: Not be an extremely common password" was displayed in the format shown in Figure 1. This message appeared irrespective of the participants' assigned study condition. Participants were allowed to confirm their password after they modified the password string to not be an exact match for a string in the blacklist.

We analyze *blacklisted passwords*, which were all the intermediary candidate passwords a participant typed during password creation that were at least eight characters long but were rejected by the meter because they were blacklisted; and the *final passwords* participants submitted, which met the requirements of containing at least eight characters and not appearing on the blacklist. Below, we describe the study conditions relevant to our analyses; specifically, meter feedback features and meter scoring stringency.

[3] Amazon's Mechanical Turk. https://www.mturk.com



Create Your Password

Username
user

Password
••••••••

Show Password □

Don't reuse a password from another account! (Why?)

Your password must:

✔ Contain 8+ characters

□ Not be an extremely common password

How to make strong passwords

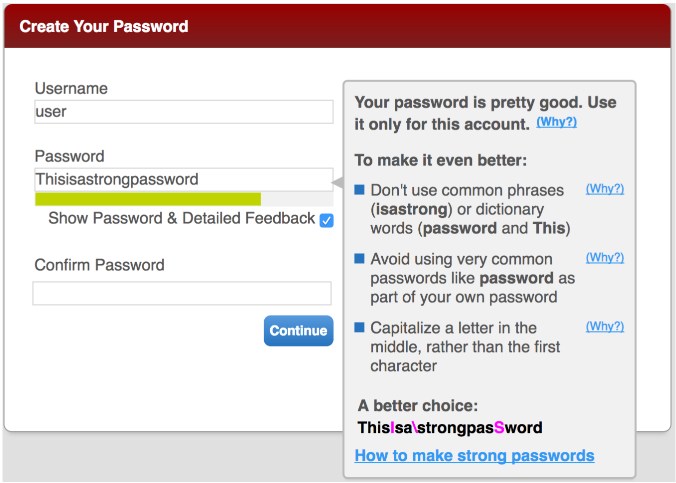
Fig. 1. Requirements text shown to participants in all study conditions during a blacklisted password attempt.

a) Feedback Type Conditions: The first dimension of the password meter was the type or types of feedback participants were shown about the password they had typed. The feedback type conditions were:

- **Standard (Std)**, shown in Figure 2, includes a visual bar and text feedback. Text feedback encompasses both bullet-point feedback about how the password could be improved and a concrete suggested improvement to the user's candidate password;
- **Standard, No Bar (StdNB)** is the same as Standard, but without the visual bar;
- **No Suggested Improvement (StdNS)** is the same as Standard, but without the suggested improvement;
- **Public (Pub)** is the same as Standard, except it does not show the suggested improvement and provides less specific advice on how the password can be improved;
- **Bar Only (Bar)** shows a visual bar that displays password strength, but does not provide any text feedback (other than which composition requirements have been met);
- **No Feedback (None)** gives no feedback on the participant's password (other than which composition requirements have been met).

b) Scoring Stringency Conditions: In conditions other than None and Standard, No Bar, in which scoring was not relevant, participants' passwords were scored at three different stringency levels. These stringency levels determined the mapping between the estimated number of guesses the password could withstand, how much of the bar was filled, and to which color. Blacklisted passwords filled the meter bar to the same level and color in all three scoring stringencies. For participants who saw feedback without a bar, or no feedback at all, we consider a fourth stringency condition (None). Our analyses divide the stringency conditions as:

- **None** if participants did not see a password meter bar (the No Feedback and Standard, No Bar conditions);
- **Low (L)** where the bar is one-third full at 10^4 estimated guesses and two-thirds full at 10^8;
- **Medium (M)** where the bar is one-third full at 10^6 estimated guesses and two-thirds full at 10^{12};
- **High (H)** where the bar is one-third full at 10^8 estimated guesses and two-thirds full at 10^{16}.



Create Your Password

Username
user

Password
Thisisastrongpassword

Show Password & Detailed Feedback ☑

Confirm Password

Your password is pretty good. Use it only for this account. (Why?)

To make it even better:

- Don't use common phrases (**isastrong**) or dictionary words (**password** and **This**) (Why?)
- Avoid using very common passwords like **password** as part of your own password (Why?)
- Capitalize a letter in the middle, rather than the first character (Why?)

A better choice:
ThisIsa\strongpasSword

How to make strong passwords

Fig. 2. The standard meter condition, which includes all feedback elements. In particular, these are the visual *bar* that fills up and changes color to display estimated password strength and *text feedback*. The latter encompasses both up to three bullet points of text pointing out predictable elements of the password and a concrete *suggested improvement*, or automatic modification of the user's password with changes displayed in magenta.

A. *Analysis*

We next describe our approach in analyzing the differences in composition and strength of passwords created by different behavior and experimental groups (defined in this section), the common techniques used to alter a blacklisted password, and the effect on user sentiment of being told a password attempt was blacklisted.

We first post-processed the study data to evaluate all collected keystrokes and tag exact matches to a password on the blacklist as a blacklisted password attempt. In some cases, participants had multiple blacklisted attempts because they replaced one blacklisted password with another that was also blacklisted. In these cases, we used the participant's final blacklisted attempt in our analyses as it was likely intended to be submitted by the participant. For example, a participant who attempted to submit "12345678" and then tried "123456789" more than likely intended to use "123456789" as their final password. This simplifies our analysis, as we consider only one blacklisted password attempt per participant. However, this approach does not consider earlier blacklisted password attempts that participants may have intended as their final password until they were rejected.

To measure password strength, we used the guessability numbers of each final and blacklisted password, calculated by Carnegie Mellon University's Password Guessability Service [8]. In analyzing the use of blacklisted passwords, and subsequent behaviors and modifications, participants were grouped into one of the four following categories:

- Participants whose password-creation session *did not include any passwords that were tagged as blacklisted*;
- Participants who attempted to create a password that was blacklisted, but *did not reuse in any way* the blacklisted password as part of their final password (e.g., 'baggins1' → 'lord1of2the3rings4');

- Participants who attempted to create a password that was blacklisted, and who exhibited *exact reuse* of the full blacklisted password string as a part of their final password (e.g., 'happyday' → 'happyday!');
- Participants who attempted to create a password that was blacklisted, and who exhibited *modified reuse* of their blacklisted password. This behavior was characterized as using parts of, but not the exact, blacklisted password string in their final password (e.g., 'greenpeace' → 'green66peace'), or as the blacklisted password being recognizably transformed (e.g, 'stewart7' → 's1t9e9w8art');

Related to the effect of feedback on password composition and strength, for our analyses we grouped the feedback conditions listed above as:

- Participants who did not see any text feedback (conditions None and Bar);
- Participants who saw text feedback (all others).

To understand how the final password and blacklisted password attempts differed in their composition, we ran paired samples t-tests to analyze the length of the passwords and number of symbols, capital letters, and digits they contained; and a Wilcoxon Signed Ranks test to compare the number of character classes used. Independent samples t-test were used to analyze differences between participant groups in total characters, symbols, capital letters, and digits used in final passwords; and a Mann-Whitney U test compared number of character classes used. The effect of stringency and feedback conditions on the characteristics of participants' blacklisted password attempt and final password were evaluated using two-way ANOVA tests adjusted for post-hoc comparisons with Bonferroni corrections. Analyses were run on the square roots of password length, number of capital letters, digits, and symbols used, as this transformation corrected for gross violations of the assumption of normality such that they were within the bounds acceptable for performing these statistical tests.

We performed a Cox Proportional-Hazards Regression, a survival analysis that was previously used to compare password guessability [26], to evaluate the differences in password strength between participant groups and feedback and stringency conditions. As the starting point of guessing is known but not the endpoint, we use a right-censored model [15]. In a traditional survival analysis, each data point is marked "deceased" (or not "alive") at different times of observation, depending on whether an event has occurred to transition the data point from "alive" to "deceased." For password guessing, analogous to "deceased" and "alive" at a given point in time is whether a password is "guessed" or "not guessed" at a given guess number. We first fit a model with the covariates of stringency, text feedback, and participant group, and included the full factorial interaction terms. To build a parsimonious model, the regression was run again with all three main effects but excluding the interaction terms that were not statistically significant. We used α = 0.05 for all statistical analyses.

We then manually analyzed the blacklisted passwords and final password of the 350 participants who had a blacklisted attempt to understand the techniques used to modify passwords once they were tagged as blacklisted. First, a researcher categorized each password pair of blacklisted password (or final blacklisted attempt if a participant had multiple) and final password as one of three categories involving blacklisted attempts (no reuse, modified reuse, exact reuse). The researcher developed a code book for modification behaviors based on common mangling rules [31], [38] and other behaviors observed in the data set, and then coded the blacklisted/final password pairs as applicable. The researcher's coding was then verified by another researcher who also coded the password pairs using the same codebook. Discrepancies between the codings were resolved after a second review by both researchers. In the end, the only contentious point was whether one particular password pair ('peanutss' → 'pbLE$uanuwt$s') should be considered 'modified reuse' or 'no reuse.' We ultimately decided that the modifications were too prominent and, as such, this pair was classified as 'no reuse.'

Lastly, to evaluate sentiment related to password creation, we analyzed participants' agreement, on a 5-point Likert scale ("strongly disagree," "disagree," "neutral," "agree," "strongly agree"), with statements about whether password creation was difficult, annoying, and fun. This analysis was completed using an ordinal regression, grouping participants by their use, modification, or reuse of blacklisted passwords.

B. *Limitations*

As the design of the original study, in which the passwords we analyze were collected, was based on previous studies used to examine different aspects of passwords [18], [24], [35], [43], a primary limitation shared by our work is that participants were not creating passwords for a real account they would use on the Internet, let alone one of high value. We cannot guarantee that participants put as much consideration into this password as they would for an actual account of high importance. However, prior research by Mazurek et al. [26], and (independently) by Fahl et al. [12] has found this experimental methodology produces reasonable approximations of "real" passwords.

Also, since the meter analyzed and provided feedback as the participant typed in the password, we cannot be sure if the blacklisted passwords captured by the study were ever meant to be submitted as final passwords in the cases where a blacklisted string was a prefix of the final password. In these situations, it could have been that the participant was only typing part of a different (and not blacklisted) password (e.g., "password" as part of "passwordsarefun!"). However, as we will demonstrate, the mere fact that a substring of the final password was on the blacklist led to the password being significantly weaker and, as such, the original intention becomes less of a concern.

Lastly, the wording of the feedback related to blacklisted passwords ("Not be an extremely common password") was subtle and did not directly mention the existence of a blacklist. Different content and formatting choices for messaging regarding the blacklist were not studied, so it is unknown whether the implemented design would be the most effective in conveying to users the reason their password was not accepted by the task. Despite these limitations, we believe that this study has value in examining the composition and strength of passwords created in the presence of a blacklist, as well as in giving initial recommendations to the type of feedback that is more inducive to stronger passwords after a blacklisted attempt.

IV. Participants

In the original study, the source of our data [42], 4,509 participants created a password. Because the current draft of the NIST standard related to passwords recommends password composition policies in which passwords must contain eight or more characters [16], we study only the data collected from the 2,280 participants assigned to create passwords under such a password composition policy. 172 people participated in the study from a mobile device, as determined through their user agent string. Their use of blacklisted passwords did not differ significantly from those not using a mobile device (χ^2 = 0.191, $df = 1$, $p = 0.662$). 52% of participants identified as female, 48% identified as male, and six participants identified as another gender or preferred not to answer. The age of participants ranged from 18 to 80 years old, with a median of 32 and mean of 34.7. Additionally, 82% of participants indicated that they did not major in or have a degree or job in computer science, computer engineering, information technology, or a related field. While there was a significant difference in the distribution of genders across stringency conditions (χ^2 = 15.6, $df = 6$, $p = 0.016$) and age groups across feedback conditions (χ^2 = 9.01, $df = 2$, $p = 0.011$), we found there to be no difference between demographics in use of blacklisted passwords. Therefore, we believe this unequal distribution had minimal effect on our analyses.

V. Results

From the 2,280 participants, 350 participants typed in passwords that were on our blacklist during the password creation process. Of these 350 participants, 228 attempted to use one unique blacklisted password, 75 attempted to use two, and 25 three; the other 22 participants typed in between four and nine different strings that were on the blacklist. Furthermore, of the 350 participants with blacklisted password attempts, 180 exactly reused a blacklisted password as part of their final password, while 106 created significantly different passwords and 64 participants modified their blacklisted password, such as by capitalizing a letter or inserting a digit, before reusing it as part of their final password.

A. *Differences in Password Composition*

We observed differences in length, number of capital letters, symbols, and digits used in composing passwords across different behavioral and experimental groupings of participants. These composition characteristics significantly differ, as later described, between final passwords of participants who attempted a blacklisted password and those who did not, as well as between feedback types and stringency conditions.

Table I shows the average length and number of character classes, capital letters, symbols, and digits used to compose the final passwords submitted by participants and the set of all blacklisted password attempts. Comparing blacklisted passwords with final passwords revealed significant differences for each of the password characteristics tested. Final passwords included more character classes ($Z = -13.7$, $p < 0.001$) and on average were 3.92 characters longer ($t = 16.0$, $df = 349$, $p < 0.001$) and contained 1.57 more digits ($t = 13.1$, $df = 349$, $p < 0.001$) than blacklisted passwords.

TABLE I. Means of password composition characteristics for final and blacklisted passwords and stringency and feedback conditions.

	Length	**Character Classes**	**Capital Letters**	**Symbols**	**Digits**
Final Password					
No blacklisted password attempts	12.1	2.95	1.49	0.76	2.99
With blacklisted password attempts	12.5	2.60	0.89	0.56	2.63
Blacklisted Passwords	8.61	1.63	0.29	0.01	1.14
Stringency					
None	11.0	2.59	0.99	0.34	2.61
Low	11.8	2.85	1.25	0.63	2.54
Medium	12.1	2.89	1.33	0.69	2.95
High	12.6	2.97	1.58	0.86	3.10
Feedback					
Without text feedback	11.31	2.72	0.99	0.49	2.70
With text feedback	12.38	2.94	1.47	0.78	2.99

TABLE II. Statistical results showing composition differences between those who did and did not attempt a blacklisted password.

Characteristic	**Statistic**	***df***	***p*-value**	**95% C.I.**	
Char. classes	$Z = -6.78$		< 0.001		
Length	$t = -1.71$	2,278	0.087	-0.115	0.008
Capital letters	$t = 8.48$	2,278	< 0.001	0.293	0.469
Symbols	$t = 3.64$	510*	< 0.001	0.062	0.206
Digits	$t = 2.63$	2,278	0.009	0.381	0.045

*equal variances not assumed

1) Between Participant Groups: Table I also shows that composition characteristics of the final passwords themselves also differed between participants who had a blacklisted password attempt and those who did not. Specifically, participants who did not attempt a blacklisted password on average used 67.4% more capital letters, 13.7% more digits, and 35.7% more symbols in their final passwords. Interestingly, the length of the final password did not differ significantly between those who attempted a blacklisted password and those who did not, even though blacklisted passwords were found to be significantly shorter than final passwords. Table II summarizes the results of these statistical comparisons.

2) Between Stringency and Feedback Conditions: There were no significant differences in the password composition of blacklisted passwords between different stringency and feedback conditions. This is likely due to how the password meter was implemented, since participants were not shown additional feedback until after their passwords passed the blacklist check. Additionally, blacklisted passwords were scored equally low for all stringency conditions in which a bar was shown.

However, participants' stringency condition significantly impacted the length and the number of capital letters, numerical digits, and symbols in their final password, as shown in Table III. Pairwise comparisons revealed that those in the High stringency conditions created significantly different passwords than those in the Low stringency conditions, using, on average,

TABLE III. STATISTICAL RESULTS SHOWING COMPOSITION DIFFERENCES BETWEEN STRINGENCY AND FEEDBACK CONDITIONS.

	Characteristic	F-Statistic	df	p-value
STRINGENCY	Length	6.01	3	< 0.001
	Capital Letters	4.69	3	0.003
	Symbols	6.50	3	< 0.001
	Digits	5.65	3	< 0.001
FEEDBACK	Length	14.7	1	< 0.001
	Capital Letters	7.09	1	0.008
	Symbols	11.5	1	0.001
	Digits	6.53	1	0.011

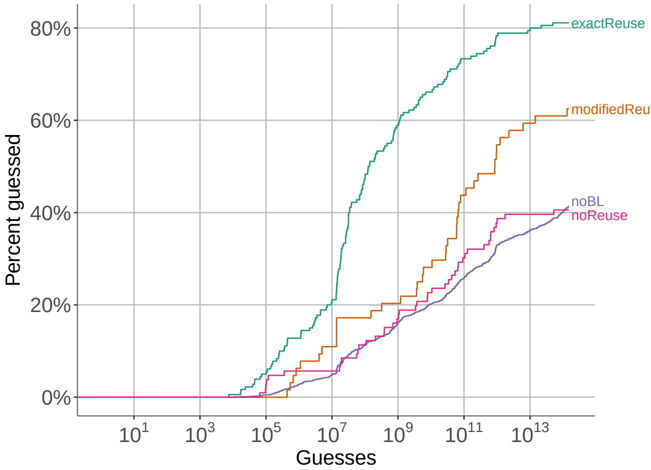


Fig. 3. Guessability of 1class8 passwords created without any blacklisted attempts ("noBL"), with blacklisted attempt but no reuse ("noReuse"), with modified reuse ("modifiedReuse"), and with exact reuse ("exactReuse") (all participant groups).

6.78% more characters, 26.4% more capital letters, 36.5% more symbols, and 22.0% more digits. Those in the High stringency conditions also differed significantly than those in the Medium stringency condition, using 18.8% more capital letters and 24.6% more symbols in their final password.

Furthermore, Table III also shows that whether or not a participant saw text feedback was significant for each composition characteristic. Participants who were in conditions in which they saw text feedback had higher means for each characteristic, as seen in Table I, creating final passwords that were, on average, 9.46% longer and containing 48.4% more capital letters, 59.2% more symbols and 10.7% more digits. Interaction between participants' stringency condition and the presence of text feedback was not significant for any of the tested composition characteristics.

B. Differences in Password Strength

Considering the different behaviors observed in relation to the use of blacklisted passwords, not reusing a blacklisted password attempt in the final password led participants to create stronger passwords, both when analyzing the full participant pool and when limiting it to only those that had a blacklisted attempt. Password strength was also impacted by participants' stringency and feedback conditions.

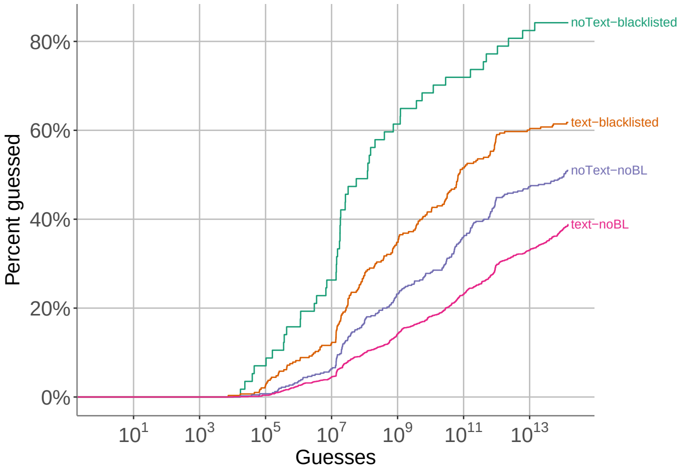


Fig. 4. Guessability of 1class8 passwords created without guidance from text feedback (noText) and those created with guidance from text feedback (text) associated with the presence of a blacklisted password.

1) Between Participant Groups: As Figure 3 shows, when compared to participants who did not have any blacklisted attempts, those who reused a previously blacklisted attempt as part of their final password (modified or not) created weaker passwords. In particular, results from a Cox regression showed that those who exactly reused the blacklisted password as a part of their final password created passwords that were 3.89 times more likely ($p < 0.001$) to be guessed than those who never had a blacklisted attempt. Those who modified the blacklisted password before reusing it created passwords that were 1.91 times more likely ($p < 0.001$) to be guessed in the same comparison. Lastly, those who created a completely new password after a blacklisted attempt created final passwords that were not significantly different in strength ($p = 0.602$) from those who never had a blacklisted attempt.

Finally, since there was no significant difference between participants with no blacklisted attempt and those that did not reuse their attempt, we ran another Cox regression, excluding participants who did not have a blacklisted attempt, to compare both reuse groups to those that did not reuse their blacklisted attempt. Participants who exhibited modified reuse of their blacklisted password created final passwords that were 1.69 times more likely ($p = 0.018$) to be guessed than those who did not reuse their blacklisted attempt. Even more significant, those who exactly reused their blacklisted attempt created passwords that were 3.31 times more likely ($p < 0.001$) to be guessed.

2) Between Stringency and Feedback Conditions: As can be seen in Table IV, when considering all participants, increasing stringency levels and providing text feedback led participants to create stronger passwords. When considering only the participants who had a blacklisted attempt, the stringency of the password meter bar is no longer significant, but the presence of text feedback has a stronger effect.

In general, passwords created in conditions with text feedback were 30.3% less likely ($p < 0.001$) to be guessed than those created in conditions with no text feedback. When considering only participants who had a blacklisted attempt, those who created passwords with text feedback had final passwords that were 41.8% less likely (p $=$ 0.005) to be

TABLE IV. COX REGRESSION RESULTS FOR STRINGENCY, PRESENCE OF TEXT FEEDBACK AND PARTICIPANT GROUPS DIVIDED BY ANALYSIS OVER ALL PARTICIPANT GROUPS OR ONLY THOSE WITH BLACKLISTED ATTEMPTS.

	Variable (baseline)	Effect	p-value	95% C.I.	
ALL PARTICIPANT GROUPS	**Stringency (None)**				
	Low	0.876	0.403	0.642	1.20
	Medium	0.757	0.038	0.583	0.985
	High	0.732	0.020	0.563	0.951
	Text feedback (No text feedback)				
	With text feedback	0.697	< 0.001	0.586	0.830
	Participant group (No blacklisted attempt)				
	No reuse	1.09	0.602	0.789	1.48
	Modified reuse	1.91	< 0.001	1.39	2.62
	Exact reuse	3.89	< 0.001	3.25	4.65
BLACKLISTED GROUPS	**Stringency (None)**				
	Low	1.07	0.858	0.508	2.25
	Medium	1.09	0.806	0.560	2.11
	High	0.969	0.924	0.508	1.85
	Text feedback (No text feedback)				
	With text feedback	0.582	0.005	0.400	0.846
	Participant group (No reuse)				
	Modified reuse	1.69	0.018	1.09	2.60
	Exact reuse	3.31	< 0.001	2.34	4.69

guessed. Figure 4 shows the impact that text feedback had in password strength when considering passwords created by those without a blacklisted attempt and those who had at least one blacklisted attempt.

Furthermore, participants who created their password in the stringency None condition, in which they did not see a visual bar, created significantly weaker passwords than those in the Medium stringency ($p = 0.038$) and High stringency ($p = 0.02$) conditions; passwords created under Medium stringency were 24.3% less likely to be guessed, while those created under High stringency were 26.8% less likely to be guessed.[4]

There was no significant difference between passwords created in stringency conditions None and Low ($p = 0.403$). However, when considering only participants who had a blacklisted attempt, stringency is no longer significant to password strength.

C. How Blacklisted Passwords Are Changed

As mentioned before, 180 of the 350 participants who typed in a blacklisted password reused their attempt as part of their final password. Exact reuse of a blacklisted attempt occurred 175 times at the beginning of the final password, four times in the middle, and one time at the end of a final password. Table V summarizes the the distribution of modification techniques used by participants who reused their blacklisted password attempts, displays the average number of characters appended or changed using each technique, and presents examples to illustrate each technique.

The total number of modifications used to modify the blacklisted attempt to the final password significantly differed (χ^2 = 44.0, p < 0.001) between those who exactly reused their blacklisted attempt and those who did not. 70.6% of participants who exactly reused their blacklisted attempt used only one type of modification technique to create their final password, while another 24.4% used two and 5.00% used three. In comparison, 36.5% of participants who reused a modified version of their blacklisted attempt used one modification technique, 30.3% used two, and 22.2% used three.

For participants who had exact reuse of their blacklisted attempt, the majority of the modifications were made to the end of the password, such that the blacklisted password was kept as a prefix of the final password (175 out of 180). For participants who reused a modified version of their blacklisted attempt, 17 out of 29 of the capitalizations occurred at the beginning of the blacklisted attempt and all character transformations occurred in the middle. Deletions mostly occurred at the end of a blacklisted attempt, as did the addition letters and digits. The placement of additional words and symbols, however, were more varied.

Lastly, participants also engaged in using general patterns and keyboard patterns as part of their modifications. Participants modified their original keyboard pattern by changing directions (e.g. q1w2e3r4 → 1q2w3e3w2q1), keyboard lines used (e.g. 7890uiop → uiophjkl), or the order (e.g. abc123xyz → xyz123abc). Another way participants modified patterns was by continuing the pattern (e.g. p1a2s3s4 → p1a2s3s4w5o6r7d8), adding numbers (e.g. alskdjfhg → alskdjfhg1029384756), or changing to capitalized letters and symbols by pressing the shift key (e.g. 1qazxsw2 → 1qazxsw2!QAZXSW@). Only one participant used a pattern as an additional extension to the original blacklisted password (Computer → 1qazcomputer@WSX).

D. Effect of Blacklisted Passwords on Sentiment

The use and reuse of blacklisted passwords also had an impact on sentiment toward password creation. Participants who expended the effort to differentiate the final password from a blacklisted attempt found the task more difficult and annoying, measured on a 5-point Likert scale. As shown in Figure 5, there was no difference in their opinion of the task being fun compared to those who did not change their password attempt.

More specifically, when compared to participants who did not have a blacklisted attempt, those who created a new password after a blacklisted attempt and those who modified it before reusing it were significantly more likely to agree that the experience was annoying (both $p < 0.001$). However, participants who directly reused the blacklisted attempt did not find the task significantly more annoying ($p = 0.891$) than those who did not have a blacklisted attempt, but did find the task less difficult ($p = 0.010$). On the other hand, participants who modified or created a new password after a blacklisted attempt

[4] In the original analysis of the experiment whose data we use [42], the researchers found that stringency did not have a statistically significant effect for the password-composition policy we study. Two factors cause this discrepancy. First, we use the None condition as our baseline for comparison, whereas the original analysis used Low stringency conditions as the baseline. Second, the original analysis analyzed many more research questions, which required that p-values be corrected for multiple testing.

TABLE V. Modifications applied by participants to their blacklisted attempt divided by reuse type (exact or modified). Modification techniques were applied either only once or more than once per password. Averages represent the average number of characters appended or changed using each modification technique.

	Modified Reuse				*Exact Reuse*			
Modification Type	**Only one**	**More than one**	**Avg.**	**Example**	**Only one**	**More than one**	**Avg.**	**Example**
Added digits	5	32	2.97	dartmouth → dart1mouth	45	84	2.19	harrypotter → harrypotter0
Added symbols	16	2	1.11	rileydog → riley&dog30	35	11	1.26	SanDiego → SanDiego$$
Added words	3	2	1.60	missions → missionnew30	29	9	1.42	familyguy → familyguyguy
Added letters	6	4	1.90	test12345 → testz12345	9	14	3.04	67876787 → i67876787
Continued a pattern					6			one1two2 → one1two2three3
Capitalized letters	25	4	1.21	purple88 → Purple88				
Transformed letters	5	4	1.78	roflcopter → r0flcopter				
Deleted digits	3	5	2.13	lakers88 → lakers40324				
Deleted letters	2	4	2.67	password123 → pword123				
Lower cased letters	3	0	1.00	Stallion → Tst@llions!				
Added a pattern	1			Computer → 1qazcomputer@WSX				
Shifted digits	1			purple15 → pur15ple				
Changed pattern	6			qawsedrf → 1qwsdcf				

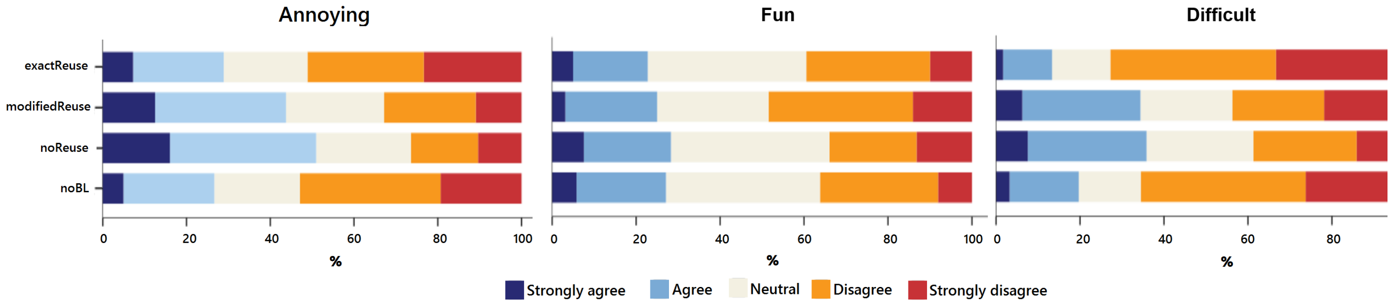


Fig. 5. Participants' agreement to whether the task was annoying, fun, and difficult, rated on a 5-point Likert scale from strongly disagree to strongly agree.

were more likely to agree that the experience was difficult and annoying ($p = 0.006$ and $p < 0.001$, respectively).

VI. Discussion

Previous work has shown that using a blacklist can be effective at forcing participants to create stronger passwords [22], [23], [34]. However, what has been missing is an evaluation and understanding of how participants behave under a policy with blacklists. In this section, we use our findings to provide recommendations to website administrators on how to best leverage password blacklists.

A. Use Blacklists, but Check for Reuse

In our analysis, 51.4% of participants who attempted a blacklisted password reused the entire blacklisted password as a part of their final password. An additional 18.0% reused their blacklisted attempt in a modified form. This is not surprising, as reuse of passwords across accounts is common among Internet users [9], [50]. Since blacklisted passwords are so common, they are targeted by password cracking tools [31], [38]. This is why the final passwords created by those who exactly reused their blacklisted attempt were over three times more easily guessed than those created by participants who did not reuse a blacklisted attempt in any form and those created by participants with no blacklisted attempts. Participants who created final passwords by reusing their blacklisted passwords with more significant modifications also had weaker passwords, but to a lesser degree.

Echoing previous work [22], we found that including a blacklist in the password creation process leads users to create stronger passwords. As such, we build upon previous recommendations of checking candidate passwords against a blacklist and further advise that system administrators put in place checks to guarantee that no simple variations of blacklisted passwords are being used as part of a final password.

Based on our analysis of how blacklisted passwords were modified to be reused in final passwords, we recommend that these checks strip all candidate passwords of digits and symbols, and perform case-insensitive searches for the string in the website's blacklist to prevent the use of easily guessed modifications to a blacklisted password. While character transformations can also be used to modify a blacklisted attempt, we observed only very few such modifications, so it is likely this behavior is not as common as inserting digits and symbols to modify a password.

B. Provide Feedback on How to Make Passwords Stronger

Our analyses support a previous conclusion [42] that users can be nudged into creating stronger passwords. The presence of text feedback advising participants on how to make their password stronger led to stronger, more complex passwords across all participant groups. However, this is more pronounced when analyzing only participants who had at least one blacklisted attempt. In such cases, the presence of text feedback had an even stronger effect on password strength, suggesting that users who attempt a blacklisted password can especially benefit from guidance on how to make a better one.

Furthermore, our findings suggest that the content of this text feedback should be tailored to discourage users from reusing their blacklisted password in any form. Specifically, feedback discouraging users from adding digits, symbols, and dictionary words to the end of their blacklisted password attempt would address the majority of the modifications made by our participants. Additionally, as these users have already demonstrated an inclination toward choosing a simple password, the feedback could more strongly recommend the creation of a complex password that does not use common patterns or phrases. This presence of actionable feedback might also mitigate any increase in negative feelings caused by the added work of creating an unrelated password after a blacklisted attempt.

At the same time, website operators should strive not to overwhelm their users with too much feedback. If a large amount of text is displayed, users might not read it, and continue relying on harmful practices such as using common passwords. The balance between the length and utility of password encouragement or feedback text is an area to be further explored.

VII. Conclusion

In this paper we analyzed 2,280 passwords created during a previous study of password creation in which participants were prohibited from using passwords appearing on a blacklist. We found that participants who initially tried to use a blacklisted password ultimately created passwords with fewer characters, capital letters, digits, and symbols. Additionally, those who reused a blacklisted password in their final password created passwords that were significantly easier to guess.

The addition of a blacklist to a password policy and text feedback to guide users in improving their passwords are features that have been proven to help users make stronger passwords [22], [34], [42], and are ones that are not difficult to implement. With the additional understanding our analyses provide of how users react to failure of a password creation attempt due to blacklisting, feedback and guidance can be more tailored to nudge users toward better behaviors.

Blacklist checks should go beyond mere exact comparisons and look for any form of reuse of blacklisted passwords. In particular, stripping passwords of digits and symbols, and performing case-insensitive searches of the string in the blacklist, were identified as techniques that would have prevented participants from making only simple modifications to a blacklisted password. Furthermore, text feedback should be used to help users understand that reuse and trivial modifications of blacklisted attempts are harmful to the strength of their password.

Acknowledgments

This work was supported in part by a gift from the PNC Center for Financial Services Innovation, NSF grant CNS-1012763, and by John & Claire Bertucci.

References

[1] A. Adams, M. A. Sasse, and P. Lunt, “Making passwords secure and usable,” in *Proc. HCI on People and Computers*, 1997.

[2] A. Biryukov, D. Dinu, , and D. Khovratovich, “Version 1.2 of Argon2,” https://password-hashing.net/submissions/specs/Argon-v3.pdf, July 8, 2015.

[3] J. Bonneau, “The science of guessing: Analyzing an anonymized corpus of 70 million passwords,” in *Proc. IEEE Symposium on Security and Privacy*, 2012.

[4] J. Bonneau and E. Shutova, “Linguistic properties of multi-word passphrases,” in *Proc. USEC*, 2012.

[5] M. Burnett, “Today I am releasing ten million passwords,” https://xato.net/today-i-am-releasing-ten-million-passwords-b6278bbe7495#.s11zbdb8q, February 9, 2015.

[6] W. E. Burr, D. F. Dodson, , and W. T. Polk, “Electronic authentication guideline,” http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-63ver1.0.2.pdf, National Institute of Standards and Technology, Tech. Rep., 2006, accessed Dec. 2016.

[7] W. E. Burr, D. F. Dodson, E. M. Newton, R. A. Perlner, W. T. Polk, S. Gupta, and E. A. Nabbus, “Nist special publication 800-63-2 - electronic authentication guideline,” National Institute of Standards and Technology, Tech. Rep., 2013.

[8] Carnegie Mellon University, “Password guessability service,” https://pgs.ece.cmu.edu, 2015.

[9] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang, “The tangled web of password reuse,” in *Proc. NDSS*, 2014.

[10] X. de Carné de Carnavalet and M. Mannan, “From very weak to very strong: Analyzing password-strength meters,” in *Proc. NDSS*, 2014.

[11] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, “Does my password go up to eleven? the impact of password meters on password selection,” in *Proc. CHI*, 2013.

[12] S. Fahl, M. Harbach, Y. Acar, and M. Smith, “On the ecological validity of a password study,” in *Proc. SOUPS*, 2013.

[13] D. Florêncio and C. Herley, “A large-scale study of web password habits,” in *Proc. WWW*, 2007.

[14] D. Florêncio, C. Herley, and P. C. van Oorschot, “An administrator’s guide to internet password research,” in *Proc. USENIX LISA*, 2014.

[15] J. Fox and S. Weisberg, *An R Companion to Applied Regression (Online Appendix)*, 2nd ed. Sage Publications, 2011, https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf.

[16] P. A. Grassi, J. L. Fenton, E. M. Newton, R. A. Perlner, A. R. Regenscheid, W. E. Burr, J. P. Richer, N. B. Lefkovitz, J. M. Danker, Y.-Y. Choong, K. K. Greene, and M. F. Theofanos, “Draft nist special publication 800-63b - digital authentication guideline,” https://pages.nist.gov/800-63-3/sp800-63b.html, National Institute of Standards and Technology, Tech. Rep., 2016, accessed Dec. 2016.

[17] C. Herley, “So long, and no thanks for the externalities: the rational rejection of security advice by users,” in *Proc. NSPW*, 2009, pp. 133–144.

[18] J. H. Huh, S. Oh, H. Kim, K. Beznosov, A. Mohan, and S. R. Rajagopalan, “Surpass: System-initiated user-replaceable passwords,” in *Proc. CCS*, 2015.

[19] P. Inglesant and M. A. Sasse, “The true cost of unusable password policies: password use in the wild,” in *Proc. CHI*, 2010.

[20] B. Ives, K. R. Walsh, and H. Schneider, “The domino effect of password reuse,” *C. ACM*, vol. 47, no. 4, pp. 75–78, 2004.

[21] M. Jakobsson and M. Dhiman, “The benefits of understanding passwords,” in *Proc. HotSec*, 2012.

[22] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez, “Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms,” in *Proc. IEEE Symposium on Security and Privacy*, May 2012.

[23] S. Komanduri, R. Shay, L. F. Cranor, C. Herley, and S. Schechter, “Telepathwords: Preventing weak passwords by reading users’ minds,” in *Proc. USENIX Security*, 2014.

[24] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman, "Of passwords and people: Measuring the effect of password-composition policies," in *Proc. CHI*, 2011.

[25] C. Kuo, S. Romanosky, and L. F. Cranor, "Human selection of mnemonic phrase-based passwords," in *Proc. SOUPS*, 2006.

[26] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur, "Measuring password guessability for an entire university," in *Proc. CCS*, 2013.

[27] W. Melicher, D. Kurilova, S. M. Segreti, P. Kalvani, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek, "Usability and security of text passwords on mobile devices," in *Proc. CHI*, 2016.

[28] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor, "Fast, lean, and accurate: Modeling password guessability using neural networks," in *Proc. USENIX Security*, 2016.

[29] B. Obama, "Protecting U.S. innovation from cyberthreats," Available at http://www.wsj.com/articles/protecting-u-s-innovation-from-cyberthreats-1455012003, February 2016, accessed on Dec, 2016.

[30] C. Percival, "Stronger key derivation via sequential memory-hard functions," http://www.tarsnap.com/scrypt/scrypt.pdf, 2009.

[31] A. Peslyak, "John the ripper," http://www.openwall.com/john/, 1996.

[32] N. Provos and D. Mazieres, "A future-adaptable password scheme," in *Proc. USENIX ATC*, 1999.

[33] S. Schechter, C. Herley, and M. Mitzenmacher, "Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks," in *Proc. USENIX HotSec*, 2010.

[34] R. Shay, L. Bauer, N. Christin, L. F. Cranor, A. Forget, S. Komanduri, M. L. Mazurek, W. Melicher, S. M. Segreti, and B. Ur, "A spoonful of sugar? The impact of guidance and feedback on password-creation behavior," in *Proc. CHI*, 2015.

[35] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, L. F. Cranor, and N. Christin, "Can long passwords be secure and usable?" in *Proc. CHI*, 2014.

[36] A. Sotirakopoulos, I. Muslukov, K. Beznosov, C. Herley, and S. Egelman, "Motivating users to choose better passwords through peer pressure," in *SOUPS (Poster)*, 2011.

[37] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton, "Analysis of end user security behaviors," *Comp. & Security*, vol. 24, no. 2, pp. 124–133, 2005.

[38] J. Steube, "Rule-based attack," https://hashcat.net/wiki/doku.php?id=rule_based_attack, 2017.

[39] E. Stobert and R. Biddle, "The password life cycle: User behaviour in managing passwords," in *Proc. SOUPS*, 2014.

[40] D. Terdiman, "Google security exec: 'passwords are dead'," Available at https://www.cnet.com/news/google-security-exec-passwords-are-dead/, sep 2013, accessed on Dec, 2016.

[41] J. Titcomb, "Do you have one of the most common passwords? They're ridiculously easy to guess," Available at http://www.telegraph.co.uk/technology/2016/01/26/most-common-passwords-revealed---and-theyre-ridiculously-easy-to/, mar 2016, accessed on Dec, 2016.

[42] B. Ur, F. Alfieri, M. Aung, L. Bauer, N. Christin, J. Colnago, L. F. Cranor, H. Dixon, P. E. Naeini, H. Habib, N. Johnson, and W. Melicher, "Design and evaluation of a data-driven password meter," in *Proc. CHI*, 2017.

[43] B. Ur, P. G. Kelly, S. Komanduri, J. Lee, M. Maass, M. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor, "How does your password measure up? The effect of strength meters on password creation," in *Proc. USENIX Security*, August 2012.

[44] 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 *Proc. SOUPS*, 2015.

[45] R. Veras, C. Collins, and J. Thorpe, "On the semantic patterns of passwords and their security impact," in *Proc. NDSS*, 2014.

[46] R. Veras, J. Thorpe, and C. Collins, "Visualizing semantics in passwords: The role of dates," in *Proc. VizSec*, 2012.

[47] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. B. Tai, J. Cook, and E. E. Schultz, "Improving password security and memorability to protect personal and organizational information," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 744–757, 2007.

[48] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proc. CCS*, 2010.

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

[50] Y. Zhang, F. Monrose, and M. K. Reiter, "The security of modern password expiration: An algorithmic framework and empirical analysis," in *Proc. CCS*, 2010.