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
{htq, jcolnago, wmelicher, ssegreti, lbauer, nicolasc, lorrie}@andrew.cmu.edu
†University of Chicago
blase@uchicago.edu

***Abstract*—Attackers often target common passwords in guessing attacks. Some website administrators have reacted to this by making these passwords ineligible for use on their site. While past research has shown that adding a blacklist 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 the impact on sentiment toward password creation that occurs when a user attempts to create 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. However, our results indicate that text feedback provided by a password meter mitigated this effect.**

I. Introduction

Some Internet services, including Microsoft[1] and Google[2], attempt to reduce the predictability of passwords on their systems by rejecting users' attempts to create passwords that are on a blacklist of ones commonly used. While past research has studied how people create and use passwords [27], [33], [34], [43] and has found that robust blacklists can reduce how easily user-chosen passwords can be guessed [22], [33], 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.

In the past, system administrators have looked to the National Institute of Standards and Technology (NIST) for guidance on password policy [52]. 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) [15]. 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 [40]. As such, listing these values in 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 [16], [19], [24], [48].

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" [15]. However, the proposal does not include a recommendation to provide guidance that could assist users in creating a better password, even though previous work has found that providing such feedback leads to stronger passwords [41]. Therefore, it becomes important to determine the common modifications users apply to blacklisted password attempts—and their impact on password strength—so that the feedback can nudge users away from them.

In this paper, we examined these questions using 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 blacklisted password attempts and final passwords participants created. We manually inspected 350 candidate passwords that were rejected because they matched passwords in our blacklist, as well as the passwords that 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 create a blacklisted password affected the composition and strength

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

In submission to USEC'17.

of participants' final passwords and participants' sentiment toward the password-creation task.

Our analyses found that passwords created by 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 70% 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 directly reused a blacklisted attempt as a substring. However, participants who spent effort to more comprehensively change their blacklisted attempt found password creation to be more difficult and annoying than those who did not.

The primary contribution of this work is the analysis of passwords created under a blacklist in relation to the consequent composition, strength, and sentiment differences between different groupings of participants based on their experimental conditions and use or reuse of blacklisted passwords. With this, we provide data-driven recommendations on the best way to leverage blacklists and the feedback website operators and system administrators can provide to users who attempted blacklisted passwords. These recommendations take into account common techniques used to alter a blacklisted password and the effect on user sentiment of being told a password attempt was blacklisted.

The remainder of this paper unfolds 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. Following, 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 [43]. This has led to a move toward multi-factor authentication, where factors belonging to the categories "something you know" (e.g., passwords), "something you have" (e.g., token), "something you are" (e.g., biometrics), are combined to provide a more secure authentication process [6]. This move has been reinforced by technology companies claiming that "passwords are dead" [39] 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 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. Online guessing attacks, in which an attacker tries to authenticate to a live system by guessing users' passwords, are a major threat to practical security [13]. Because systems that follow security best practices will rate-limit authentication attempts and may require secondary authentication following a number of incorrect attempts, attackers rely on guessing two types of passwords that have a relatively high probability of success.

Commonly used passwords are the first source of high-probability password guesses in an offline guessing attack, in which an attacker has no limit to the number of guesses they can try [13]. Users sometimes create passwords that are easy to guess [3], [18], [44], [47] even for important accounts [11], [26]. Passwords frequently contain words and phrases [4], [25], [43], 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], [41]. 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 such high-probability 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 [8], [12], [20], [38]. 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 [13], 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 different 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 listing only dozens of extremely common passwords [9], [42] 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 can take different approaches to determining what constitutes a password being on the blacklist, varying factors such as case-sensitivity, whether the full password or only a substring must be on the blacklist, and similar factors [9], [22], [42].

Some 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 [49]. Because they 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 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) [22]. 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.

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) [33]. 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. The authors found that having a blacklist increased security without making password recall significantly more difficult, yet decreased other aspects of usability in password creation.

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 [9], [13], [42] and the use of blacklists features prominently in current NIST draft password guidance [15].

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], [16], [19], [24], [36], [48]. 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 to guide users towards stronger passwords for accounts the user deems meaningful [10], [42]. Different meters rely on client-side heuristics [9], [51], 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 [41], 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 [35].

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 [41]. Recruitment occurred on Amazon's Mechanical Turk[3] and was limited to those aged 18 and older, located within the United States. Participants were required to complete the task on 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 [50], 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 regardless of the participants' assigned study condition. Participants were allowed to submit 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 that 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

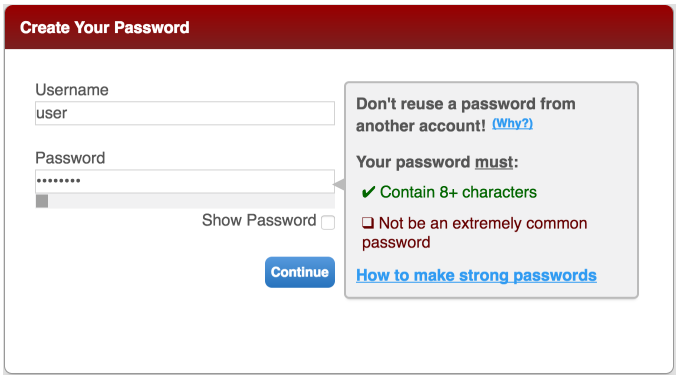

Fig. 1. Feedback shown to participants during a blacklisted password attempt

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)** includes the password meter bar and text feedback. Text feedback includes both (public) feedback about how the password could be improved and a suggested improvement to the user's candidate password;
- **Standard, No Bar (StdNB)** is the same as Standard, without the password meter bar. Only text feedback is provided;
- **No Suggested Improvement (StdNS)** is the same as Standard, without the suggested improvement;
- **Public (Pub)** is the same as Standard, except it does not show the suggested improvement and only provides general advice on how the password can be improved;
- **Bar Only (Bar)** shows the password meter bar to gauge 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: Participants who saw the password meter bar 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. For participants who saw feedback without a bar, we consider a fourth stringency condition. Our analyses divide the stringency conditions as:

- **None** if participants did not have any feedback type (feedback condition None);
- **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}.

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 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 this was likely intended to be submitted by the participant. For example, a participant who attempted to submit "12345678" and then tried "abcdefgh" more than likely intended to use "abcdefgh" as their final password.

To measure password strength we used the guessability numbers of each final and blacklisted password, calculated by Carnegie Mellon University's Password Guessability Service [7]. In analyzing the use of blacklisted passwords and subsequent behaviors and modifications, participants were grouped into one of the four following categories:

P1 Participants whose password-creation session did not include any passwords that were tagged as blacklisted;

P2 Participants who attempted to create a password that was prohibited because it was blacklisted, but did not reuse the blacklisted password as part of their final password;

P3 Participants who attempted to create a password that was prohibited because it was blacklisted, and whose blacklisted candidate password was modified to produce their final password such that it did not occur as a substring of the final password;

P4 Participants who attempted to create a password that was prohibited because it was blacklisted, and whose blacklisted candiate password was a substring of their final password.

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

F1 Participants who did not see any text feedback (conditions **None** and **Bar**);

F2 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 to compare 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 [14]. 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 different points in time is whether a password is "guessed" or "not guessed" at different guess numbers. We first fit a model with the covariates of stringency, text feedback, and participant group (P1-P4) 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 (P2, P3, P4). They also developed a code book for modification behaviors based on common mangling rules [31], [37] 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. Any conflicts between the codings were resolved.

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 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 [17], [24], [34], [42], 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 in 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 Fahl et al. [11] has studied this limitation and has found this methodology to be a reasonable means of attaining intermediaries for this type of password.

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

The password creation task was completed by a total of 4,509 participants. Our analyses utilized the data collected from 2,280 participants assigned to "1class8" conditions, and the results we report from here on examined only those 2,280. 172 people participated in the study from a mobile device, as determined through their user agent string. 52% of participants identified as female, 48% identified as male, and 6 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 engineer, 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. From 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, from the 350 participants with blacklisted password attempts, 180 directly reused a blacklisted password as part of their final password, while 106 created significantly different passwords

TABLE I. MEANS OF PASSWORD COMPOSITION CHARACTERISTICS FOR FINAL AND BLACKLISTED PASSWORDS

Password Type	Length	Character Classes	Capital Letters	Symbols	Digits
Final Password					
Those without a blacklisted password attempt	12.1	2.95	1.49	0.76	2.99
Those with a blacklisted password attempt	12.5	2.60	0.89	0.56	2.63
Blacklisted Passwords	8.61	1.63	0.29	0.01	1.14

and 64 participants modified the 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 shows 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.

1) Between Participant Groups: Composition characteristics of the final passwords themselves also differed between participants who had at least one blacklisted password attempt (P2-P4) and those who did not (P1). These groups significantly differed in the number of character classes used. 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 I summarizes the means of each password characteristic for both groups, while Table II summarizes the results of the 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 IV. Pairwise comparisons revealed that those in the High stringency conditions created significantly different passwords than those in the Low stringency conditions, using, on average, 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 IV also shows 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 III, 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.

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

TABLE III. MEANS OF PASSWORD COMPOSITION CHARACTERISTICS FOR STRINGENCY AND FEEDBACK CONDITIONS

Condition	Length	Character Classes	Capital Letters	Symbols	Digits
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 IV. 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

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

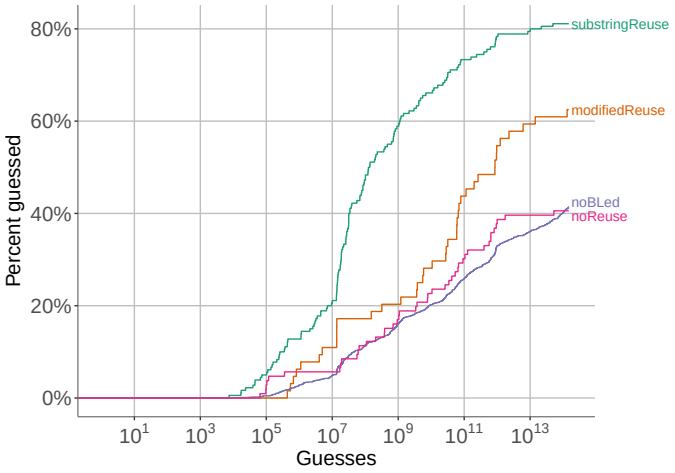

Fig. 2. Guessability of 1class8 passwords created without any blacklisted attempts ("noBL"), with blacklisted attempt but no reuse ("noReuse"), with modified reuse ("modifiedReuse"), and with exact reuse ("substringReuse") (all participant groups).

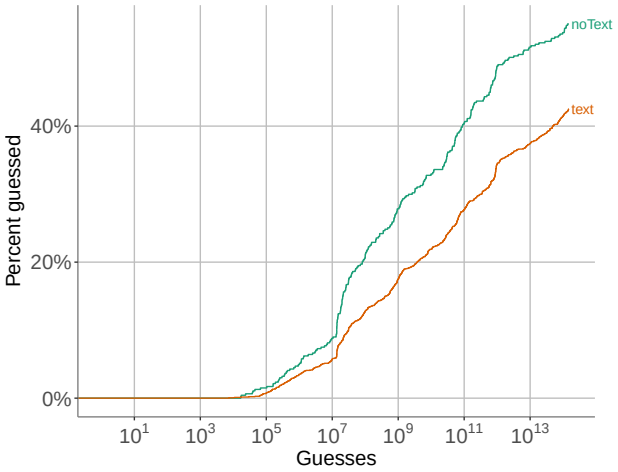

Fig. 3. Guessability of 1class8 passwords created without guidance from text feedback (noText) and those created with guidance from text feedback (text) (all participant groups).

pool and when limiting it to only those that had a blacklisted attempt. Password strength was also impacted by participants' stringency and feedback conditions.

1) Between Participant Groups: As Figure 2 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 reused the blacklisted password as a substring of their final password (P4) created passwords that were 3.89 times more likely ($p < 0.001$) to be guessed than those who never had a blacklisted attempt the blacklisted password before reusing it (P3) created passwords that were 1.91 times more likely ($p < 0.001$) to be guessed in the same comparison. Those who created a completely new password after a blacklisted attempt (P2) did not show a significant difference ($p = 0.602$) from those who never had one.

Finally, since there was no significant difference between participants with no blacklisted attempt and those that did not

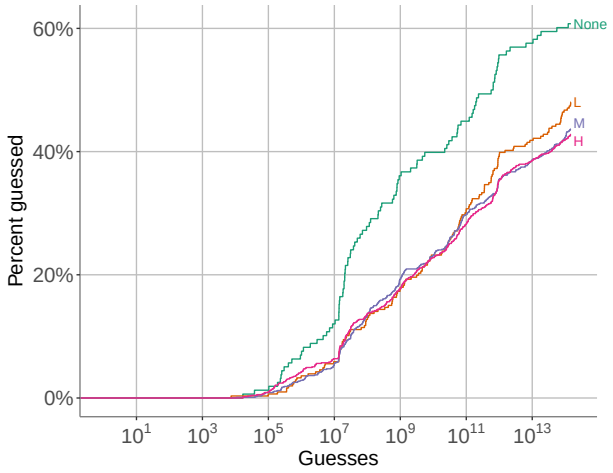

Fig. 4. Guessability of 1class8 passwords created without any scoring stringency ("None"), and with low ("L"), medium ("M"), and high stringency levels ("H") (all participant groups).

reuse their attempt, we ran another Cox regression excluding participants who did not have a blacklisted attempt to compare the different reuse groups (P3 and P4) to those that did not reuse their blacklisted attempt (P2). Having a blacklisted attempt led participants to create significantly weaker passwords if they reused a modified version of their blacklisted attempt (Effect = 1.69, $p = 0.018$) or reused the blacklisted attempt as a direct substring of their final password (Effect = 3.31, $p < 0.001$).

2) Between Stringency and Feedback Conditions: Increasing stringency levels and providing text feedback led participants to create stronger passwords, as can be seen in Figures 3 and 4. 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 (see Table V).

More specifically, 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. Additionally, when considering 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 guessed.

Furthermore, participants who created their password in the None stringency condition, without the aid of the password meter 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. 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, a large proportion of participants who typed in a blacklisted password reused their attempt as part of

TABLE V. 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 (P1-P4)	**Stringency (None)**		0.035		
	High	0.732	0.020	0.563	0.951
	Low	0.876	0.403	0.642	1.20
	Medium	0.757	0.038	0.583	0.985
	Text Feedback (No Text Feedback)				
	With Text Feedback	0.697	< 0.001	0.586	0.830
	Participant Group (No Blacklisted Attempt)		< 0.001		
	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 (P2-P4)	**Stringency (None)**		0.888		
	High	0.969	0.924	0.508	1.85
	Low	1.07	0.858	0.508	2.25
	Medium	1.09	0.806	0.560	2.11
	Text Feedback (No Text Feedback)				
	With Text Feedback	0.582	0.005	0.400	0.846
	Participant Group (No Reuse)		< 0.001		
	Modified Reuse	1.69	0.018	1.09	2.60
	Exact Reuse	3.31	< 0.001	2.34	4.69

their final password. Table VI summarizes where participants reused their blacklisted attempt in their final password, as well as the techniques used to change blacklisted passwords so that they could pass the blacklist check.

By far the most commonly used modification was adding or inserting at least one digit into the blacklisted password attempt, which was applied by 173 participants. Other common techniques were adding or inserting a symbol, adding an entire word or letter, and capitalizing at least one letter in the password. Additionally, 50% of participants with blacklisted attempts directly reused their blacklisted attempt as a prefix of their final password (i.e., they appended at least one character to the end). Interestingly, the few participants who attempted a blacklisted keyboard pattern were not deterred from trying other, less obvious patterns that were not on the blacklist.

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. 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 (P1), those who created a new password after a blacklisted attempt (P2) and those who modified it before reusing it (P3) were significantly more likely to agree that the experience was annoying (both $p < 0.001$). However, participants who directly reused the blacklisted attempt (P4) 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 were more likely to agree that the experience was difficult ($p = 0.006$ and $p < 0.001$, respectively).

TABLE VI. MODIFICATION USED TO SUBMIT BLACKLISTED PASSWORDS

Reuse Position of Blacklisted Substring	**Number of Participants**
Prefix	175
Middle	4
Suffix	1
Modification Type	**Number of Participants**
Added at least one digit	173
Added at least one symbol	68
Added at least one word	46
Added at least one letter	37
Capitalized at least one letter	30
Used a character transformation (e.g. s to $)	9
Added at least one word	7
Changed keyboard pattern	6
Deleted at least one digit	6
Deleted at least one letter	5
Shifted Digits	1

VI. DISCUSSION

A. Use Blacklists, But Check for Reuse

Previous work has shown that using a blacklist can be effective at forcing participants to create stronger passwords [22], [23], [33]. However, what has been missing is an evaluation and understanding of how participants behave under a policy with blacklists.

The majority of participants we examined whose initial attempt at creating a password was rejected because the password was blacklisted used only small modifications to the blacklisted password to create their final passwords. This is not surprising, as reuse of passwords across accounts is common among Internet users [8], [53]. Since blacklisted passwords are so common, they are targeted by password cracking tools [31], [37], which is why the final passwords that reused blacklisted attempts were significantly weaker than those that did not reuse a blacklisted attempt or those that were created by participants with no blacklisted attempts.

In alignment with 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 their recommendation of using 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, the observed number of such modifications was quite small so it is likely this behavior is not as common as inserting digits and symbols to modify a password.

B. Provide Feedback and Encouragement

Our analyses support a previous conclusion [41] 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 specifically tailored to discourage users from reusing their blacklisted password as part of their new one. 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 includes more capital letters, digits, and symbols.

Lastly, participants who reused a blacklisted password found the password creation task to be less difficult than those who did not have a blacklisted attempt, but equally annoying. To mitigate any increase of negative feelings caused by the added work of creating an unrelated password after a blacklisted attempt, the text feedback content can also provide positive encouragement, in addition to the advice guiding users on how to create stronger passwords.

VII. Conclusion

In this paper we analyzed 2,280 passwords, created during a previous study to evaluate users' password creation behaviors in settings where the password composition policy includes a prohibition against using blacklisted passwords. We found that participants who created 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], [33], [41], and are ones that are not difficult to implement. With the additional understanding our analyses provide of how users react to password creation attempts failing because of a blacklist, 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, as well as to provide positive encouragement to counteract negative feelings associated with the extra effort required in making a stronger password.

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, 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.

[7] Carnegie Mellon University, "Password guessability service," https://pgs.ece.cmu.edu, 2015.

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

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

[10] 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.

[11] S. Fahl, M. Harbach, Y. Acar, and M. Smith, "On The Ecological Validity of a Password Study," in *Proc. SOUPS*, 2013.

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

[13] D. Florêncio, C. Herley, and P. C. van Oorschot, "An administrator's guide to internet password research," in *Proc. USENIX LISA*, 2014.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] T. Hunt, "The science of password selection," Blog Post, July 2011, http://www.troyhunt.com/2011/07/science-of-password-selection.html.

[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] 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.

[34] 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.

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

[36] 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.

[37] J. Steube, "Rule-based attack," https://hashcat.net/wiki/doku.php?id=rule_based_attack, 2009-.

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

[39] 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.

[40] J. Titcomb, "Do you have one of the most common passwords? theyre 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.

[41] B. Ur, "Supporting password-security decisions with data," Ph.D. dissertation, Carnegie Mellon University, September 2016, CMU-ISR-16-110.

[42] 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.

[43] 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.

[44] A. Vance, "If your password is 123456, just make it HackMe," New York Times, http://www.nytimes.com/2010/01/21/technology/21password.html, 2010.

[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] E. von Zezschwitz, A. De Luca, and H. Hussmann, "Survival of the shortest: A retrospective analysis of influencing factors on password composition," in *Proc. INTERACT*, 2013.

[48] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. B. Tai, and J. Cook, "Improving password security and memorability to protect personal and organizational information," *IJHCS*, vol. 65, no. 8, pp. 744–757, 2007.

[49] 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.

[50] D. Wheeler, "zxcvbn: Realistic password strength estimation," https://blogs.dropbox.com/tech/2012/04/zxcvbn-realistic-password-strength-estimation/, 2012.

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

[52] D. F. D. William E. Burr 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.

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