Using Context-Based Password Strength Meter to Nudge Users' Password Generating Behavior: A Randomized Experiment

Warut Khern-am-nuai
McGill University
warut.khern-am-nuai@mcgill.ca

Weining Yang
Purdue University
yang469@purdue.edu

Ninghui Li
Purdue University
ninghui@cs.purdue.edu

Abstract

Encouraging users to create stronger passwords is one of the key issues in password-based authentication. It is particularly important as prior works have highlighted that most passwords are weak. Yet, passwords are still the most commonly used authentication method. This paper seeks to mitigate the issue of weak passwords by proposing a context-based password strength meter. We conduct a randomized experiment on Amazon MTurk and observe the change in users' behavior. The results show that our proposed method is significantly effective. Users exposed to our password strength meter are more likely to change their passwords after seeing the warning message, and those new passwords are stronger. Furthermore, users are willing to invest their time to learn about creating a stronger password, even in a traditional password strength meter setting. Our findings suggest that simply incorporating contextual information to password strength meters could be an effective method in promoting more secure behaviors among end users.

1. Introduction

Passwords have been an essential user authentication method for years despite the availability of stronger authentication mechanisms [1]. However, there is one major dilemma associated with password-based authentication. That is, a password must be easy for the owner to remember but it must be hard for others to guess. Naturally, many users prefer memorability over security, choosing weak passwords. According to a survey conducted by TeleSign in 2015 [2], 3 out of 4 users chose weak passwords, and 40 percent of them experienced an issue with their account security in the past year.

One popular approach to mitigate the issue of weak passwords is to create a visual feedback in the password generation process through a "password strength meter". The meter calculates password strength using an algorithm, and displays the strength information to the user. Previous works in this area have focused on improving the underlying algorithm by seeking to find the better (faster or more accurate) ways to calculate password strength [3, 4]. Yet, little research has been conducted to understand how users perceive password strength meters and the implication of the meters' component, particularly the warning message, despite the fact that most users do not understand the implication of warning messages due to their limited technical background (e.g., what does "weak" password really mean?). This lack of understanding could significantly impact the effectiveness of password strength meters.

In this study, we draw from theories in psychology, human-computer interaction, and warning sciences to develop a theoretical foundation of how password strength meter works. We consider password strength meters to be an interactive warning and adopt the Communication-Human Information Processing Model (C-HIP) [5] as a framework to explain why traditional password strength meters are ineffective. We hypothesize that incorporating contextual information into warning messages could draw users' attention, positively affect their understandings and beliefs, and act as a stimulus to nudge their password generating behavior. Hence, the context-based password strength meter would be more effective in nudging users to think more about their passwords, and promoting their secure behaviors.

We test our hypotheses by analyzing data from a human subject study conducted on Amazon Mechanical Turk. We introduce hypothetical situations where participants are required to create an online account in different scenarios. We examine the effects of three experimental treatments (different context-based warning messages) on the effectiveness of the password strength meter. More specifically, we evaluate the password strength, how often a user changes her password immediately after seeing the warning message, and how often a user wants to learn more about how to create stronger passwords. We find that the context-based password meter enhances password security. It nudges users to change the password more often, and the new passwords they pick are stronger. We also find that incorpo-

URI: http://hdl.handle.net/10125/41220
ISBN: 978-0-9981331-0-2

rating a link that points to learning more about how to create a stronger password is effective even for the traditional password strength meter. Our findings suggest that incorporating contextual information into password strength meters' warning message could be one effective method to promote more secure behaviors among end users.

The rest of this paper is organized as follows. In Section 2, we discuss related work and develop our hypotheses. We then describe the design of our study and the methodology of our human subject study in Section 3. In Section 4, we present our analysis. Finally, we discuss our results and conclude our study in Section 5.

2. Conceptual Background

In this section, we first describe previous works related to password security, password strength meters, users' perception of warning messages and warning models, and contextual information. Following that, we develop our main hypotheses.

One line of research on password security studies the property of the passwords. In general, it is found that end users tend to choose passwords that are easy to remember but also easy to guess. It has been suggested that organizations should use tools, such as password validation software, to mitigate this issue [1, 7]. It has been argued that users could be motivated to adopt secure behavior through well-planned security mechanisms or punishment threats [6]. Another line of research investigated users' password creation behavior when facing different stimuli such as password policy designs [8], and training techniques [9]. This paper follows this line by investigating password meter designs that could affect users' password generating behavior.

2.1 Password Strength Meters

The concept of password strength meters has been discussed in the literature for decades [11]. It has been shown to be effective in leading users to create stronger passwords [12, 13]. However, the design and implementation of these meters are usually ad-hoc, and operate like a black-box (i.e., without explanations or justifications of design choices) [14]. Hence, the password strength meters on most websites are found to be inconsistent [15], which are confusing and may weaken the purpose of the password strength meter itself.

Most research works that aim to enhance the effectiveness of password strength meters have focused on the algorithm. It has been recognized that password strength meters are less accurate than entropy measurement of an ideal case [16, 17]. And many techniques have been proposed to overcome this limitation, including the use of probabilistic context-free grammars [3, 18], and Markov model [19, 4]. However, despite the advancement in developing algorithms, relative little research has been done to investigate how password meters interact with human users. We consider password strength meters as a form of warning and adopt models and theories from the warning science literature, which is reviewed next.

2.2 Warning Models

The warning science literature has identified two critical factors of warning, "hazard matching", and "arousal strength" [20]. The first term refers to the ability of a warning message to convey potential risks to users. If the message cannot convince a user regarding the level of risks involved, she may choose to ignore it. Meanwhile, the latter term refers to how users perceive the urgency of the warning [21]. If the warning message is perceived to be nonessential, then the user may choose to ignore it as well.

In this paper, we adopt the Communication-Human Information Processing (C-HIP) Model, which has been adopted for identifying potential reasons of warning ineffectiveness by several prior works [20, 22, 23]. The full model description is shown in Figure 1.

The C-HIP model consists of nine phases. It begins when a source delivers a warning message to a receiver through a channel and ends with a change of receiver's behavior. At the time a receiver receives the warning message, she would also receive other environmental stimuli, which might distract her from paying attention to the warning. The essential phases of this model for our study are within the information processing phases, which start after the receiver receives the warning message. As these phases are recognized and processed by each receiver, they essentially determine the effectiveness of the warning (i.e., whether the warning results in a change of a receiver's behavior or not). In this paper, we adopt a practice used by literature in human-computer interaction, which suggests to use a set of questions to evaluate the effectiveness of the warning in each phase [20, 24].

First, we start at the *Attention Switch* phase and the *Attention Maintenance* phase, which correspond to the question "Do users notice the indicators?" Although indicators of password strength meters are found to be inconsistent across websites, they are generally well-designed to sufficiently catch users' attention [14]. In addition, previous works have shown that they are generally recognized among end users [25]. Therefore, attention switch is unlikely to be a significant factor that con-

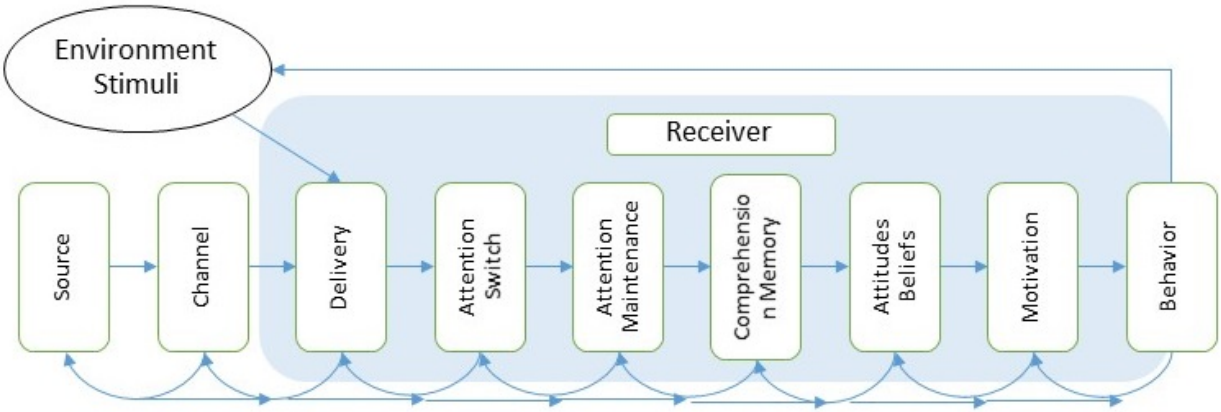

Figure 1: Diagram of the C-HIP model

tributes to the ineffectiveness of password strength meters. The next phase is *Comprehension/Memory*, which corresponds to the question "Do users know what the indicators mean?" To the best of our knowledge, there is no previous research that looks into this specific question in the context of password strength meters. However, empirical evidences have shown that end users are unaware of information security issues in general [26, 27]. Furthermore, end-users often do not recognize the issue of weak passwords [1]. In addition, a previous study in warning science has concluded that *novice* users usually make mistake in terms of warning recognition, such as considering the wrong variables, missing some variables, or considering the right variables but in the wrong order [28]. Therefore, it can be inferred that not understanding the warning messages could be one of the factors that lead to password strength meters' ineffectiveness. This lack of understanding could also lead to the ineffectiveness of other phases down the line. For instance, with a lack of understanding, users might not believe the warning in the *Attention/Beliefs* phase and thus have no motivation to change their behavior in the *Motivation* phase.

2.3 Contextual Information

The practice of using contextual information to remedy lack of understanding is common in the literature [29, 30, 31]. Vance et al. find that incorporating "fear appeals", which could be considered as one type of contextual information, can increase final password strengths [32]. However, that paper only investigates the changes in final password strength, which could be a misleading dependent variable as described later. Our work is significantly different from theirs as we seek to examine the changes in password meters' effectiveness with three different measurements along with other contextual information.

In the context of warning design, the use of contextual information is outline by Wogalter et al. [33]. In addition, Bauer et al. [28, 34] proposes that relevant contextual information should be presented in warning design as they observe that users tend to ignore warning messages as they do not understand them. Base on these guidelines, we predict that adding contextual information to password strength meters would enhance users' understanding in the warning messages, resulting in more effective *Comprehension/Memory* phase in the C-HIP model, and leading to improved security in users' behaviors. Desired behaviors of users who are exposed to password strength meters include changing their passwords and selecting stronger passwords. We also anticipate that users who are exposed to contextual information would want to seek additional information regarding password security, if available. Therefore, we hypothesize the following:

Hypothesis 1a *Participants in the context-based warning message treatment choose a stronger password than their original password after seeing the warning.*

Hypothesis 1b *New passwords chosen by participants in the context-based warning message treatment are stronger than those chosen by participants in the control group.*

Hypothesis 2a *Participants in the context-based warning message treatment change their passwords after seeing the warning.*

Hypothesis 2b *Participants in the context-based warning message treatment are more likely to change their passwords compared to those in the control group.*

Hypothesis 3a *Participants in the context-based warning message treatment try to seek additional information regarding how to create a stronger password after seeing the warning.*

Hypothesis 3b *Participants in the context-based warning message treatment are more likely to seek additional information compared to those in the control group.*

3. Methodology

To test our hypotheses, we conduct a human subject study on Amazon Mechanical Turk involving hypothetical situations where users need to create an online account. Participants are informed of the experimental study and asked for their consent to participate. By agreeing, participants allow researchers to collect and analyze their passwords in unencrypted, but anonymized manner.

3.1 Dependent Variable

There are three dependent variables that our study seeks to investigate. First, the increase in strength of passwords generated. Second, the number of occasions participants change their passwords after seeing the password strength meter. Third, the number of occasions participants invest their time to learn how to create strong passwords.

The increase in password strength is a common dependent variable in password studies. It represents the ultimate goal of password strength meters, which is to encourage a user to change his or her password and pick a password that is stronger than the original one. In our study, the strength of passwords is measured using the Backoff Markov Model with end symbol normalization proposed in [4]. The model is trained with the RockYou dataset, which contains over 32 million passwords leaked from the social application site Rockyou in December 2009. In the model, we only maintain substrings with frequency no less than 1,000 and drop the ones appear less than 1,000 times (meaning the frequency threshold is set to be 1,000). We then assign a strength label for each password as the following. We label a password as *Weak* if when it is the among the $300,000$ most frequent passwords according to the model. This means that if an attacker tries passwords following a descending order of the probability generated by the model, the account will be compromised in 300,000 attempts. We label a password as *Medium* if it is among the top 5,000,000 passwords but not in the top 300,000. If the probability of a password is not in the top 5,000,000, we label the password as *Strong*.

Second, we are interested in how password strength meters affect the number of occasions where users change their password after seeing the warning message. This variable is usually overlooked in password strength meter studies. However, without fully understanding the effect of password meters on this variable, the interpretation of the final password strength, which is commonly used as a dependent variable in many password strength meter studies, could be missing important information. For instance, suppose there are two users. Users A initially picked a strong password and never changes it. Meanwhile, users B initially picked a weak password but changes to a strong one after seeing a password strength assessment. If the strength of the final password is the only measurement (as commonly used in previous password strength meter studies), we would not be able to distinguish these two users and hence treat them as the same group of users (e.g., users with a strong password) in the analysis. Our study provides additional insights in that regard.

Third, as we present contextual information to participants, we anticipate that they might be involved as they understand more about the passwords they use. As a result, they might be willing to invest their time to learn more about how to create a strong password, which will help them improve security in the future. Therefore, together with the strength of the password, we present a link at the bottom of the password meter labeling as "Tips towards strong passwords". We are interested in the number of occasions participants click this link as it presents an opportunity to educate users regarding how to create stronger passwords and promote security awareness among them. To the best of our knowledge, Our study is the first that investigates the effectiveness of a password strength meters as a tool to educate users and promote security awareness.

3.2 Experimental Design

At the beginning of the experiment, each participant is randomly presented with one of the following scenarios: *Bank*, *Restaurant*, and *Forum*. Following that, the participant is asked to create an online account corresponding to her scenario. *Bank* is a scenario where users have to store both financial information and personal information in the account. Meanwhile, she needs to store only a part of her personal information (e.g., name and delivery address) in *Restaurant* and only her email in *Forum*.

Next, the participant is asked to create a password for her account. The interface of the password input form is in Figure 2. Once the participant finishes typing a password, we calculate the strength of the password. The password strength meter then displays a warning message to the right of the password input form. The system randomly assigns one of four experimental treatments: *Control*, *Time*, *Rank*, or *Probability*. They correspond to the type of warning messages displayed with the pass-

word strength meter. Note that we impose a password length requirement of at least six characters for all treatments. No other requirements (e.g., policies such as using both letters and digits) are imposed.

In the*control* treatment, users receive a warning message that contains only the labels derived from calculated strength of the passwords they pick (e.g., Weak, Medium, or Strong). It is a standard practice in many major websites [15].

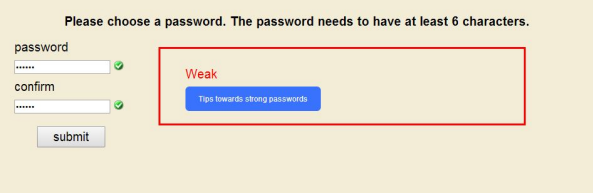

Figure 2: The Interface of Password Generation

In the *time* treatment, users receive a warning message that contains the strength of the password they pick, along with the time that a hypothetical hacker could use to crack that password. We calculate the estimated time to crack by assessing the time needed if an attacker continuously attacking the account by trying the most popular passwords first. We assume the attacker can try 100 passwords per second, which is a conservative assumption considering that it could be much faster in a real-world scenario [35]. An example of this warning message is: "*Weak. We estimate that it takes 10 seconds to crack your password, assuming that the attacker can try 100 passwords every second.*".

In the *rank* treatment, users receive a warning message that contains the strength and the rank of the password. We calculate the rank as the following. First, we generate a list of passwords using the Backoff Markov Model trained on an existing password dataset. The passwords are generated in descending order of the probability. Therefore, in the generation process, weaker passwords (i.e., those with higher probability) are output earlier. The rank of passwords chosen by participants are estimated based on the list of passwords generated. Specifically, we pick passwords of the ranks that have only one significant digit (e.g., 10, 20, ..., 100, 200, ...) from this list and calculate their corresponding probability, resulting in rank-probability pair for each one. Then, once a participant enters a password, we calculate the probability of the password using the same password model, and match the probability with the closest rank-probability pair generated earlier. Thus, The rank of the password is estimated as the highest rank whose corresponding probability is smaller than the probability of the chosen password. In this way, we essentially round up the actual rank of the password and display one significant digit. For example, if the rank of a password is 24,321, we show: "*Weak. We estimate that the password you chose is among the 30,000 weakest passwords*".

In the *probability* treatment, users receive a warning message that contains the strength of the password along with an estimated number of accounts that use the same password if there are 10 billion accounts globally. The estimated value is calculated by the probability of the password generated by the Backoff Markov Model multiplies by 10 billion. For example, if the probability calculate by the model is 0.001, we show: "*Weak. We estimate that about 10,000,000 other accounts will have the same password as you within 10 billion accounts*".

In all treatments, we first calculate the probability of passwords utilizing the Backoff Markov Model when the webpage is loaded. Then, the strength of the passwords and the context-based warning message are generated and shown when the active cursor in the input box becomes inactive. Specifically, we listen to the event "focusout" of the password input field. Once an event is triggered, the script estimating and displaying the strength of passwords is executed. Note that the model is implemented in Javascript and all necessary data are transmitted to users' browser by AJAX calls. Hence, there is no communication between the browser and our server until the passwords are finally submitted.

Lastly, below the warning message, we include a clickable link with a caption "Tips towards strong password". The password generation tips are displayed to the user when the link is clicked. (We obtain password generation tips from http://windows.microsoft.com/en-us/windows-vista/tips-for-creating-a-strong-password.) Note that we record the entire history of passwords generated on the webpage. All user interactions, including the timestamp of each event that shows when the event occurs, are recorded as well.

Once the users finish creating their passwords for their hypothetical account, we conduct a post-test survey to address one of the major concerns in password generation studies. That is, the password that users generated in the experiment might not be usable [36]. For example, participants might generate a random password that they do not intend to remember. To alleviate such as issue, we ask participants to complete a survey regarding their password generation strategy after they submit their passwords. The survey is about how users generate their passwords in general (not only the password they just created for this study). It takes about 30-60 minutes to complete the survey. After users finish answering the survey, we ask them to re-enter the password they created for their account. Participants who fail to recall the previously generated password after three attempts

are excluded from our analysis. Following that, we ask users to answer demographic questions including age, gender, and education level, for the purpose of statistical controls.

3.3 Data Collection

We first conducted a pilot test involving 10 participants. The primary purpose of the pilot test is to ensure that our system works as intended. We also collected and analyzed the data to evaluate our experimental design. Some texts displayed were updated as necessary. We also started recording the entry history of passwords in the password generation process.

Our primary sample consists of 500 participants. None of them fails the post-test password recall check. Participants are compensated $0.75 after they complete the experiment. We limit the age of participants to be no less than 18.74 of the 500 participants are from 18 to 22; 202 of them are from 23 to 30; 133 of them are from 31 to 40; 64 of them are from 41 to 50; 25 of them are older than 51 and the remaining 2 refuse to disclose their age. For the 497 participants who disclose their gender, 283 of them are male and 214 are female. Regarding the education level, the majority of them either have some college credit (no degree) or bachelor's degree. These two categories account for 192 and 205 participants, respectively. For the rest of the participants, 41 have high school diploma, 47 have a master's degree, 8 have a professional degree, 6 have a doctoral degree, and 3 refuse to reveal their education level.

4. Analysis

In this section, we present the results of our experiment. We first use the regression analysis to examine the effect of our control variables, which are age, gender, and education level, on the dependent variables. We find that none of them are significantly correlated with our DVs, which are the increase in password strength, the number of occasions users change their password, and the number of occasions users click "Tips towards strong password". That is, these control variables explain the variance of our dependent variables only marginally and none of them is statistical significant at $p < 0.10$.

Next, we present our main results. Note that among 500 participants, 116 are assigned to the control group, 133 are in *time* treatment, 131 are in *rank*, and 120 are in *probability*. As per the scenario assignment, 180 are in *bank*, 166 are in *restaurant*, and 150 are in *forum*. The details of the treatment/scenario assignment are presented in Table 1.

Table 1: Total number of participants in each treatment/scenario

	Time	Rank	Probability	Control	Total
Bank	41	55	47	41	184
Restaurant	51	36	43	36	166
Forum	41	40	30	39	150
Total	133	131	120	116	500

4.1 Password Strength

We begin with an analysis of the password strength. We leverage the Probability Threshold Graph proposed in [4]. Each line on the graph represents the guessibility of a password dataset calculated based a password model (the Backoff Markov Model in our study.) A point (x,y) on a line means that y percent of passwords in the dataset have the probability of at least 2^{-x}. For the purpose of comparing the strength of password datasets, we fix the x value, which is $-\log_2 prob$, where *prob* is the probability assigned to a password by the password model, and compare y values of different curves. A higher y value indicates the password dataset is weaker because it means that more passwords can be cracked if passwords with less than the probability *prob* are attempted.

The graph represents the strength of passwords generated by participants given different scenarios is illustrated in Figure 3. We display the curve of each scenario as well as two curves generated from two existing leaked password datasets. The PhpBB dataset includes about 250,000 passwords leaked from Phpbb.com in January 2009. The Yahoo dataset includes around 450,000 passwords published in July 2012. It is clear that passwords generated in our study are much stronger than those in the Yahoo and PhpBB datasets. We believe such a phenomenon results from 1) users in general are more aware of password security compared to users 3 years ago 2) participants in MTurk are more involved in the digital world and thus have more knowledge about cyber security. Among 3 scenarios, the final strength of passwords created for online banks dominates, indicating that people tend to be more careful when facing financial or monetary related situations. However, none of the differences of final password strengths among 3 scenarios and 4 treatments is statistically significant.

Next, we investigate the *increase* in password strength rather than the *final* password strength. Note that the distribution of the password strength is highly skewed by nature. Therefore, we follow previous works in the literature by applying the natural logarithms to the password strength [32]. The average increase of log password strength in different scenarios and types

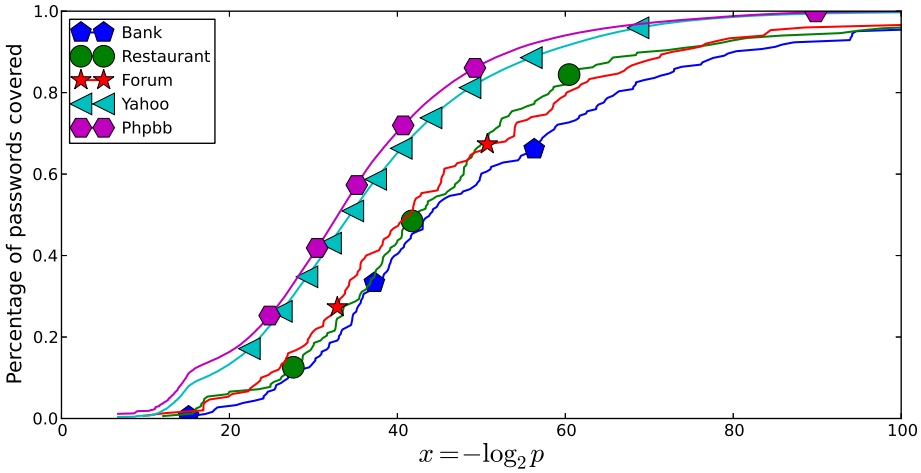

Figure 3: The strength of passwords generated by participants given different scenarios

of contextual information is shown in Table 2. Overall, context-based password meters outperform the traditional password meter in terms of password strength increase. Interestingly, users under the *Forum* scenario (where users do not have to store any personal information) have the highest average increase in password strength compare to other two scenarios where some personal information and financial information are involved. This behavior could be explained through the initial password strength as the strength of initial passwords in *Forum* are far weaker than those in *Restaurant* and *Bank*. In addition, we also observe that a few users change their passwords to *weaker* ones after they see the password strength meters.

Table 2: The average increase of password strength score (calculated by the Backoff Markov Model) under different scenarios and contextual warning messages

	Time	Rank	Probability	Control	Total
Bank	0.0277	0.0373	0.0269	-0.0439	0.0144
Restaurant	0.0353	0.1082	0.0501	0.0410	0.0562
Forum	0.0510	0.1664	0.0718	0.0243	0.0790
Total	0.0378	0.0962	0.0465	0.0054	0.0477

We then employ the t-test analysis to test if the password strength increase in each treatment group is higher than zero. We find that the increase in password strength under 3 types of contextual information are all statistically significant ($t = 2.7667, p = 0.0032$; $t = 3.7886, p = 0.0001$; and $t = 3.0906, p = 0.0012$ for *Time*, *Rank*, and *Probability* respectively). Therefore, Hypothesis 1a is supported. Following that, we use ANOVA to test whether the difference in password strength increase between at least one treatment group and control group is statistically significant. The results indicate that the difference is significant ($F = 3.5234, p = 0.0075$). However, our pairwise comparison (using Tukey HSD Test) reveals that even though the average increase in password strength of all 3 types of contextual information are higher than the control one, only the difference of *Rank* is statistically significant ($t = 3.18, p = 0.004$). Hence, Hypothesis 1b is partially supported. It is also worth noting that the password strength increase in *Rank* treatment significantly outweighs that of other experimental treatments, followed by the *Probability* treatment while the *Time* treatment performs the worst.

4.2 Password Reset

Our second variable of interest is the number of occasions where users change their password after seeing a warning message provided by password strength meters. As we argue earlier, this variable is particularly important to measure the effectiveness of password strength meters but it is typically overlooked by prior studies.

Table 3 reports the numbers of times when users change their passwords in different settings. Context-based password strength meters also outperform the traditional password strength meter in terms of nudging users to change passwords. The user behavior in the *Bank* scenario here is also consistent with our observation in the previous section. They change passwords much less often compared to users in the other two scenarios. We believe that this behavior is consistent with the fact that their initial passwords are strong (e.g., they might be confident about the strength of their initial passwords).

For hypothesis testing, the average number of occa-

Table 3: The average number of occasions where users change their password under different scenarios and contextual warning messages

	Time	Rank	Probability	Control	Total
Bank	0.1220	0.2182	0.1277	0.1463	0.1576
Restaurant	0.1961	0.3333	0.2326	0.1111	0.2169
Forum	0.1220	0.3750	0.3667	0.0769	0.2267
Total	0.1504	0.2977	0.2250	0.1121	0.1980

sions where users change their password after observing three types of the contextual information we propose are all statistically significant ($t = 3.5790, p = 0.0002$; $t = 4.4163, p < 0.0001$; and $t = 4.3096, p < 0.0001$ for *Time*, *Rank*, and *Probability* respectively). Therefore, Hypothesis 2a is statistically supported. However, as in the case with the increase in password strength, ANOVA reports that the difference in the average between treatment and control is significant ($F = 2.4658, p = 0.0308$) but the post-hoc pairwise comparison shows that only the difference between the *Rank* type and the control group is statistically significant ($t = 2.48, p = 0.032$). Hence, Hypothesis 2b is partially supported. In this evaluation, the order of the effectiveness of our three experimental treatments remains unchanged. *Rank* still performs the best while the *Time* treatment remains the worst. However, the difference between *Rank* and *Probability* treatments is shortened.

4.3 Password Generation Tips

Third, we measure the number of times users click on the "Tips towards strong password" link and view password generation tips. If the context-based password strength meters can draw attentions and promote understanding among users as theorized, this would be an excellent opportunity to improve their awareness regarding the issue of weak passwords.

The results are displayed in Table 4. Surprisingly, it appears that users in the control group click the password generation tips more often than users in all of our treatment groups. This finding is particularly interesting since it suggests that incorporating additional information that can promote information security awareness can be effective even for the traditional password strength meter. As for the hypothesis testing, the average numbers of occasions where users click "Tips towards strong password" under three types of contextual information we provided are all significantly greater than zero ($t = 4.0963, p < 0.0001$; $t = 4.5506, p < 0.0001$; and $t = 4.1231, p < 0.0001$ for *Time*, *Rank*, and *Probability* respectively). Therefore, our Hypothesis 3a is statistically supported. Meanwhile, our Hypothesis 3b is not supported as the ANOVA analysis yields $F = 0.6790, p = 0.5652$. In other words, incorporating password generation tips into context-based password generation meters is significantly effective. However, the effectiveness is at the same level as incorporating these tips in the traditional password strength meter as the difference in effectiveness between them is statistically insignificant. Also, none of the difference between three experimental treatments is statistically significant as well.

Table 4: The average number of occasions users click "Tips towards strong password" under different scenarios and contextual warning messages

	Time	Rank	Probability	Control	Total
Bank	0.0976	0.1455	0.1489	0.1951	0.1467
Restaurant	0.1373	0.1389	0.1163	0.1944	0.1446
Forum	0.0976	0.1250	0.1000	0.1282	0.1133
Total	0.1128	0.1374	0.1250	0.1724	0.1360

5. Discussions and Conclusions

Nudging users to create stronger passwords is one important goal of information security managers and researchers. Our study shows that providing additional contextual information along with warning messages displayed by password strength meters could enhance understanding among users, resulting in improved password generating behaviors. We draw theories from psychology, human-computer interaction, and warning sciences to identify potential weaknesses in the traditional password strength meter. Following that, we conduct a human subject study on Amazon Mechanical Turk to test our hypotheses that adding contextual information could enhance the effectiveness of password strength meters. We find that the contextual information induces users to pick stronger passwords. In addition, users change their password more often. Furthermore, we also find that adding a link that leads to password security awareness training is significantly effective even in the traditional password strength meter setting.

Our findings have implications for the use of contextual information and password strength meters to promote secure behaviors among end users and make significant contributions to the literature in behavioral information security. As most of previous works that study password strength meters focus on understanding how the underlying algorithm and appearance of password strength meters affect user behavior, our work is among the first to show that given the same algorithm and appearance, the effectiveness of the password strength meter can be significantly improved by tweaking the warn-

ing message. Particularly, we show that adding contextual information could help improving users' password generating behaviors. More importantly, we measure the effectiveness of password strength meters by both the increase in password strength, and the increase in number of occasions users change their password after seeing the password meter, which are usually overlooked in prior studies. In addition, we find that different types of contextual information can affect the effectiveness of password strength meters differently. In our study, although our three types of contextual information positively impact the effectiveness of password strength meters, only the benefits from password rank is statistically significant. This finding is a potential future research avenue to find the optimal contextual information that can nudge users' password generating behavior. Theories in psychology, human-computer interaction, and usable security, among others, could be used to draw a conceptual framework to develop the optimal contextual information. Furthermore, we find that adding a link that leads to password security awareness training is effective even for the traditional password strength meter. This finding is crucial as creating security awareness is one of the most important parts in every security programs. Nevertheless, it is important to note that one possible reason that the traditional password strength meter leads more users to click the link is because they want to know more about the reason of the assessment. Lastly, our results are also relevant to practitioners. For one, adding contextual information to password strength meters is relatively simple. In our study, the traditional password meter calculates the probability using the Backoff Markhov Model. We derive three types of contextual information out of that probability by estimating the time to crack, the rank of the password, and the number of accounts that share the same password, which could be done without significant computation resources. Making these minor changes to the existing password strength meter could potentially lead to stronger passwords in general.

Our research is not without limitations. First, participants in Amazon MTurk are known to be tech-savvy, which might bias our study. As we show earlier, passwords in this study are significantly stronger than passwords in other datasets, which might indicate that our observations are bias. Therefore, replicating our experiment in a stricter control setting (e.g., laboratory with student samples), or a more realistic setting (e.g., field experiments) to validate our findings could be a great avenue for future research. Second, although we find that "Tips towards strong password" is significantly effective in terms of *quantity* (number of visits), we do not focus on the *quality* part. For instance, what happened after users click that link? Do they really read the provided information? Can they remember the information and apply it later? Future research which employs additional post-test surveys to gauge participants' understanding before and after the study, or research that leverages a laboratory equipped with eyes-tracking devices could be conducted to improve our understandings.

6. References

[1] M. Zviran and W. J. Haga, "Password security: an empirical study," *Journal of Management Information Systems*, pp. 161–185, 1999.

[2] "Consumers lose faith in passwords," 2015. https://www.telesign.com/resources/research-and-reports/telesign-consumer-account-security-report/.

[3] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," in *IEEE Symposium on Security and Privacy*, pp. 391–405, 2009.

[4] J. Ma, W. Yang, M. Luo, and N. Li, "A study of probabilistic password models," in *Security and Privacy (SP), 2014 IEEE Symposium on*, pp. 689–704, IEEE, 2014.

[5] M. Wogalter, "Communication-human information processing model," *Handbook of warnings*, pp. 51–61, 2006.

[6] A. Adams and M. A. Sasse, "Users are not the enemy," *Communications of the ACM*, vol. 42, no. 12, pp. 40–46, 1999.

[7] J. Zhang, X. Luo, S. Akkaladevi, and J. Ziegelmayer, "Improving multiple-password recall: an empirical study," *European Journal of Information Systems*, vol. 18, no. 2, pp. 165–176, 2009.

[8] M. Keith, B. Shao, and P. Steinbart, "A behavioral analysis of passphrase design and effectiveness," *Journal of the Association for Information Systems*, vol. 10, no. 2, p. 2, 2009.

[9] D. Charoen, M. Raman, and L. Olfman, "Improving end user behaviour in password utilization: An action research initiative," *Systemic Practice and Action Research*, vol. 21, no. 1, pp. 55–72, 2008.

[10] P. B. Goes, "Editor's comments: information systems research and behavioral economics," *MIS quarterly*, vol. 37, no. 3, pp. iii–viii, 2013.

[11] M. Bishop and D. V. Klein, "Improving system security via proactive password checking," *Computers & Security*, vol. 14, no. 3, pp. 233–249, 1995.

[12] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, *et al.*, "How does your password measure up? the effect of strength meters on password creation.," in *USENIX Security Sympo-*

sium, pp. 65–80, 2012.

[13] 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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2379–2388, ACM, 2013.

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

[15] S. Furnell, "Assessing password guidance and enforcement on leading websites," *Computer Fraud & Security*, vol. 2011, no. 12, pp. 10–18, 2011.

[16] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 162–175, ACM, 2010.

[17] S. Houshmand and S. Aggarwal, "Building better passwords using probabilistic techniques," in *Proceedings of ACSAC*, pp. 109–118, 2012.

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

[19] C. Castelluccia, M. Dürmuth, and D. Perito, "Adaptive password-strength meters from Markov models," in *Proceedings of NDSS*, 2012.

[20] S. Egelman, L. F. Cranor, and J. Hong, "You've been warned: an empirical study of the effectiveness of web browser phishing warnings," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1065–1074, ACM, 2008.

[21] E. Hellier, D. B. Wright, J. Edworthy, and S. Newstead, "On the stability of the arousal strength of warning signal words," *Applied Cognitive Psychology*, vol. 14, no. 6, pp. 577–592, 2000.

[22] V. Visschers, R. Ruiter, M. Kools, and R. Meertens, "The effects of warnings and an educational brochure on computer working posture: a test of the c-hip model in the context of rsi-relevant behaviour," *Ergonomics*, vol. 47, no. 14, pp. 1484–1498, 2004.

[23] L. F. Cranor, "A framework for reasoning about the human in the loop.," *UPSEC*, vol. 8, pp. 1–15, 2008.

[24] L. F. Cranor, "What do they indicate?: evaluating security and privacy indicators," *Interactions*, vol. 13, no. 3, pp. 45–47, 2006.

[25] 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 *Proceedings of CHI*, pp. 2379–2388, 2013.

[26] E. Albrechtsen, "A qualitative study of users' view on information security," *Computers & security*, vol. 26, no. 4, pp. 276–289, 2007.

[27] K. H. Guo, "Revisiting the human factor in organizational information security management," *ISACA Journal*, vol. 6, pp. 1–5, 2013.

[28] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri, "Bridging the gap in computer security warnings: A mental model approach," *IEEE Security & Privacy*, no. 2, pp. 18–26, 2010.

[29] E. Tulving and C. Gold, "Stimulus information and contextual information as determinants of tachistoscopic recognition of words.," *Journal of Experimental Psychology*, vol. 66, no. 4, p. 319, 1963.

[30] J. D. Bransford and M. K. Johnson, "Contextual prerequisites for understanding: Some investigations of comprehension and recall," *Journal of verbal learning and verbal behavior*, vol. 11, no. 6, pp. 717–726, 1972.

[31] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 1, pp. 103–145, 2005.

[32] A. Vance, D. Eargle, K. Ouimet, and D. Straub, "Enhancing password security through interactive fear appeals: A web-based field experiment," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pp. 2988–2997, IEEE, 2013.

[33] M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson, "Research-based guidelines for warning design and evaluation," *Applied ergonomics*, vol. 33, no. 3, pp. 219–230, 2002.

[34] L. Bauer, C. Bravo-Lillo, L. F. Cranor, and E. Fragkaki, "Warning design guidelines (cmucylab-13-002)," 2013.

[35] D. Mirante and J. Cappos, "Understanding password database compromises," tech. rep., Technical Report TR-CSE-2013-02, Department of Computer Science and Engineering Polytechnic Institute of NYU, 2013.

[36] J. K. Goodman, C. E. Cryder, and A. Cheema, "Data collection in a flat world: The strengths and weaknesses of mechanical turk samples," *Journal of Behavioral Decision Making*, vol. 26, no. 3, pp. 213–224, 2013.