Measuring Password Guessability for an Entire University

Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer,

Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur

October 22, 2013

CMU-CyLab-13-013

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Measuring Password Guessability for an Entire University

Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer,
Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley*, Richard Shay, and Blase Ur

Carnegie Mellon University
Pittsburgh, PA
{mmazurek, sarangak, tvidas, lbauer,
nicolasc, lorrie, rshay, bur}@cmu.edu

*University of New Mexico
Albequerque, NM
pgk@cs.unm.edu

ABSTRACT

Despite considerable research on passwords, empirical studies of password strength have been limited by lack of access to plaintext passwords, small data sets, and password sets specifically collected for a research study or from low-value accounts. Properties of passwords used for high-value accounts thus remain poorly understood.

We fill this gap by studying the single-sign-on passwords used by over 25,000 faculty, staff, and students at a research university with a complex password policy. Key aspects of our contributions rest on our (indirect) access to plaintext passwords. We describe our data collection methodology, particularly the many precautions we took to minimize risks to users. We then analyze how guessable the collected passwords would be during an offline attack by subjecting them to a state-of-the-art password cracking algorithm. We discover significant correlations between a number of demographic and behavioral factors and password strength. For example, we find that users associated with the computer science school make passwords more than 1.8 times as strong as those of users associated with the business school. In addition, we find that stronger passwords are correlated with a higher rate of errors entering them.

We also compare the guessability and other characteristics of the passwords we analyzed to sets previously collected in controlled experiments or leaked from low-value accounts. We find more consistent similarities between the university passwords and passwords collected for research studies under similar composition policies than we do between the university passwords and subsets of passwords leaked from low-value accounts that happen to comply with the same policies.

Categories and Subject Descriptors

D.4.6 [**Management Of Computing and Information Systems**]: Security and Protection—*Authentication*

Keywords

Passwords; authentication; password security

This is a version of work published at ACM CCS 2013. Please cite:
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. ACM CCS, 2013.

1. INTRODUCTION

Researchers have documented the numerous problems of text passwords for decades — passwords are easy to guess, hard to remember, easily stolen, and vulnerable to observation and replay attacks (e.g., [28, 38]). The research community has invested significant effort in alternatives including biometrics, graphical passwords, hardware tokens, and federated identity; however, text passwords remain the dominant mechanism for authenticating people to computers, and seem likely to remain that way for the foreseeable future [5,23]. Better understanding of text passwords therefore remains important.

Considerable effort has been spent studying the usage and characteristics of passwords (e.g., [13, 17, 34, 35, 45]), but password research is consistently hampered by the difficulty in collecting realistic data to analyze. Prior password studies all have one or more of the following drawbacks: very small data sets [36], data from experimental studies rather than from deployed authentication systems [31], no access to plaintext passwords [3], self-reported password information [47], leaked data of questionable validity, or accounts of minimal value [26, 53]. As a result, the important question of whether the results apply to real, high-value passwords has remained open.

In this paper, we study more than 25,000 passwords making up the entire user base of Carnegie Mellon University (CMU). Notably, these passwords are the high-value gatekeeper to most end-user (i.e., non-administrative) online functions within the university, including email, grading systems, transcripts, financial data, health data, payroll, and course content. Furthermore, these passwords were created under a password-composition policy among the stricter of those in common use [18], requiring a minimum of eight characters and four different character classes. Using indirect access to the plaintext of these passwords, we measure their strength. In addition, we obtain contextual information from personnel databases, authentication logs, and a survey about password creation and management, and correlate these factors with password strength. To acquire this data, we established a partnership with the CMU information technology division; the research was also vetted by our Institutional Review Board (IRB). Our approach to analyzing this sensitive data securely provides a blueprint for future research involving security-sensitive data in the wild.

Using this data, we make two important and novel contributions to the field of password research. First, we identify interesting trends in password strength, measured as resistance to offline guessing attacks, in which an attacker attempts to recover plaintext passwords from their hashes [2, 6, 44]. Using statistical methods adopted from survival analysis, we find that users associated with science and technology colleges within the university make

passwords more than 1.8 times as strong as those of users associated with the business school. Perhaps unsurprisingly, strong passwords are correlated with higher rates of failed login attempts due to password errors. Users who report annoyance with CMU's complex password-composition policy made weaker passwords. For the first time, we are also able to directly investigate whether insights from work based on lower-value passwords also apply to high-value passwords. For example, we confirm Bonneau's finding that men's passwords are slightly stronger than women's [3]. We also confirm that passwords with more digits, symbols, and uppercase letters are stronger, and that digits and symbols are least effective when placed at the end of a password, while uppercase letters are least effective placed at the beginning.

Our second major contribution is a comparison of our real, high-value password data with password sets more typically used in password research. We compare the CMU passwords to passwords collected in an online study simulating CMU's password-creation process, as well as to data from online studies and a self-reported survey of CMU users discussed in prior work [30,47]. We also consider plaintext and cracked leaked passwords from low-value accounts [2, 19, 32, 43, 52]. We show that simulated password sets designed to closely mirror real authentication conditions consistently provide reasonably accurate substitutes for high-value real-world passwords, while leaked passwords vary widely in their effectiveness. This has important implications for passwords research, as most researchers must choose between leaked sets and experimental data. In the past, many researchers have chosen leaked sets; our results show this may be the wrong choice. Taken together, our approach and results provide a unique understanding of frequently used, high-value passwords as well as insight for improving future password research.

2. RELATED WORK

In this section, we review background in two key areas: the types of password corpora that have been analyzed previously, and efforts to define metrics for password strength.

2.1 Password corpora

Due to the sensitive nature of passwords, acquiring high-quality password corpora for analysis can be difficult. Data sets that have been used in previous work on passwords have all been non-ideal in at least one dimension.

Many researchers use password corpora collected from various security leaks [14, 24, 26, 35, 53, 55]. These corpora tend to be very large (tens of thousands to millions), and they represent in-use passwords selected by users. While this approach has many benefits, these passwords come with no contextual information about how they were made or used, and the released lists are difficult to verify. Furthermore, the largest leaks thus far have come from low-value accounts with weak password-composition policies, such as the RockYou gaming website. In addition, if the password file is encrypted or hashed, only those passwords that have been successfully cracked can be analyzed, biasing the data toward the more guessable. Other researchers obtain an organization's hashed or encrypted password file with permission and attempt to crack it [10, 57]. As with leaked password sets, the results are biased toward more guessable passwords.

Researchers who want to control the circumstances under which passwords are created often use lab studies. Some of these studies are small and focus on targeted populations, such as undergraduates [9,41,54]. Others are larger online studies with a more diverse population [30, 51]. In some cases, users are asked to create passwords for a low-value account associated with the study [9, 31]. Other studies have asked students to create passwords for accounts tied to a class [21, 29]. In contrast to these studies, the passwords used in our paper are created for high-value accounts, and are used frequently over a longer period of time.

Rather than collect passwords expressly for an experiment, some researchers ask users to self-report password information, including both password composition details and user sentiment information [7, 33, 37, 47, 49, 58]. While self-reported data can be very useful and can provide a lot of context, it cannot always be considered reliable, particularly with regard to a sensitive topic like passwords.

Finally, a small number of researchers have been able to work with large organizations to collect authentic data. Florêncio and Herley used an opt-in component of the Windows Live toolbar to collect information when users log into websites [17]. Bonneau worked with Yahoo! to analyze plaintext passwords entered by their users [3]. Both studies include very large, reliable samples, as well as good contextual information. Due to security concerns, however, in both studies researchers were able to record only extremely limited information about the content of the passwords, precluding many types of interesting analyses. In contrast, our paper connects information about each user with analysis of that user's plaintext password. Perhaps closest to our work, Fahl et al. use a within-subjects study to manually compare passwords users made in lab and online studies with their actual passwords; they found that while study passwords do have some problems, they can provide a reasonable approximation for real passwords if used carefully [16]. While Fahl et al. compare real password data with online and lab studies, we also compare real password data with commonly used leaked password sets.

In this paper, we overcome many limitations of past studies. Our password corpus includes more than 25,000 real passwords, created by users for frequently used, high-value accounts unrelated to our research context. We have indirect, yet extensive, access to plaintext passwords, allowing us to perform more complex and thorough analyses than was possible for other similarly authentic corpora. We also collect a significant amount of contextual information, including demographics, behavioral data, and user sentiment.

2.2 Password cracking and strength metrics

Accurately judging the strength of a password is crucial to understanding how to improve security. Historically, information entropy has been used to measure password strength, but it may insufficiently capture resistance to intelligent guessing attacks [17, 55]. More recently, researchers have suggested using password *guessability*, or ability to withstand guessing by a particular password cracker with particular training data, as a security metric [55]. This metric has the advantages of modeling knowledge a real-world adversary might have, as well as of bounding the attempts an adversary might make to guess a password, but its results are dependent on the chosen setup. *Guess numbers*, as defined by Kelley et al., measure how many guesses it would take a particular cracking algorithm and training setup to reach a given password [30]. Working without access to plaintext passwords, Bonneau suggests a guessing metric that reflects how many passwords an optimal attacker, who knows exactly which passwords to guess and what order to guess them in, will successfully break before guessing the password under consideration [4]. Several researchers have also used machine-learning techniques to classify passwords as weak or strong, based on labeled training data, but these techniques are only as good as the original classification [1, 27, 50].

The guessability metric of password strength dovetails with recent advances in password cracking. In contrast to prior brute-force or dictionary-based approaches [48], researchers have begun

to use deeper insights about the structure of passwords in cracking. For instance, Narayanan and Shmatikov substantially reduce the search space for attacks by modeling passwords as a character-level Markov chain using the probabilities of letters in natural language [39]. Using passwords leaked from websites like RockYou and others as training data, Weir creates a probabilistic context-free grammar for password cracking [56]. In this approach, guesses are ordered according to their likelihood, based on the frequency of their character-class structures in the training data, as well as the frequency of their digit and symbol substrings. This approach has been shown to be efficient in password cracking [30,57]. In this paper, we primarily use the guessability metric, simulating cracking using a modified version of Weir's algorithm [30].

3. DATA COLLECTION

In this section, we discuss our data sources. First, we review the university data collected for this study. Second, we discuss our procedures for ensuring security while working with real data, as well as the challenges these procedures create for analysis. Third, we briefly review other data sets that we use as supplemental data in our analysis. Fourth, we discuss the composition of our sample and the generalizability of our results.

3.1 University data

We study the passwords used by all of the more than 25,000 faculty, staff, and students of Carnegie Mellon University. These passwords are used as part of a single-sign-on system that allows users to access resources like email, tax and payroll statements, personnel directories, health information, grades and transcripts, course information, and other restricted university resources.

CMU's password-composition policy is a complex one, requiring at least one each of upper- and lowercase letters, digits, and symbols, as well as forbidding a dictionary word. All non-letter characters are removed and the password is lowercased before it is checked against a 241,497-word dictionary, unless the password is greater than 19 characters. The minimum length is eight characters, and no character can appear more than four times unless the password is greater than 19 characters in length.

We collect data from four sources: logs from the university's single-sign-on web authentication service, demographic data from the university's two personnel databases, responses to a survey advertised to users immediately after changing their passwords, and the plaintext passwords themselves. The web logs represent the period from January 1, 2012 through July 27, 2012. On July 28, the university's authentication infrastructure was replaced, and the logs from the two systems are incomparable. The personnel databases, as with most large organizations, are subject to bureaucratic errors that may cause some data to be incorrect.

Data from all four sources can be correlated using hashed user IDs (salted with a salt unknown to us, as described below). Plaintext passwords are divided into two groups: 25,459 passwords belonging to users with active CMU accounts, and 17,104 passwords belonging to users whose accounts were deactivated after they left the university, but which have not yet been deleted. Hereafter we refer to these as the CMUactive and CMUinactive sets respectively. Some of the CMUinactive accounts were created under an earlier composition policy that required only that passwords contain at least one character; as result, 1,635 CMUinactive passwords do not conform to the strict policy described above.

3.2 Working with real data securely

To get access to this hard-to-acquire real data, we spent months negotiating a process, vetted by both our IRB and information security office, that would allow the university to remain comfortable about security while also allowing us to perform useful analyses.

Plaintext passwords were made indirectly available to us through fortunate circumstances, which may not be reproducible in the future. The university was using a legacy credential management system (since abandoned), which, to meet certain functional requirements, reversibly encrypted user passwords, rather than using salted, hashed records. Researchers were never given access to the decryption key.

We were required to submit all the analysis software needed to parse, aggregate, and analyze data from the various data sources for rigorous code review. Upon approval, the code was transferred to a physically and digitally isolated computer accessible only to trusted members of the university's information security team. Throughout the process, users were identified only by a cryptographic hash of the user ID, created with a secret salt known only to one information technology manager.

We were able to consult remotely and sanity-check limited output, but we were never given direct access to passwords or their guess numbers. We did not have access to the machine on which the passwords resided — information security personnel ran code on our behalf. Decrypted plaintext passwords were never stored in non-volatile memory at any point in the process, and the swap file on the target machine was disabled. All analysis results were personally reviewed by the director of information security to ensure they contained no private data. We received only the results of aggregate analyses, and no information specific to single accounts. After final analysis, the source data was securely destroyed. The information security staff, who are not authors of the paper, represented an independent check on the risks of our analysis.

There was also concern that analyses might open up small segments of the population to risk of targeted attack. To address this, categories for demographic factors were combined so that the intersection of any two groups from different factors always contained more than 50 users. In some cases, this required the creation of an "other" group to combine several small, unrelated categories.

This approach helped to ensure that users were not put at risk, but it did create some challenges for analysis. We were never allowed to explore the data directly. For the most part, decisions about what data to collect and how to analyze it were made far in advance, without benefit of exploratory data analysis to guide our choices. To compensate for this, we selected a large set of possibly useful statistical comparisons; the correspondingly high chance of false positives forced us to apply strong statistical correction, somewhat reducing the statistical power of our analysis. To avoid wasting the time of the information security team members who performed the analysis, we automated as much of it as possible. The combination of complex calculations with long automation scripts unsurprisingly led to many bugs; the inevitable differences between anonymized sample data provided by the information technology division and the real data led to many more. The result was many iterations of remote debugging and subsequent code re-audit.

Further, our inability to examine the passwords directly masks some aspects of the data. We rely on an algorithm to learn and exploit common strings and structures within the CMU passwords, but we cannot know which patterns it exploits. It is possible that the data contains commonalities a determined attacker could exploit, but which the algorithm is not sophisticated enough to recognize.

3.3 Supplemental data sets

We use guess numbers generated using a modified version of the Weir algorithm to measure password strength, applying the approach of Kelley et al. [30]. Guess numbers depend on the amount

and quality of training data available to the guessing algorithm. We use training data that includes publicly available dictionaries (including the Google web corpus and the Openwall cracking dictionary); leaked password sets that were previously made public (including MySpace and RockYou), and data from online studies (using Amazon's Mechanical Turk service, or MTurk) in which participants created passwords under various conditions. For some tests, the algorithm is also trained on a subset of the CMU passwords (training and test sets are always kept disjoint for a given experiment).

We also compare the CMU passwords with 11 other data sets from various sources, as follows:

MTsim. 1,000 passwords collected from an MTurk experiment designed to simulate CMU password creation as closely as possible, both in policy requirements and in the website design.

MTbasic8. 1,000 passwords collected from MTurk [30]. The only requirement is a minimum length of 8 characters.

MTbasic16. 1,000 passwords collected from MTurk [30]. The only requirement is a minimum length of 16 characters.

MTdictionary8. 1,000 passwords collected from MTurk [30]. Minimum length 8. Discarding non-alphabetic characters, the password cannot be found in the 2.9-million-word free Openwall dictionary,[1] an order of magnitude larger than the dictionary used by CMU.

MTcomp8. 1,000 passwords collected from MTurk [30]. Same as MTdictionary8, but also requiring at least one lowercase letter, one uppercase letter, one digit, and one symbol.

SVcomp8. 470 self-reported responses to a previously published survey of CMU users [47].

We also compared our results with data from five real websites. In each case, we use a subset of the website passwords that meet CMU's requirements. Where more then enough conforming passwords were available, we draw the test set at random. Three of these leaked sets were leaked in plaintext; the other two come from the subset of the original leak that was successfully cracked.

RYcomp8. 1,000 plaintext passwords from RockYou (42,496 conforming, 32,603,144 total).

Ycomp8. 1,000 plaintext passwords from Yahoo! Voices (2,693 conforming, 453,488 total).

CSDNcomp8. 1,000 plaintext passwords from the Chinese Developer Network (12,455 conforming, 6,428,285 total).

SFcomp8. 1,000 cracked passwords from Strategic Forecasting, Inc., also known as *Stratfor*. (8,357 conforming, 804,034 total).

Gcomp8. 896 cracked passwords from Gawker (896 conforming, 694,064 total). All eight characters long.

Hereafter, we refer to all the leaked password sets, MTcomp8, MTsim, SVcomp8, and the real university passwords (CMUactive and CMUinactive) collectively as the *comprehensive-policy* passwords, as each includes four character classes and a dictionary check.

3.4 Experimental validity

CMU's complex password policy meets guidelines established by the InCommon Federation, which provides "a common framework for trusted shared management of access to on-line resources"[2] for educational and research institutions across the United States. As such, it is representative of similarly complex policies at other institutions. InCommon relies on NIST guidelines, which influence security standards at organizations across the United States [8].

We believe our results are reasonably representative of medium-sized research universities in the United States, and may be applicable to universities of other sizes or to other organizations with similar demographic profiles. CMU personnel represent a broad cross-section of ages, ethnic backgrounds, and nationalities, as well as a broad spectrum of geographic regions of origin within the United States. Although the sample is broad, its proportions do not match the general population.

Overall, the sample is considerably wealthier and more educated than the general American population. Most members of the sample currently live and work in or near Pittsburgh, where the main campus is located, but a fraction do live and work at other locations around the United States and internationally. We include some demographic factors as covariates, but many were not available from the university's personnel databases.

Compared to existing password research, we have significantly more knowledge of demographic factors than is available for most sets collected in the wild, and the CMU population overall is more diverse than the typical group of undergraduates used in lab studies. As a result, we believe that our results, if considered judiciously, can be applied to broader populations.

The guessing algorithm we use to measure password strength may not be optimal. While the algorithm has been used successfully in the past [30, 55], a more sophisticated algorithm might guess passwords more efficiently and produce different results.

4. UNDERSTANDING CMU PASSWORDS

In this section, we correlate the password strength of subsets of the CMU passwords with various demographic, password-composition, behavioral, and sentiment factors.

4.1 Analysis approach

For our correlation analysis in this section, we use guess numbers as our metric of password strength. We use a guess calculator for a modified version of Weir's guessing algorithm (see Section 2). We separate the CMUactive passwords into three folds for cross-validation, with each fold used once as a test set. The guess calculator is trained on the other two folds of CMUactive, all of CMUinactive, and a *Public* training set consisting of public and leaked password data. The *Public* set is composed of passwords from the MySpace, RockYou, Yahoo, CSDN, Stratfor, Gawker, and paid Openwall sets, as well as strings from the standard Unix dictionary and an inflection list[3] that includes various grammatical transformations. The set of alphabetic strings also includes unigrams from the Google Web N-Gram corpus.[4] The *Public* set was pruned so that only passwords containing at least eight characters and four character classes would be guessed.

Because of limitations in processing power, guess numbers can only be computed up to a given threshold. For each password, we either calculate a specific guess number, conclude that the password would have a guess number greater than the threshold n, or conclude that the password cannot be guessed given the current training data. The guessing threshold depends on the content of the training data as well as the experimental setup; experiments with higher guessing thresholds take longer to process. For this analysis, the guessing threshold is approximately 3.8 E14, or more than 380 trillion guesses; on our hardware, calculating guess numbers for each fold (about 8,000 passwords) takes about a day.

Once guess numbers are calculated for all CMUactive passwords, they are joined with data from the other university sources by matching hashed user IDs. Regressions are then performed on the re-

[1] http://download.openwall.net/pub/wordlists/

[2] http://www.incommon.org

[3] http://wordlist.sourceforge.net

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

sulting table to find correlations between guess numbers and other factors. Some of the factors we measure do not have useful values for all users, so we perform three separate regression analyses on appropriate subsets of users. These subsets are described in Sections 4.1.1-4.1.4.

The main regression technique we use is Cox regression, a technique adapted from the survival analysis literature, which typically studies factors that affect mortality [11]. The outcome variable is a pair of values: an observed state and the time of observation. If a password is guessed before the threshold, we mark the observed state as "dead" and the guess number as the "time of death." Otherwise, the observed state is "alive," and the guess threshold is the last time of observation. In the parlance of survival analysis, this is called right-censored data.

Using Cox regression, we are able to incorporate all the available data, over the range of guess numbers. This is an improvement over prior work, in which guessing success is examined at arbitrary points, such as the percentage guessed after a certain number of guesses [30, 56] or the amount of effort required to crack some percentage of passwords [3]. As with ordinary linear regression models, Cox regression estimates a linear term for each factor. This assumes that factors affecting the probability of survival have a linear effect over the guessing range; this is a common simplification often used to represent factors which might, in reality, have nonlinear effects.

To counteract overfitting, we use the standard backward elimination technique, removing one factor from the model at a time, until we minimize the Bayesian information criterion (BIC) [42]. We only report the final model. Before the analysis, we centered all the quantitative variables around zero by subtracting the mean from each value; this standard technique makes the regression output easier to interpret [15].

In addition to the main analysis, we check for interactions between factors — that is, factors whose effects are not independent — by performing the above survival analysis with all two-way interaction terms, again using the BIC for model selection. Since this project required all code to be reviewed before the passwords were analyzed, and survival analysis is a novel approach to measuring password strength, we supplemented the Cox regressions with logistic regressions on a binary outcome: guessed before the cutoff threshold (success) or not (failure). We found that both regression approaches generally agreed.

We next describe the subsets of data used in this analysis and the factors included in the corresponding regressions.

4.1.1 Model 1: All personnel

This data set contains all current users with complete demographic information (17,088). We consider the following factors:

Gender. Male or female.

Age. Birth year as recorded in the personnel database.

Status. Undergraduate, graduate student, non-degree-seeking student, faculty, or staff. As with all regression techniques, categorical factors are represented as a set of binary factors, with only the appropriate category (e.g., "undergraduate" for an undergraduate student) coded as true for each user. One arbitrarily selected category, known as the baseline, is left out: users belonging to that category are coded false for every binary factor. For this factor, faculty is the baseline.

College. Personnel are divided into eight colleges within the university, including a catch-all "other" category.

Location. Because the vast majority of university personnel are based at the main campus, we use only two groups: main campus and other location. This unfortunately groups locations from several different areas of the world (e.g., Silicon Valley, Qatar, Australia) together, despite possibly important cultural and organizational differences.

4.1.2 Model 2: All personnel plus composition

This data set contains the 17,088 users from Model 1. We consider Models 1 and 2 separately because all the factors in Model 1 are included in each subsequent model, providing a baseline. For Model 2, we add the following factors related to the composition of passwords:

Number of digits, lowercase, uppercase, and symbols. Four separate factors that measure the number of occurrences of each type of character in the user's password.

Location of digits, uppercase, and symbols. For each of the three special character classes, we identified the location of characters of this type in the user's password. This location is categorized as either all at the beginning; all at the end; all in a single group in the middle of the password; or spread out in some other pattern.

4.1.3 Model 3: Personnel with stable passwords

This data set includes the 12,175 users who did not change their passwords throughout the log measurement period. This allows us to conclude that the password for which we calculate a guess number was in use during all behavioral measurements. Factors include everything from Model 1, plus the following factors extracted from the authentication service logs:

Login count. The number of times this user successfully logged in during the measurement period. We hypothesized that users who log in more often might use stronger passwords, either because they are confident that repetition will allow them to remember, or because they simply value the account more highly.

Median interlogin time. The median time elapsed between each pair of this user's successive web logins. We hypothesized that users who go a shorter time between logins might be able to choose and remember more complex (stronger) passwords.

Password and username failures. The number of login failures attributable to an incorrect password or username (treated separately), normalized by login count. To avoid counting potentially malicious guessing, we count only failures during what was eventually a successful login session. We hypothesized that users with stronger passwords might find them more difficult to type, leading to more errors; alternatively, users who know they have difficulty typing might choose weaker passwords as a mitigation.

Median time elapsed during authentication. The median time elapsed between when the user arrives at the login page and when she successfully logs in, taken from server log data. This measure is imperfect; long elapsed times may represent users who open an authentication session in their browser and then ignore it for a long time before logging in, and different authentication servers may not have globally consistent timestamps. We hypothesized that users who take longer to log in might have passwords that are more difficult to remember or type in.

Wired login rate. The number of successful logins originating from an IP address associated with the main campus wired network, normalized by login count. This excludes logins made on the university's wireless network and those made from other campuses, as well as remote logins (such as from a user's home). We hypothesized that users who access their accounts only from organizational infrastructure on the main campus might think about passwords differently from those who frequently connect remotely or via wireless, since they are connecting over a more trusted medium. Unfortunately, we were unable to distinguish mobile devices like phones

or tablets from other wireless devices like laptops using the available log data.

Non-web authentication rate. The number of successful logins that do not correspond to web authentication events, normalized by login count. Because of incomplete log data, this value is only an approximation. We hypothesized that users who routinely access their accounts via an email program or other tools that store passwords might choose stronger passwords.

Personnel who did not log in at least twice are excluded, as they have no interlogin time.

4.1.4 *Model 4: Survey participants*

This data set includes 694 users who completed the survey after changing their passwords. This group is disjoint from the *stable passwords* group, as all members of this group have changed their passwords at least once since the start of the logging period.

The survey sample is a subset of the overall university population. New users, who must all change their system-assigned starter passwords, made up 22% of the sample (164). The sample is of course also biased toward people who are willing to take a survey in exchange for an entry in a drawing for an Amazon.com gift card. All data is self-reported.

We include all the factors from Model 1, plus the following:

Method of creation. Each user's selections from a list including: reused a password from another account; added or substituted numbers or symbols within a dictionary word; used the name of someone or something; and other choices. Participants were asked to choose all that apply; each possible answer was represented as a factor and coded true if the user selected it or false otherwise.

Reason for change. Each user's choice from among changing the default starter password, feeling the password was too old, resetting a forgotten password, and suspicion of a security breach. Participants were allowed to choose only one option. We hypothesized that those who forget their passwords might select simpler passwords to reduce cognitive load.

Storage. True if the user indicated she stores her password, either electronically or on paper; otherwise false. Prior work suggests that users who are asked to make more complex passwords are more likely to write them down [25]. Self-reporting for this category may be an undercount, as users who write down passwords are contravening common (although not necessarily correct [46]) advice and may be embarrassed to admit to it.

Sentiment during creation. Three factors, coded as true for users who agree or strongly agree (on a five-point Likert scale) that it was difficult, annoying, or fun to create a password conforming to the university's policy. Users who indicated disagreement or neutrality were coded false for that factor. We hypothesized that users who struggle to create a conforming password might be more likely to give up and choose something simple out of frustration.

4.2 Results

We find interesting correlations between password strength and other factors across each of the subpopulations we investigate. We describe the results for each model separately. Note that while we interpret the results for the reader, this was not a controlled experiment with random assignment, so we make no formal attempt at a causal analysis — any observed effect may involve confounds, and many independent variables in our data set are correlated with other independent variables.

4.2.1 *Model 1: All personnel*

For these users, we find password strength to be correlated with gender and college. Men have slightly stronger passwords than women: men's passwords are only 92% as likely as women's to be guessed at any point. Among colleges, users from the business school are associated with the weakest passwords: 24% more likely to be guessed than users in the arts school (the baseline in the regression). Computer science users have the strongest passwords, 68% as likely as the arts school and 55% as likely as the business school to be guessed. Every college except public policy is significantly different from the baseline. The full regression results, after backward elimination, are shown in Table 1. No significant interactions between factors were found in the final model, meaning the effects of the various factors are independent.

Pairwise comparisons reveal science, engineering, and computer science to be associated with stronger passwords than humanities, business, and public policy; computer science is also associated with stronger passwords than arts (all Holm-corrected Wilcoxon test, $p < 0.05$). Figure 1 illustrates these relationships. Figure 2 shows guess number results by gender. Our findings agree with Bonneau's result that men's passwords are slightly more resistant to offline guessing [3].

It is perhaps equally interesting to note the factors that do not appear in the final regression, including age, primary campus, and status as faculty, staff, or student. While we cannot positively con-

Factor	Coef.	Exp(coef)	SE	p-value
gender (male)	-0.085	0.918	0.023	**<0.001**
engineering	-0.218	0.804	0.042	**<0.001**
humanities	-0.106	0.899	0.048	**0.028**
public policy	0.081	1.084	0.051	0.112†
science	-0.286	0.751	0.055	**<.001**
other	-0.102	0.903	0.045	**.025**
computer science	-0.393	0.675	0.047	**<.001**
business	0.211	1.235	0.049	**<.001**

Table 1: Final Cox regression results for all personnel. Negative coefficients indicate stronger passwords. The exponential of the coefficient (exp(coef)) indicates how strongly that factor being true affects the probability of being guessed over a constant guessing range, compared to the baseline category. The baseline category for gender is female and for college is fine arts. For example, the second line indicates that engineering personnel are 80.4% as likely to have their passwords guessed as fine arts personnel. Results that are not statistically significant with $p < 0.05$ are grayed out and indicated by †.

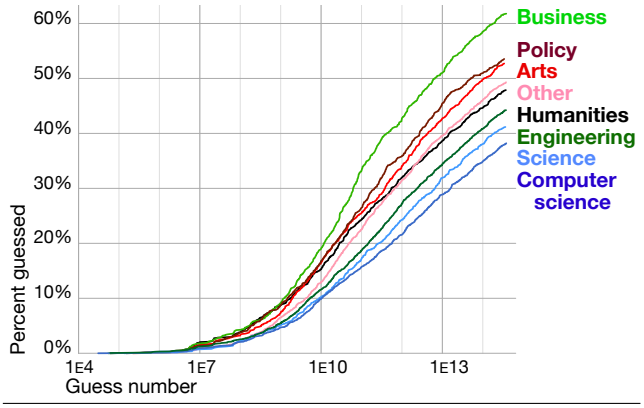


Figure 1: The percentage of passwords guessed after a given number of guesses (shown in log scale), by college within the university.

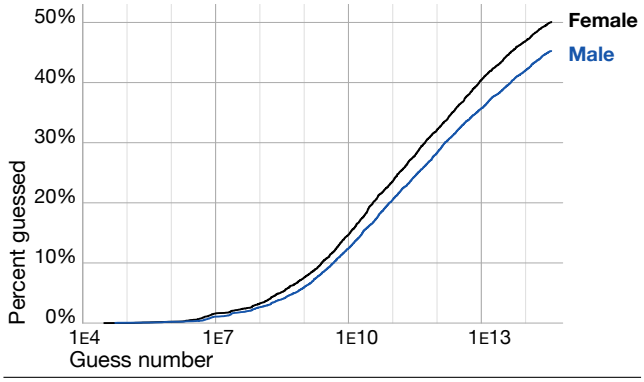


Figure 2: The percentage of passwords guessed after a given number of guesses (shown in log scale), by user gender.

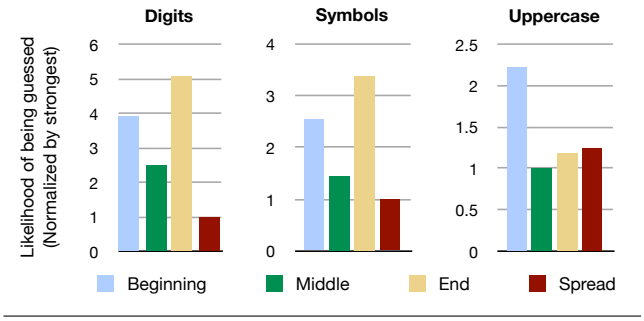


Figure 3: The relative likelihoods of passwords with digits, symbols, or uppercase letters in a given location being cracked. For example, a password with all its digits at the end is five times as likely to be cracked as a password with its digits spread throughout, other things being equal. The values are derived from the exponent of the regression coefficient, for the non-interaction model (Table 2). Each character class is normalized independently.

clude that these factors have no association with password strength, given our large sample size it seems likely that any effect is either a small one or is accounted for by the other factors in the regression.

4.2.2 Model 2: All personnel plus composition

In this section, we find that password composition is strongly correlated with password strength. In the non-interaction model, increasing the number of characters of any type is correlated with stronger passwords (Table 2). With the addition of each lowercase character or digit, a password becomes an estimated 70% as likely to be guessed. Additional symbols and uppercase characters have a stronger effect, reducing the likelihood of guessing to 56% and 46% per added character respectively. Placing digits and symbols anywhere but at the end, which is the baseline for the regression, is also correlated with stronger passwords. Multiple characters spread out in more than one location are associated with the strongest passwords — only 20% and 30% as likely to be guessed as passwords with digits and symbols, respectively, at the end. Placing uppercase characters at the beginning instead of at the end of a password is associated with much weaker passwords: 88% more likely to be guessed. Figure 3 illustrates the relative likelihood of being guessed based on placement for each character class.

Factor	Coef.	Exp(coef)	SE	p-value
number of digits	-0.343	0.709	0.009	<**0.001**
number of lowercase	-0.355	0.701	0.008	<**0.001**
number of uppercase	-0.783	0.457	0.028	<**0.001**
number of symbols	-0.582	0.559	0.037	<**0.001**
digits in middle	-0.714	0.490	0.040	<**0.001**
digits spread out	-1.624	0.197	0.051	<**0.001**
digits at beginning	-0.256	0.774	0.066	<**0.001**
uppercase in middle	-0.168	0.845	0.105	0.108†
uppercase spread out	0.055	1.057	0.114	0.629†
uppercase at beginning	0.631	1.879	0.105	<**0.001**
symbols in middle	-0.844	0.430	0.038	<**0.001**
symbols spread out	-1.217	0.296	0.085	<**0.001**
symbols at beginning	-0.287	0.751	0.070	<**0.001**
gender (male)	-4.4 E-4	1.000	0.023	0.985†
birth year	0.005	1.005	0.001	<**0.001**
engineering	-0.140	0.870	0.042	<**0.001**
humanities	-0.078	0.925	0.049	0.108†
public policy	0.029	1.029	0.051	0.576†
science	-0.161	0.851	0.055	**0.003**
other	-0.066	0.936	0.046	0.154†
computer science	-0.195	0.823	0.047	<**0.001**
business	0.167	1.182	0.049	<**0.001**

Table 2: Final Cox regression results for all participants, including composition factors. For an explanation, see Table 1.

Adding composition factors seems to account for some of the results from Model 1. Gender is no longer a significant factor, and the effects of all colleges are reduced. This indicates that simple password features such as length and the number and location of special characters might partially explain differences in guessability between these populations. In this model, younger users are associated with weaker passwords. This result agrees with that of Bonneau [3], and is also fairly small: each additional year is estimated to increase the likelihood of guessing by 0.5%.

We find several significant terms in the model with interactions (Table 3). In most cases, these are superadditive interactions, where two factors that were correlated with stronger passwords in the no-interactions model are associated with a stronger-than-expected effect when combined. For example, having both digits and symbols in the middle of a password has a much stronger impact on reducing guessability than one would expect given a model with no interactions — 35% as likely to be guessed. In contrast, two significant interactions are subadditive. First, adding lowercase when uppercase characters are spread out is 25% less effective than one would expect. While we do not have the data to investigate this particular result, one possible explanation is that users are adding capitalized words in predictable ways. Second, although additional digits and lowercase letters are correlated with stronger passwords, the benefit of adding a letter decreases with each digit already present, and vice versa. Passwords with more digits than average receive 3% less benefit than expected, per extra digit, from adding lowercase letters. For example, if a password has three more digits than average, adding lowercase letters is 9% less effective than expected.

4.2.3 Model 3: Personnel with stable passwords

We next consider users who kept the same password throughout the analysis period, for whom we have additional behavioral data. In addition to college and gender, several behavioral factors correlate with password strength for these users. In the model without interaction, we find that users who make more password errors have stronger passwords than other users, and users who log in more often have slightly weaker passwords (Table 4). An additional password error per login attempt is associated with a password only

Factor	Coef.	Exp(coef)	SE	p-value
number of digits	-0.309	0.734	0.011	<**0.001**
number of lowercase	-0.349	0.705	0.085	<**0.001**
number of uppercase	-0.391	0.676	0.099	<**0.001**
number of symbols	-0.632	0.531	0.037	<**0.001**
digits in middle	-0.130	0.878	0.296	0.660†
digits spread out	-1.569	0.208	0.294	<**0.001**
digits at beginning	0.419	1.520	0.304	0.168†
uppercase in middle	-0.006	0.994	0.158	0.970†
uppercase spread out	0.540	1.717	0.175	**0.002**
uppercase at beginning	0.854	2.349	0.160	<**0.001**
symbols in middle	-0.319	0.727	0.296	0.281†
symbols spread out	-1.403	0.246	0.339	<**0.001**
symbols at beginning	0.425	1.530	0.296	0.151†
gender (male)	0.007	1.007	0.023	0.773†
birth year	0.007	1.007	0.001	<**0.001**
engineering	-0.137	0.872	0.042	**0.001**
humanities	-0.071	0.931	0.049	0.144†
public policy	0.032	1.033	0.051	0.530†
science	-0.170	0.844	0.055	**0.002**
other	-0.081	0.922	0.046	0.079†
computer science	-0.193	0.825	0.048	<**0.001**
business	0.167	1.182	0.049	<**0.001**
(# dig.:# lower.)	0.032	1.032	0.004	<**0.001**
(# lower.:dig. middle)	-0.110	0.896	0.027	<**0.001**
(# lower.:dig. spread)	-0.237	0.789	0.035	<**0.001**
(# lower.:dig. begin.)	0.045	1.046	0.036	0.216†
(# lower.:upper. middle)	0.029	1.030	0.073	0.688†
(# lower.:upper. spread)	0.222	1.249	0.076	**0.004**
(# lower.:upper. begin.)	0.134	1.143	0.074	0.071†
(# lower.:sym. middle)	-0.146	0.864	0.026	<**0.001**
(# lower.:sym. spread)	-0.164	0.849	0.051	**0.001**
(# lower.:sym. begin.)	0.019	1.019	0.041	0.638†
(# lower.:birth year)	0.002	1.002	<0.001	<**0.001**
(# upper.:upper. middle)	-0.310	0.733	0.111	**0.005**
(# upper.:upper. spread)	-0.613	0.542	0.134	<**0.001**
(# upper.:upper. begin.)	-0.528	0.590	0.106	<**0.001**
(dig. middle:sym. middle)	-1.042	0.353	0.300	<**0.001**
(dig. spread:sym. middle)	-0.137	0.872	0.293	0.640†
(dig. begin.:sym. middle)	-0.314	0.730	0.307	0.306†
(dig. middle:sym. spread)	0.207	1.230	0.341	0.545†
(dig. spread:sym. spread)	0.225	1.253	0.379	0.552†
(dig. begin.:sym. spread)	-0.602	0.548	0.559	0.282†
(dig. middle:sym. begin.)	-0.604	0.547	0.306	**0.048**

Table 3: Final Cox regression results for all participants, including composition factors, with interactions. Interaction effects, shown in parentheses, indicate that combination of two factors is associated with stronger (negative coefficient) or weaker (positive coefficient) passwords than would be expected simply from adding the individual effects of the two factors.

58% as likely to be guessed. Each additional login during the measurement period is associated with an estimated increase in the likelihood of guessing of 0.026%. Though this effect is statistically significant, we consider the effect size to be negligible. No significant interactions between factors were found in the final model.

Notable behavioral factors that do not appear in the final regression include median time between login events, wired login rate (as opposed to wireless), and non-web authentication rate (e.g., using an email client to retrieve email without using the web interface).

4.2.4 Model 4: Survey participants

Among survey participants, we find correlations between password strength and responses to questions about compliance strategies and user sentiment during creation. As before, college also appears in the final model.

Factor	Coef.	Exp(coef)	SE	p-value
login count	<0.001	1.000	<0.001	<**0.001**
password fail rate	-0.543	0.581	0.116	<**0.001**
gender (male)	0.078	0.925	0.027	**0.005**
engineering	-0.273	0.761	0.048	<**0.001**
humanities	-0.107	0.898	0.054	**0.048**
public policy	0.079	1.082	0.058	0.176†
science	-0.325	0.722	0.062	<**0.001**
other	-0.103	0.902	0.053	0.051†
computer science	-0.459	0.632	0.055	<**0.001**
business	0.185	1.203	0.054	<**0.001**

Table 4: Final Cox regression results for personnel with consistent passwords, using a model with no interactions. For an explanation, see Table 1.

Factor	Coef.	Exp(coef)	SE	p-value
annoying	0.375	1.455	0.116	**0.001**
substituted numbers	-0.624	0.536	0.198	**0.002**
gender (male)	-0.199	0.820	0.120	0.098†
engineering	0.523	1.693	0.342	0.124†
humanities	0.435	1.545	0.367	0.235†
public policy	1.000	2.719	0.394	**0.011**
science	0.432	1.541	0.416	0.299†
other	0.654	1.922	0.334	0.051†
computer science	0.681	1.976	0.351	0.052†
business	1.039	2.826	0.376	**0.006**

Table 5: Final Cox regression results for survey participants. For an explanation, see Table 1.

Perhaps unsurprisingly, users who report that complying with the university's password policy was annoying have weaker passwords, 46% more likely to be guessed than those who do not report annoyance. This suggests that password policies that annoy users may be counterproductive. Users who substitute numbers for some of the letters in a word or name, by contrast, make passwords only 54% as likely to be guessed. We do not know whether or not these are typical "l33t" substitutions. Figures 4-5 illustrate these findings and full details appear in Table 5. For this subpopulation, there are not enough data points for a model with interaction to be valid.

Factors that do not appear in the final model include responses that complying with the password policy was difficult or fun; about twice as many users (302) agreed that it was annoying as agreed that it was difficult (162), and only 74 users found it fun. In addition, self-reported storage and the reason why the password was changed are not significant factors.

5. COMPARING REAL AND SIMULATED PASSWORD SETS

Acquiring high-quality password data for research is difficult, and may come with significant limitations on analyses. As a result, it is important to understand to what extent passwords collected in other settings — e.g., from data breaches or online studies — resemble high-value passwords in the wild. In this section, we examine in detail similarities and differences between the various password sets to which we have access. We first compare guessability, then examine other properties related to password composition. Overall, across several measures, passwords from online studies are consistently similar to the real, high-value CMU passwords. In contrast, passwords leaked from other sources prove to be close matches in some cases and by some metrics but highly dissimilar in others.

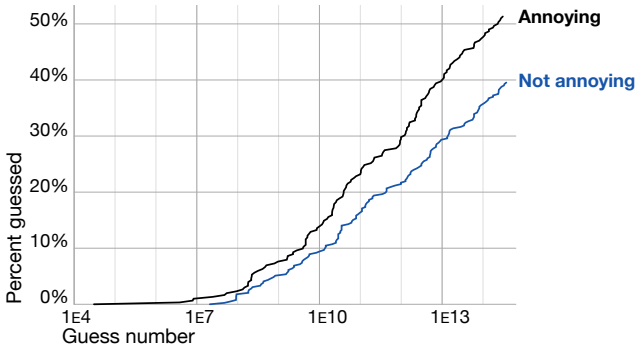


Figure 4: The percentage of passwords guessed after a given number of guesses (shown in log scale), by whether the user found password-creation annoying.

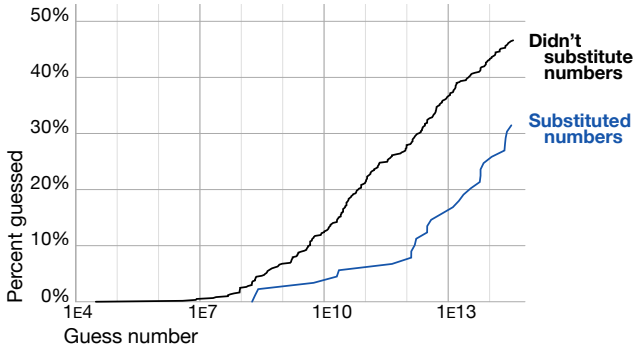


Figure 5: The percentage of passwords guessed after a given number of guesses (shown in log scale), by whether the user created the password by substituting numbers into a word.

5.1 Comparing guessability

We compare password sets primarily using guessability results. First, we calculate guess numbers for two attackers. The *limited-knowledge* attacker trains on publicly available data: the *Public* set described in Section 4.1. The *extensive-knowledge* attacker trains on the same public data, plus 20,000 CMUactive and 15,000 CMUinactive passwords. In each case, all data sources are weighted equally during training. Because these trainings are optimized for guessing passwords under the comprehensive policy, we cannot use this approach to compare university passwords to MTbasic8, MTbasic16, or MTdictionary8. We do compare CMUactive passwords to the other comprehensive-policy conditions: MTsim and MTcomp8 (online studies), RYcomp8, Ycomp8, and CSDNcomp8 (leaked plaintext sets), and Gcomp8 and SFcomp8 (leaked cracked sets). In all cases, we calculate and compare guess numbers only for passwords that are not used in training.

After calculating guess numbers, we compare guessability across password sets using another technique from the survival analysis literature: the Peto-Peto generalization of the Wilcoxon test [40], also known as a G^ρ test with $\rho = 1$ [20]. This test is designed to compare two survival data sets under the null hypothesis that both data sets were drawn from the same distribution. It has the additional property of weighting differences in early parts of the curve more heavily than later parts. As passwords are guessed and the population dwindles, the power of the test decreases. Unlike Cox regression, it does not assume that differences should occur at a constant rate across the entire curve. Table 6 shows percentages guessed at several guessing thresholds, as well as results of the G^1 significance test. In this table, differences in p-values indicate relative similarity to CMUactive; smaller p-values indicate greater divergence. Figure 6 shows guessability results for both attackers.

Online studies. Overall, the online studies provide more consistently similar matches to the CMU passwords than the leaked sets do. For both attackers, the CMUactive passwords are weaker than MTcomp8 and stronger than MTsim, but closer to MTcomp8. While the MTsim passwords were restricted to exactly match CMU policy, the MTcomp8 passwords were collected under a policy that, while similar, includes a notably harder dictionary check. As a result, it is unsurprising that MTcomp8 might produce more guess-resistant passwords. In fact, instrumentation from the MTurk studies shows that more than twice as many MTcomp8 participants as MTsim participants failed the dictionary check at least once during password creation (35% to 14%), suggesting the harder dictionary check did make an important difference.

The real CMUactive passwords were produced under the easier dictionary check, but they more closely resemble MTcomp8 than MTsim. We hypothesize that deployed passwords are harder to guess than the simulated version because online studies can only partially reproduce the effort users make to create strong passwords for high-value accounts.

Cracked password sets. As might be expected, cracked password sets provide especially poor points of comparison. Because they consist of a subset of the original data that was easiest to crack, they are guessed much more quickly than the CMU passwords, with 62% (SFcomp8) and 79% (Gcomp8) guessed before the cutoff.

Plaintext leaked password sets. The three plaintext leaked password sets are a more complicated case. Although the RYcomp8 subset appears highly similar to the CMU passwords under both attackers, CSDNcomp8 is only similar for the public attacker, and Ycomp8 is far off under both. Subsetted passwords from Ycomp8 and CSDNcomp8 are harder to guess than CMU passwords created under the same policy, which agrees with a previous finding [30]. Although this pattern does not hold for RYcomp8, it is important to note that there is much more RockYou data than data for any other set available in the training data. This advantage may partially compensate for subsets otherwise tending to be harder to guess.

To further examine our hypothesis that online studies provide a reasonably good proxy for real passwords, we obtain CMUactive guess numbers for two additional attackers: one trained on *Public* plus 3,000 CMUactive passwords, and another trained on *Public* plus 3,000 MTsim passwords. The distribution of guess numbers in the two data sets is not significantly different (G^1, uncorrected $p = 0.583$). This suggests that using MTsim passwords for cracking CMUactive passwords is a viable strategy. These results are shown in Figure 7.

5.2 Comparing other password properties

In addition to guessability, we compare several other properties of our data sets, including mean password length and quantity of characters per password from various character classes. We also consider estimated entropy, calculated as described in prior work [31]. For length, composition, and entropy, we can also compute confidence intervals using the statistical technique known as bootstrapping. Specifically, we use the "basic" bootstrap technique as identified by Davison and Hinkley [12].

We also compare the diversity of password structures, which correspond to high-level representations of passwords in the Weir grammar [56]. For example, the structure of "PassW0rd!" is "UL-

Attacker	Password set	N	1 E6	1 E9	1 E12	Cutoff	G^1 p-value compared to CMUactive
public	CMUactive	25,459	0.1	4.4	27.0	44.0	–
	MTsim†	1,000	0.2	6.3	30.5	47.5	0.005
	MTcomp8†	1,000	0.0	3.7	26.3	42.8	0.453
	RYcomp8†	1,000	0.1	3.6	26.6	47.3	0.250
	Ycomp8	1,000	0.0	2.7	20.8	37.7	**9.02 E-6**
	CSDNcomp8†	1,000	0.4	2.3	20.9	42.5	0.007
	SFcomp8	1,000	0.0	8.4	41.9	62.0	**0**
	Gcomp8	896	0.3	8.3	44.1	79.1	**0**
knowledgeable	CMUactive	5,459	0.4	6.4	30.7	48.7	–
	MTsim	1,000	0.1	11.5	34.7	54.1	**5.26 E-5**
	MTcomp8 †	1,000	0.0	7.7	27.9	43.1	0.008
	RYcomp8 †	1,000	0.0	5.0	29.4	49.4	0.573
	Ycomp8	1,000	0.0	4.9	23.3	39.7	**4.08 E-8**
	CSDNcomp8	1,000	0.4	2.9	24.1	42.2	**8.66 E-7**
	SFcomp8	1,000	0.2	10.4	44.2	63.0	**0**
	Gcomp8	896	0.2	10.7	49.4	73.3	**0**

Table 6: Guessing results for comprehensive-policy password sets. The columns provide the number of passwords in the set (N), the percentage of passwords guessed at various guessing points, and the results of the G^1 test comparing the guessing distributions. Rows in bold have guessing distributions that are statistically significantly different from CMUactive, with Bonferroni-corrected $p < 0.00139$; rows that are not significantly different are marked with †. The guessing cutoff is 3.6 E14 for the limited-knowledge attacker and 3.8 E14 for the extensive-knowledge attacker.

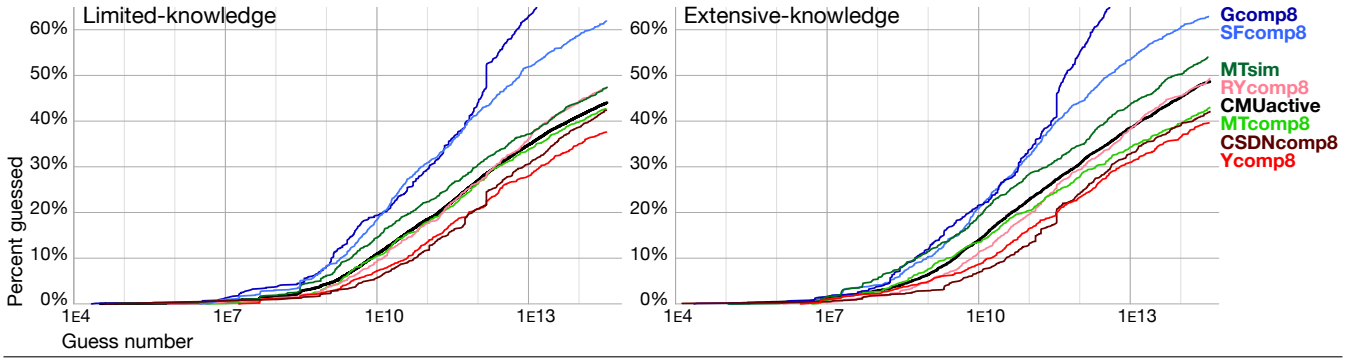


Figure 6: The percentage of passwords guessed after a given number of guesses (shown in log scale), by password set. The guessability results on the left are for the *limited-knowledge* attacker, who has only publicly available data. The guessability results on the right are for the *extensive-knowledge* attacker, who has access to some data from the same password sets for training.

LLUDLLS" (U = uppercase, L = lowercase, D = digit, and S = symbol). We measure diversity by randomly sampling 1,000 passwords from a data set, identifying their structures, and counting the number of distinct structures in the sample; we repeat this sampling process 100 times and use the average structure count. The choice of 1,000 for sample size is arbitrary and any reasonable sample size might be used.

Finally, we compare password sets using their probability distributions, an essential component of many password-strength metrics [3]. We use empirical probabilities as observed in each data set. We can only consider the most popular passwords, since almost all passwords are unique in sets as small as many of our sources.

For each of these measurements, we use all conforming passwords from each original data set, rather than the 1,000-password samples that were used for consistency in the guessability results.

Results for length, composition, entropy, and structural diversity are given in Table 7, and a subset are also shown in Figure 8. For the most part, using these metrics, comprehensive-policy passwords more closely resemble each other than passwords from other policies. As expected, passwords from policies that did not require them have fewer symbols and uppercase letters.

Perhaps more interesting is to consider how the other password sets within the comprehensive-policy group relate to CMU passwords, which protect high-value accounts. In length and composition, the passwords from online studies are consistently similar to the real CMU passwords, while passwords from leaked sets show more variance, sometimes appearing very similar and other times very different. It is particularly interesting to note that although RYcomp8 appeared very similar to CMUactive in guessability, its composition features are highly dissimilar, suggesting that it may not make a good proxy for real high-value passwords.

Self-reported survey responses from comprehensive-policy users are perhaps surprisingly similar in length and composition to other comprehensive-policy responses; the small sample size makes it difficult to ascertain precisely how similar.

Using entropy as a metric, passwords taken from Yahoo! are most similar to CMU passwords, while RYcomp8 and CSDNcomp8 passwords are most different. In structural diversity MTsim and MTcomp8 are closest, while Ycomp8 and SFcomp8 are farthest; perhaps unsurprisingly, the cracked SFcomp8 set shows by far the least structural diversity of any comprehensive-policy set.

	N	Length	# Digits	# Symbols	# Uppercase	Entropy	# Structures
CMUactive	25,459	10.7 [10.67–10.74]	2.8 [2.77–2.81]	1.2 [1.20–1.21]	1.5 [1.44–1.47]	36.8 [36.20–37.40]	689
MTsim	3,000	10.7 [10.54–10.77]	2.6 [2.56–2.67]	1.2 [1.17–1.22]	1.5 [1.41–1.50]	35.1 [34.50–35.60]	624
MTcomp8	3,000	10.7 [10.60–10.77]	2.2 [2.15–2.25]	1.2 [1.14–1.17]	1.5 [1.48–1.56]	34.2 [33.75–34.67]	630
RYcomp8	42,496	12.6 [12.35–12.80]	2.6 [2.56–2.61]	1.9 [1.89–1.99]	1.8 [1.79–1.82]	40.3 [38.55–42.31	769
Ycomp8	2,693	10.4 [10.27–10.44]	2.5 [2.41–2.50]	1.6 [1.52–1.58]	1.8 [1.76–1.84]	36.7 [36.29–37.19]	811
CSDNcomp8	12,455	11.1 [11.01–11.11]	3.8 [3.78–3.86]	1.5 [1.44–1.47]	2.0 [1.96–2.00]	41.2 [40.47–41.91]	782
SFcomp8	8,357	11.0 [10.88–11.05]	2.4 [2.39–2.45]	1.3 [1.26–1.29]	1.5 [1.51–1.57]	34.7 [34.13–35.18]	585
Gcomp8	896	8.0	1.9 [1.80–1.93]	1.2 [1.13–1.18]	1.3 [1.29–1.38]	†	†
SVcomp8	470	10.5 [10.18–10.78]	2.7 [2.39–2.98]	1.4 [1.23–1.53]	1.5 [1.35–1.72]	†	†
MTbasic8	1,000	9.7 [9.52–9.82]	2.4 [2.23–2.56]	0.1 [0.09–0.25]	0.4 [0.33–0.52]	29.6	322
MTdictionary8	1,000	9.7 [9.57–9.83]	2.6 [2.39–2.77]	0.2 [0.11–0.20]	0.4 [0.30–0.46]	29.1	317
MTbasic16	1,000	17.9 [17.76–18.11]	3.8 [3.46–4.05]	0.2 [0.12–0.20]	0.5 [0.41–0.67]	44.7	391

Table 7: Comparing password properties. Shown are mean values with 95% confidence intervals for various password properties. The Structures column gives the number of unique structures found in 1,000 passwords (average of 100 samples). † Because fewer than 1,000 passwords were available for Gcomp8 and SVcomp8, comparable entropy values and structure counts could not be calculated.

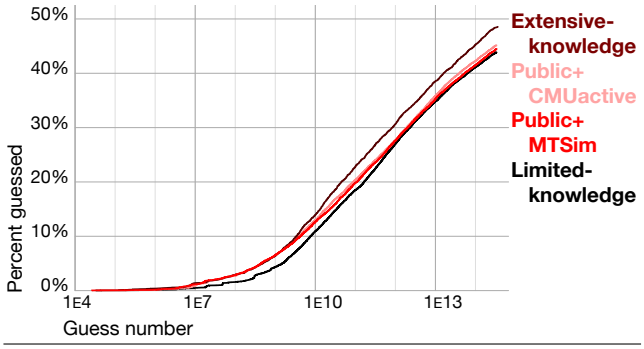


Figure 7: Results from a simulated attack from four different attackers, each with different training data, against CMUactive. *Public + 3K CMUactive* is trained on *Public* plus 3,000 CMUactive passwords. *Public + 3K MTsim* is trained on *Public* plus 3,000 MTsim passwords. The distributions of guess numbers for these two attackers do not differ significantly. *limited-knowledge* and *extensive-knowledge* are discussed in Figure 6.

Results from comparing probability distributions are given in Table 8. Based on the empirical probability of just the most popular password, CMUactive is the strongest policy, and RYcomp8, Ycomp8, CSDNcomp8, Gcomp8, MTbasic8, and MTdictionary8 are all significantly weaker. Among all sets considered, only the empirical probabilities of MTsim and MTcomp8 are not significantly different from CMUactive for passwords of any rank (Bonferroni-corrected χ^2, $p < 0.05$), though this could be attributed to small sample size. Surprisingly, SFcomp8 is not significantly different from CMUactive at first, but it becomes significantly different when comparing passwords of rank greater than one. In addition, the empirical probabilities of SFcomp8 do not drop off from ranks one to four, unlike every other set. If this is a byproduct of how the set was cracked, this provides further evidence against the use of cracked password sets in research.

6. CONCLUSIONS

The CMU information technology division agreed to work with us on this research in part to gain improved understanding of the current state of password security at CMU. We expect that future updates to the university's password policies and procedures will take our results into account. Beyond that, we believe our results provide guidance to users, system administrators, and information security personnel, as well as informing future password research.

We find that some elements of the university population create more vulnerable passwords than others. It is possible that some of these users, such as personnel in the business and arts schools, would create stronger passwords if they received instruction about how to do so. On the other hand, if these users are creating weak passwords because they don't feel that creating a strong password is in their interest [22], an education campaign could focus on password strength as a community issue.

In line with prior work, we find that male users, older users, and users who log in less frequently are correlated with slightly stronger passwords; however, in each case the effect size is small, and using different models we find that other factors may partially account for the effects. We also confirm patterns previously held as folk wisdom: passwords with more digits, symbols, and uppercase letters are harder to crack, but adding them in predictable places is less effective.

Using personnel databases, server logs, and surveys, we extend our analysis to include user behavior, sentiment, and additional demographic factors. We find that users who expressed annoyance with CMU's complex password policy were associated with weaker passwords; again, targeted education about the rationale for the policy requirements might help. Our findings also suggest further research into the usability and security implications of password managers as an aid to these users.

It is important to note that our analysis centers on passwords created under CMU's comprehensive password policy. While our results suggest that users who go beyond the minimum requirements of this policy have stronger passwords, our analysis does not allow us to draw conclusions about how the various requirements of the policy contribute to password strength. Our analysis suggests it would be useful to find policies that would be less annoying to users and that would discourage users from complying with the policy in predictable ways. Further work is needed to determine whether the CMU policy might be improved by relaxing some requirements and replacing them with others; for example, reducing the number of required character classes but requiring longer passwords, prohibiting special characters at the beginning or end of the password, or changing the dictionary check to permit dictionary words with symbols or digits in the middle.

Our research also provides guidance for future password studies. For researchers who may have an opportunity to gain limited access

	Observed probability of nth most popular password										%
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Unique**
CMUactive	0.094%	0.051%	0.043%	0.039%	0.035%	0.035%	0.031%	0.027%	0.024%	0.024%	97.910%
MTsim	0.200%†	0.100%†	0.100%†	0.100%†	0.100%†	0.067%†	0.067%†	0.067%†	0.067%†	0.067%†	99.067%
MTcomp8	0.233%†	0.100%†	0.067%†	0.067%†	0.067%†	0.067%†	0.067%†	0.067%†	0.067%†	0.067%†	99.133%
RYcomp8	0.513%	0.304%	0.242%	0.214%	0.134%	0.115%	0.101%†	0.099%†	0.068%†	0.066%†	87.877%
Ycomp8	0.520%	0.149%†	0.149%†	0.149%†	0.111%†	0.111%†	0.111%†	0.111%†	0.111%†	0.111%†	93.427%
CSDNcomp8	2.529%	1.429%	0.715%	0.426%	0.241%	0.233%	0.225%	0.217%	0.161%	0.128%	78.667%
SFcomp8	0.191%†	0.191%	0.191%	0.191%	0.179%	0.168%	0.096%†	0.096%†	0.084%†	0.084%†	95.058%
Gcomp8	4.911%	0.893%	0.893%	0.893%	0.670%	0.670%	0.670%	0.558%	0.558%	0.558%	79.464%
MTbasic8	1.300%	0.700%	0.600%	0.200%†	0.200%†	0.200%†	0.200%†	0.200%†	0.100%†	0.100%†	96.400%
MTdictionary8	2.300%	0.800%	0.400%	0.300%	0.200%†	0.200%†	0.200%†	0.200%†	0.200%†	0.200%†	94.600%
MTbasic16	0.600%	0.500%	0.500%	0.400%	0.300%	0.300%	0.300%	0.200%†	0.200%†	0.200%†	95.700%

Table 8: Empirical probabilities for the 10 most popular passwords and the total probability mass of unique passwords in each set. Probabilities that are not significantly different from CMUactive for a given password rank are grayed out and marked with a † (Bonferroni-corrected χ^2 test, $p < 0.05$).

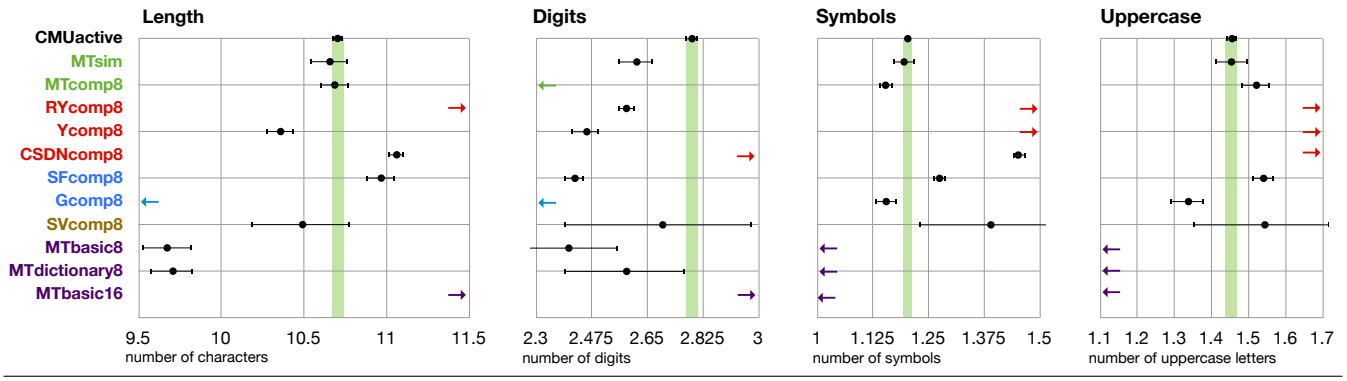


Figure 8: Password-composition characteristics, by password set, with 95% confidence intervals. The confidence interval for CMUactive is shaded. By these metrics, MTsim is generally the closest match for CMUactive.

to genuine passwords, we discuss our procedure for analyzing those passwords while respecting users' privacy and security.

For researchers restricted to more traditional mechanisms of password collection — such as lab or online studies and subsetting from leaked password data — we provide insight into similarities and differences between those passwords sets and frequently used passwords protecting real-world, high-value accounts. Consistent with previous work [30], we find that subsetting passwords from those created under one policy to approximate passwords created under another policy is not an optimal solution to gathering good password data for analysis. While these passwords are sometimes similar to the targeted passwords on some metrics, their high variance makes them unreliable as proxies.

We find that passwords created on MTurk are not a perfect substitute for high-value passwords either; the simulated passwords we collected were slightly weaker than the genuine ones. However, the simulated passwords do seem to be reasonably close in several respects, including length and character composition. Further, when used as training data for guessing genuine passwords, passwords from MTurk were just as effective as genuine passwords. These results indicate that passwords gathered from carefully controlled experimental studies may be an acceptable approximation of real-world, high-value passwords, while being much easier to collect.

7. ACKNOWLEDGMENTS

We gratefully acknowledge the many people at CMU who made this work possible, especially the identity services team and information security office. We thank Howard Seltman for statistics advice. This research was supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office, by NSF grants DGE-0903659 and CNS-1116776, by Air Force Research Lab Award No. FA87501220139, by the DoD through the NDSEG Fellowship Program, by the Facebook graduate fellowship program, and by a gift from Microsoft Research.

8. REFERENCES

[1] F. Bergadano, B. Crispo, and G. Ruffo. Proactive password checking with decision trees. In *Proc. CCS*, 1997.

[2] J. Bonneau. The Gawker hack: how a million passwords were lost. *Light Blue Touchpaper* blog, December 2010. `http://www.lightbluetouchpaper.org/2010/12/15/the-gawker-hack-how-a-million-passwords-were-lost/`.

[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. Statistical metrics for individual password strength. In *Proc. SPW*, 2012.

[5] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proc. IEEE Symposium on Security and Privacy*, 2012.

[6] P. Bright. Sony hacked yet again, plaintext passwords, e-mails, DOB posted. *Ars Technica*, June 2011. http://arstechnica.com/tech-policy/2011/06/sony-hacked-yet-again-plaintext-passwords-posted/.

[7] K. Bryant and J. Campbell. User behaviours associated with password security and management. *Australasian Journal of Information Systems*, 14(1):81–100, 2006.

[8] W. E. Burr, D. F. Dodson, and W. T. Polk. Electronic authentication guideline. Technical report, NIST, 2006.

[9] D. S. Carstens, L. C. Malone, and P. McCauley-Bell. Applying chunking theory in organizational password guidelines. *Journal of Information, Information Technology, and Organizations*, 1:97–113, 2006.

[10] J. A. Cazier and B. D. Medlin. Password security: An empirical investigation into e-commerce passwords and their crack times. *Information Systems Security*, 15(6):45–55, 2006.

[11] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[12] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, 1997.

[13] M. Dell'Amico, P. Michiardi, and Y. Roudier. Password strength: An empirical analysis. In *Proc. INFOCOM*, 2010.

[14] M. M. A. Devillers. *Analyzing password strength*. PhD thesis, Radboud University Nijmegen, 2010.

[15] J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Thomson Learning Brooks/Cole, 2004.

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

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

[18] D. Florêncio and C. Herley. Where do security policies come from? In *Proc. SOUPS*, 2010.

[19] D. Goodin. Hackers expose 453,000 credentials allegedly taken from Yahoo service. *Ars Technica*, July 2012. http://arstechnica.com/security/2012/07/yahoo-service-hacked/.

[20] D. P. Harrington and T. R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[21] K. Helkala and N. Svendsen. The security and memorability of passwords generated by using an association element and a personal factor. In *Proc. NordSec*, 2011.

[22] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proc. NSPW*, 2009.

[23] C. Herley and P. Van Oorschot. A research agenda acknowledging the persistence of passwords. *IEEE Security and Privacy*, 10(1):28–36, 2012.

[24] T. Hunt. The science of password selection. *TroyHunt.com* blog, July 2011. http://www.troyhunt.com/2011/07/science-of-password-selection.html.

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

[26] M. Jakobsson and M. Dhiman. The benefits of understanding passwords. In *Proc. HotSec*, 2012.

[27] K. S. Jamuna, S. Karpagavalli, and M. S. Vijaya. A novel approach for password strength analysis through support vector machine. *International Journal on Recent Trends in Engineering*, 2(1):79–82, 2009.

[28] D. L. Jobusch and A. Oldehoeft. A Survey of Password Mechanisms. *Computers and Security*, 8(8):675–689, 1989.

[29] M. Keith, B. Shao, and P. Steinbart. A behavioral analysis of passphrase design and effectiveness. *Journal of the Association for Information Systems*, 10(2):63–89, 2009.

[30] 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*, 2012.

[31] 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.

[32] M. Kumar. China Software Developer Network (CSDN) 6 million user data leaked. *The Hacker News*, December 2011. http://thehackernews.com/2011/12/china-software-developer-network-csdn-6.html.

[33] N. Kumar. Password in practice: An usability survey. *Journal of Global Research in Computer Science*, 2(5):107–112, 2011.

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

[35] D. Malone and K. Maher. Investigating the distribution of password choices. In *Proc. WWW*, 2012.

[36] B. D. Medlin and J. A. Cazier. An empirical investigation: Health care employee passwords and their crack times in relationship to hipaa security standards. *International Journal of Healthcare Information Systems and Informatics*, 2(3):39–48, 2007.

[37] B. D. Medlin, J. A. Cazier, and D. P. Foulk. Analyzing the vulnerability of US hospitals to social engineering attacks: how many of your employees would share their password? *International Journal of Information Security and Privacy*, 2(3):71–83, 2008.

[38] R. Morris and K. Thompson. Password security: a case history. *Communications of the ACM*, 22(11):594–597, 1979.

[39] A. Narayanan and V. Shmatikov. Fast dictionary attacks on passwords using time-space tradeoff. In *Proc. CCS*, 2005.

[40] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207, 1972.

[41] R. W. Proctor, M.-C. Lien, K.-P. L. Vu, E. E. Schultz, and G. Salvendy. Improving computer security for authentication of users: Influence of proactive password restrictions. *Behavior Research Methods, Instruments & Computers*, 34(2):163–169, 2002.

[42] A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

[43] S. Ragan. Report: Analysis of the Stratfor password list. *The Tech Herald*, January 2012. http://www.thetechherald.com/articles/Report-Analysis-of-the-Stratfor-Password-List.

[44] Rapid7. LinkedIn passwords lifted. http://www.rapid7.com/resources/infographics/linkedIn-passwords-lifted.html, retrieved September 2012.

[45] D. A. Sawyer. *The Characteristics of User-Generated Passwords*. PhD thesis, Naval Postgraduate School, 1990.

[46] B. Schneier. Write down your password. *Schneier on Security* blog, 2005. http://www.schneier.com/blog/archives/2005/06/write_down_your.html.

[47] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proc. SOUPS*, 2010.

[48] Solar Designer. John the Ripper, 1996-present. http://www.openwall.com/john/.

[49] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton. Analysis of end user security behaviors. *Computers and Security*, 24(2):124–133, 2005.

[50] G. Suganya, S. Karpgavalli, and V. Christina. Proactive password strength analyzer using filters and machine learning techniques. *International Journal of Computer Applications*, 7(14):1–5, 2010.

[51] B. Ur, P. G. Kelley, 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*, 2012.

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

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

[54] 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. *International Journal of Human-Computer Studies*, 65(8):744–757, 2007.

[55] 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.

[56] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. In *Proc. IEEE Symposium on Security and Privacy*, 2009.

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

[58] M. Zviran and W. J. Haga. Password security: an empirical study. *Journal of Management Information Systems*, 15(4):161–185, 1999.