Password Correlation: Quantification, Evaluation and Application

Shouling Ji[†], Shukun Yang[‡], Anupam Das[§], Xin Hu[♯], and Raheem Beyah[‡]
[†] College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China
[‡] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0765, USA
[§] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[♯] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
Email: {sji, syang87}@gatech.edu, anupamd@cs.cmu.edu, huxinsmail@gmail.com, rbeyah@ece.gatech.edu

***Abstract*—In this paper, we study the correlation between passwords across different datasets which quantitatively explains the success of existing *training-based* password cracking techniques. We also study the correlation between a user's password and his/her social profile. This enabled us to develop the first social profile-aware password strength meter, namely *SocialShield*. Our quantification techniques and SocialShield have meaningful implications to system administrators, users, and researchers, e.g., helping them quantitatively understand the threats posed by a password leakage incident, defending against emerging profile-based password attacks, and facilitating the research of countermeasures against existing and newly developed training-based password attacks. We validate our proposed quantification techniques and SocialShield through extensive experiments by leveraging real-world leaked passwords. Experimental results demonstrate that our quantification techniques are accurate in measuring correlation among different leaked datasets and that although SocialShield is light-weight, it is effective in defending against profile-based password attacks.**

I. Introduction

Text based authentication (*password* for convenience) is the most widely used user authentication method in modern computer systems [1]–[5]. Although it has been observed that passwords have several shortcomings [3], [5]–[8], e.g., vulnerable to password cracking attacks, they are likely to remain as the most dominating means of authentication for the foreseeable future. This is because passwords have several important advantages over their alternatives, e.g., scalability, simplicity, and high performance–price ratio [4].

In the past decade, several powerful password cracking algorithms have been presented, e.g., *Markov model-based schemes* [9]–[11], *structure-based schemes* [12], [13], John the Ripper (JtR) [14], and Hashcat [15]. Generally, when conducting an attack, an adversary employs these cracking techniques to train a password cracker leveraging some known passwords (e.g., leaked passwords), and then use the cracker to generate password guesses to attack the target dataset. A natural question is *what is the underlying reason for the success of these password cracking techniques?* The answer to this question is that user-chosen passwords (specifically, those susceptible to training-based password cracking attacks) are "*similar*" with respect to password structure and semantics, i.e., passwords exhibit demographic, behavioral, cultural, lingual, and regional correlations as demonstrated by many empirical observations [3]–[5], [7], [16]. Therefore, trained password crackers are powerful in making accurate guesses when attacking a target password dataset.

Furthermore, recently, it has been shown that users' social profiles can be utilized to facilitate the password attacking process [11]. Leveraging users' education, address, and other profile information, a trained Markov model-based cracker can crack 5% - 30% more passwords [11]. The improvement is due to the fact that many user-chosen passwords are correlated with those users' profile. Thus, users' profile can be employed to improve the password cracking process.

Contributions and Implications: Although password dataset correlation and the correlation between user-chosen password and his/her social profile have important implications on password security (e.g., significantly making existing password systems more vulnerable), to the best of our knowledge, there is no existing work that has quantitatively studies these issues. This motivates us to study and quantify the correlation between password datasets (which we term as *password–password* correlation). We also study how one's password might be related to one's social profile (which we term as *password–profile* correlation). Such an investigation quantitatively specifies how a password (dataset) is compositionally, linguistically, and/or semantically correlated to other passwords and/or users-profile information. Specifically, we make the following contributions in this paper:

(i) We propose the first *password–password* correlation quantification framework, under which we can quantify *structure-based*, n-*gram–based*, and *dictionary-based* correlation of passwords. Our quantification explains the theoretical foundation of existing password correlation based observations [3]–[5], [7], [16] as well as the success of existing *training-based* password cracking techniques [9]–[15]. We also examine the performance of our password–password correlation quantification via a comprehensive *attack-based evaluation* leveraging real-world passwords. The experimental results demonstrate that our quantification can accurately indicate the correlation among different password datasets.

(ii) We propose the first *password–profile* correlation quantification framework, which provides the theoretical foundation for the success of emerging *profile-based* password attacks (e.g., [11], [14]). Based on our password–profile correlation

quantification, we develop the first light-weight *social profile-aware password strength meter*, namely *SocialShield*. Through extensive evaluations leveraging real-world passwords and their profile information, we validate that SocialShield is very accurate and effective in measuring the password strength in terms of their associated profile information.

Our correlation quantification techniques and profile-aware password meter have meaningful implications to password system administrators, users, and researchers. For system administrators, they can employ our password–password correlation quantification techniques to quantitatively understand the threat to their password datasets caused by the leakage of other password datasets (which have become common nowadays, e.g., the recent Gmail password leakage [17] and Yahoo! password leakage [18]). Furthermore, they can also plug in our light-weight SocialShield as an *add-on* to their password systems to defend against emerging profile-based password attacks. For users, SocialShield can provide *real-time* feedback of the correlation between their chosen passwords and profile information (information used for registration) which they can use to improve their password's resistance against emerging profile-based password attacks. For researchers, our quantification techniques enable them to quantitatively understand the correlation of passwords and then develop effective countermeasures to defend against existing and newly developed password attacks.

Roadmap: The rest of this paper is organized as follows. In Section II, we summarize the related work. In Section III, we describe the password datasets used in our evaluations. In Section IV, we propose the password–password correlation quantification framework along with experimental evaluation and analysis. In Section V, we present the password–profile correlation quantification framework, the design of SocialShield, and some evaluation results. The limitations and future work of this work are discussed in Section VI. Finally, we conclude this paper in Section VII.

II. Related Work

A. Password Measurement

In [19], Weir et al. evaluated testing metrics for password creation policies by attacking revealed passwords using their Probabilistic Context-Free Grammar (PCFG) based cracking algorithm. Another work employing the password cracking idea to measure password strength is [7], where Kelley et al. analyzed 12K passwords collected under seven composition policies via an online study. Komanduri et al. implemented another tool, namely *Telepathwords*, to help users create strong passwords [20]. In [8], Ur et al. studied the effect of strength meters on password creation. Another work studying existing password meters is [6], where Carnavalet and Mannan analyzed 11 commercial meters. To improve the accuracy of password strength measurement, Castelluccia et al. presented *adaptive password strength meters* [10]. In [5], Ma et al. conducted a study of probabilistic password models. In [21], [22], Ji et al. conducted a large-scale cracking-based password security measurement. They also developed an open-source and modular Password Analysis and Research System (PARS) in [1], [23].

B. Password Cracking

In [9], Narayanan and Shmatikov proposed to use *standard Markov modeling techniques* to dramatically reduce the search size of password space. In [10], Castelluccia et al. improved the Markov model proposed in [9]. They proposed to construct an n-*gram* based Markov model to generate password guesses. Dürmuth et al. proposed an improved password cracking algorithm, namely *Ordered Markov ENumerator* (OMEN) in [11], which can make password guesses in the decreasing order of likelihood. Furthermore, they also extended OMEN to OMEN+, where users' social profiles are considered in password cracking. Taking another approach, Weir et al. proposed a password cracking algorithm using Probabilistic Context-Free Grammars (PCFGs) [12]. Veras et al. in [13] proposed an improved PCFG based password cracking algorithm, denoted by VCT, where the grammars take into account *structures, syntactics, and semantics* of passwords. In [24], Zhang et al. studied the effect of expired passwords on the security of current passwords. Another similar scheme is presented in [25], where Das et al. studied the password reuse problem.

There are also many password cracking tools available, among which the most popular one is John the Ripper (JtR) [14]. JtR supports multiple modes: *Wordlist mode* (JtR-W), *Single mode* (JtR-S), *Incremental mode* (JtR-I), and *Markov mode* (JtR-M). In JtR-W, a dictionary and a password hash file serve as inputs. JtR will try each word in the dictionary as a seed to perform cracking. In JtR-S, each password hash serves as an input along with an auxiliary string, e.g., username. Then JtR-S applies a set of mangling rules to the auxiliary string to generate password guesses. JtR-I is an intelligent *brute force* cracking method. JtR-M is a similar Markov model based cracking strategy as described in [9].

C. Password Habits

In [26], Florêncio and Herley conducted a large scale study of web password habits. In [27], Bonneau et al. evaluated two decades of text-password alternatives. Leveraging the single-sign-on passwords used by 25K faculty, staff, and students at CMU, Mazurek et al. measured the password guessability of university passwords [16]. In [3], Li et al. conduct an empirical analysis of Chinese web passwords. In [4], Bonneau analyzed of 70M Yahoo! passwords.

In [28], Bonneau and Schechter challenged the conventional wisdom that users cannot remember cryptographically-strong secrets. In [29], Chiasson et al. presented a usability study of two recent password managers *PwdHash* and *Password Multiplier*.

III. Datasets

In this section, we briefly describe the leaked password datasets used for our evaluations. Table I presents the 4 datasets used which consists in total of about 60.5M real-world

TABLE I
DATASET SUMMARY.

name	size	unique	username	email	language	website	type
CSDN	6.4M	4M	✓	✓	Chinese	www.csdn.net	programmer
Duduniu	16.1M	10M		✓	Chinese	www.duduniu.cn	Internet Cafe
LinkedIn	5.4M	4.9M			English	www.linkedin.com	social networks
Rockyou	32.6M	14.3M			English	www.rockyou.com	game

passwords and covers several forms of web applications. The datasets were leaked due to various password leakage incidents [3], [5], [25]. In Table I, CSDN is a resource sharing website for programmers; Duduniu is a website of Internet cafe service softwares; LinkedIn is a social networking service; and Rockyou is a popular gaming information website. According to [3], [5], [25], most users of CSDN and Duduniu are Chinese speaking users, and most users of LinkedIn and Rockyou are English speaking users. Furthermore, from Table I, we can also see that some datasets were leaked with usernames and/or emails, e.g., CSDN, Duduniu.

Standard Datasets: For our following quantification and evaluation, in order to guarantee *fairness* and to reduce possible *bias* caused by dataset size differences, we randomly and uniformly sample 3 million *unique* passwords as a *standard dataset* from each original dataset. Consequently, we obtain 4 standard datasets: CSDN, Duduniu, LinkedIn, and Rockyou. In the rest of this paper we use the standard datasets for our evaluations, unless specified otherwise.

Ethical Discussion: Note that all the datasets in Table I are now publicly available. Further, these datasets have been extensively used for multi-purpose and meaningful academic research [3], [5], [11]–[13], [25], [30]. Although these real world passwords provide valuable resources to researchers, they were initially leaked illegally. Therefore, in this paper, we only use these data for research purposes. Rather than causing additional harm, our research is expected to be helpful to the community by promoting security awareness of passwords.

IV. PASSWORD CORRELATION: QUANTIFICATION AND EVALUATION

As observed in [3]–[5], [7], [16], passwords exhibit demographic, behavioral, cultural, lingual, and regional correlations. However, all the existing works only show such correlation by experiments, e.g., the CSDN-trained crackers are more effective than Rockyou-trained crackers when cracking Duduniu, since both CSDN and Duduniu are Chinese password datasets while Rockyou is an English password dataset, and thus CSDN and Duduniu are more correlated in terms of language, culture, and behavior. To date, *how to quantify the correlation of two password datasets is still an open problem*. We address this open problem by proposing a *password correlation quantification* framework, which can quantify the correlation of two password datasets from multiple perspectives. Leveraging 10 *correlation quantification functions* developed in this section, we can conduct *structure-based* correlation quantification, *n-gram-based* correlation quantification, and *dictionary-based* correlation quantificationfor two given password datasets.

A. Structure-based Correlation Quantification

Inspired by structure-based password cracking algorithms like PCFG [12] and VCT [13], we first quantify the correlation of two password datasets based on their password structural similarity. Theoretically, one dataset will be more vulnerable if it is cracked by a structure-based cracking algorithm that is trained with another structurally similar dataset.

For any password, according to the techniques in [12], it can be assigned a structure, e.g., '*$$password123*' has a structure of $S_2L_8D_3$ (where L, S, and D represent letters, symbols, and digits respectively). Let U be the possible password space. Then, given two password datasets $V, W \subseteq U$, we use $\mathcal{S}^v$ and $\mathcal{S}^w$ to denote the sets of *password structures* obtained from passwords in V and W, respectively, i.e. $\mathcal{S}^v = \{s^v | s^v$ is a password structure that appeared in $V\}$ and $\mathcal{S}^w = \{s^w | s^w$ is a password structure that appeared in $W\}$. Let $\mathcal{S} = \mathcal{S}^v \cup \mathcal{S}^w$, be the union of password structures of V and W, and $\Gamma = |\mathcal{S}|$. Furthermore, for $s_i \in \mathcal{S}$ $(i \in [1, \Gamma])$, we define two functions $f_s^v(s_i)$ and $f_s^w(s_i)$ which are the *appearance frequencies* of structure s_i in V and W, respectively, i.e., the fraction of passwords having structure s_i in V and W, respectively. Based on $f_s^v(\cdot)$ and $f_s^w(\cdot)$, we define two vectors $\mathbf{V}_s^v =< f_s^v(s_i) >$ and $\mathbf{V}_s^w =< f_s^w(s_i) >$ where $i = 1, 2, \cdots, \Gamma$.

Now, we are ready to quantify the structural correlation of V and W by measuring their structural similarity. Mathematically, the *Jaccard index* (a.k.a. *Jaccard similarity coefficient*) and *cosine similarity* are two efficient means to measure the *element-wise similarity* (i.e., how many common elements) and *distribution similarity* (i.e., how similar the elements are distributed) of two sets, respectively. Therefore, we extend the traditional Jaccard index and cosine similarity to measure the structural correlation of V and W with respect to element similarity and distribution similarity. Formally, the Jaccard index-based structural correlation of V and W is quantified as[1]

$$\psi_s^J(V, W) = \frac{\sum_{i=1}^{\Gamma} \min\{f_s^v(s_i), f_s^w(s_i) | s_i \in \mathcal{S}\}}{\sum_{i=1}^{\Gamma} \max\{f_s^v(s_i), f_s^w(s_i) | s_i \in \mathcal{S}\}}, \quad (1)$$

and the cosine similarity-based structural correlation of V and W is quantified as

$$\psi_s^c(V, W) = \frac{\mathbf{V}_s^v \bullet \mathbf{V}_s^w}{\| \mathbf{V}_s^v \| \times \| \mathbf{V}_s^w \|}. \quad (2)$$

[1]Note that, it is possible to design some alternative Jaccard index-based correlation metric, e.g., the weighted Jaccard index. Also, it is possible to employ other metrics, e.g., the Chi-squared test, to quantify the element similarity. Here, the defined Jaccard index based correlation metric is useful for our purpose. We take the research of further improving this metric as one of our future research directions.

From the above quantification, ψ_s^J indicates how many common password structures are shared by two datasets, while ψ_s^c indicates how similar the structure distributions of two datasets are.

B. n-gram-based Correlation Quantification

Inspired by Markov model based cracking algorithms, e.g., NS [9], OMEN [11], n-gram [5], [10], we now quantify the correlation of V and W based on n-grams. Let l^v and l^w be the *maximum lengths* of the passwords in V and W, respectively. Then, we partition all the passwords in V into n-grams ($n = 1, 2, \cdots, l^v$), denoted by $\mathcal{G}^v = \{g^v | g^v$ is a gram of some password in $V\}$. Similarly, we partition all the passwords in W into n-grams ($n = 1, 2, \cdots, l^w$) and denote the set of the grams as $\mathcal{G}^w = \{g^w | g^w$ is a gram of some password in $W\}$. Let $\mathcal{G} = \mathcal{G}^v \cup \mathcal{G}^w$ be all the possible grams of the passwords in V and W, and $\Lambda = |\mathcal{G}|$. For $g_i \in \mathcal{G}$ ($i \in [1, \Lambda]$), we define two functions $f_g^v(g_i)$ and $f_g^w(g_i)$ to indicate the *appearance frequencies* of gram g_i in V and W, respectively. Then, we can define two vectors $\mathbf{V}_g^v =< f_g^v(g_i) >$ and $\mathbf{V}_g^w =< f_g^w(g_i) >$, where $i = 1, 2, \cdots, \Lambda$.

Similar to the structure-based correlation quantification, we quantify the n-gram-based correlation of V and W by measuring their n-gram similarity using the Jaccard index and n-gram distribution similarity using cosine similarity. Formally, the Jaccard index-based n-gram correlation can be quantified as

$$\psi_g^J(V, W) = \frac{\sum_{i=1}^{\Lambda} \min\{f_g^v(g_i), f_g^w(g_i) | g_i \in \mathcal{G}\}}{\sum_{i=1}^{\Lambda} \max\{f_g^v(g_i), f_g^w(g_i) | g_i \in \mathcal{G}\}}, \quad (3)$$

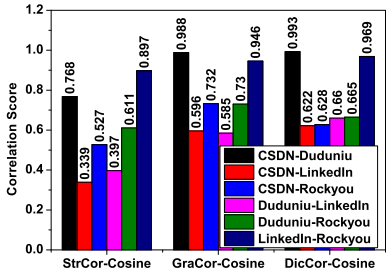
and the cosine similarity-based n-gram correlation can be quantified as

$$\psi_g^c(V, W) = \frac{\mathbf{V}_g^v \bullet \mathbf{V}_g^w}{\| \mathbf{V}_g^v \| \times \| \mathbf{V}_g^w \|}. \quad (4)$$

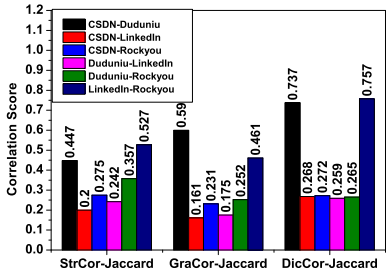
From the above quantification, ψ_g^J indicates the common n-grams shared by the passwords in V and W, while ψ_g^c indicates the distribution similarity of the n-grams in the passwords of V and W. Theoretically, if these two values are high, the corresponding two datasets are more similar with respect to n-grams, and thus the Markov model based cracking algorithms will be more effective in cracking one dataset when it is trained by the other dataset.

C. Dictionary-based Correlation Quantification

Based on the password cracking results in [3]–[5], the employed dictionaries can significantly affect the performance of password cracking algorithms in practice. To understand the fundamental reason for this fact, we quantify the correlation of password datasets with respect to dictionaries. Theoretically, if V and W are highly correlated with respect to a dictionary $\mathbb{D}$, then V is more vulnerable to password cracking algorithms that are trained by W and generate guesses leveraging $\mathbb{D}$, e.g., PCFG [12], VCT [13]. Therefore, we propose to conduct dictionary-based correlation quantification of two password datasets.



(a) Cosine similarity-based



(b) Jaccard index-based

Fig. 1. Password correlation quantification.

Now, given two datasets V and W, and a dictionary $\mathbb{D}$, we can segment the passwords in V and W in terms of $\mathbb{D}$ using the *natural language processing* based password segmentation method proposed by Veras et al. in [13]. We denote the password segmentation results of V and W by $\mathcal{D}^v = \{d^v | d^v$ is a segment of passwords in V with respect to $\mathbb{D}\}$ and $\mathcal{D}^w = \{d^w | d^w$ is a segment of passwords in W with respect to $\mathbb{D}\}$, respectively. Let $\mathcal{D} = \mathcal{D}^v \cup \mathcal{D}^w$ and $\Pi = |\mathcal{D}|$. Similarly, for $d_i \in \mathcal{D}$ ($i \in [1, \Pi]$), we define two functions $f_d^v(d_i)$ and $f_d^w(d_i)$ to indicate the *appearance frequencies* of segment d_i in V and W, respectively. Then, we define two vectors $\mathbf{V}_d^v =< f_d^v(d_i) >$ and $\mathbf{V}_d^w =< f_d^w(d_i) >$ where $i = 1, 2, \cdots, \Pi$.

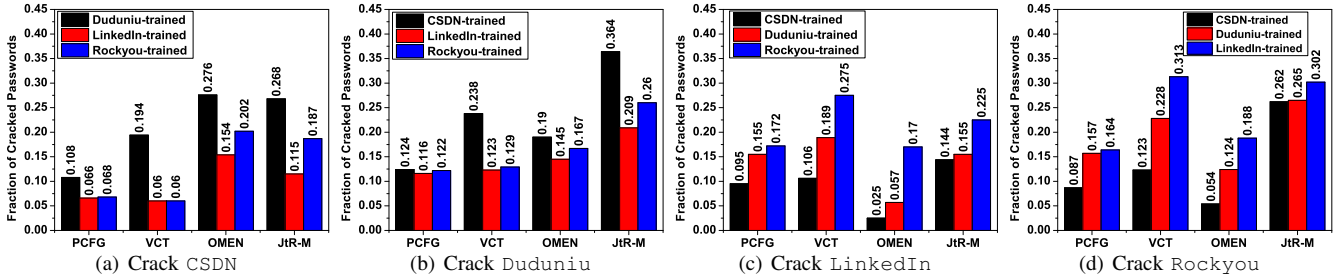
Similar to the previous quantification techniques, we can quantify the dictionary-based correlation of V and W by measuring the *element-wise similarity* and *element distribution similarity* of their dictionary-based segmentation results using the Jaccard index and cosine similarity, respectively. Formally, the element similarity of the dictionary-based correlation is quantified as

$$\psi_d^J(V, W) = \frac{\sum_{i=1}^{\Pi} \min\{f_d^v(d_i), f_d^w(d_i) | d_i \in \mathcal{D}\}}{\sum_{i=1}^{\Pi} \max\{f_d^v(d_i), f_d^w(d_i) | d_i \in \mathcal{D}\}}, \quad (5)$$

and the element distribution similarity of the dictionary-based correlation is quantified as

$$\psi_d^c(V, W) = \frac{\mathbf{V}_d^v \bullet \mathbf{V}_d^w}{\| \mathbf{V}_d^v \| \times \| \mathbf{V}_d^w \|}. \quad (6)$$

Here, ψ_d^J measures the common segments shared by two datasets and ψ_d^c indicates how similar the two segment distributions are.



(a) Crack CSDN (b) Crack Duduniu (c) Crack LinkedIn (d) Crack Rockyou

Fig. 2. Password correlation evaluation: from the attacker's perspective.

D. Evaluation

1) Structure/n-gram/Dictionary-based Correlation Evaluation: First, we quantitatively examine the structure, n-gram, and dictionary base correlation of the 4 standard datasets. When conducting the dictionary-based correlation evaluation, we employ the combination of the widely used *Dic-029* [12], [13] and *Pinyin* [3]. The results are shown in Fig.1, where StrCor, GraCor, and DicCor represent the structure, n-gram, and dictionary based correlation of two datasets, respectively. For instance, the n-gram correlation score of LinkedIn and Rockyou is 0.946 with respect to cosine similarity and is 0.461 with respect to Jaccard index. From Fig.1, we observe that:

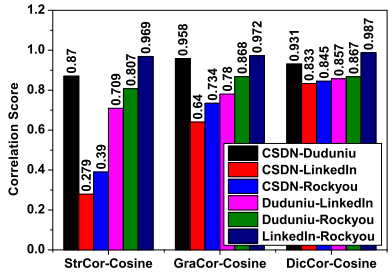
(i) In all the correlation quantification scenarios, CSDN and Duduniu have higher correlation scores with each other than with LinkedIn or Rockyou, while LinkedIn and Rockyou have higher correlation scores with each other than with CSDN or Duduniu. In other words, CSDN and Duduniu are more correlated with each other while LinkedIn and Rockyou are more correlated with each other. This is because most of the users of CSDN and Duduniu are Chinese-speaking users while most of the users of LinkedIn and Rockyou are English-speaking users, and thus CSDN and Duduniu (similarly, LinkedIn and Rockyou) are more demographically, behaviorally, and lingually similar. Our quantification results are consistent with the password correlation observations in [3]–[5], [7], [16]. Therefore, our correlation quantification provides the *theoretical foundation* of the empirical observations found in [3]–[5], [7], [16] and enables quantitative measurement of password dataset correlation.

(ii) Rockyou has higher correlation scores with CSDN and Duduniu than what LinkedIn has with CSDN and Duduniu in most of the quantification scenarios. This implies that Rockyou is more similar with CSDN and Duduniu with respect to password structure, n-gram, and dictionary compared to LinkedIn's similarity with CSDN and Duduniu. According to this fact, theoretically, we can conclude that the leakage of Rockyou will cause a greater threat to CSDN and Duduniu than LinkedIn and vice versa, i.e., Rockyou-trained password crackers will be more powerful than LinkedIn-trained crackers when cracking CSDN and Duduniu. We will validate this assertion in the next section.

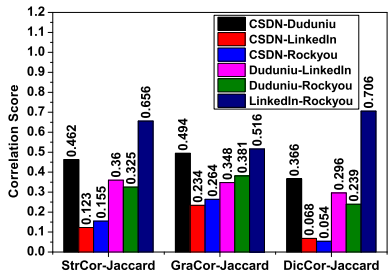
2) Attack-based Evaluation: We further validate our password correlation quantification framework by an *attack-based evaluation*. The methodology is that we first employ a password dataset to train a password cracker; then, we use this password cracker to attack the other three password datasets. Intuitively, *if the testing dataset is more similar with the training dataset with respect to structure, n-gram, and/or dictionary-based segmentation, more passwords of the testing dataset will be cracked* (note that, this is because existing password cracking techniques are implemented based on password structure, n-gram, and/or dictionary based segmentation as discussed in Section II). To conduct a comprehensive attack-based evaluation, we use the latest password cracking algorithms and tools: PCFG [12], VCT [13], OEMN [11], and JtR-M [14], which cover existing structure-based, semantics-based, and Markov model-based password cracking techniques. For each password cracking technique that requires a dictionary input, we use *Dic-029* [12], [13] and *Pinyin* [3]. Furthermore, when cracking a dataset, we limit each trained cracker to generate two billion guesses. Note that, it is possible to generate more guesses. However, two billion is sufficient to validate the accuracy of our correlation quantification framework. We show the password cracking results in Fig.2. Comparing the results of Fig.2 and Fig.1, we have the following observations:

(i) Generally, the quantification results in Fig.1 agree with the cracking results in Fig.2. If two datasets are highly correlated, a cracking algorithm trained by one dataset will be more effective when cracking the other dataset. For instance, CSDN is highly correlated with Duduniu in both cosine similarity-based and Jaccard index-based correlation quantifications. Then, from Fig.2 (a) and (b) we see that CSDN can be more effectively cracked by Duduniu-trained algorithms and vice versa. For instance, when cracking CSDN, using Duduniu-trained PCFG, VCT, OMEN, and JtR-M are more powerful than crackers trained on LinkedIn or Rockyou. Similarly, Rockyou is more correlated with LinkedIn than other datasets. Accordingly, from Fig.2 (c) and (d), LinkedIn is more crackable by Rockyou-trained algorithms and vice versa. This demonstrates that our quantification can accurately characterize the correlation of two password datasets, and is helpful in understanding the security of one password dataset given another dataset from multiple perspectives.

(ii) According to our quantification results in Fig.1, Rockyou is more correlated with CSDN and Duduniu



(a) Cosine similarity-based



(b) Jaccard index-based

Fig. 3. Correlation quantification of the cracked passwords by JtR-I (Table II).

compared to LinkedIn. From Fig.2 (a) and (b), we can see that Rockyou-trained crackers are more powerful than LinkedIn-trained crackers when cracking CSDN and Duduniu. For instance, when attacking Duduniu, Rockyou-trained PCFG, VCT, OMEN, and JtR-M can crack 12.2%, 12.9%, 16.7% and 26% of the passwords, respectively, while the same crackers trained by LinkedIn can crack 11.6%, 12.3%, 14.5%, and 20.9% of the passwords, respectively. This further proves the accuracy of our correlation quantification framework.

3) Correlation Quantification of Hashed Datasets: Our quantification technique is also helpful in analyzing the security of hashed password datasets. Specifically, for future attack-based password security evaluation, our quantification is meaningful in guiding the selection of proper training data and cracking algorithms. When multiple datasets are available for training, password cracking is more effective if the training dataset is highly correlated with the target dataset. For instance, if two datasets are highly structurally correlated, a structure-based cracking algorithm trained by one dataset is likely to be more powerful when cracking the other dataset. So far, based on our discussion, the availability of plain-text passwords is necessary to conduct structure-based, n-gram-based, and dictionary-based password correlation quantification. Then, a natural question is – *when the password datasets are hashed, how can system administrators evaluate the threat of leaked datasets on their (hashed) password dataset through password correlation quantification?*

To address this issue and to quantify the correlation of two hashed password datasets, we employ an *attack-based* approach: we use JtR-I (a training-free and intelligent *brute-force* password cracking mode of JtR [14]) to attack the two hashed password datasets and obtain a small portion of cracked passwords; we then quantify the correlation on the cracked passwords[2].

To validate this *attack-based* approach, we first use JtR-I to attack CSDN, Duduniu, LinkedIn, and Rockyou, and the percentages of cracked passwords are shown in Table II (the number of guesses is limited to two billion). From Table II, we see that a considerable portion of each dataset can be cracked by JtR-I. If we compare the results from Table II with those in Fig.2, we find that training-based password crackers are more powerful than the training-free JtR-I in most scenarios, especially when the training-based crackers are properly trained. This further demonstrates the importance of understanding the correlation between passwords. Specifically, *system administrators can evaluate the threat caused by other password leakage incidents on their password datasets by quantifying the correlation of their datasets with leaked datasets* (which in practice may be utilized by adversaries to train a password cracker).

TABLE II
CRACKING RESULTS OF JTR-I.

	CSDN	Duduniu	LinkedIn	Rockyou
JtR-I	6.3%	11.7%	17.0%	24.9%

According to the cracked passwords in Table II, we can quantify the structure-based, n-gram-based, and dictionary-based correlation of the 4 password datasets and the quantitative results are shown in Fig.3, where (a) and (b) show the *cosine similarity-based* and *Jaccard index-based* correlation, respectively. From Fig.3, we see that the 4 password datasets exhibit similar correlation distribution to that in Fig.1 even when we only consider the cracked passwords by JtR-I, i.e., we use a small portion of plain-text passwords. This implies that *our quantification technique is stable and accurate even if the underlying database is small*, and thus our quantification technique in general is applicable in practical scenarios. Therefore, when we try to quantify the correlation of two hashed password datasets (or the correlation between one hashed password dataset and one plain-text password dataset), we can employ the training-free JtR-I to crack a small portion of the passwords first and then quantify their correlation using the cracked passwords.

V. PASSWORD–PROFILE CORRELATION

A. Status Quo and Preliminary Analysis

In Table I, several datasets were leaked with potentially useful auxiliary information. Specifically, CSDN was leaked with corresponding username and email information and Duduniu was leaked along with corresponding email information. Therefore, we take the username and email as auxiliary information and employ JtR-S to crack CSDN and Duduniu.

[2]System administrator can use expired user-account passwords instead of active user passwords for this purpose.

Surprisingly, we find that 17.2% of CSDN passwords can be cracked within 1130 guesses based on the associated email information, 7.4% of CSDN passwords can be cracked within 789 guesses based on the associated username information, and 33.2% of Duduniu passwords can be cracked within 706 guesses based on the associated email information.

On the other hand, according to our summarization in Section II, no existing password strength meter comprehensively considers a user's social profile, e.g., username and email, when measuring the password strength. Although some commercial password meters/checkers (14 out of the top 150 ranked sites by http://www.alexa.com/) do reject a password that is exactly the same as the username or that contains the username as a substring, they do little to prevent users from selecting social profile-related passwords. This is because they can be easily bypassed with a password that is constructed by appending a number to the username or changing one character in the username.

Furthermore, instead of being related to usernames/emails, user-chosen passwords may be related to other forms of social profiles, which cannot be detected by these meters. As demonstrated in a recent study [11], the password cracking process can be accelerated and improved by leveraging user profile information (about 5% - 30% more passwords can be cracked leveraging users' social profile information). Therefore, to protect user-chosen passwords and to understand the security of passwords given users' social profiles, it is important to use *password–profile* correlation quantification techniques to develop Social Profile-aware Strength Meters (SPSMs) in the near future.

Leveraging our *password–profile* correlation quantification technique, system administrators and users can understand the impacts of social profiles (which are widely and easily available by crawling online social networks, data mining applications, etc. [11]) on password security. Furthermore, by developing a social profile-aware password strength meter, the password–profile correlation score (an indicator of the threat posed by a user's profile) can be provided to users in *real-time* during the registration process, which can help them choose more secure passwords.

B. Social Profile-aware Password Meter

1) Challenges and Solutions: When leveraging social profile information to improve password cracking, an algorithm such as JtR-S, OMEN+, actually exploits the structure, n-gram, and/or dictionary based similarity between a user's social profile and the user's chosen password. Therefore, inspired by the design idea of existing attack-based academic meters, we develop a SPSM, named *SocialShield*, that quantifies the correlation between a user-chosen password and that user's social profile. If a user-chosen password is highly correlated with his/her profile (e.g., username, email, education, address, and other social information), a high *Social Profile Correlation Score* (SPCS) (with value in [0, 1]) will be assigned to that password, implying that the password is more vulnerable to a social profile-aware cracking algorithm. However, we have two challenges in implementing SocialShield: First, the base data is small. The available data consists of one password and a limited amount of social information, e.g., username, email. Therefore, it is a challenge to accurately measure the correlation between the password and the corresponding social profile. Second, even if we have enough base data, how should the correlation/similarity be quantified.

To address the first challenge, we employ *mangling* and *transformation* techniques, which is inspired by existing password cracking schemes. In existing cracking schemes, password guesses are generated through mangling rules and transformation rules [14], [24], [25], e.g., "*password*" may be mangled/transformed to "*pa$$word*", "*password1*", "*Password*", "*1password*", "*pass_word*", "*Tompassword*", etc. Therefore, we take a similar idea to pre-process the password and the corresponding social profile to enlarge the base data for correlation quantification. Let ζ be the input password and Ω be the set of available social profile information. To address the first challenge, we apply the mangling and transformation rules proposed in [24], [25] to both ζ and each element in Ω to generate two sets of base data, denoted by $\mathcal{P} = \{\pi | \pi$ is a transformed/mangled data item generated from $\zeta\}$ and $\mathcal{S} = \{\xi | \xi$ is a transformed/mangled data item generated from one item in $\Omega\}$, respectively. Then, we define the SPCS of ζ and Ω as the correlation between $\mathcal{P}$ and $\mathcal{S}$.

Since we have two sets of based data, to address the second challenge, we can employ our password correlation quantification framework described in Section IV to measure the *structure-based*, n*-gram-based*, and/or *dictionary-based* correlation between $\mathcal{P}$ and $\mathcal{S}$ (assuming $\mathcal{P}$ and $\mathcal{S}$ are two password datasets).

2) SocialShield Design: SocialShield's design is shown in Fig.4, which consists of four steps: (1) a user enters his/her profile information (such as email and username) and the chooses a password; (2) the password-base dataset and the profile-base dataset are generated from the user-input password and profile information (information entered during the registration process) by applying mangling and transformation rules from [24], [25]; (3) the password–profile correlation is quantified by applying our quantification techniques proposed in Section IV; and (4) the generated password–profile correlation score (SPCS) is provided to the user.

Note that, after applying the mangling and transformation rules, the password/profile based dataset usually consists of tens or hundreds of data items according to our experiments (more details in next section). Therefore, SocialShield can quantify the password–profile correlation and provide users with the SPCS in real-time. Furthermore, considering that existing academic/commercial password strength meters do not take into account the security impacts of social profiles, they cannot effectively defend against emerging social profile-based password cracking techniques (as we analyzed in Section V-A). Therefore, as the first countermeasure of its sort, SocialShield can serve as a light-weight *add-on* to existing academic/commercial password strength meters that can defend against emerging social profile-based password

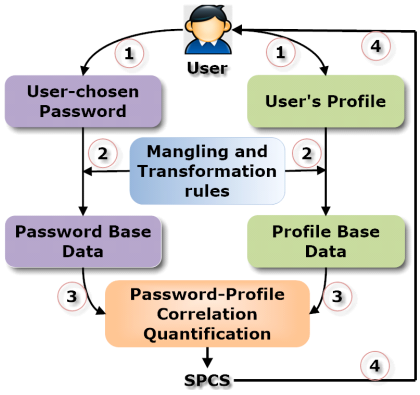


Fig. 4. Architecture of SocialShield.

attacks as shown in the following subsection.

C. Evaluation

In this subsection, we examine the effectiveness of SocialShield in terms of an *attack-based evaluation*. As we demonstrated before, a significant number of passwords in CSDN and Duduniu can be cracked by JtR-S based on the associated email/username information. Based on this ground truth, we evaluate the performance of SocialShield. Here, for the correlation quantification phase of SocialShield (step 3 in Fig.4), we employ the n-gram-based correlation quantification framework (from Section IV) as an example to compute the SPCS of $\mathcal{P}$ and $\mathcal{S}$.

First, we use SocialShield to measure the SPCS of each password in CSDN and Duduniu with respect to the associated email/username. The distribution of the results is shown in Fig.5 (a), where we equally partition the entire correlation score domain [0, 1] into ten windows (each window has a length of 0.1), i.e., for each W_i $(0 \leq i \leq 9)$, it corresponds to the range $[\frac{i}{10}, \frac{i}{10} + 0.1]$. We further employ JtR-S to crack the passwords in each window based on their email/username information. The fraction of passwords that are cracked in each window is shown in Fig.5 (b).

From Fig.5 (a), most passwords have their SPCSs in W_0 (i.e., [0, 0.1]) (88.72% of CSDN passwords with respect to the email information, 85.47% of CSDN passwords with respect to the username information, and 76.45% of Duduniu passwords with respect to the email information), which implies that they are minimally correlated with their corresponding emails/usernames. Compared to CSDN, Duduniu has more email-aware SPCSs in W_9 (10.61% of all the Duduniu passwords). Therefore, compared to CSDN, the passwords of Duduniu are more correlated with their email information. Further, for the passwords of CSDN in W_9, they are more correlated with their corresponding usernames (3.2%) than with their corresponding emails (1.8%).

From Fig.5 (b), (1) with the increase of SPCS, more passwords become crackable based on their associated social information. For instance, in W_9, 75.2% to 79.2% of CSDN passwords and 92.8% of Duduniu passwords can be cracked using the associated email/username information. On the other hand, in W_0, only 0.1% to 0.3% of CSDN passwords and 0.7% of Duduniu passwords can be cracked (for the cracked passwords in W_0, although they are less correlated with the associated profile information, they are more likely to be weak passwords like "123456", "password123" and hence easily guessable). Therefore, if the passwords are labeled as more correlated with their social profiles by SocialShield, they are more likely to be crackable; and (2) Based on the results in Fig.5 (a), passwords in Duduniu are more correlated with its email information, and thus it is more crackable based on the email information as shown in Fig.5 (b). Furthermore, the CSDN passwords are more guessable based on the username information in the windows where they are more correlated with their username (e.g., W_9). Therefore, the *attack-based evaluation results validate that SocialShield is accurate and effective in measuring the strength of passwords with respect to their social profiles*.

In summary, although SocialShield is a light-weight implementation, it is very accurate and effective in measuring the strength of passwords given users' profile information. Therefore, SocialShield can serve as an *add-on* to existing academic and commercial meters. Furthermore, the design of SocialShield can shed light on developing powerful social profile-aware password strength meters in the future.

VI. Limitation and Future Work

A. Limitation

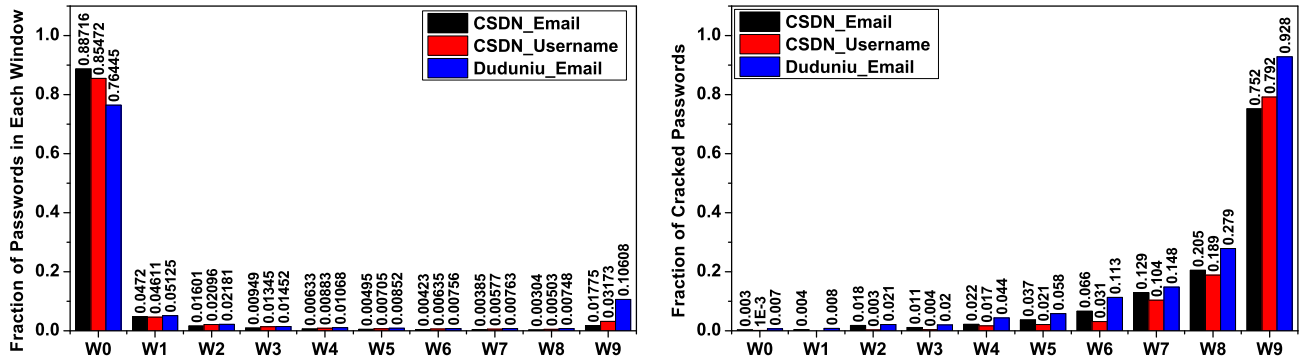
When evaluating SocialShield, the employed profile information is username and email address. Although SocialShield is capable of quantifying the correlation between a user-chosen password and that user's other profile information such as education, company, phone number and address, we did not conduct such evaluation due to the lack of real-word data. It would be possible for us to crawl users' profiles online using the email information in CSDN and Duduniu. However, that will raise legal concerns.

B. Future Work

The future research work of this paper includes: (1) In our password–password and password–profile correlation quantification, we did not take into account the semantic information carried by passwords/profiles. In the future, we will improve the quantification accuracy by incorporating the semantic information of passwords/profiles. (2) To further evaluate the performance of SocialShield, we will try to collect more passwords along with social information. Furthermore, it would be meaningful to conduct a user study on the performance and usability of SocialShield. (3) We will implement SocialShield in a real password system (e.g., our university's password system) and evaluate its practical performance.

VII. Conclusion

In this paper, (1) we propose the first *password–password* correlation quantification framework, which enabled us to quantify the correlation between password datasets in terms



(a) SPCS distribution

(b) Cracked results distribution

Fig. 5. SocialShield evaluation.

of *structure*, *n-gram*, and *dictionary* (words).Theoretically, our quantification results explain the success of existing training-based password cracking techniques. Leveraging an attack-based evaluation, our password–password correlation quantification is demonstrated to be accurate; (2) we propose the first *password–profile* correlation quantification framework, which explains the success of emerging *profile-based password attacks*. Furthermore, based on our quantification, we develop the first social profile-aware password strength meter, namely *SocialShield*. By experiments, we demonstrate that SocialShield is a light-weight, yet effective means to defend users against profile-based password attacks. The developed correlation quantification techniques and SocialShield have meaningful implications to password system administrators, users, and researchers in helping them understand the threats posed by leaked passwords and profile information.

ACKNOWLEDGMENT

This work was partly supported by the Provincial Key Research and Development Program of Zhejiang under No. 2016C01G2010916 and by the CCF-Tencent Open Research Fund under No. CCF-Tecent AGR20160109.

Shouling Ji is the corresponding author of this paper.

REFERENCES

[1] S. Ji, S. Yang, T. Wang, C. Liu, W.-H. Lee, and R. Beyah. Pars: A uniform and open-source password analysis and research system. *ACSAC*, 2015.

[2] S. Yang, S. Ji, X. Hu, and R. Beyah. Effectiveness and soundness of commercial password strength meters. *NDSS poster session*, 2015.

[3] Z. Li, W. Han, and W. Xu. A large-scale empirical analysis on chinese web passwords. *Usenix Security*, 2014.

[4] J. Bonneau. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. *S&P*, 2012.

[5] J. Ma, W. Yang, M. Luo, and N. Li. A study of probilistic password models. *S&P*, 2014.

[6] X. C. Carnavalet and M. Mannan. From very weak to very strong: Analyzing password-strength meters. *NDSS*, 2014.

[7] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. López. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. *S&P*, 2012.

[8] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. 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. *USENIX Security*, 2012.

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

[10] C. Castelluccia, M. Dürmuth, and D. Perito. Adaptive password-strength meters from markov models. *NDSS*, 2012.

[11] M. Dürmuth, A. Chaabane, D. Perito, and C. Castelluccia. When privacy meets security: Leveraging personal information for password cracking. *CoRR abs/1304.6584*, 2013.

[12] M. Weir, S. Aggarwal, B. Medeiros, and B. Glodek. Password cracking using probabilistic context-free grammars. *S&P*, 2009.

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

[14] John the Ripper-bleeding jumbo. https://github.com/magnumripper/johntheripper.

[15] Hashcat v0.47. http://hashcat.net/hashcat/.

[16] 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. *CCS*, 2013.

[17] Gmail password leakage. http://lifehacker.com/5-million-gmail-passwords-leaked-check-yours-now-1632983265.

[18] Yahoo! password leakege. http://www.cnet.com/news/yahoos-password-leak-what-you-need-to-know-faq/.

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

[20] S. Komanduri, R.Shay, L. F. Cranor, C. Herley, and S. Schechte. Telepathwords: Preventing weak passwords by reading users' minds. *USENIX Security*, 2014.

[21] S. Ji, S. Yang, X. Hu, W. Han, Z. Li, and R. Beyah. Zero-sum password cracking game: A large-scale empirical study on the crackability, correlation, and security of passwords. *IEEE Transactions on Dependable and Secure Computing*, 2015.

[22] S. Ji, S. Yang, W. Li, X. Liao, X. Hu, and R. Beyah. Password cracking: A large-scale empirical study. *USENIX Security poster session*, 2014.

[23] S. Ji, S. Yang, T. Wang, C. Liu, W.-H. Lee, and R. Beyah. Pars: A uniform and open-source password analysis and research system. *http://www2.ece.gatech.edu/cap/PARS/*.

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

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

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

[27] J. Bonneau, C. Herley, P. C. Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. *S&P*, 2012.

[28] J. Bonneau and S. Schechter. Towards reliable storage of 56-bit secrets in human memory. *USENIX Security*, 2014.

[29] S. Chiasson, P. C. V. Oorschot, and R. Biddle. A usability study and critique of two password managers. *USENIX Security*, 2006.

[30] M. Dell' Amico, P. Michiardi, and Y. Roudier. Password strength: An empirical analysis. *Infocom*, 2010.