Understanding Passwords of Chinese Users: A Data-Driven Approach

Ding Wang *Student Member, IEEE*, Haibo Cheng, Ping Wang *Senior Member, IEEE*

Abstract—Textual passwords remain the dominant authentication mechanism over the Internet and are likely to persist in the foreseeable future. Much attention has been paid to passwords chosen by English users, yet only a few studies have examined how *non*-English users select passwords. In this paper, we fill the gap by employing both user-survey data and real-life data. We conduct so far *the first user survey* on the password behaviors of Chinese web users, the world's largest Internet population (710 million). Our survey reveals a number of similarities and differences in coping strategies between these two user groups. We also explore some new questions regarding user sentiments and perceptions of security: 45% of Chinese users report that they insert (birth)dates into passwords, 55% believe their current passwords are more secure than previous ones and 69% consider password management as a bother.

We further perform *an extensive, empirical analysis* of 73.1 million Chinese web passwords in a comparison with 33.2 million English counterparts. We identify a number of interesting structural and semantic characteristics in Chinese passwords, confirm many findings from our survey and also examine password security by employing two state-of-the-art cracking techniques. Particularly, our cracking results reveal the bifacial-security nature of Chinese passwords: when the guess number allowed is small, they are much weaker than their English counterparts, yet this relationship will be reversed when the guess number becomes large. This well reconciles two conflicting claims about the strength of Chinese web passwords made by Bonneau in IEEE S&P'12 and Li et al. in Usenix SEC'14, respectively. At 10^7 guesses, our improved PCFG-based attack can crack 33.2%$\sim$49.8% of the Chinese datasets, indicating that our attack can crack 92%$\sim$188% *more* passwords than the best record reported by Li et al. We also discuss some implications of our findings.

Keywords—User authentication, Password security, Semantic pattern, Probabilistic context-free grammar, Markov model.

1 Introduction

Password-based authentication is the dominant form of access control in almost every web service today. Though the security pitfalls of textual passwords were revealed as early as four decades ago [28] and various alternative authentication methods (e.g., biometrics and multi-factor authentication) have been proposed since then, passwords stubbornly survive and proliferate with almost every new web service. Passwords offer many advantages, such as low deployment cost, easy recovery and remarkable simplicity, that cannot always be matched by their alternatives [15]. Users are also in favor of passwords—A 2016 survey [36] on 1,119 US users show that 58% would prefer passwords as online login credentials, while only 16% prefer biometrics, 10% prefer other ways and the remaining 16% prefer not to answer. Thus, passwords are likely to persist in the foreseeable future.

Despite its ubiquity, password authentication is stuck with an inherent tension [43]: truly random passwords are hard for users to memorize, while user-memorable passwords tend to be highly predictable. To eliminate this "security-usability" dilemma, researchers have devoted significant efforts [3], [7], [35] to the following two types of studies. *Type-1* research aims at evaluating the strength of a password dataset (distribution) by gauging its Shannon entropy as in NIST [6], or by measuring its "guessability" [27]. Guessability characterizes the fraction of passwords that, at a given number of guesses, can be cracked by cracking algorithms like the Markov-Chains [26] and probabilistic context-free grammars (PCFG) [38], [42]. As with most of these previous studies, we mainly consider *trawling guessing* [38], while other attacking vectors (e.g., shoulder-surfing, phishing and targeted guessing [40]) are out of our focus. Hereafter whenever the term "guessing" is used, it means trawling guessing.

Type-2 research attempts to reduce the use of weak passwords, and mainly two approaches have been utilized: proactive password checking [22] and password strength meter [38]. The former checks the user-selected passwords and only accepts the ones that comply with the system policy (e.g., at least 8 characters long and contain two three types of characters). The latter is typically a visual feedback of password strength, often presented as a colored bar to help users create stronger passwords [10]. Most of today's leading sites employ a combination of these two approaches to restrict users from choosing weak passwords. In this work, though we mainly focus on type-1 research, one can see that our results are also helpful for type-2 research.

Existing literature mainly focuses on passwords chosen by English users, or more precisely, by netizens using the Latin alphabet, yet little attention has been paid to the characteristics and strength of passwords generated by those who use other native languages. For instance, password "`wanglei123`" is currently deemed "Strong" by password strength meters (PSMs) of many high-profile services like AOL, Google, IEEE, and Sina weibo. However, as shown in [39], this password is highly prone to guessing, for "`wanglei`" is a very popular Chinese Pinyin (i.e., the Chinese phonetic alphabet) name but not a random string of length seven. Failing to catch this would overlook the weaknesses of Chinese passwords, posing the corresponding web accounts at high risks.

1.1 Motivations

There are 710 million Chinese netziens by June, 2016 [1], which account for over 25% (the largest fraction) of the world's Internet population. However, to our knowledge,

• *D. Wang, H. Cheng and P. Wang are with the School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China. Email: {wangdingg, chenghaibo, pwang}@pku.edu.cn*

Manuscript received 16 Nov. 2016; revised .

so far there has been no satisfactory answer to the following important questions: *(i) What are the password management behaviors of Chinese users (e.g., creation, memorization and sentiment)? (ii) Are there structural or semantic characteristics that differentiate Chinese passwords from English ones? (iii) How about the security strength of Chinese passwords? (iv) Are they weaker or stronger than English passwords?* It is of great importance to answer these basic questions to provide both security engineers and Chinese users with necessary security guidance. For instance, if the answer to the second question is affirmative, then it highly indicates that the password policies (see [39]) and strength meters (e.g., [38]) originally designed for English users cannot be readily applied to Chinese users.

A few password studies (e.g., [19], [26], [38], [40]) have employed some Chinese datasets, yet they mainly deal with the effectiveness of various probabilistic cracking models, and relatively little attention is given to the differences between Chinese passwords and English passwords. Besides, most existing works (e.g., [19], [26], [38]) on Chinese passwords mainly employ an empirical approach. Yet, this empirical approach is inherently *unable* to reveal many important user password behaviors when users are confronted with the demanding tasks of keeping track of many accounts, such as what fraction of users write down passwords, how many different passwords users have, to what extent users accept password manager and what are the sentiments (e.g., satisfactory or not) about their own current management of passwords.

As far as we know, Li et al.'s work [24] may be the closest to the current paper, but *our work differs from it in five aspects*. First, we uses not only empirical data but also user survey data, and it enables us to reveal many important Chinese user behaviors in managing passwords. Second, we for the first time explore a number of fundamental characteristics, such as the extent of language dependence, length distribution, frequency distribution and Pinyin-name-based semantic patterns. Third, our improved PCFG-based algorithm can achieve success rates from 29.41% to 39.47% at *just* 10^7 guesses, while the best success rate of their improved PCFG-based cracking algorithm is only 17.3% at 10^{10} guesses (i.e., an underestimation of attackers). Fourth, based on more comprehensive analysis, we report that both Li et al.'s and Bonneau's [4] conclusions about the strength of Chinese password are biased. Last but not the least, we demonstrate that two of Li et al.'s five Chinese datasets are improperly pre-processed when performing data cleansing,[1] and thus the effectiveness of the results reported in [24] might be impaired. For example, Li et al. reported that there are 3,639,517 (11.78%) passwords in Tianya with consecutive exactly eight digits, yet the actual value is 2.4 times larger: 8,664,708 (28.04%).

B. Contributions

To answer the above four fundamental questions, we first conduct an online user survey on the password management behaviors of Chinese users. To *complement* our user survey, we further perform a large-scale empirical analysis by leveraging over 73.1 million real-world passwords from six popular Chinese sites and more than 33.2 million passwords from three English sites, one of the largest corpuses of user-chosen passwords ever studied. Benefiting from the plain-text form of these 106M passwords, we seek for fundamental properties of user-chosen passwords and systematically measure their structural/semantic characteristics and security. In summary, we make the following key contributions:

- **A user survey**. To reveal user behaviors in managing passwords, we carry out a user survey with 442 effective participants. Our survey shows that, among Chinese users, 48.6% have no more than 5 *different* passwords, 74.2% have no more than 5 *different types* of passwords; 47.3% have ever recorded passwords on paper or electronic media, 13.8% have ever used a password manager; 41.6% have developed their own coping strategies and 69.0% feel bothered. We identify a number of similarities and differences between these two user groups, e.g., Chinese users directly reuse passwords less but write more on paper. *This is the first password survey* that targets Chinese users and reveals many user behaviors that are otherwise difficult to be explored by an empirical study.
- **An empirical analysis**. By leveraging 73.1 million real-life Chinese passwords, *we for the first time*: (1) provide a quantitative measurement of to what extent user passwords are influenced by their native language; (2) explore the name semantics in passwords, and show that there is an every one in nine probability that Chinese users insert their Pinyin names into passwords, while the fraction of English passwords that include an English name (with length$\geq$5) reaches 24.3%, i.e., one in four; (3) reveal that every one in six Chinese users inserts a 6-digit date (e.g., birthdate) into their password; and (4) show that passwords chosen by these two distinct user groups are of quite similar Zipf-like frequency distributions.
- **An reversal principle**. We employ two state-of-the-art password-cracking algorithms (i.e., PCFG-based and Markov-based [26]) to measure the strength of Chinese web passwords. Based on the identified characteristics, we improve the PCFG-based algorithm to more accurately capture passwords that are of a monotonically long structure (e.g., "`1qa2ws3ed`"). At 10^7 guesses, our algorithm can crack 92% to 188% more passwords than the results in [24]. Particularly, *we reveal a "reversal principle"*, i.e. the the bifacial-security nature of Chinese passwords: when the guess number allowed is small, they are much weaker than their English counterparts, yet this relationship is reversed when the guess number is large, thereby reconciling the contradictory claims made in [4], [24].
- **Some insights**. We highlight some useful insights for password cracking and creation policies. To our knowledge, we are *the first to provide a large-scale empirically grounded evidence* that supports for the hypothesis proposed in [10], [33]: users rationally choose more secure passwords for accounts associated with higher value, and knowingly select weaker passwords for lower-value web service even though the latter imposes a stricter policy.

[1] We have reported this issue to the authors of [24], they acknowledged this flaw in their journal version [13]. However, as we will show in Sec. IV-B, in [13] they still fail to pre-process (clean) the datasets properly. As their journal version [13] is essentially a verbatim of their conference version [24], in this work we mainly use [24] for comparison and discussion.

II. Related work

In this section, we briefly review prior research on password characteristics and security to facilitate later discussions.

A. Password characteristics

Basic statistics. In 1979, Morris and Thompson [28] analyzed a corpus of 3,000 passwords and reported that 71.12% of passwords are no more than 6 characters long and 13.93% of passwords are non-alphanumeric characters. In 1990, Klein [22] collected 13,797 computer accounts from his friends and acquaintances around US and UK, and observed that users tend to choose passwords that can be easily derived from dictionary words: a dictionary of 62,727 words is able to crack 24.21% of the collected accounts and 51.70% of the cracked passwords are shorter than 6 characters long. In 2004, Yan et al. [43] found that user-chosen passwords are likely to be dictionary words since users have difficulty in memorizing random strings, and that the lengths of passwords in their study (involving 288 participants) are on average between 7 and 8.

In 2012, Bonneau [4] conducted a systematic analysis of 70M Yahoo private passwords. This work examined dozens of subpopulations based on demographic factors (e.g., age, gender and language) and site usage characteristics (e.g., email and retail), and found that "even seemingly distant language communities choose the same weak passwords". This research was recently reproduced in [2] by using differential privacy techniques. Particularly, Chinese passwords are found among the most difficult ones to crack [4]. In 2014, however, Li et al. [24] argued that Bonneau's 70 million Yahoo dataset is not representative of general Chinese users, because most Yahoo users are familiar with English. Accordingly, Li et al. leveraged a corpus of five datasets from Chinese sites and observed that Chinese users like to use digits when creating passwords, as compared to English users who like to use letters to build passwords. However, as an elementary defect, two of their Chinese datasets have not been cleaned properly (see Section IV-B), which might lead to inaccurate measures and biased comparisons. More importantly, several critical password properties (such as the length and frequency distributions of passwords, and semantics) remain to be explored.

In 2014, Ma et al. [26] investigated password characteristics about the length and the structure of six datasets, three of which are from Chinese websites. Nonetheless, this work mainly focuses on the effectiveness of probabilistic password cracking models and pays little attention to the deeper semantic patterns of passwords (e.g., no information is provided about the role of Pinyins, names or dates).

Semantic patterns. In 1989, Riddle et al. [30] found that birth dates, personal names, nicknames and celebrity names are popular in user-chosen passwords. In 2004, Brown et al. [5] confirmed this by conducting a thorough survey that involved 218 participants and 1,783 passwords, and they reported that the most frequent entity in passwords is the self (accounts for 66.5%), followed by relatives (7.0%), lovers and friends; also, names (32.0%) were found to be the most common information used, followed by dates (7.2%). In 2012, Veras et al. [37] examined the 32 million RockYou dataset by employing visualization techniques and observed that 15.26% of passwords contain sequences of 5-8 consecutive digits, 38% of which could be further classified as dates. They also found that repeated days/months and holidays are popular, and when non-digits are paired with dates, they are most commonly single-characters, or names of months.

In 2014, Li et al. [24] showed that Chinese users tend to insert Pinyins and dates into their passwords. However, many other important semantic patterns (e.g., Pinyin name and mobile number) are left unexplored. In addition, due to an imprudent processing of data cleansing (see Sec. IV-B) and improper tuning of cracking algorithms (see Sec. V), Li et al.'s measurement of the strength of Chinese passwords will be shown to be biased. In 2015, Ji et al. [19] noted that user-IDs and emails have a great impact on password security. For instance, 53.24% of Dodonew passwords can be guessed by using user-IDs within, on average, 706 guesses. This motivates us to investigate to what extent the Pinyin names and Chinese-style dates impact the security of Chinese passwords.

B. Password security

A crucial research area in password cryptography is to evaluate password strength. To avoid using the brute-force attack, earlier works (e.g., [22], [30]) use a combination of ad hoc dictionaries and mangling rules to model the common password generation practice and see whether user passwords can be successfully rebuilt in a period of time. This technique greatly improves the cracking efficiency and has given rise to automated tools such as John the Ripper (JTR).

Borrowing the idea of Shannon entropy, the NIST-800-63-2 guide [6] attempts to use the concept of *password entropy* for estimating the strength of password creation policy underlying a password system. Password entropy is calculated mainly according to the length of passwords and augmented with bonus for special checks. Florencio and Herley [11], and Egelman et al. [10] improved this approach by adding the size of the alphabet into the calculation and called the resulting value $log_2((alpha.size)^{pass.len})$ the bit length of a password.

However, previous ad hoc metrics (e.g., password entropy and bit length) have recently been shown far from accurate by Weir et al. [41], and they suggested that the approach based on simulating password cracking sessions is more promising. They also developed a novel method that first automatically derives word-mangling rules from password datasets, and then instantiates the derived grammars by using string segments from external input dictionaries to generate guesses in decreasing probability order [42]. This PCFG-based cracking approach is able to crack 28% to 129% more passwords than JTR when given the same number of guesses. It has been considered as one of the state-of-the-art password cracking techniques and used in a number of recent works [26], [35], [40].

Differing from the PCFG-based approach, Narayanan and Shmatikov [29] introduced the Markov-Chain theory for assigning probabilities to letter segments, and it substantially reduces the password search space. This approach was tested in an experiment against 142 real-life user passwords and was able to break 67.6% of them. In 2014, by using various normalization and smoothing techniques from the natural language processing (NLP) domain, Ma et al. [26] systematically evaluated the Markov-based model and found that, in some cases, it performs significantly better than the PCFG-based model at large guesses (e.g., 2^{30}) when parameterized

appropriately. Based on [26], [42], some useful automated password analysis tools have been developed in [20], [35]. In this work, we will perform extensive experiments by using both attacking models (combined with the identified characteristics of passwords) to evaluate the strength of Chinese passwords.

III. A Survey on Password Management Behaviors of Chinese Web Users

To reveal Chinese user behaviors and sentiment in the process of password management (e.g., creation and memorization), we carried out an anonymized user survey by using Sojump, the most popular Chinese online survey platform. A copy of our questions (translated from Chinese into English for review purpose) can be found at https://sojump.com/jq/7005139.aspx, the third part of which relates to this work. We got approval from our center's IRB. The participants are surveyed anonymously and the results are all aggregated so that no user identifiable info can be inferred. By comparing with existing password surveys on English users (see [8], [14], [32], [33]), we identify a number of *similarities* and *differences* in password behaviors between these two user groups.

A. Survey design

Though our survey is an anonymized one and conducted on a well-known third-party platform, we fear that users may have privacy concerns and refuse to answer or provide untruthful answers, because passwords are highly sensitive. To minimize such potential adverse effects and ensure ecological validity, we optimize the survey questions before the release of the survey by employing 15 students in other teams of our lab as pilot testers. We get feedbacks from them and revise the questions accordingly. This process is performed in an iterative manner. Finally, a few questions (e.g., "Will you reuse an existing email password for this new email account?" and "How many PIN codes do you maintain?") that are inclined to be offensive are abandoned. To avoid user drop out, as with [32], we include an "I prefer not to answer" option for questions that might make participants un-comfortable.

Our survey is expected to take about 22 to 35 minutes to complete. Each member in our team is requested to invite 10 to 20 reliable friends, relatives or colleagues to fill out the questionnaire. If the participants are willing to further disseminate the survey, they are asked to provide us with the number of invitations they send. In all, a total of 983 invitations are sent and we acquire 442 effective responses.

We initiate our survey by asking participants to create a password for an email account that will be frequently used. The first 7 questions explore the transformation rules in password reuse (e.g., "how similar the password created for this new account is similar to your existing passwords?", "what transformation rules will be used?" and "what's their popularity?"), and detailed results are referred to [38]. The next 13 questions are devoted to explore user password behaviors regarding semantics composition, overall-reuse, storage, evolution and sentiment, which are the focuses of this work.

B. Survey results

Our first question aims to figure out what's the semantics of the letters in a user's password created for this new email account. As shown in Fig. 1, the most popular choice is "Chinese Pinyin names" (about 40%)[2], followed by "Chinese Pinyin words" (33.71%). These two figures well accord with what we have obtained from real-life datasets, i.e. $\frac{11.24\%}{22.42\%} \approx 50.13\%$ and $\frac{9.73\%}{22.42\%} \approx 30.02\%$ (see Table IV), respectively. Interestingly, many Chinese users also tend to include English words (24.43%) or their English names (22.17%) into their passwords, while these two figures obtained from the large-scale empirical data (see Table IV) are $10.74\%(=\frac{2.41\%}{22.42\%})$ and $23.86\%(=\frac{5.35\%}{22.42\%})$, respectively. This suggests that users in our survey are two times more likely to use English words than ordinary users do. A plausible reason is that our users are more educated (i.e., 92% have at least a bachelor's degree, see Sec. III-C), and they are more familiar with English than normal user population.

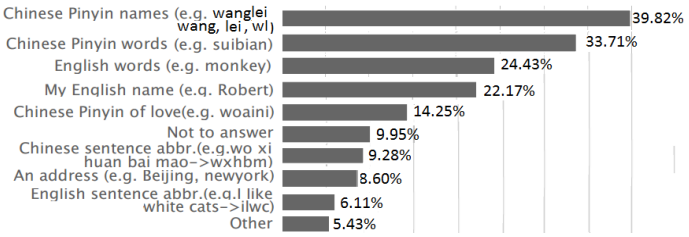


Fig. 1. If you use letters in your password, what are they related to (multiple-choice)? Most users prefer meaningful strings like Pinyin names and words.

It has been shown in [24], [26] that Chinese users like to build passwords with digits. Are there semantics in these digits? Fig. 2 shows that 44.8% of users tend to use dates, which well consists with what we have empirically obtained from real-life datasets, i.e. $\frac{31.60\%}{78.04\%} \approx 40.49\%$ (see Table IV). In addition, 42.3% use phone numbers (either mobile or fixed-line), 16.74% use homophone (e.g., `5201314`) and 3.62% use the time of site registration. What's most interesting is that, 14.03% Chinese users include their PINs of various financial cards into their passwords, and this figure is greatly consistent with Bonneau's PIN survey on English users (i.e., 15%). This, for the first time, provides more than anecdotal evidence of to what extent Chinese users reuse PINs of financial cards in passwords of web accounts. There are also 27.31% of users choose random digit sequences. Further considering that there are about 1.5 times more Chinese users include digits in passwords than English users do (see Table III), this partially explains why Chinese web passwords are more difficult to crack at large guess numbers as compared to English passwords (see the "reversal principle" in Sec. V). Besides dates, the other six semantics are inherently difficult to be captured from empirical datasets, *highlighting the superior aspect of a user survey study over an empirical analysis*.

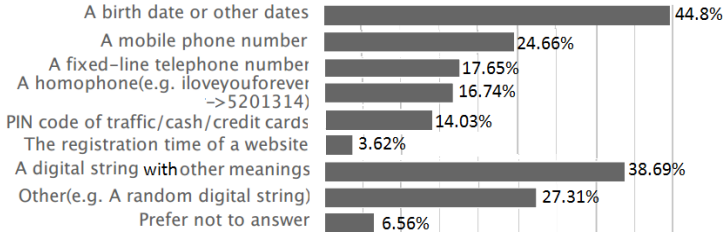


Fig. 2. If you use digits in your password, what are they related to (multiple-choice)? Dates and phone numbers are most prevalent, PINs also show up.

[2]Chinese Pinyins are made of 26 Latin characters in the Chinese phonetic alphabet, and they are used to latinize the Chinese hieroglyphic characters. For instance, Chinese Pinyin "woaini" means "iloveyou".

Nowadays the number of web accounts we need to maintain constantly increases, yet human working memory is limited and stable [5]. Thus, users cope by reusing passwords [8], [38]. It is interesting to find out *how many unique passwords* and *how many different types of passwords* they maintain. Our survey shows that 81.22% of users (see Fig. 3) maintain no more than 10 unique passwords. This suggests that frequent password expiration not only is ineffective [45], but also would potentially increase user fatigue/frustration. In comparison, only 20.59% of Chinese users hold fewer than 4 unique passwords for all their web services, while this figure for English users is as high as 46% [8]; only 48.64% of Chinese users hold less than 6 distinct passwords, while this figure for English users is 72% [8]. This suggests that Chinese users reuse passwords less.

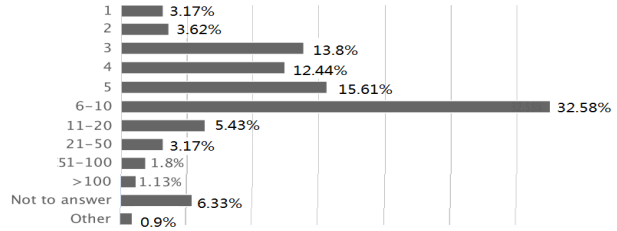


Fig. 3. How many different passwords do you maintain? Note that `password1` and `password2` shall be seen as different ones. Over 80% of Chinese users are with 10 or fewer unique passwords.

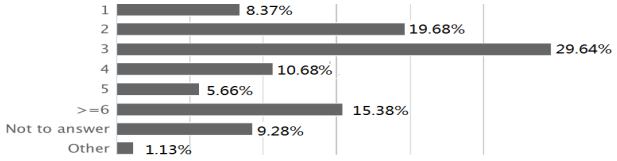


Fig. 4. How many different types of passwords do you maintain? Note that similar passwords fall into the same type. The majority have $\leq$3 types.

As our participants are more diversified than that of [8] and our participant number is also two times higher, the observation that Chinese users reuse passwords less is unlikely due to survey bias. A plausible reason may be that, a large number of breached Chinese sites (e.g., over 100 popular sites had leaked passwords during the past few years [44]) offer security advices and request users not to reuse passwords. This climax came when China Central Television screened some "best password tips" at peak time in Dec. 2014 [25]. Such advices may increase users' security consciousness and nudge users towards less *direct* password reuse, yet the rate of *indirect* password reuse is still staggering: 57.69% and 74.05% of Chinese users (see Fig. 4) keep no more than 3 and 5 different types of passwords, respectively. In average, each Chinese user keeps 9.74 unique passwords and 3.15 different types of passwords. This is roughly comparable to English users (e.g. 5 unique passwords in [33] and 6.5 unique passwords in [11]).

Fig. 5 illustrates whether there are relationships between users' different types of passwords. As expected, the majority of users (60.63%) acknowledge some relationships. In the remains, 21.72% show very good practice, while 17.65% are at high risks because all their passwords are created from a single basic password. We further examine whether Chinese users rationally use different types of passwords for different types of online accounts as recommended [25]. Fig. 6 shows that, 28.05% conform to the good practice, about half of users report they only sometimes conform to the good practice, while 15.84% acknowledge a "definitely no". This is comparable to Haque et al.'s survey on 80 English students [14]: 19.1% of users reuse (or simple modify) an existing password from a low-value account for a higher-value account.

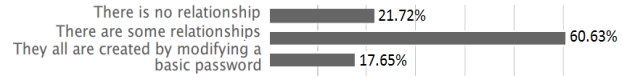


Fig. 5. Is there any relationship among your different types of passwords? 80% of users admit there are at least some relationships.

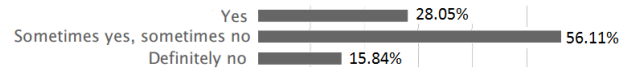


Fig. 6. Do you use different types of passwords for different types of web services? Less than one third say definitely yes.

The next two questions relate to user behaviors in password storage. Regarding how users saving their passwords, Fig. 7 shows that the most popular way is saving passwords in mind (54.3%), successively followed by writing on paper (25.57%), recording electronically (21.72%), saving in browser (12.9%) and using password manager (7.24%). Komanduri et al. [23] found that in their email survey scenario (1000 English participants), about 19% store passwords on paper and 25% store electronically. In comparison, the difference in writing on paper between their and our user group is significant (two-proportion Z-test, p-value $= 0.0048 < 0.05$), while the difference in recording electronically is not significant (p-value$= 0.18024 > 0.05$).

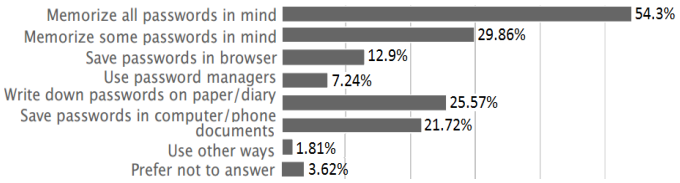


Fig. 7. How do you save your passwords (multiple-choice)? Over half of users memorize *all* passwords in mind, while 30% memorize *some* in mind.

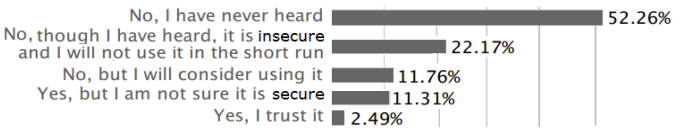


Fig. 8. Have you ever used password managers (e.g. KeePass and LastPass)? Less than 14% have used them, and many users have security concerns.

While password managers are deemed promising tools for alleviating the password management burden on users by security researchers [17], it remains a question of to what extent they have been accepted by the normal users in reality. Fig. 8 shows that, somewhat surprisingly, over 50% of Chinese users even have not heard about password managers. While 22.17% know password managers, yet they will not use them because of security concerns. Only 13.8% have used them. What's most staggering is that, *only 2.49% of Chinese users have used password managers and also trust their security*. Further considering that our participants are more educated and younger than normal Chinese users (see Sec. III-C), the acceptance of password managers in China will be even much lower (probably less than 1%). In a survey on English users, Stobert and Biddle were also "surprised to find that none of our participants used a dedicated password manager" [33]. All this suggests that there is still a long way to go before password managers can be widely accepted, especially when considering users' serious security concerns.

The next three questions relate to how users' passwords evolve as time goes on. As shown in Fig. 9, 43.22% of users acknowledge little or no difference, while 54.53% show there are tangible differences. The latter well accords with the statistics shown in Fig. 10 that, 58.6% of users acknowledge a tangible strength increase in their current passwords over previous ones. Disturbingly, 28.96% admit no strength increase, while the password cracking techniques (e.g., [26], [40]) are evolving rapidly as time goes by. Fig. 11 shows how Chinese users cope with diversified password requirements from various sites: 47.74% of users cope in a temporary manner, while 41.63% have developed their own coping method as time goes by.

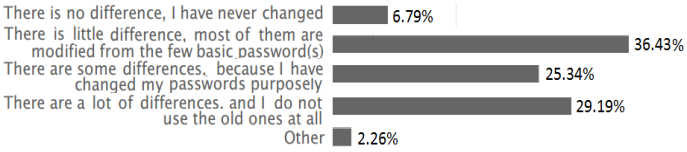


Fig. 9. Is there any difference between the passwords you have used before and the ones currently in use? 54.53% show there are tangible differences.

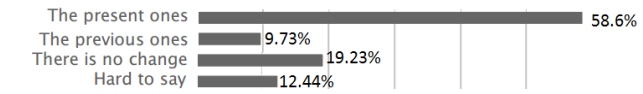


Fig. 10. Which are more secure? Your previous passwords or your present passwords? 58.6% acknowledge a tangible strength increase in present ones.

The large fraction of users coping with password requirements in a passive manner partially explains why 69.01% of users (see Fig. 12) feel bothered (upset/very upset) in their management of passwords. Shay et al.'s study on a password policy shift in Carnegie Mellon University found that, the shift to a more complex one annoyed 61.91% of users [32]. Interestingly, there are 15.38% of Chinese users show a casual emotion and in the meantime, there are also 11.31% show a positive emotion. This is mainly because there are some users who do not care their online security at all [31] and some users are competent in managing security affairs [33], [34].

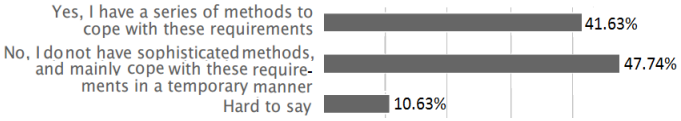


Fig. 11. Have you ever developed your own method to create passwords to satisfy the varied requirements from various sites? Their are more users who cope in a temporary manner than users who have developed their method.

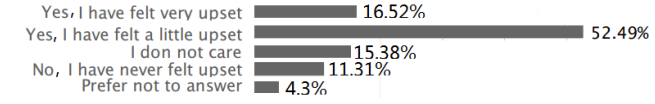


Fig. 12. Have you ever felt upset while managing passwords? (e.g. creation, memorization and change)? 69.01% see password management as a bother.

Fig. 13. Suppose you have to create a password with the following constraints: 1) at least 8 characters; and 2) at least one letter, one number and one symbol. Do you think this policy reasonable? 41.86% of users show a rational emotion.

The final question explores whether Chinese users are rational when confronted with stringent password policies from a site. Fig. 13 shows that 41.86% of users are rather rational: the necessity of strict policies depends on the importance of the value (service) to be protected. This figure well accords with English users (46.60% [32]). This implies that, from the user prospective, it is unwise for unimportant sites to impose strict password policies. Unfortunately, our recent large-scale investigation (see [39]) into the leading sites shows quite the opposite: important services (e.g., email and e-commerce) often impose loose policies, while less important services (e.g., IT corporation and web portal) often enforce strict policies.

C. Demographics of participants

We also obtained some basic demographic info (see Figs. 14∼16) about our participants: two-thirds are male. 80.55% are between 18∼34 years old, and 15.67% are older than 35; 80.55% are pursuing or with at least a bachelor's degree, and 43.44% are pursuing or with at least a master's degree. This is as expected, for the majority of our team members are young teachers and PhD candidates, and thus their personal relationships are mainly young and educated people. Comparatively, our participants are larger in size and more diversified than that of [8], [14], [33]. For instance, in [8], 93% of its 220 participants are 18∼34 years old and 92% have at least a bachelor's degree. Nevertheless, our participants are younger and more educated than the normal population, and they may be more security-conscious and technically savvy. Hence, it is likely that *we are underestimating many problems* such as password reuse, slow evolution and negative emotion.

Fig. 14. What's you gender? About two thirds are male.

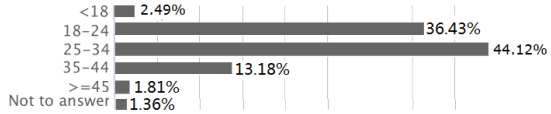


Fig. 15. What's you age? About 80% are between 18 and 34.

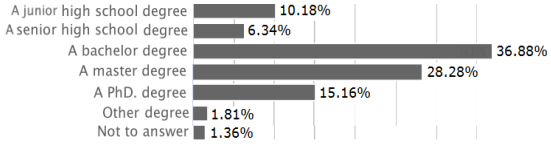


Fig. 16. Which is you highest degree (or you are pursuing)?

Limitations. Though our participants are superior to earlier studies [8], [14], [33] in terms of both user size and diversity, yet password-related user surveys are inherently subject to a limitation: users may provide false data and the ecological validity is hard to be assured. Still, in many aspects (e.g., the rate of usages of Pinyin names/words, and birthdates) our results comply with real-life datasets (see Sec. IV).

Summary. To our knowledge, this survey is the first one that targets Chinese users, the world's largest Internet population. From the 442 responses, we have revealed a wide variety of useful info: 14.03% reuse PINs in passwords, over 80% maintain 10 or fewer unique passwords, 54.3% memorize all their passwords, only 7.24% use a password manager and as high as 70% consider password management as a bother. *This provides a deeper understanding of Chinese user password behaviors. Such info is inherently difficult to be obtained from an empirical analysis*. As compared to previous surveys on English users [8], [14], [32], [33], there are both *similarities* (e.g., the rates of PIN reuse and recording passwords electronically, and popularity of password managers) and *differences* (e.g., the rates of direct password reuse and recording passwords on paper) in password behaviors between these two user groups.

TABLE I. DATA CLEANSING OF THE PASSWORD DATASETS LEAKED FROM NINE HIGH-PROFILE WEB SERVICES ("PWS" STANDS FOR PASSWORDS)

Dataset	Web service	Location	Language	Original PWs	Miscellany	Length>30	All removed	After cleansing	Unique PWs
Tianya	Social forum	China	Chinese	31,761,424	860,178	5	2.71%	30,901,241	12,898,437
7k7k	Gaming	China	Chinese	19,138,452	13,705,087	10,078	71.66%*	5,423,287	2,865,573
Dodonew	E-commerce&Gaming	China	Chinese	16,283,140	10,774	13,475	0.15%	16,258,891	10,135,260
178	Gaming	China	Chinese	9,072,966	0	1	0.00%	9,072,965	3,462,283
CSDN	Programmer forum	China	Chinese	6,428,632	355	0	0.01%	6,428,277	4,037,605
Duowan	Gaming	China	Chinese	5,024,764	42,024	10	0.83%	4,982,730	3,119,060
Rockyou	Social forum	USA	English	32,603,387	18,377	3140	0.07%	32,581,870	14,326,970
Yahoo	Portal(e.g., E-commerce)	USA	English	453,491	10,657	0	2.35%	442,834	342,510
Phpbb	Programmer forum	USA	English	255,421	45	3	0.02%	255,373	184,341

*We remove 13M duplicate accounts from 7k7k, because we identify that they are copied from Tianya as we will detail in Section IV-B.

IV. CHARACTERISTICS OF CHINESE PASSWORDS

We now empirically explore Chinese password characteristics, most of which so far have received little attention. Besides, we point out weaknesses in previous studies [13], [24].

A. Dataset descriptions and ethics consideration

Our empirical analysis employs six Chinese web password datasets and three English password datasets. In total, these nine datasets consist of 130 million passwords, one of the largest corpuses ever studied. As summarized in Table I, these nine datasets are different in terms of service, language/culture, size and user localization. They were hacked by external attackers or disclosed by anonymous insiders, and were subsequently made public on the Internet.

We realize that though publicly available, these datasets are private data. Therefore, we only report the aggregated statistical information, and treat each individual account as confidential so that using it in our research will not increase risk to the corresponding victim, i.e., no personally identifiable information can be learned. Furthermore, these datasets may be exploited by attackers as cracking dictionaries, while our use of them is both beneficial for the academic community to understand password choices of Chinese netizens and for security administrators to secure user accounts. Also note that, since the datasets we employed are all publicly available, all the results in this work are reproducible.

B. Data cleansing

We note that some datasets (e.g., Rockyou and Tianya) consist of *un*-necessary headers/descriptions/footnotes, password strings with len>100, etc. Thus, before any exploration, we first launch a data cleansing. We get rid of email addresses and user names from the original data. As with [26], we remove strings that include characters beyond the 95 printable ASCII symbols. We further remove strings with len>30, because after having manually scrutinized the original datasets, we find that these long strings do *not* seem to be generated by human beings, but more likely by password managers or simply junk info. Moreover, such unusually long passwords are often beyond the scope of attackers who specially care about cost-effectiveness [4], [35]. In all, the fraction of excluded passwords is *negligible*, yet this cleansing step unifies the input of cracking algorithms and simplifies the later data processing.

Besides, we observe that there is a non-negligible overlap between the datasets of Tianya and 7k7k. We provide compelling evidence (see Appendix A at http://bit.ly/2fI2mAP) that the joint accounts were copied from Tianya to 7k7k, but not the reverse as done in [24]. After removing 7.82M joint accounts from 7k7k, we further note that, half of the remaining 7k7k accounts are of the form {user name, password}, while the rest are of the form {email address, password}. It follows that both forms are directly derived from the form {user name, email address, password}. For instance, both {wanglei, wang123} and {wanglei@gmail.com, wang123} are derived from the single account {wanglei, wanglei@gmail.com, wang123}. Thus, we further divide 7k7k into two equal parts and discard one part. The detailed info on data cleansing is listed in Table I.

C. Password characteristics

Language dependence. It is widely hypothesized that user-generated passwords are greatly influenced by their native languages, yet to our knowledge, no quantitative measurement has ever been given. To fill this gap, here we first illustrate the character distributions of the nine password datasets, and then measure the closeness of passwords with their native languages in terms of inversion number of the character distributions.

As expected, passwords from different language groups have significantly varied letter distributions (see Fig. 17). What's unexpected is that, even though generated and used in vastly diversified service types, passwords from the same language group have quite similar letter distributions. This inherently suggests that, when given a user's password to crack/evaluate, one shall exploit a training set in the the victim's native language. Arranged in descending order, the letter distribution of all Chinese passwords is `aineohglwuy szxqcdjmbtfrkpv`, while this distribution for all English passwords is `aeionrlstmcdyhubkgp jvfwzxq`. While some letters (e.g., 'a', 'e' and 'i') occur frequently in both groups, some letters (e.g., 'q' and 'r') only occur frequently in one group. Such info can be exploited by attackers to reduce the search space and optimize cracking strategies. Note that, here all the percentages are handled case-insensitively.

While users' passwords are greatly affected by their native languages, the letter frequencies in general language usages may be somewhat different from the frequencies of letters in passwords. *To what extent do they differ?* According to Huang et al.'s work [16], the letter distribution of Chinese language (i.e., written Chinese texts like literary work, newspapers and academic papers), when converted into Chinese Pinyin, is `inauhegoyszdjmxwqbctlpfrkv`. This shows that some letters (e.g., 'l' and 'w'), which are popular in Chinese passwords, appear much less frequently in written Chinese texts. A plausible reason may be that 'l' and 'w' is the first letter of the family name `li` and `wang` (which are the top-2 family names in China), respectively, while Chinese users, as we will show, love to use names to build their passwords.

Similar observation also holds for English passwords. The letter distribution of English language (i.e., `etaoinshrdl cumwfgypbvkjxqz`) is obtained from www.cryptograms.org/letter-frequencies.php. For instance, the letter 't' is common in English texts, but not so in English passwords. A plausible reason may be that 't' is used in popular words

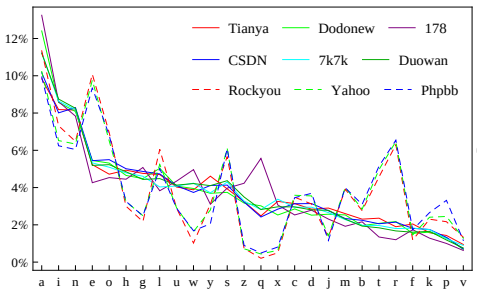


Fig. 17. Letter distributions of passwords

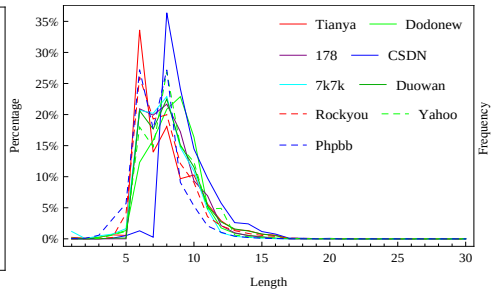


Fig. 18. Length distributions of passwords

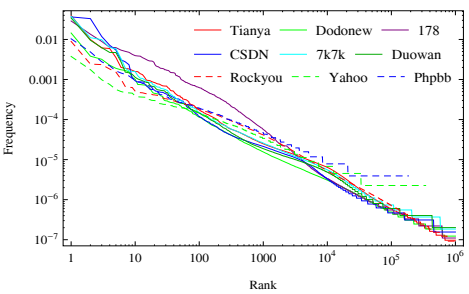


Fig. 19. Frequency distributions of passwords

TABLE II. INVERSION NUMBER OF THE LETTER DISTRIBUTIONS (IN DESCENDING ORDER) BETWEEN THE DATASETS

	Tianya	7k7k	178	CSDN	Dodonew	Duowan	All Chinese PWs	Chinese language	Pinyin fullname	Pinyin word	Rockyou	Yahoo	Phpbb	All English PWs	English language
Tianya	0	15	22	42	15	17	14	40	32	37	100	100	113	100	99
7k7k	15	0	23	31	14	10	13	41	39	38	105	101	112	105	96
Dodonew	22	23	0	42	21	15	12	52	40	49	94	92	105	94	99
178	42	31	42	0	41	35	32	56	48	47	134	130	141	134	125
CSDN	15	14	21	41	0	12	15	45	39	42	95	95	106	95	96
Duowan	17	10	15	35	12	0	9	49	39	44	99	97	110	99	98
All_Chinese_PWs	14	13	12	32	15	**9**	0	44	34	43	104	102	115	104	101
Chinese_language	40	41	52	56	45	49	44	0	38	27	118	114	123	118	113
Pinyin_fullname	32	39	40	48	39	39	34	38	0	31	124	122	135	124	123
Pinyin_word	37	38	49	47	42	44	43	27	31	0	115	113	124	115	112
Rockyou	100	105	94	134	95	99	104	118	124	115	0	12	23	0	47
Yahoo	100	101	92	130	95	97	102	114	122	113	12	0	15	12	39
Phpbb	113	112	105	141	106	110	115	123	135	124	23	15	0	23	44
All_English_PWs	100	105	94	134	95	99	104	118	124	115	**0**	12	23	0	47
English_language	99	96	99	125	96	98	101	113	123	112	47	39	44	47	0

TABLE III. TOP-3 STRUCTURAL PATTERNS IN TWO USER GROUPS (THE % IS TAKEN BY DIVIDING THE CORRESPONDING TOTAL ACCOUNTS.)

Top-3 patterns in Chinese PWs	Chinese password datasets: Tianya	7k7k	Dodonew	178	CSDN	Duowan	Average of Chinese PWs	Top-3 patterns in English PWs	English password datasets: Rockyou	Yahoo	Phpbb	Average of English PWs
D(e.g., 123456)	**63.77%**	**59.62%**	**30.76%**	**48.07%**	**45.01%**	**52.84%**	**52.93%**	L (e.g., passowrd)	**41.69%**	33.03%	**50.07%**	**41.59%**
LD (e.g., a123456)	14.71%	17.98%	43.50%	31.12%	26.14%	23.97%	23.72%	LD (e.g., abc123)	27.70%	38.27%	19.14%	28.37%
DL(e.g., 123456a)	4.12%	3.91%	7.55%	6.25%	5.88%	5.83%	5.25%	D (e.g., 123456)	15.94%	5.89%	12.06%	11.30%
Sum of top-3	82.61%	81.51%	81.80%	85.45%	77.03%	82.64%	**81.90%**	Sum of top-3	85.33%	77.19%	81.25%	**81.26%**

such as the, it, this, that, at, to, while such words are not common in passwords.

To further explore the closeness of passwords with their native languages and with the passwords from other datasets, we measure the inversion number of the letter distribution sequences (in descending order) between two password datasets (as well as languages), and the results are summarized in Table II. "Pinyin_fullname" is a dictionary consisting of 2,426,841 unique Chinese full names (e.g., wanglei and zhangwei), "Pinyin_word" is a dictionary consisting of 127,878 unique Chinese words (e.g., chang and cheng), and these two dictionaries will be detailed later. Note that the inversion number of sequence A to sequence B is equal to that of B to A. For instance, the inversion number of inauh to aniuh is 3, which is equal to that of aniuh to inauh.

Table II illustrates that, the inversion number of letter distributions between passwords from the same language group is generally much smaller than that of passwords from different language groups. This value is also distinctly smaller than that of the letter distributions between passwords and their native language (see the bold values in Table II). All these indicate that passwords from different languages are intrinsically different from each other in letter distributions, and that passwords are close to their native language yet the distinction is still noticeable (measurable).

Note that, among all Chinese datasets, Duowan has the least inversion number (i.e., 9 in dark gray) with the dataset "All_Chinese_PWs". This indicates that *Duowan passwords are likely to best represent general Chinese web passwords, and thus Duowan will be selected as the training set for attacking other Chinese datasets (see Sec V).* For a similar reason, Rockyou will be selected as the training set for attacking English passwords.

Length distribution. Fig. 18 depicts the length distributions of passwords. Irrespective of the web service, language and culture differences, the most common password lengths of every dataset are between 6 and 10, among which length-6 or 8 takes the lead. Merely passwords of these five lengths can account for more than 75% of every entire dataset, and this value will rise to 90% if we consider passwords with lengths of 5 to 12. As expected, very few users prefer passwords longer than 15 characters. Notably, people seem to prefer even length over odd length. Another interesting observation is that, CSDN exhibits only one peak in its length distribution curve and has much fewer passwords (i.e., only 2.16%) in length below 8. This is likely due to the fact that this site has enforced a minimal length-8 policy since an early stage.

Frequency distribution. Fig. 19 portrays the frequency vs. the rank of passwords from different datasets in a log-log scale. We first sort each dataset according to the password frequency in descending order, and then each individual password will be associated with a frequency f_r and a rank r. Interestingly, the curve for each dataset closely approximates a straight line, and this trend will be more pronounced if we take all the nine curves as a whole. This well accords with the Zipf's law [40]: f_r and r follow a relationship of the type $f_r = C' \cdot r^s - C \cdot (r-1)^s \approx C \cdot s \cdot r^{s-1}$, where $C \in$[0.01, 0.06] and $s \in$[0.15,

0.30] are constants. Particularly, $1 - s$ is the absolute value of slope of the Zipf linear regression line and slightly less than 1.0. The Zipf theory indicates that *the popularity of user-generated passwords decreases polynomially with the increase of their rank*. This further implies that a few passwords are overly popular (and thus online guessing can be effective, even if security mechanisms like rate-limiting and suspicious login detection [9] are implemented at the server), while the least frequent passwords are very sparsely scattered in the password space (and thus the offline guessing attackers need to consider cost-effectiveness and weigh when to stop).

Top popular structures. As we have seen that digits are popular in top-10 passwords of Chinese datasets, are they also popular in the whole datasets? To answer this question, we investigate the frequencies of password patterns that involve *digits*, and results on the top 3 most frequent ones are shown in the left hand of Table III. The first column of the table denotes the pattern of a password as in [42] (i.e., L denotes a lower-case sequence, D for digit sequence, U for upper-case sequence, S for symbol sequence, and the structure pattern of password "`Wanglei123`" is ULD). We see that an average of more than 50% of Chinese web passwords are only composed of digits, while this value of English datasets is only 11.30%. In contrast, English users prefer the patterns L and LD. .

It is somewhat surprising to see that, the sum of merely the three digit-based patterns (i.e., D, LD and DL) accounts for an average of 81.90% for Chinese datasets. In contrast, English users favor letter-related patterns, and on average, their top-3 patterns (i.e., L, LD and D) also account for slightly over 80%. This indicates that, unlike English users, Chinese users are inclined to employ digits to build their passwords — digits serve the role of letters that play in English passwords, while letters mainly come from Chinese Pinyin words/names. This is probably due in large part to the fact that most Chinese users are unfamiliar with English language (and Roman letters on the keyboard). If this is the case, is there any meaningful (semantic) information underlying these digit sequences?

Semantics in passwords. As there is little existing work, to gain an insight into the underlying semantic patterns, we have to construct semantic dictionaries from scratch by ourselves. Finally, we construct 22 dictionaries of different semantic categories and investigate their prevalence, and detailed info is referred to Appendix B at http://bit.ly/2fI2mAP. Table IV shows the various semantic patterns existing in Chinese and English passwords. We can see that, a large fraction of English users tend to use English words as their password building blocks. More specially, 25.88% English users insert a 5-letter or longer (denoted by 5^+-letter) *word* into their passwords, and this figure accounts for over a third of the total passwords with a 5^+-letter substring. In comparison, few Chinese users choose raw Pinyin words or English words to build passwords, yet they prefer Pinyin names, especially *full names*.

Particularly, of all the Chinese passwords (22.42%) that include a 5^+-letter substring, more than half (11.24%) include a 5^+-letter Pinyin full name. This well complies with that of our user survey. There is also a non-negligible proportion (i.e., 4.10%) of English passwords that contain a 5^+-letter full Pinyin name, and a reasonable explanation is that many Chinese users have created accounts in these English sites. For instance, the popular Chinese name "`zhangwei`" appears in both Rockyou and Yahoo. We also note that English names are also widely used in English passwords, yet full names are less popular than last names and first names. As far as we know, *for the first time we have explored name-based semantic patterns in a large-scale empirical password study*.

Equally interestingly, we find that, on average, 16.99% of Chinese users insert a six-digit date into their passwords. Further considering Fig. 2, such dates are likely to be users' birthdays. Besides, about 30.89% of Chinese users employ a 4^+-digit date as their password building blocks, which is 3.59 times higher than that of English users (i.e. 8.61%); there are 13.49% of Chinese users inserting a four-digit year into their passwords, which is about 3.55 times higher than that of English users (3.80%, which is comparable to the results reported in [8]). We note that there might be some overestimates, for there is no way to definitely tell apart whether some digit sequences are dates or not, e.g., `010101` and `520520`. These two sequences may be dates, yet they are also likely to be of other semantic meanings (e.g., `520520` sounds like "I love you$\cdots$"). As discussed later, we have devised reasonable ways to address this issue. In all, dates play a vital role in Chinese user passwords.

Another interesting observation is that, 2.91% of Chinese users just use their 11-digit mobile numbers as passwords, making up 39.59% of all passwords with a 11^+-digit substring. In average, 12.39% of Chinese passwords are longer than 11. Thus, if an attacker can determine (e.g., by shoulder-surfing) that the victim uses a long password, she is likely to succeed with a high chance of 23.48%($= \frac{2.91\%}{12.39\%}$) by just trying the victim's 11-digit mobile number. *This reveals a practical attacking strategy against long Chinese passwords.*

Note that there are some un-avoidable ambiguities when determining whether a text/digit sequence belongs to a specific dictionary, and improper resolution of these ambiguities would lead to an overestimation or underestimation of human choices. Here we take the dictionary "YYMMDD" for illustration. For example, both `111111` and `520521` fall into "YYMMDD" and are highly popular, yet it is more likely that users choose them simply because they are easily memorable repetition numbers or meaningful strings, and counting them as dates would lead to an *overestimation*. Yet, they can really be dates (e.g., `111111` stands for "Jan. 1th, 2011" and `520521` stands for "May 21th, 1952"), and completely excluding them from "YYMMDD" would result in an *underestimation* of dates.

Thus, we assume that user birthdays are randomly distributed, and assign the expectation of frequency of dates (denoted by E), instead of zero, to the frequency of these abnormal dates. We manually identify 17 abnormal dates in the dictionary "YYMMDD", each of which is originally with a frequency$> 10E$ and appears in every top-1000 list of the six Chinese datasets. In this way, the dilemma can be largely resolved. We similarly tackle 21 abnormal items in "MMDD". As for the other 19 dictionaries in Table IV, few abnormal items can be identified, and thus they are processed as usual.

Summary. We have measured nine password datasets in terms of letter distribution, length distribution, frequency distribution and semantic patterns. To our knowledge, most of these fundamental characteristics have *not* been examined in the literature (see [13], [19], [24], [26], [38]). We have identified

TABLE IV. THE PREVALENCE OF VARIOUS SEMANTIC WORDS IN PASSWORDS (THE PERCENTAGE IS TAKEN BY DIVIDING THE TOTAL ACCOUNTS.)

Semantic dictionary	Tianya	7k7k	Dodonew	178	CSDN	Duowan	Avg. Chinese	Rockyou	Yahoo	Phpbb	Avg. English
English_word_lower(len ≥ 5)	2.08%	2.05%	3.69%	0.83%	3.41%	2.37%	2.41%	**23.54%**	**29.49%**	**24.60%**	**25.88%**
English_firstname(len ≥ 5)	1.11%	0.93%	2.23%	0.53%	1.47%	1.19%	1.24%	18.80%	15.21%	9.20%	14.40%
English_lastname(len ≥ 5)	2.16%	2.34%	4.48%	1.93%	3.65%	2.77%	2.89%	**20.16%**	**20.82%**	**15.22%**	**18.73%**
English_fullname(len ≥ 5)	4.03%	4.30%	6.14%	4.99%	6.58%	5.07%	5.18%	13.05%	11.35%	8.25%	10.88%
English_name_any(len ≥ 5)	4.60%	4.65%	6.32%	5.20%	6.87%	5.18%	5.35%	**27.67%**	**26.51%**	**18.71%**	**24.30%**
Pinyin_word_lower(len ≥ 5)	7.34%	8.56%	10.82%	10.24%	11.51%	9.92%	9.73%	3.33%	2.99%	2.50%	2.94%
Pinyin_familyname(len ≥ 5)	1.35%	1.64%	2.34%	2.24%	2.47%	1.88%	1.99%	0.05%	0.07%	0.07%	0.06%
Pinyin_fullname(len ≥ 5)	**8.39%**	**9.87%**	**12.91%**	**11.81%**	**13.14%**	**11.29%**	**11.24%**	4.79%	4.17%	3.35%	**4.10%**
Pinyin_name_any(len ≥ 5)	8.56%	10.05%	13.31%	12.11%	13.46%	11.53%	11.50%	4.80%	4.18%	3.36%	4.11%
Pinyin_place(len ≥ 5)	1.24%	1.27%	1.64%	1.58%	2.12%	1.48%	1.55%	0.20%	0.18%	0.16%	0.18%
PW_with_a_5^+-letter_substring	**18.51%**	**19.99%**	**26.95%**	**19.38%**	**28.03%**	**21.70%**	**22.42%**	**71.69%**	**75.93%**	**68.66%**	**72.09%**
Date_YYYY	14.38%	12.82%	12.45%	10.06%	16.91%	14.33%	13.49%	4.34%	4.30%	2.77%	3.80%
Date_YYYYMMDD	6.06%	5.42%	3.93%	3.94%	8.78%	6.17%	**5.72%**	0.10%	0.05%	0.09%	**0.08%**
Date_MMDD	24.99%	19.97%	17.08%	16.46%	24.45%	22.59%	20.92%	7.53%	4.46%	3.59%	5.20%
Date_YYMMDD	**21.29%**	**15.89%**	**12.70%**	**13.09%**	**20.67%**	**18.28%**	**16.99%**	3.24%	1.23%	1.55%	2.01%
Date_any_above	**36.61%**	**30.39%**	**26.66%**	**27.07%**	**35.30%**	**33.58%**	**31.60%**	11.33%	8.77%	6.45%	8.85%
PW_with_a_digit	89.49%	88.42%	88.52%	90.76%	87.10%	89.26%	88.93%	54.04%	64.74%	46.14%	54.97%
PW_with_a_4^+-digit_substring	81.64%	76.98%	71.90%	78.76%	78.38%	80.60%	78.04%	24.72%	21.85%	19.33%	21.97%
PW_with_a_6^+-digit_substring	**75.59%**	**68.32%**	**61.16%**	**70.02%**	**69.87%**	**73.10%**	**69.68%**	17.77%	8.48%	11.28%	12.51%
PW_with_a_8^+-digit_substring	28.04%	27.56%	26.53%	26.37%	49.73%	31.03%	31.54%	6.88%	2.50%	3.73%	4.37%
Mobile_Phone_Number(11-digit)	**2.90%**	**1.76%**	**2.63%**	**3.97%**	**3.75%**	**2.44%**	***2.91%***	0.07%	0.01%	0.02%	0.03%
PW_with_a_11^+-digit_substring	**4.71%**	**2.09%**	**3.39%**	**5.08%**	**7.57%**	**3.35%**	**4.36%**	0.75%	0.17%	0.18%	0.37%

a number of *similarities* (e.g., frequency distribution and the theme of love) and *differences* (e.g., letter distribution, structural patterns, and semantic patterns) between passwords of these two user groups. Particularly, the popular usages of (birth)dates and long Pinyin names in Chinese passwords would pose a potential vulnerability for an attacker to largely reduce her search space, as we will establish in what follows.

V. STRENGTH OF CHINESE WEB PASSWORDS

Now we employ two state-of-the-art password attacking algorithms (i.e., PCFG-based [42] and Markov-based [26]) to evaluate the strength of Chinese web passwords. We further investigate whether the characteristics identified in Sec. IV-C can be practically exploited to reduce password space and facilitate guessing. We note that probability-threshold graphs may provide cracking results on the full spectrum of passwords, yet their theoretical basis is left as "interesting future research" [26]. Moreover, they only approximate the likelihood of passwords and thus cannot yield precise results. Thus, as with [35], [40], [42], we employ guess-number graphs.

Necessity of pairing passwords by service. There are a number of confounding factors that impact password security, among which language, service type and password policy are the three most important ones [18], [38]. As shown in [26], [38], except for CSDN that imposes a length 8^+ policy, all our datasets (Table I) reflect no explicit policy requirements. It has recently been revealed that users often rationally choose robust passwords for accounts perceived to be important [33], while knowingly choose weak passwords for unimportant accounts [10]. This can also be seen from our survey (Fig. 13). Since accounts of the same service generally would have the same level of value for users, we divide datasets into three pairs according to their types of services (i.e., Tianya vs. Rockyou, Dodonew vs. Yahoo, and CSDN vs. Phpbb) for fairer strength comparison. We emphasize that, it is less reasonable if one compares Dodonew passwords (from an e-commerce site) with Phpbb passwords (from a low-value programmer forum): Even if Dodonew passwords are stronger than Phpbb passwords, one can not conclude that Chinese passwords are more secure than English ones, because there is a potential that Dodonew passwords will be weaker than Yahoo e-commerce passwords. *As far as we know, we, for the first time, show the necessity of pairing two password datasets on the basis of their service type when comparing security between two different user groups*, as opposed to existing works [4], [13], [24].

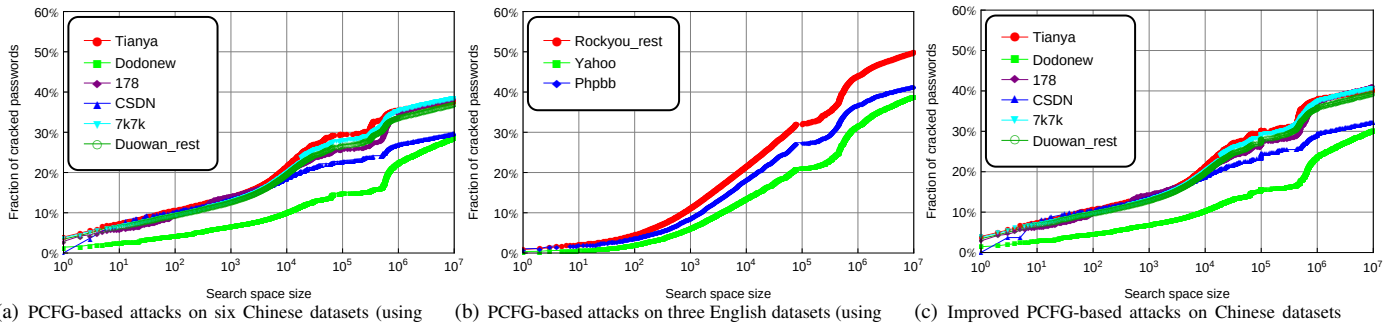
A. PCFG-based attacks

The PCFG-based model [42] is one of the state-of-the-art cracking models. Firstly, it divides all the passwords in a training set into segments of similar character sequences and obtains the corresponding base structures and their associated probabilities of occurrence. For example, "`wanglei@123`" is divided into the L segment "wanglei", S segment "@" and D segment "123", and its base structure is $\mathsf{L}_7\mathsf{S}_1\mathsf{D}_3$. The probability of $\mathsf{L}_7\mathsf{S}_1\mathsf{D}_3$ is $\frac{\#\text{of } \mathsf{L}_7\mathsf{S}_1\mathsf{D}_3}{\#\text{of base structures}}$. Such information is used to generate the probabilistic context-free grammar.

Then, one can derive password guesses in decreasing order of probability. The probability of each guess is the product of the probabilities of the productions used in its derivation. For instance, the probability of "`liwei@123`" is computed as P("`liwei@123`")= $P(\mathsf{L}_5\mathsf{S}_1\mathsf{D}_3)\cdot P(\mathsf{L}_5 \rightarrow \text{liwei})\cdot P(\mathsf{S}_1 \rightarrow \texttt{@})\cdot P(\mathsf{D}_3 \rightarrow 123)$. In Weir et al.'s original proposal [42], the probabilities for D and S segments are learned from the training set by counting, yet L segments are handled either by learning from the training set or by using an external input dictionary. Ma et al. [26] revealed that PCFG-based attacks with L segments directly learned from the training set generally perform better than using an external input dictionary. Thus, we prefer to instantiate the PCFG L segments of password guesses directly learning from the training set.

We divide the nine datasets into two groups by language. For the Chinese group of test sets, we randomly select 1M passwords from the Duowan dataset as the training set (denoted by "Duowan_1M"). The reason why we select Duowan as the training set has been discussed in Sec. IV-C. For the English group of test sets, for a similar reason we select 1M passwords from Rockyou as the training set. Since we have only used part of Duowan and Rockyou, their remaining passwords as well as the other seven datasets are used as the test sets. The attacking results on the Chinese group and English group are depicted in Fig. 20(a) and Fig. 20(b), respectively.

Bifacial-security. When the guess number (i.e., search space size) allowed is below about 3,000, Chinese passwords are generally much *weaker* than English passwords from the same service (i.e., Tianya vs. Rockyou, Dodonew vs. Yahoo, and CSDN vs. phpbb).For example, at 100 guesses, the success rate against Tianya, Dodonew and CSDN is 10.2%, 4.3% and 9.7%, respectively, while their English counterparts are 4.6%,



(a) PCFG-based attacks on six Chinese datasets (using 1 million (1M) Duowan passwords as the training set)

(b) PCFG-based attacks on three English datasets (using 1M Rockyou passwords as the training set)

(c) Improved PCFG-based attacks on Chinese datasets (using 1M Duowan passwords as the training set)

Fig. 20. General and our improved PCFG-based attacks on different groups of datasets. Our algorithm gains tangible increase in success rate.

1.9% and 3.7%, respectively. However, when the search space size is above 10,000, Chinese web passwords are generally much *stronger* than their English counterparts. For example, at 10 million guesses, the success rate against Tianya, Dodonew and CSDN is 37.5%, 28.8% and 29.9%, respectively, while their English counterparts are 49.7%, 39.0% and 41.4%, respectively. The strength gap between these two groups of datasets will be even wider when the guess number further increases. This reveals a reversal principle, i.e., the bifacial-security nature of Chinese passwords: they are more vulnerable to online guessing attacks (i.e., when the guess number allowed is small), yet more secure against offline guessing attacks. This well reconciles two drastically conflicting claims (see Section I-A) made about the security strength of Chinese web passwords. The bifacial-security is highly because that top Chinese passwords are more concentrated (see Table VIII of [38]), and that Chinese passwords include more digits (see Table III) while digits are generally more random than letters.

A weakness in PCFG. We observe that, the original PCFG-based algorithm [26], [42] inherently gives extremely low probabilities to password guesses (e.g., "1q2w3e4r" and "1a2b3c4d") that are of a monotonically long structure (e.g., $\mathsf{D}_1\mathsf{L}_1\mathsf{D}_1\mathsf{L}_1\mathsf{D}_1\mathsf{L}_1\mathsf{D}_1\mathsf{L}_1$, or $(\mathsf{D}_1\mathsf{L}_1)_4$ for short). For example, $P(\text{"1q2w3e4r"}) = P((\mathsf{D}_1\mathsf{L}_1)_4) \cdot P(\mathsf{D}_1 \rightarrow 1) \cdot P(\mathsf{L}_1 \rightarrow \mathtt{q}) \cdot P(\mathsf{D}_1 \rightarrow 2) \cdot P(\mathsf{L}_1 \rightarrow \mathtt{w}) \cdot P(\mathsf{D}_1 \rightarrow 3) \cdot P(\mathsf{L}_1 \rightarrow \mathtt{e}) \cdot P(\mathsf{D}_1 \rightarrow 4) \cdot P(\mathsf{L}_1 \rightarrow \mathtt{r})$ can hardly be larger than 10^{-9}, for it is a multiplication of *nine* probabilities. Thus, some guesses (e.g., "1q2w3e4r" and "a12b34c56") will never appear in the top 10^7 guesses generated by the original PCFG-based algorithm, even though they are popular (e.g, "1q2w3e4r" appears in the top-200 list of every dataset). The essential reason is that the PCFG-based algorithm simply assumes that each segment in a structure is independent. Yet, in many situations this is *not* true. For instance, the four D_1 segments and L_1 segments in the structure $(\mathsf{D}_1\mathsf{L}_1)_4$ of password "1q2w3e4r" are evidently interrelated with each other (i.e., D_4: 1234, and L_4: qwer).

Our simple yet effective solution. To address this problem, we specially tackle a few structures that are long but simple alternations of short segments by treating them as short structures, e.g., $(\mathsf{D}_1\mathsf{L}_1)_4$ and $(\mathsf{D}_1\mathsf{L}_2)_3$ are converted to $\mathsf{D}_4\mathsf{L}_4$ and $\mathsf{D}_3\mathsf{L}_6$, respectively. In this way, the probability of "1q2w3e4r" now is computed as $P(\text{"1q2w3e4r"}) = P((\mathsf{D}_1\mathsf{L}_1)_4) \cdot P((\mathsf{D}_1\mathsf{L}_1)_4 \rightarrow \mathsf{D}_4\mathsf{L}_4) \cdot P(\mathsf{D}_4 \rightarrow 1234) \cdot P(\mathsf{L}_4 \rightarrow \mathtt{qwer})$. *Our approach is language-irrelevant and constitutes a* general *amendment to the state-of-the-art PCFG-based algorithm in [26].*

To further exploit the characteristics of Chinese passwords, we insert the "Pinyin_name_any" dictionary and the six-digit date dictionary (see Sec. IV-C) into the original PCFG L-segment dictionary and D-segment dictionary, respectively. The resulting changes to the original PCFG grammars learned from the training sets are shown in Table V.

TABLE V. CHANGES CAUSED TO THE ORIGINAL PCFG GRAMMARS

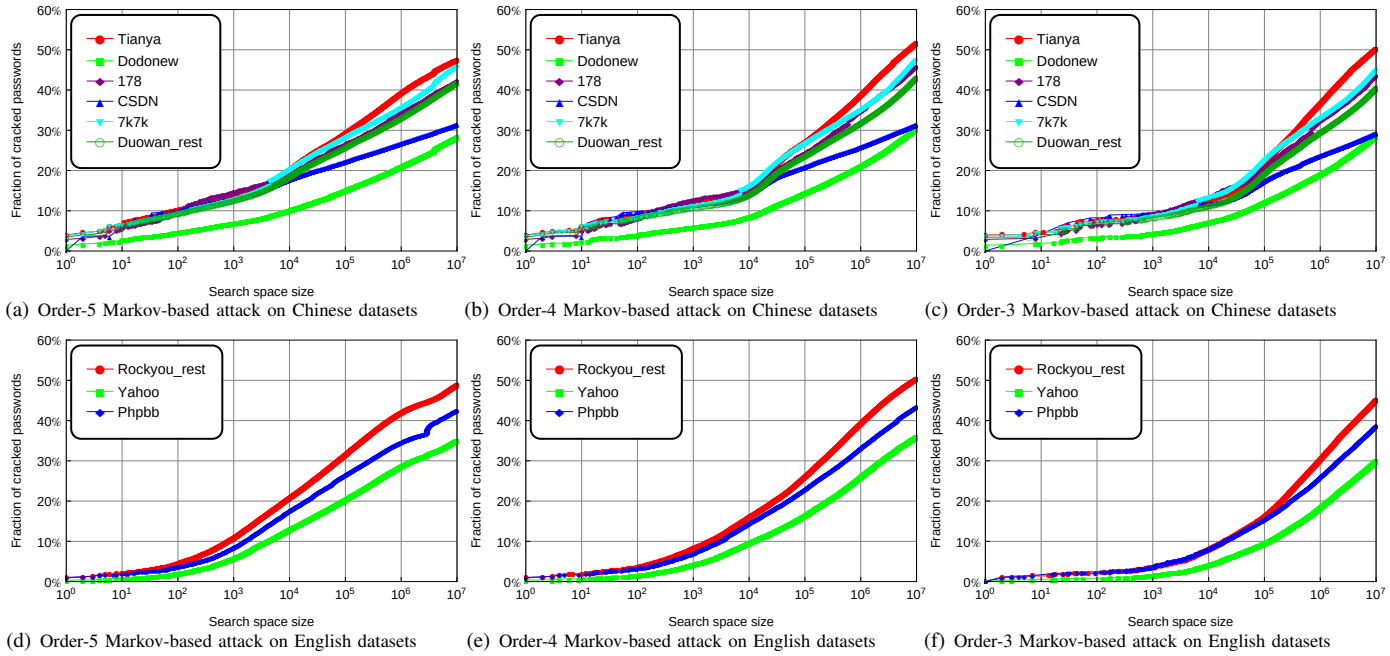
Training set	Base structures	L segments	D segments	S segments
Duowan_1M	8905+0	155693+24416	465157+20341	865+0
Duowan_All	20961+0	559017+98654	1824404+9744	2417+0

Fig. 20(c) shows that, when the guess number is small (e.g., 10^3), our improved attack exhibits little improvement; while the guess number grows, the improvement increases. For example, at 10^5 guesses, there is 0.09%∼0.85% improvement in success rate; at 1M guesses, this figure is 1.32∼4.32%; at 10M guesses, this figure reaches 1.70%∼4.29%. This indicates that, the popular usages of Pinyin names and birthdays facilitate an attacker to reduce her search space, and this issue is more serious when large guesses are allowed.

Comparison. In 2014, Li et al. [24] reported that using 2M Dodonew passwords as the training set and at 10 billion guesses, their best cracking record is about 17.30% (see Fig. 5 of [24]). However, our improved attack, which uses only 1M passwords as the training set and at merely 10M guesses, is able to achieve success rate from 29.41% to 39.47%. This means that we can crack 70% to 128% more passwords than Li et al.'s best record (i.e., 17.3%, see Fig. 5 of [24]).

The role of Names. In our improved PCFG-based attacks, external name segments are added into the PCFG L-segment dictionary during training, and we get gladsome increases in success rates (see Fig. 20(c)). However, such improvements are still not so prominent as compared to the prevalence of names in Chinese passwords. *To explicate this paradox, we scrutinize the internal process of PCFG-based guess generation and manage to identify its crux.* Here we take the improved PCFG-based attack against Tianya (training on Duowan) as an example. During training, we have added 98K name segments (see Table V) into the L-segment dictionary.

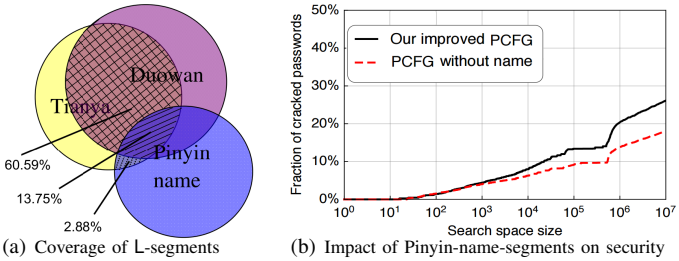
As shown in Fig. 22(a), these 98K name segments *only* cover 2.88% of the total L segments of the Tianya test set. However, the original L segments trained from Duowan can cover 13.75% of the name segments and 60.59% of the non-name L segments in the Tianya test set. This suggests that Duowan is able to *well* cover the name segments in the test set Tianya, and thus the addition of some extra names would be of limited yields. Note that this does *not* contradict our conjecture that Pinyin names pose a serious vulnerability, and actually, it *does* suggest that when the training set is selected properly, the name segments in passwords can be well represented. Still, when there is no proper training set available, our improved



(a) Order-5 Markov-based attack on Chinese datasets (b) Order-4 Markov-based attack on Chinese datasets (c) Order-3 Markov-based attack on Chinese datasets

(d) Order-5 Markov-based attack on English datasets (e) Order-4 Markov-based attack on English datasets (f) Order-3 Markov-based attack on English datasets

Fig. 21. Markov-chain-based attacks on different groups of datasets (scenario #1: *Laplace Smoothing* and *End-Symbol Normalization*). Attacks (a)∼(c) use 1 million Duowan passwords as the training set, while attacks (d)∼(f) use 1 million Rockyou passwords as the training set. The reversal principle also holds. The other four scenarios #2∼#5 show similar cracking results, and due to space constraints, they are omitted.



(a) Coverage of L-segments (b) Impact of Pinyin-name-segments on security

Fig. 22. Coverage and security impact of Pinyin-name-segments in the test set Tianya with L-segments involved (using Duowan as the training set and Pinyin_name as an extra input dictionary in our improved PCFG-based attack).

attack would show its advantages (see Fig. 22(b)). Moreover, although our improved PCFG-based algorithm might not be optimal, its cracking results represent a new benchmark that any future algorithm should aim to decisively clear.

B. Markov-Chain-based attacks

To establish the robustness of our findings about Chinese password security, we further conduct a series of Markov-Chain-based experiments. As recommended in [26], we consider two smoothing techniques (i.e., Laplace Smoothing and Good-Turing Smoothing) to deal with the data sparsity problem, two normalization techniques (i.e., distribution-based and end-symbol-based) to deal with the unbalanced length distribution problem of passwords. This brings about four attacking scenarios as listed in Table VI. In each scenario we consider three types of markov order (i.e., order-5, 4 and 3) to investigate which order performs best. It is reported in [26] that another scenario (i.e., backoff with end-symbol normalization) performs "slightly better" than the aforementioned four scenarios, yet it is "approximately 11 times slower, both for guess generation and for probability estimation" [26]. We also investigate this scenario and observe similar results in our experiments. Therefore, attackers, who particularly care about the cost-effectiveness, are highly unlikely to exploit this scenario. Due to space constraints, the detailed password guess

TABLE VI. FIVE MARKOV-BASED ATTACKING SCENARIOS

Attacking scenario	Smoothing	Normalization	Markov order
#1	Laplace	End-symbol	3/4/5
#2	Laplace	Distribution	3/4/5
#3	Good-Turing	End-symbol	3/4/5
#4	Good-Turing	Distribution	3/4/5
#5	Backoff	End-symbol	Backoff

generation procedure for Scenario #1 is referred to Algorithm 1 in the supplemental material, and the generation procedures for the other scenarios are quite similar.

Experimental setup. As with PCFG-based attacks, in our implementation we use a max-heap to store the interim results to maintain efficiency. To produce $k = 10^7$ guesses, we employ the strategy of first setting a lower bound (i.e., 10^{-9}) for the probability of guesses generated, then sorting all the guesses and finally selecting the top k ones. In this way, we are able to reduce the time overheads by 170% at the cost of about 110% increase in storage overheads, as compared to the strategy of producing exactly k guesses. In Laplace Smoothing, it is required to add δ to the count of each substring and we set $\delta = 0.01$ as suggested by Ma et al. [26]. In Good-Turing Smoothing, we reveal and address a subtlety in Appendix C (see http://bit.ly/2fI2mAP). The cracking results for Scenario #1 are included in Fig. 21. The experiments for attacking scenarios #2∼#5 show similar results with Scenario #1. Due to space constraints, they are omitted.

Results. Our experiments show that for both Chinese and English test sets: (1) At large guesses (i.e., $>2*10^6$), order-4 markov-chain evidently performs better than the other two orders, while at small guesses (i.e., $<10^6$) the larger the order, the better the performance will be; (2) There is little difference in performance between Laplace and GT Smoothing at small guesses, while the advantage of Laplace Smoothing gets greater as the guess number increases; (3) End-symbol normalization always performs better than the distribution-based approach, while at small guesses its advantages will be more obvious. This suggests that at large guesses, the attacks

with order-4, Laplace Smoothing and end-symbol normalization (see Figs. 21(b) and 21(e)) perform best; At small guesses, the attacks preferring order-5, Laplace Smoothing and end-symbol normalization (see Figs. 21(a) and 21(d)) perform best.

Note that, *the bifacial-security nature found it in our PCFG-based attacks (see Sec. V-A) also applies in all the Markov-based experiments*. For example, in order-4 Markov-based attacks (see Figs. 21(b) and 21(e)), we can see that, when the guess number is below 7000, Chinese passwords are generally much *weaker* than their English counterparts. For example, at 10^3 guesses, the success rate against Tianya, Dodonew and CSDN is 11.8%, 6.3% and 11.6%, respectively, while their English counterparts (i.e., Rockyou, Yahoo and Phpbb) are merely 8.1%, 4.3% and 7.1%, respectively. However, when the guess number allowed is over 10^4, Chinese passwords are generally *stronger* than their English counterparts. This suggests that, *for Chinese users/websites, more attention shall be paid to the online guessing threat*.

C. Two implications

For password creation policies. It is interesting to see that, 97.83% passwords in CSDN are of length 8^+, suggesting that CSDN enforces a minimum length-8 policy. In contrast, as shown in Tables IX and X of [38], Dodonew enforces no apparent rule (i.e., neither minimum length nor character set requirement). However, Figs. 20 and 21 indicate that, given any guess number below 10^7, passwords from CSDN are significantly weaker than passwords from Dodonew. A plausible reason is that Dodonew provides e-commerce services, and users perceive it as more important. As a result, users "rationally" [10], [33] choose more complex passwords for it. As for CSDN, since it is only a technology community, users knowingly choose weaker passwords for it.

In 2012, Bonneau [4] cast doubt on the hypothesis that users will rationally select more secure passwords to protect their more important accounts. In 2013 Egelman et al. [10] initiated a field study involving 51 students and confirmed this hypothesis. In 2014, Stobert and Biddle [33] interviewed 27 participants to investigate user behaviour in managing passwords, and their results also corroborate this hypothesis. As far as we know, *here we for the first time provide a large-scale empirical evidence (i.e., on the basis of 6.43 million CSDN passwords and 16.26 million Dodonew passwords) that supports for the hypothesis that users rationally select more secure passwords for more important accounts*.

We also note that though the overall security of Dodonew passwords are higher than the five other Chinese sites, many seemingly complex yet popular passwords (e.g., `5201314`, `321654a` and `love4ever`) dwelling in Dodonew also appear in other less sensitive sites. This is because that users inadvertently choose popular passwords, for they have no idea of what passwords other users have chosen [34], and that many users reuse the same password across multiple sites (see Figs. 3 and 4 in our survey). What's even more disturbing is that many users fail to recognize different categories of accounts, because achieving this goal is not easy [12]. Further considering the over-constrained nature of Web authentication and the finite-effort user, we suggest that when designing password policies, instead of merely insisting on stringent rules, administrators should put more efforts on helping users gain more *accurate perceptions of the importance* of the accounts to be protected and on guiding users towards *better ability of recognizing different categories of accounts*. Both efforts would help enhance user internal impetus and are essential for normal users to responsibly allocating passwords (i.e., selecting one from their limited pool of *passwords memorized* [12], [33]).

In addition, the "reversal principle" revealed in our work shows that Chinese passwords are more vulnerable to online guessing attacks. This is highly due to the fact that top popular Chinese passwords are more concentrated (see Table VIII of [38]). Thus, a special blacklist that includes a moderate number (e.g., 50K as suggested in [41]) of most common Chinese passwords would be very helpful for Chinese users to avoid trivial online guessing attacks. Such a blacklist can be learned from various leaked Chinese datasets (see one list in http://bit.ly/2aL38vU as constructed according to [40]). Any password falling into this list shall be deemed weak. However, it is well known that if some popular passwords are banned, new popular ones will arise. These new popular passwords may be complex and subtle to detect. Hence, whenever possible, besides password creation policies (rules), password strength meters shall be further employed by critical services (e.g., email and e-commerce) to detect weak user-chosen passwords.

For password cracking. Our extensive cracking experiments show that, as compared to Markov-based attacks, PCFG-based ones are simpler to implement (31% less computation and 70% less memory cost), yet they perform equally well (even better, see Fig. 20(a) to Fig. 21(a)) when the guess number is small (e.g., 10^3). For large guess numbers, order-4 Markov-based attacks are the best choices over order- 3 and 5. As far as we know, these observations have *not* been previously elucidated (see [26], [35]). We have only shown the Markov-based results when the guess number is below 10^7, there is a potential that order-3 Markov-based attacks will outperform order-4 and 5 ones at larger guess numbers (e.g., 10^{14}).

Summary. Both PCFG-based and Markov-based cracking results reveal the bifacial-security nature of Chinese passwords: they are more prone to online guessing yet more secure against offline guessing as compared to English passwords. This reconciles the conflicting claims made in [4], [24]. Alarmingly high cracking rates (40%∼50%) highlight the urgency of developing effective countermeasures (e.g., practical mnemonic techniques [3] and cracking-resistant honeywords [21]) to alleviate the situation. We also highlight two important implications. To the best of knowledge, *we, for the first time, provide a large-scale empirical evidence* for the hypothesis raised by the HCI community [10], [33], [34]: users rationally choose more robust passwords for accounts with higher value.

VI. Conclusion

In this paper, we have carried out the first survey on password behaviors of Chinese users, and performed a large-scale empirical analysis of 73.1 million real-world Chinese web passwords. Our survey got data from 442 effective participants, revealed a number of Chinese users' unique coping strategies for managing passwords, and found that 69% of them feel bothered. In our empirical analysis, we, for the first time, explored several fundamental properties (e.g., the distance between passwords and languages, frequency distribution and various semantic patterns) that characterize user passwords.

Particularly, we evaluated Chinese password strength by using two state-of-the-art cracking algorithms and our improved PCFG-based algorithm, and uncovered their nature of bifacial-security: Chinese passwords are more susceptible to online guessing attacks, yet more secure against offline guessing attacks than English ones. This well reconciles two conflicting claims made in [4], [24]. Our measurement results provide a deeper understanding of Chinese web passwords, while our systematic methodologies would facilitate understanding of passwords in other languages. Such understanding, we hope, will help us better design password policies and guidelines, thereby maintaining usability while improving security.

REFERENCES

[1] *China now has 656m mobile web users, and 710m total Internet users*, Aug. 2016, http://bit.ly/2avZdlK.

[2] J. Blocki, A. Datta, and J. Bonneau, "Differentially private password frequency lists," in *Proc. NDSS 2016*, pp. 1–15.

[3] J. Blocki, S. Komanduri, L. Cranor, and A. Datta, "Spaced repetition and mnemonics enable recall of multiple strong passwords," in *Proc. NDSS 2015*, pp. 1–15.

[4] J. Bonneau, "The science of guessing: Analyzing an anonymized corpus of 70 million passwords," in *IEEE S&P 2012*, pp. 538–552.

[5] A. S. Brown, E. Bracken, and S. Zoccoli, "Generating and remembering passwords," *Applied Cogn. Psych.*, vol. 18, no. 6, pp. 641–651, 2004.

[6] W. Burr, D. Dodson, R. Perlner, S. Gupta, and E. Nabbus, "NIST SP800-63-2: Electronic authentication guideline," National Institute of Standards and Technology, Reston, VA, Tech. Rep., Aug. 2013.

[7] X. Carnavalet and M. Mannan, "A large-scale evaluation of high-impact password strength meters," *ACM Trans. Inform. Syst. Secur.*, vol. 18, no. 1, pp. 1–32, 2015.

[8] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang, "The tangled web of password reuse," in *Proc. NDSS 2014*, pp. 1–15.

[9] M. Dürmuth, D. Freeman, and B. Biggio, "Who are you? A statistical approach to measuring user authenticity," in *Proc. NDSS 2016*.

[10] S. Egelman, A. Sotirakopoulos, K. Beznosov, and C. Herley, "Does my password go up to eleven?: the impact of password meters on password selection," in *Proc. ACM CHI 2013*, pp. 2379–2388.

[11] D. Florencio and C. Herley, "A large-scale study of web password habits," in *Proc. WWW 2007*, pp. 657–666.

[12] D. Florêncio, C. Herley, and P. C. Van Oorschot, "Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts," in *Proc. USENIX SEC 2014*, pp. 575–590.

[13] W. Han, Z. Li, L. Yuan, and W. Xu, "Regional patterns and vulnerability analysis of chinese web passwords," *IEEE Trans. Inform. Foren. Secur.*, vol. 11, no. 2, pp. 258–272, 2016.

[14] S. T. Haque, M. Wright, and S. Scielzo, "Hierarchy of users? web passwords: Perceptions, practices and susceptibilities," *Int. J. Human-Comput. Stud.*, vol. 72, no. 12, pp. 860–874, 2014.

[15] C. Herley and P. Van Oorschot, "A research agenda acknowledging the persistence of passwords," *IEEE Secur. Priv.*, vol. 10, pp. 28–36, 2012.

[16] J. Huang, H. Jin, F. Wang, and B. Chen, "Research on keyboard layout for chinese pinyin ime," *J. Chin. Inf. Process.*, vol. 24, no. 6, pp. 108–113, 2010.

[17] I. Ion, R. Reeder, and S. Consolvo, ""... no one can hack my mind": Comparing expert and non-expert security practices," in *Proc. SOUPS 2015*, pp. 327–346.

[18] M. Jakobsson and M. Dhiman, "The benefits of understanding passwords," in *Proc. HotSec 2012*, pp. 1–6.

[19] 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 Trans. Depend. Secur. Comput.*, 2015, doi: 10.1109/TDSC.2015.2481884.

[20] S. Ji, S. Yang, T. Wang, and et al., "PARS: A uniform and open-source password analysis and research system," in *ACSAC 2015*, pp. 321–330.

[21] A. Juels and R. L. Rivest, "Honeywords: Making password-cracking detectable," in *Proc. ACM CCS 2013*, pp. 145–160.

[22] D. V. Klein, "Foiling the cracker: A survey of, and improvements to, password security," in *Proc. of USENIX SEC 1990*, pp. 5–14.

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

[24] Z. Li, W. Han, and W. Xu, "A large-scale empirical analysis on chinese web passwords," in *Proc. USENIX SEC 2014*, pp. 559–574.

[25] Y. Liu, *12306.cn cracked*, Dec. 2014, http://bit.ly/2aUB3VR.

[26] J. Ma, W. Yang, M. Luo, and N. Li, "A study of probabilistic password models," in *Proc. IEEE S&P 2014*, pp. 689–704.

[27] M. L. Mazurek, S. Komanduri, T. Vidas, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur, "Measuring password guessability for an entire university," in *Proc. ACM CCS 2013*, pp. 173–186.

[28] R. Morris and K. Thompson, "Password security: A case history," *Comm. ACM*, vol. 22, no. 11, pp. 594–597, 1979.

[29] A. Narayanan and V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," in *Proc. ACM CCS 2005*, pp. 364–372.

[30] B. L. Riddle, M. S. Miron, and J. A. Semo, "Passwords in use in a university timesharing environment," *Comput. Secur.*, vol. 8, no. 7, pp. 569–579, 1989.

[31] S. Riley, "Password security: What users know and what they actually do," *Usability News*, vol. 8, no. 1, pp. 2833–2836, 2006.

[32] R. Shay, S. Komanduri, P. Kelley, P. Leon, M. Mazurek, L. Bauer, N. Christin, and L. Cranor, "Encountering stronger password requirements: user attitudes and behaviors," in *Proc. SOUPS 2010*, pp. 1–20.

[33] E. Stobert and R. Biddle, "The password life cycle: user behaviour in managing passwords," in *Proc. SOUPS 2014*, pp. 243–255.

[34] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, and A. Deepak, "Do users' perceptions of password security match reality?" in *Proc. ACM CHI 2016*, pp. 1–10.

[35] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. Cranor, S. Komanduri, and et al., "Measuring real-world accuracies and biases in modeling password guessability," in *Proc. USENIX SEC 2015*, pp. 463–481.

[36] L. Vaas, *People like using passwords way more than biometrics*, Aug. 2016, http://www.hardwareheaven.com/news.php?newsid=526.

[37] R. Veras, J. Thorpe, and C. Collins, "Visualizing semantics in passwords: The role of dates," in *Proc. ACM VizSec 2012*, pp. 88–95.

[38] D. Wang, D. He, H. Cheng, and P. Wang, "fuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars," in *Proc. IEEE/IFIP DSN 2016*, pp. 595–606, http://bit.ly/2ahJ8CO.

[39] D. Wang and P. Wang, "The emperor's new password creation policies," in *Proc. ESORICS 2015*, pp. 456–477.

[40] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang, "Targeted online password guessing: An underestimated threat," in *Proc. ACM CCS 2016*, pp. 1242–1254, available at http://bit.ly/2cNsfCC.

[41] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proc. ACM CCS 2010*, pp. 162–175.

[42] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," in *Proc. IEEE S&P 2009*, pp. 391–405.

[43] J. Yan, A. F. Blackwell, R. J. Anderson, and A. Grant, "Password memorability and security: Empirical results." *IEEE Secur. Priv.*, vol. 2, no. 5, pp. 25–31, 2004.

[44] E. Yu, *An incomplete list of Chinese websites that have publichly leaked passwords*, Feb. 2016, http://www.liu16.com/post/476.html.

[45] Y. Zhang, F. Monrose, and M. K. Reiter, "The security of modern password expiration: an algorithmic framework and empirical analysis," in *Proc. ACM CCS 2010*, pp. 176–186.

Ding Wang received his B.S. Degree in Information Security from Nankai University in 2008. Currently, he is pursuing his Ph.D. at Peking University. His works appear in prestigious venues like ACM CCS, IEEE/IFIP DSN, ESORICS and IEEE TDSC, and two of them are selected as "ESI highly cited papers". He focuses on password cryptography.

Ping Wang received his Ph.D. degree from University of Massachusetts in 1996. Now he is a Professor in Peking University. He served as TPC co-chairs of several conferences. He focuses on network and system security.

APPENDIX A
CLEANSING OF OUR DATASETS

Contaminated datasets. Of particular interest is our observation that, there is a non-negligible overlap between the original Tianya dataset and 7k7k dataset. We were first puzzled by the fact that the password "`111222tianya`" originally lay in the top-10 most popular list of both datasets. We manually scrutinize the original datasets (i.e., before removing the email addresses and user names) and are surprised to find that there are around 3.91 million (actually 3.91*2 million due to a split representation of 7k7k accounts, as we will discuss later) joint accounts in both datasets. We realize that someone probably have copied these joint accounts from one dataset to the other.

Our cleansing approach. Now, a natural question arises: *From which dataset have these joint accounts been copied?* It is highly likely that these joint accounts were copied from Tianya to 7k7k, *mainly for two reasons*. Firstly, it is unreasonable for 0.34% users in 7k7k to insert the string "`tianya`" into their 7k7k passwords, while users from tianya.cn are natural to include the site name "`tianya`" into their passwords for convenience. The following second reason is quite subtle yet convincing. In the original Tianya dataset, we find that the joint accounts are of the form {user name, email address, password}, while in the original 7k7k dataset such joint accounts are divided into two parts: {user name, password} and {email address, password}. The password "`111222tianya`" occurs 64822 times in 7k7k and 48871 times in Tianya, and one gets that $64822/2 < 48871$. Therefore, it is more plausible for someone to copy *some* (i.e., $64822/2$ of a total of 48871) accounts using "`111222tianya`" as the password from Tianya to 7k7k, rather than to copy all the accounts (i.e., $64822/2$) using "`111222tianya`" as the password from 7k7k to Tianya and further reproduces $16460 (= 48871 - 64822/2)$ such accounts.

After removing 7.82 million joint accounts from 7k7k, we found that all of the passwords in the remaining 7k7k dataset occur even times (e.g., 2, 4 and 6). This is expected, for we observe that in 7k7k half of the accounts are of the form {user name, password}, while the rest are of the form {email address, password}, and it is likely that both forms are directly derived from the form {user name, email address, password}. For instance, both {wanglei, wanglei123} and {wanglei@gmail.com, wanglei123} are actually derived from the single account {wanglei, wanglei@gmail.com, wanglei123}. Consequently, we further divide 7k7k into two equal parts and discard one part. The detailed information on data cleansing is summarized in Table I of the main text.

Weaknesses in existing studies. In 2014, Li et al. [7] has also exploited the datasets Tianya and 7k7k. However, contrary to what we have done above, they think that the 3.91M joint accounts are copied from 7k7k to Tianya. Their main reason is that, when dividing these two datasets into the reused passwords group (i.e., the joint accounts) and the not-reused passwords group, they find that "the proportions of various compositions are similar between the reused passwords and the 7k7k's not-reused passwords, but different from Tianya's not-reused passwords". However, they have *never* explained what these "various compositions" are. Their explanation also cannot answer the critical question: why are there so many 7k7k users using "`111222tianya`" as their passwords? Hence, it would be more reasonable that they had removed 3.91*2 million joint accounts from 7k7k but not 3.91 million ones from Tianya. In addition, they did not observe the extremely *abnormal* fact that all the passwords in 7k7k occur even times. *Such contaminated data would highly lead to inaccurate results and unreliable comparisons*. For example, Li et al. [7] reported that there are 9,477,069 (30.67%) passwords in Tianya with consecutive exactly six digits, yet the actual value is 2.5 times larger: 23,358,248 (75.59%). For another example, Li et al. reported that there are 32.41% of passwords in 7k7k containing dates in "YYYMMDD", yet the actual value is 6 times lower: 5.42%.

We have reported this issue to the authors of [7], they responded to us and acknowledged this flaw in their journal version [6]. Unfortunately, Han et al. [6] still fail to clean the datasets properly in the journal version and address our revealed issue in an oversimplified (and crude) way: "we removed these duplicate passwords from both websites" [6]. As their journal version [6] is essentially a verbatim of [7], we mainly use [7] for comparison and discussion.

APPENDIX B
DETAILED INFO ABOUT OUR 22 SEMANTIC DICTIONARIES

We now detail how to construct our 22 semantic-based dictionaries, in order to make our work reproducible as well as to facilitate the community. "English_word_lower" is from http://bit.ly/2b2uPBX and it contains about 58,000 popular lower-case English words. "English_lastname" is a dictionary consisting of 18,839 last names with over 0.001% frequency in the US population during the 1990 census, according to US Census Bureau [3]. "English_firstname" contains 5,494 most common first names (1,219 male and 4,275 female names) in US [3]. "English_fullname" is a cartesian product of "English_firstname" and "English_lastname", consisting of 1.04 million most common English full names.

To get a Chinese full name dictionary, we employ the 20 million hotel reservations dataset [5] leaked in Dec. 2013. The Chinese family name dictionary includes 504 family names which are officially recognized in China. Since the first names of Chinese users are widely distributed and can be almost any combinations of Chinese words, we do not consider them in this work. As the names are originally in Chinese, we transfer them into Pinyin without tones by using a Python procedure from https://pypinyin.readthedocs.org/en/latest/ and remove the duplicates. We call these two dictionaries "Pinyin_fullname" and "Pinyin_familyname", respectively.

"Pinyin_word_lower" is a Chinese word dictionary known as "SogouLabDic.dic", and "Pinyin_place" is a Chinese place dictionary. Both of them are from [9] and also originally in Chinese, and we translate them into Pinyin in the same way as we tackle the name dictionaries. "Mobile_number" consists of all potential Chinese mobile numbers, which are 11-digit strings with the first seven digits conforming to predefined values and the last four digits being random. Since it is almost impossible to build such a dictionary on ourselves, we instead write a Python script and automatically test each 11-digit string against the mobile-number search engine http://ku.13131313131.com/.

As for the birthday dictionaries, we use date patterns to match digit strings that might be birthdays. For example,

"YYYYMMDD" stands for a birthday pattern that the first four digits indicate years (from 1900 to 2014), the middle two represent months (from 01 to 12) and the last two denote dates (from 01 to 31). "PW with a l^+-letter substring" is a subset of the corresponding dataset and consists of all passwords that include a letter substring *no shorter than* l, and similarly for "PW with a l^+-digit substring".

APPENDIX C
A SUBTLETY ABOUT GOOD-TURING SMOOTHING ON PASSWORD CRACKING

There is a subtlety to be noted when implementing the Good-Turing (GT) smoothing technique. We denote f to be the frequency of an event, and N_f to be the frequency of frequency f. According to the basic GT smoothing formula, the probability of a string "$c_1c_2\cdots c_l$" in a Markov model of order n is denoted by

$$P(\text{“}c_1c_2\cdots c_{l-1}c_l\text{”}) = \prod_{i=1}^{l} P(\text{“}c_i|c_{i-n}c_{i-(n-1)}\cdots c_{i-1}\text{”}), \quad (1)$$

where the individual probabilities in the product are computed empirically by using the training sets. More specifically, each empirical probability is given by

$$P(\text{“}c_i|c_{i-n}\cdots c_{i-1}\text{”}) = \frac{S(count(c_{i-n}\cdots c_{i-1}c_i))}{\sum_{c\in\Sigma} S(count(c_{i-n}\cdots c_{i-1}c))}, \quad (2)$$

where the alphabet Σ includes 10 printable numbers on the keyboard plus one special end-symbol (i.e., c_E) that denotes the end of a password, and $S(\cdot)$ is defined as:

$$S(f) = (f+1)\frac{N_{f+1}}{N_f}. \quad (3)$$

It can be confirmed that this kind of smoothing works well when f is small, yet it fails for passwords with a high frequency because the estimates for $S(f)$ are not smooth. For instance, `12345` is the most common 5-character string in the Rockyou dataset and occurs $f = 490,044$ times. Since there is no 5-character string that occurs 490,045 times, N_{490045} will be zero, implying the basic GT estimator will give a probability 0 for $P(\text{“}12345\text{”})$. A similar problem regarding the smoothing of frequency of passwords has been identified in [2].

There have been various improvements suggested in linguistics to cope this problem, among which is Gale and Hill's "simple Good-Turing smoothing" [4]. This improvement is famous for its simplicity and accuracy. This improvement (denoted by SGT) takes two steps of smoothing. Firstly, SGT performs a smoothing for N_f:

$$SN(f) = \begin{cases} N(1) & \text{if } f = 1 \\ \dfrac{2N(f)}{f^+ - f^-} & \text{if } 1 < f < \max(f) \\ \dfrac{2N(f)}{2f - f^-} & \text{if } f = \max(f) \end{cases} \quad (4)$$

where f^+ and f^- stand for the next-largest and next-smallest values of f for which $N_f > 0$. Then, SGT performs a linear regression for all values SN_f and obtains a Zipf distribution: $Z(f) = C\cdot(f)^s$, where C and s are constants resulting from regression. Finally, SGT conducts a second smoothing by replacing the raw count N_f from Eq.3 with $Z(f)$:

$$S(f) = \begin{cases} (f+1)\dfrac{N_{f+1}}{N_f} & \text{if } 0 \le f < f_0 \\ (f+1)\dfrac{Z(f+1)}{Z(f)} & \text{if } f_0 \le f \end{cases} \quad (5)$$

where $t(f) = |(f+1)\cdot\frac{N_{f+1}}{N_f} - (f+1)\cdot\frac{Z(f+1)}{Z(f)}|$ and $f_0 = \min\left\{ f\in\mathbb{Z} \,\middle|\, N_f > 0, t(f) > 1.65\sqrt{(f+1)^2\frac{N_{f+1}}{N_f^2}(1+\frac{N_{f+1}}{N_f})} \right\}$.

In 2014, Ma et al. [8] introduced GT smoothing into Markov-based attacks to facilitate more accurate generation of password guesses, yet little attention has been paid to the unsoundness of GT for high frequency events as illustrated above. To the best of our knowledge, we for the first time well explicate the combination uses of GT and SGT in Markov-based password cracking.

REFERENCES

[1] J. Bonneau, "The science of guessing: Analyzing an anonymized corpus of 70 million passwords," in *IEEE S&P 2012*, pp. 538–552.

[2] ——, "Guessing human-chosen secrets," Ph.D. dissertation, University of Cambridge, 2012.

[3] R. A. Butler, *List of the Most Common Names in the U.S.*, Jan. 2016, http://names.mongabay.com/most_common_surnames.htm.

[4] W. Gale and G. Sampson, "Good-turing smoothing without tears," *Journal of Quantitative Linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[5] J. Goldman, *Chinese Hackers Publish 20 Million Hotel Reservations*, Dec. 2013, http://bit.ly/2aVKyBw.

[6] W. Han, Z. Li, L. Yuan, and W. Xu, "Regional patterns and vulnerability analysis of chinese web passwords," *IEEE Trans. Inform. Foren. Secur.*, vol. 11, no. 2, pp. 258–272, 2016.

[7] Z. Li, W. Han, and W. Xu, "A large-scale empirical analysis on chinese web passwords," in *Proc. USENIX SEC 2014*, pp. 559–574.

[8] J. Ma, W. Yang, M. Luo, and N. Li, "A study of probabilistic password models," in *Proc. IEEE S&P 2014*, pp. 689–704.

[9] *Sogou Internet thesaurus*, Sogou Labs, April 17 2016, http://www.sogou.com/labs/dl/w.html.