Regional Patterns and Vulnerability Analysis of Chinese Web Passwords

Weili Han, *Member, IEEE,* Zhigong Li, Lang Yuan, and Wenyuan Xu, *Member, IEEE,*

Abstract—Current research on password security pays much attention on users who speak Indo-European languages (English, Spanish, etc.), and thus the countermeasures are heavily influenced by Indo-European speakers' choices as well. However, languages have a strong impact on passwords. Analysis without considering other languages (e.g. Chinese) might lead to some biased results, such as, Chinese passwords are one of the most difficult ones to guess. We believe that such a conclusion could be biased because, to the best of our knowledge, little empirical study has examined regional differences of passwords at a large scale, especially on Chinese passwords. In this paper, we comprehensively study the differences between passwords from Chinese and English-dominant users, leveraging over 100 million leaked and publicly available passwords from Chinese and international websites in recent years. We find that Chinese prefer digits when composing their passwords while English-dominant users prefer letters, especially lowercase letters. However, their strength against password guessing is similar. Second, we observe that both groups of users prefer to use the patterns that they are familiar with, *e.g.*, Chinese Pinyins for Chinese and English words for English-dominant users. In particular, since multiple input methods require various sequences of letters to enter the same Chinese characters, we evaluate the impacts of various Chinese input methods, in addition to Pinyin. Third, we observe that both Chinese and English-dominant users prefer their conventional format when they use dates to construct passwords. Based on these observations, we improve two password guessing methods: PCFG (Probabilistic Context-Free Grammar)-based and Markov model-based password guessing methods. For the PCFG-based method, the guessing efficiency increases by up to 48% after inserting Pinyins (about 2.3% more entries) into the attack dictionary and inserting the observed composition rules into the guessing rule set. For the Markov-model-based method, the guessing efficiency increases by up to 4.7% after we increase the percentage of Pinyins in the training set. Our research sheds light on understanding the impact of regional patterns on passwords.

Index Terms—Authentication, Password, Chinese Users, Password Patterns, PCFG, Markov Model.

1 Introduction

Since the invention of passwords, they have remained as the most widely used credentials for authenticating Web users around the world [1], including the users that do not speak English. Researchers have long been studying how to attack or protect users' passwords. Although insightful, most existing work focuses on passwords of English-dominant (English for short) users. Little work has studied the impact of regional convention and languages on password selection. One exception is Bonneau [2], who studied password strength based on languages by performing an empirical study on `Yahoo` users and concluded that Chinese passwords are among the hardest ones to guess. We believe his finding is biased because of the dataset (*i.e.*, `Yahoo` users are familiar with English). In this paper, we analyze passwords of non-English speakers, specifically, Chinese users, which represent *649* million Internet users at the end of 2014 [3], and compare them with passwords of English users.

In this paper, we leverage over 100 million leaked and publicly available passwords from several popular Chinese websites (`CSDN` [4], `Tianya` [5], `Duduniu` [6], `7k7k` [7], and `178.com` [8]) and English websites (`RockYou` [9] and `yahoo` [10]). These Chinese websites only provide Chinese webpages, and we consider their users as *Chinese users*, just as the one defined in prior work [2][11]. In addition, English websites mainly intend to serve users who are familiar with English, and we consider the users of `RockYou` and `Yahoo` as *English users*. These leaked passwords can help us to comprehensively identify the differences between Chinese and English passwords. Note that, these websites (except `Duduniu`, which is an e-commerce website) provide similar social services, such as web portal, online communities, social networking, online forums, etc. Thus, we consider them comparable and having similar influence on their users when choosing passwords. This makes their password data corpus promising for studying the impact of languages on password composition.

We designed analysis tools and leveraged the cross-regional guessing resistance indicators (such as *α-work-factors* [12] and *β-success-rates* [13]) to find the differences between accounts of multiple websites, and find the preference of the two groups of users. Then, we improved the efficiency of the Probabilistic Context-Free Grammar (PCFG) based password guessing method [14] and the Markov model based password guessing method [11] by adding regionally preferred patterns (*i.e.* Pinyins and dates) into their corresponding sources for guessing. We summarize our findings and main contributions as follows:

- **Different Characters Sets and Patterns.**: Chinese users prefer digits in their passwords, while English users prefer letters, especially lowercase letters. The

- *W. Han, Z. Li and L. Yuan are with Software School, Fudan University, Shanghai Key Laboratory of Data Science, Fudan University.*
- *W. Xu is with Department of Electronic Engineering, Zhejiang University.*
- *W. Han is also with Key Lab of Information Network Security, Ministry of Public Security (courtesy).*

This paper is the extended version of the paper: "Zhigong Li, Weili Han, Wenyuan Xu. A Large-Scale Empirical Analysis of Chinese Web Passwords, In Proceedings of the 23rd USENIX Security Symposium, San Diego, August 20-22, 2014: 559-574."

TABLE 1
Basic information of leaked passwords of the websites that are analyzed in this paper. We removed the duplicate accounts between `Tianya` and `7k7k` from both `Tianya` dataset and `7k7k` dataset.

	Language	Site Address	Amount	Distinct Accounts
CSDN	Chinese	http://www.csdn.net/	6,428,629	6,423,483
Tianya	Chinese	http://www.tianya.cn/	30,179,474	26,223,020
Duduniu	Chinese	http://www.duduniu.cn/	16,282,969	15,131,833
7k7k	Chinese	http://www.7k7k.com/	19,138,270	12,107,865
178.com	Chinese	http://www.178.com/	9,072,824	9,072,804
RockYou	English	http://www.rockyou.com/	32,603,048	32,602,882
Yahoo	English	http://www.yahoo.com/	442,837	442,837
Total			**114,148,051**	**102,004,724**

pure digit patterns are more popular in Chinese users than the English ones. However, the password strength against guessing is similar for both groups and thus both groups share similar security concerns in protecting passwords.

- **Patterns of Languages and Date Formats.**: Both Chinese and English users prefer to use language-related patterns for passwords. That is, Chinese users prefer Chinese Pinyins, and will use special Chinese phrases (e.g., Chinese idioms, ancient poetry, and Chinese location names) to compose passwords, and Chinese users like to create their passwords in form of *acronyms*. As for dates, both groups prefer their conventional formats: Chinese users prefer dates with the year at the beginning and English users prefer dates with the year at the end.
- **Regional Patterns can Improve the Efficiency of Password Guessing.**: Based on our observations, we add 20,000 Pinyins into the dictionary and add the guessing rules, resulting in an improvement of efficiency by up to 48% in guessing Chinese passwords using a PCFG based guessing method. Moreover, by replacing 200,000 letter-only passwords in the training set with Pinyins, we manage to improve efficiency of a Markov model based guessing method by up to 4.7%.

The aforementioned findings shed light on understanding how regional conventions can influence password composition. In addition, they can improve password guessing efficiency and can serve as a guidance of web administrators from different regions for protecting their users' accounts.

The rest of the paper is organized as follows: Section 2 summarizes our observations on the differences between passwords from Chinese and English users. Section 3 presents the results of guessing using modified Bonneau's methods [2], PCFG based method [14] and Markov-based method [11]. In Section 4, we discuss the related work and conclude in Section 5.

2 Regional Differences on Passwords

2.1 Dataset Setup

To discover the differences between the passwords of Chinese and English users, we analyzed a corpus of over 100 million passwords from multiple websites that are in Chinese and English, respectively. All the leaked passwords are publicly available for downloading. During our research, we followed the ethical practice and never utilized the leaked passwords for reasons other than understanding the overall statistical observation of passwords.

TABLE 2
The most popular passwords and their occurrence percentages.

	Chinese	English
1	123456 (2.12%)	123456 (0.88%)
2	123456789 (0.66%)	12345 (0.24%)
3	111111 (0.57%)	123456789 (0.23%)
4	12345678 (0.40%)	password (0.18%)
5	000000 (0.36%)	iloveyou (0.15%)

At the end of 2011, an incident known as *CSDN Password Leakage Incident* happened, and passwords from five websites, including `CSDN`, `Tianya`, `Duduniu`, `7k7k` and `178.com`, were leaked in several consecutive days. The total number of leaked accounts is over 80 million, and all the leaked passwords are in plaintext. We summarize the website information in Table 1.

`CSDN` [4] is one of the most popular Chinese IT professional communities, similar to `MSDN`. `Tianya` [5] is one of the largest online forums and blogs in China. `7k7k` [7] and `178.com` [8] are two websites providing game information and online flash games. `Duduniu` [6] is a commercial site that mainly sells management software platforms for Internet bars. It is worth noting that all these websites are extremely popular in China, among which `CSDN` and `Tianya` have been ranked top 1,000 in Alexa Top Global Sites recently. Thus, their users cover a large percentage of Internet users in China.

Besides their popularity, the leaked password data corpus is promising for understanding the language impact on passwords because few password policies are enforced in the above five Chinese websites before the leakage according to our investigation. For example, `CSDN` allows a password with as few as five digits, and such a rule remains unchanged even after the password leakage event. Furthermore, `Tianya` allows passwords as short as six characters since it was founded. Thus, the leaked password data corpus represents the password set that was composed with little influence from password policies.

Password leakage events happened to English websites as well. In 2009, attackers broke into the database of `RockYou` and released 32 million passwords (in plaintext) to the public. `RockYou` is a developers' website for social games. It provides applications (*e.g.*, voice mails, photo stylization) for social networking sites, including `Facebook`, `MySpace`, and `Friendster`. The `RockYou` dataset has been studied extensively as a main source of real passwords in the literature [11][15][16][17], *etc.* In 2012, `Yahoo`'s accounts were leaked. A hacking group *DD3Ds Company* utilized a union-based SQL injection to obtain login details

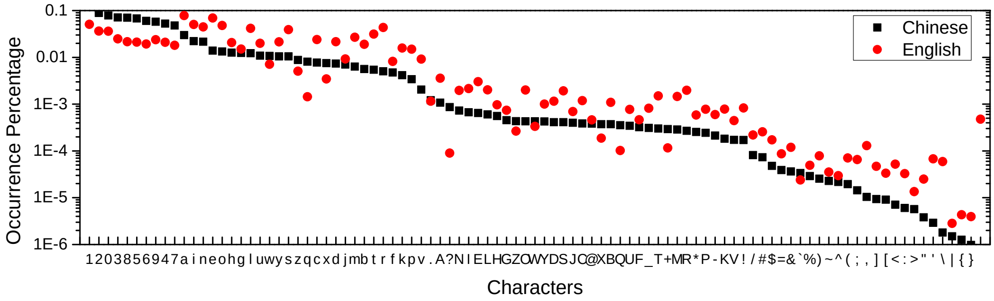


Fig. 1. Character distribution, i.e., the occurrence percentage of each character for Chinese and English passwords. The characters are arranged in descending order according to the percentages in Chinese passwords.

of about 450 thousand user accounts. Although there are many leaked passwords from other English websites, like `eHarmony`, `Gamigo`, `LinkedIn`, etc., it appears that they might have been filtered (*i.e.*, most passwords are unique), which might make the statistical patterns biased. Therefore, we just choose `RockYou` and `yahoo` as representatives of English websites. Considering that the size of `RockYou` data set is larger than `yahoo` (32,603,048 *vs* 442,837), our results could be biased towards the characteristics of `RockYou`. Nevertheless, we believe that the `RockYou` dataset can at least partially reflect the password patterns and vulnerabilities of English users, as the size of this dataset is over 32 million.

The raw files contain duplication and blank passwords that can affect the analysis. For instance, we detected that there is a large number of common entries between `Tianya` and `7k7k`. Thus, we removed these duplicate passwords from both websites. Details are described in Appendix A. After removing the accounts with blank passwords and filtering out duplicate accounts, we obtained 102,004,724 accounts, as detailed in Table 1.

2.2 Password Comparison

2.2.1 *The Most Popular Passwords*

We list the five most popular passwords of Chinese and English users in Table 2, from which we have the following observations:

- In total, the five most popular passwords constitute 4.11% of all Chinese passwords and 1.69% of all English passwords, which shows that Chinese passwords are more clustered.
- Interestingly, although in English datasets there are a larger number of letter-only passwords (see details in Section 2.2.3), the top 3 most popular passwords are digit-only. In addition, both groups share similar popular passwords, *e.g.*, 123456 and 123456789.

2.2.2 *Character Distribution*

To understand the frequency of each character, which includes letters (a-z, A-Z), digits (0-9), and symbols (all printable characters except digits and letters), we analyzed the percentage of each character for Chinese and English passwords and depict them in Figure 1, where the characters are arranged in descending order according to the percentages in Chinese passwords.

- **Digits**. In Chinese passwords, the most frequently used characters are digits. Although English users do not use digits as frequently as Chinese users do, digits are among the most frequently used characters.
- **Letters**. In general, Chinese passwords use letters less frequently than English passwords do. In addition, some letters exhibit similar usage percentages for both groups of passwords, *e.g.*, the letter *a* is the mostly used letter in both groups. Some letters show distinct usages, *e.g.*, the letter *q* is frequently used in Chinese passwords but is much less used in English passwords; the letter *r* is much more popular in English passwords than in Chinese ones. This is due to the word patterns in either language. For instance, the letters *q* and *a* are popular building blocks of Pinyins, but the letter *r* is not. We will discuss Chinese Pinyins and English words in detail in Section 2.2.5.
- **Symbols**. Symbols are less used in both Chinese and English passwords in general. Interestingly, for both groups of passwords, several symbols share the similar usage percentages: the symbol dot (.) is the most frequently used, and symbols like left brace ({) and right brace (}) are less likely to be used. However, regional differences on symbol usages do exist: the question mark (?) is more frequently used in Chinese passwords than in English passwords.

2.2.3 *Compositions and Structures of Passwords*

To understand the structures of passwords in both groups, we analyzed passwords in two ways. (1) We divided passwords according to their compositions and calculated the percentages in seven categories (shown in Table 3). The categories are pure digits, pure letters, digits and letters, letters and symbols, etc. (2) We calculated the percentages of different types of password structures utilizing representations in the Probabilistic Context-Free Grammar [14]. For example, the structure of *Johns0n!* is modeled as *ULLLLDLS* (U = uppercase, L = lowercase, D = digit, and S = symbol). The structure comparison of both password groups is shown in Table 4 where *# of Structures/10K* refers to the number of different structures in every 10,000 passwords. *The most popular structure* is the one that appears the most in the data set. From Tables 3 and 4, we can obtain the following observations:

- A majority (around 50% on average) of Chinese users prefer digit-only passwords. This could be due to their language using habits. Chinese characters cannot be entered directly as a password, and digits

TABLE 3
Compositions of passwords. The percentages outside parentheses are the ones counting both uppercase and lowercase letters, and the percentages inside parentheses are the ones counting only lowercase letters. The sum of the percentages in one row is slightly smaller than one, because symbol-only passwords are not listed, and they only account for a small percentage.

	Digit -only	**Letter-only (Lowercase-only)**	**Letter+Digit (Lowercase+Digit)**	**Letter+Symbol (Lowercase+Symbol)**	**Symbol +Digit**	**Letter+Digit+Symbol (Lowercase+Digit+Symbol)**
CSDN	45.06%	12.39% (11.68%)	39.02% (35.60%)	0.50% (0.42%)	0.61%	2.39% (2.04%)
Tianya	64.56%	10.20% (9.89%)	23.12% (21.27%)	0.25% (0.22%)	0.71%	1.14% (1.01%)
Duduniu	32.86%	11.76% (11.08%)	53.69% (50.93%)	0.52% (0.48%)	0.17%	0.92% (0.80%)
7k7k	61.67%	11.40% (11.05%)	25.24% (22.94%)	0.12% (0.11%)	0.26%	1.19% (0.37%)
178.com	48.07%	9.17% (9.00%)	42.11% (41.25%)	0.06% (0.06%)	0.31%	0.27% (0.26%)
RockYou	15.93%	44.04% (41.68%)	36.22% (33.17%)	1.91% (1.64%)	0.16%	1.71% (1.44%)
Yahoo	5.89%	34.64% (33.08%)	56.62% (50.60%)	0.62% (0.49%)	0.04%	2.18% (1.38%)

TABLE 4
Structures of passwords. # of structures/10K refers to the number of different structures in every 10,000 passwords, and the other two columns contain the structures and occurrence percentages of the most popular structures in both Chinese websites and English ones. D represents a digit, and L represents a lowercase letter.

	# of Structures/10K	**Most Popular Structure**	**Most Popular Structure%**
CSDN	884	DDDDDDDD	21.50%
Tianya	756	DDDDDD	30.10%
Duduniu	610	DDDDDD	7.25%
7k7k	590	DDDDDD	18.37%
178.com	459	DDDDDD	15.48%
RockYou	803	LLLLLL	5.40%
Yahoo	1165	LLLLLL	9.19%

TABLE 5
Percentages of passwords with different keyboard patterns. Most passwords of the *Same Row* pattern are digit-only. The numbers in the parentheses represent passwords that have the *Same Row* pattern and are not digit-only.

	Chinese	**English**
Same Row	8.38% (0.55%)	2.42% (0.25%)
Zig Zag	0.27%	0.06%
Snake	0.27%	0.08%

appear to be the best candidate when users are creating new passwords. Although Chinese users can use Pinyins as discussed in Section 2.2.5, digits seem to be more convenient. As shown in Table 4, *DDDDDD* is the dominant structure in most Chinese websites. For `CSDN`, the structure *DDDDDDDD* is the top selection, and *DDDDDD* is ranked at 14. A six-digit number may be an ATM PIN, a birthday, or the last six digits of citizen ID cards. We will discuss details in Section 2.2.7.

- For both password groups, a good portion of passwords contain both letters and digits, and no obvious differences seem to exist between these websites. The owners of the passwords in this category could be users who are concerned with password security but are unwilling to be bothered with symbols.

2.2.4 Keyboard Patterns

Sometimes, users prefer to create their passwords according to keyboard patterns [18]. Thus, we analyzed the percentages of three primary keyboard patterns. Note that Chinese users utilize standard English keyboards, i.e., they use the same ones as English users.

- **Same Row**: The same row passwords are formed by a consecutive sequence of characters in the same row on keyboard, *e.g.*, asdfgh.
- **Zig Zag**: The zig-zag passwords are formed by a sequence of characters, where each key is adjacent to the next one but not in the same row, *e.g.*, qawsxd.
- **Snake**: The snake passwords consist of a sequence of characters whose keys are adjacent on keyboards yet they are neither in the *Same Row* nor *Zig Zag*, *e.g.*, zxcfgh.

We believe these three keyboard patterns are representative because if a password consists of a sequence of adjacent characters on a keyboard, which is easy to remember, it can be categorized as one of them.

Identify Keyboard Patterns. In order to automatically classify passwords into the aforementioned three categories, we assign a coordinate to each character on the keyboard. We define that the x-axis increases from left to right and the y-axis increases from top to bottom. For example, the coordinate of 1 (and !) is (1,0), and the coordinates of q, a, and z are (1,1), (1,2), and (1,3), respectively. Provided the coordinates of the characters, we can determine if a password is in a specific keyboard pattern by determining whether two letters are adjacent in the same row or column.

Result. The statistics are shown in Table 5, in which we observe that more than 8% of Chinese passwords are composed according to keyboard patterns but fewer English passwords are. After removing all digit-only passwords, the keyboard pattern passwords reduce to about 1%. This is because most passwords of the *same row* pattern are digit only. Nevertheless, Chinese users tend to use keyboard pattern passwords more often than English users do, e.g., there are 0.21% more *Zig Zag* passwords for Chinese than English users. The reason could be that keyboard patterns are easy to create and remember for Chinese users who are unfamiliar with English.

2.2.5 Chinese Pinyins and English Words

Chinese Pinyin was developed in 1950s and is designed to represent the pronunciation of Chinese characters. Although there are lots of dialects in China, the Pinyins for characters are the same. As Chinese computer users are trained with Pinyin since primary school, they can be expected to be familiar with it. Pinyin is the most popular method to input Chinese characters into a computer because it requires almost no extra training for Chinese. Typically, a Chinese character is entered by multiple keystrokes and ignoring the tones, a word in Pinyins uses a set of 21 sounds representing the beginning of the word called initials, and a set of 37 sounds representing the end of the word called finals. These two combine to form about 420 different basic Pinyin elements [19].

Identify Pinyins or English Words. We can determine whether a password is composed of Chinese Pinyins or English words by string matching. For example, a password *nihao* is composed of Pinyins *ni* and *hao* and a password *helloworld* is composed of English words *hello* and *world*. For English words, we chose the Oxford English Dictionary [20] and extracted more than 20,000 commonly used English words. Notice that, when identifying Pinyins and English words, we chose the longest run of letters in the passwords. For example, *iloveyou* may have a higher frequency than *love* because *iloveyou* is considered as a whole.

To improve the matching efficiency, we use Trie (or prefix tree) to identify if the passwords are composed of Chinese Pinyins or English words. We first construct Trie by inserting Chinese Pinyins or English words one by one. With the Trie, we can identify if a password is composed of Chinese Pinyins or English words. In our experiments, we constructed two Tries: one is constructed out of Chinese Pinyins, and the other is built based on more than 20,000 commonly used English words. We designed an algorithm to identify if a password is composed of one or more Chinese Pinyins or English words. The algorithm will try to match the password with the known strings from Trie recursively.

Note that it is challenging to extract Chinese Pinyins or English words accurately. First, a user may intentionally compose a password that is semantically meaningless even if it appears to be a composition of Chinese Pinyins or English words. Furthermore, some letter sequences in passwords are ambiguous. They can be interpreted as either a composition of Pinyins or a composition of English words. For example, “likeyou” can be interpreted as two English words (i.e., “like” and “you”) or three Chinese Pinyins (i.e., “li”, “ke” and “you”). Another example is “linglong”, where both “ling” and “long” can be either English words or Chinese Pinyins. Granted that according to the semantics, it is more likely that “likeyou” consists of two English words and “linglong” consists of two Chinese Pinyins. To avoid overdoing, in our experiments, we didn’t perform semantic analysis, and we did not identify the ambiguity. Therefore, our statistics are, to some extent, conservative evaluations.

Result. We performed statistical analysis of the usage of Chinese Pinyins and English words in two aspects. Firstly, we calculated the percentages of passwords that are composed of Chinese Pinyins or English words out of all the letter-only passwords. Secondly, we calculated the percentages of Pinyins or English words out of all the mixed passwords (*i.e.*, passwords containing at least two types of characters with one of them being letters). The results are shown in Table 6. Table 7 lists the top five most popular Chinese Pinyins and English words. From Tables 6 and 7, we draw the following conclusions:

- Among all the letter-only passwords, Pinyins are the dominant patterns for Chinese users in composing their passwords, while English words dominate the English passwords. Even when we consider all of the passwords, these patterns are still the basic building blocks for a large portion of passwords, *i.e.*, more than 10% English passwords contain English words, and about 5% of Chinese passwords consist of Pinyins.

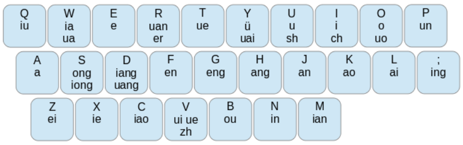


Fig. 2. One keyboard layout of Microsoft Shuangpin

- Interestingly, it seems that love is always the main theme of human beings. As shown in Table 7, *iloveyou* and *love* are ranked at the second and the third in English passwords. Meanwhile, *woaini* is the top ranked Pinyin, meaning *I love you* in Chinese.
- The Pinyins of names are widely used in Chinese passwords. Chinese surnames *li*, *wang*, *zhang* and *liu*, four of the most popular surnames in China, are ranked among top 5 of the most popular Pinyins in Chinese passwords in Table 7. Note that it is difficult to identify first names in Chinese, since they could be almost any combinations of Pinyins.
- The website names appear to be an important part of Chinese passwords. For example, *tianya*, which is a website name, is ranked at the eighth (0.57%) in Chinese Pinyins. In fact, a similar result has also been revealed for English passwords in [21], where the name of the website appears in the top ten of the `computerbits` [22] and `RockYou` lists respectively.
- We found that some passwords from `RockYou` and `Yahoo` are composed of Pinyins, and we suspect that the owners are Chinese. Most of these Pinyins do not map to meaningful expressions, and thus we suspect they are names. For example, *yaowei*, which is composed of Pinyins *yao* and *wei*, is most likely to be a name.

Chinese Shuangpin Input Methods. To reduce the time of keystrokes and speed up the character input, Chinese users can use Shuangpin where each key in a keyboard is mapped to several intials and finals as shown in Figure 2. Shuangpin allows users to type keyboard only twice for any Chinese character (e.g., Zhong (type ‘vs’) Guo (type ‘go’)). The method is popular because it is easy-to-use and efficient. Nowadays, there are several different customized layouts for Shuangpin: Shuangpin (MS, shown in Figure 2) [23], Shuangpin (ZiRanMa) [24], Shuangpin (ZiGuang) [25], Shuangpin (JiaJia) [26] and Shuangpin (ZhiNengABC) [27]. We calculated all the occurrence frequencies of passwords that are typed with input methods mentioned above. The result is shown in Table 8. Note that, the layouts of the five Shuangpin input methods share some similarities. Thus, the sum of the percentages of every individual input method is larger than their corresponding percentage of *Total Shuangpin*.

Supposing that percentages of Shuangpin in English websites represent false positive results, the difference of percentages between Chinese websites and English websites can confirm the existence of a large amount of Shuangpins in Chinese websites. However, compared with the result in

TABLE 6
Percentages of the passwords that contain Chinese Pinyins or English words. Mixed passwords refer to the ones that contain at least two types of characters with one of them being letters. The percentages inside the parentheses are the proportions out of the entire password dataset, and the percentages ahead of the parentheses are the ones out of the letter-only passwords or mixed passwords. For example, in the row of CSDN, 41.61% (5.15%) means that in the letter-only passwords, 41.61% are composed of Chinese Pinyins, and these passwords occupy 5.15% in the whole dataset of CSDN.

	Letter-only Passwords		**Mixed Passwords**	
	Chinese Pinyins%	English Words%	Chinese Pinyins%	English Words%
CSDN	41.61% (5.15%)	15.59% (1.93%)	25.49% (10.68%)	7.97% (3.34%)
Tianya	40.63% (4.15%)	10.39% (1.06%)	23.59% (5.78%)	6.05% (1.48%)
Duduniu	33.28% (3.91%)	15.35% (1.80%)	25.17% (13.87%)	6.48% (3.57%)
7k7k	45.52% (5.19%)	9.25% (1.05%)	21.24% (5.64%)	5.88% (1.56%)
178.com	57.31% (5.25%)	2.20% (0.20%)	23.49% (9.97%)	4.58% (1.94%)
RockYou	6.94% (2.99%)	25.47% (10.98%)	6.88% (2.61%)	28.11% (10.65%)
Yahoo	4.31% (1.46%)	34.92% (11.86%)	4.53% (2.59%)	27.99% (16.01%)

TABLE 7
The most popular Chinese Pinyins and English words. The percentage base for top Chinese Pinyins is all the Pinyins we extracted from letter-only and mixed passwords in five Chinese websites. Similarly, the percentage base for top English words is all the words we extracted from letter-only and mixed passwords in both English websites.

	Top Chinese Pinyins	**Top English Words**
1	woaini (1.49%)	password (1.28%)
2	li (1.11%)	iloveyou (0.98%)
3	wang (1.00%)	love (0.76%)
4	zhang(0.86%)	angel (0.59%)
5	liu (0.73%)	monkey (0.45%)

Table 6, the difference of percentages is not as apparent as that with respect to Pinyin. That is, Chinese users prefer the Pinyin input method to Shuangpin input methods when creating Chinese passwords.

2.2.6 Compositions with Special Phrases

Besides input methods, we also examine how many common phrases, such as Chinese idioms, ancient Chinese poetry and Chinese location names, are used to compose passwords. Dataset of original phrases is obtained from Sogou cell thesauruses [28]. These thesauruses are among the most popular ones to be downloaded and are officially recommended to use. In total, we collect 129,285 Chinese idioms, 22,395 ancient Chinese poetry fragments, and 195,987 Chinese location names. Then we convert these words into four different combination forms, which are "Pinyin", "Pinyin + initials", "initials" and "acronyms". Take the Pinyin input method for example, the location name "Shanghai" will be represented unchanged as "Shanghai". Form "Pinyin + initials" indicates the first word is written in Pinyin while the sequential words just be written as initials, like "Shangh". "Initials" requires the extraction of initials for every words. Consequently, "Shanghai" will appear as "shh". Last, "acronyms" means "Shanghai" will be given as "sh". Similarly, we calculate the percentage of passwords composed of words that belong to one of the four deformations in Pinyins. we find that among the four different combination forms, the two most popular ones are "initials" and "acronyms", while "Pinyin" and "Pinyin + initials" appear to be used less. Results of "initials" and "acronyms" are shown in Table 9 and 10.

From Tables 9 and 10, we can obtain the following observations:

- In general, in Chinese websites, the percentages of passwords that are composed of Chinese idioms, ancient Chinese poetry or Chinese location names are larger than those in English websites. For example, in Table 10, the percentage of Chinese idioms in CSDN is 1.28% while that in RockYou is 0.30%. Thus, we can infer that the three Chinese phrases are more or less taken into account when Chinese users choose their passwords.
- The use of Chinese idioms is more common than that of ancient Chinese poetry. Notably, the percentage of Chinese location names is always much higher than that of Chinese idioms and ancient Chinese poetry. In our experiment, the length of all idioms and poetry fragments is at least four words while the length of location names is unlimited, which to some extent, makes the location names have a higher percentage.

2.2.7 Dates

Given that digits are commonly used in passwords, we try to understand the meaning of these digits. Similar to prior work, we interpret digits as dates and analyze their format in this section. Previously, Veras *et al.* [29] investigated the usage of dates in passwords and showed that nearly 5% passwords in the RockYou [9] dataset consist of pure dates. In their later work [16], Veras *et al.* also showed that dates were popular in passwords by analyzing password semantics. Along the same line, in this subsection, we analyze the date formats used in passwords of Chinese and English users in details.

Date Format. We focused our attention on six-digit and eight-digit dates. We first extracted all consecutive sequences of exactly six or eight digits from these passwords, and then calculated the dates which are in the range from 1900 to 2099. We classified six-digit dates into three formats: *YYMMDD*, *MMDDYY*, and *DDMMYY*. Similarly, we classified eight-digit dates into *YYYYMMDD*, *MMDDYYYY* and *DDMMYYYY*. The results are shown in Tables 11 and 12. Note that there might be ambiguity when interpreting dates. For example, *11121987* may be interpreted as either *November 12, 1987* or *December 11, 1987*. In this case, we assigned the passwords to one of the formats according to the probability distribution of all the passwords that can be uniquely determined. For instance, if 20% of passwords that contain date can be uniquely identified as *MMDDYY* and

TABLE 8
Percentages of passwords composed of the forms of Chinese Shuangpin. All the letter-only passwords and the mixed passwords contribute to the denominator. Term *Total Shuangpin* means a password is counted once if it is composed of any of the five Chinese Shuangpin input methods.

	MS	ZiRanMa	ZiGuang	JiaJia	ZhiNengABC	Total Shuangpin
CSDN	14.55%	22.37%	19.69%	15.89%	18.82%	31.06%
Tianya	15.47%	23.73%	21.22%	16.78%	20.57%	33.34%
Duduniu	15.39%	22.12%	19.40%	16.12%	19.00%	29.66%
7k7k	14.24%	23.09%	20.07%	15.64%	19.42%	31.85%
178.com	20.36%	28.23%	21.12%	21.56%	18.81%	35.18%
RockYou	8.75%	11.25%	13.04%	9.96%	11.79%	21.21%
Yahoo	7.96%	9.92%	11.45%	9.13%	10.74%	20.26%

TABLE 9
Initials: Percentages of passwords that are composed of Chinese idioms', ancient Chinese poetry's and Chinese location names' initials. We only list the result of the Shuangpin(ZiGuang).

	Chinese Idioms	Ancient Chinese Poetry	Chinese Location Names
CSDN	1.39%	0.32%	45.88%
Tianya	1.46%	0.33%	47.85%
Duduniu	0.85%	0.18%	37.67%
7k7k	1.08%	0.23%	42.44%
178.com	1.50%	0.13%	44.00%
RockYou	0.37%	0.08%	18.98%
Yahoo	0.50%	0.08%	21.90%

TABLE 10
Acronyms: Percentages of passwords that are composed of Chinese idioms', ancient Chinese poetry's and Chinese location names' acronyms.

	Chinese Idioms	Ancient Chinese Poetry	Chinese Location Names
CSDN	1.28%	0.41%	37.29%
Tianya	1.30%	0.37%	38.40%
Duduniu	0.54%	0.14%	29.60%
7k7k	0.93%	0.25%	33.72%
178.com	0.85%	0.17%	35.63%
RockYou	0.30%	0.03%	16.42%
Yahoo	0.36%	0.06%	19.24%

TABLE 11
Statistics of **eight-digit** date patterns: the number of occurrences of eight consecutive digits and percentages of three date formats. The percentage bases are listed in the second column. Y=year, M=month and D=day. For example, 20130115 is in the format of *YYYYMMDD*.

	YYYYMMDD	MMDDYYYY	DDMMYYYY
CSDN (1,621,954)	29.24%	0.25%	0.43%
Tianya (3,639,517)	36.26%	0.35%	0.60%
Duduniu (1,700,329)	28.87%	0.28%	0.84%
7k7k (1,927,543)	33.52%	0.15%	0.36%
178.com (995,832)	30.46%	0.13%	0.19%
RockYou (929,987)	2.64%	7.70%	17.66%
Yahoo (6,981)	2.78%	12.00%	11.17%

TABLE 12
Statistics of **six-digit** date patterns: the number of occurrences of six consecutive digits and percentages of three date formats. The percentage bases are listed in the second column.

	YYMMDD	MMDDYY	DDMMYY
CSDN (809,050)	27.21%	4.04%	1.24%
Tianya (9,477,069)	23.93%	3.05%	1.19%
Duduniu (2,688,347)	17.84%	2.97%	1.78%
7k7k (2,858.320)	22.70%	2.27%	0.97%
178.com (2,525,254)	13.96%	1.72%	1.30%
RockYou (2,758,871)	5.63%	21.90%	18.42%
Yahoo (21,020)	4.66%	25.99%	7.77%

80% of them as *DDMMYY*. Then, we assigned 20% of the ambiguous passwords to *MMDDYY* and the other 80% to *DDMMYY*.

Furthermore, there may be false positive when considering a general six-digit number as a date. For example, *123123* could be *December 31, 1923*, but it is more likely to be just two consecutive *123*. Thus, we removed 30 six-digit numbers that might cause such type of false positive [1]. Granted that we could have introduced false negatives or cannot manage to remove all the false positives for sure, these 30 numbers represent the patterns that have special meanings or are easy to remember, and more likely they do not map to any dates. For instance, "520520" has a similar sound as "i love you i love you" in Chinese. Thus, we believe that eliminating them will increase the accuracy of our statistics.

Tables 11 and 12 show the results. For example, the 29.24% in the first row in Table 11 means that among the 1,621,954 eight-digit numbers, 29.24% of them are in the format of *YYYYMMDD*. We can conclude that Chinese users prefer to use the format *YYYYMMDD* and *YYMMDD*. This conforms with Chinese conventions where people prefer to begin dates with years. On the contrary, a majority of English users prefer to end the date with years.

Password Composition. What are the compositions of passwords that contain dates? Are they composed of pure digits or mixed with letters? We calculated the percentages of digit-only, letter+digit, symbol+digit, letter+digit+symbol passwords out of all passwords that contain dates (both six-digit and eight-digit dates). Note that, the letter+digit passwords consist of one or more letters and digits. The

[1] The 30 six-digit numbers are: *111111, 123123, 111000, 112233, 100200, 111222, 121212, 520520, 110110, 123000, 101010, 111333, 110120, 102030, 110119, 121314, 521125, 120120, 010203, 122333, 121121, 101101, 131211, 100100, 321123, 110112, 112211, 111112, 520521, 110111.*

TABLE 13
Compositions of passwords that contain dates. The percentages outside parentheses are the ones counting both uppercase and lowercase letters, and the percentage inside parentheses are the ones counting only lowercase letters.

	Digit-only	Letter+Digit (Lowercase+Digit)	Symbol+Digit	Letter+Digit+Symbol (Lowercase+Digit+Symbol)
CSDN	51.98%	45.59% (41.36%)	0.50%	1.93% (1.67%)
Tianya	78.84%	19.91% (18.69%)	0.31%	0.72% (0.65%)
Duduniu	41.28%	58.17% (54.86%)	0.24%	0.31% (0.30%)
7k7k	75.20%	24.36% (23.60%)	0.16%	0.28% (0.25%)
178.com	50.91%	48.73% (48.07%)	0.32%	0.04% (0.04%)
RockYou	82.62%	16.52% (14.99%)	0.23%	0.63% (0.54%)
Yahoo	60.94%	38.03% (34.61%)	0.16%	0.86% (0.62%)

TABLE 14
Structures of date-containing passwords. L represents a lowercase letter and T represents a six-digit or eight-digit date. *Beginning*, *Middle*, and *End* shows the position of dates.

	Dominant Structure	Beginning	Middle	End
CSDN	LLLT (23.65%)	21.68%	4.32%	74.00%
Tianya	LLLT (20.29%)	27.33%	4.75%	67.07%
Duduniu	LLT (23.05%)	24.76%	1.36%	73.88%
7k7k	LLLT (23.43%)	24.76%	2.34%	73.06%
178.com	LLLT (30.14%)	22.30%	1.03%	76.67%
RockYou	LT (13.09%)	27.40%	3.91%	68.69%
Yahoo	LLT (19.64%)	22.66%	5.00%	72.34%

same applies to the symbol+digit, letter+digit+symbol passwords. As shown in Table 13, for all Chinese and English websites except Duduniu, most dates observed in our analysis are digit-only passwords, *i.e.*, when dates are used as passwords, they are used alone. What ranks the second is the passwords containing letters and digits. Note for Duduniu, the passwords that contain dates are more likely to contain both digits and letters than digits only. This could be because Duduniu is an e-commerce website and its users tend to choose a password with stronger strength, i.e., they tend to select passwords with both digits and letters, but not digits only.

Structures of date-containing passwords To understand the structures of date-containing passwords and the position of dates in passwords, we analyzed the passwords which contain dates, excluding digit-only passwords.

In Table 14, we list the dominant structure and categorize the position of the dates as *beginning*, *middle*, and *end*. For both Chinese and English users, they prefer to have dates appear at the end of passwords and incline to put several lowercase letters in the beginning.

2.3 Resistance to Guessing

Given the huge differences between Chinese and English passwords, a fundamental question is whether those differences lead to different levels of password strength. In this section, we examine password strength against password cracking.

2.3.1 Metrics to Measure Password Sets

We evaluated how resistant those passwords are against guessing by using the measurement metrics adopted by Bonneau [2][30], which are designed to evaluate the password strength in different regions.

As shown in Table 15, we briefly introduce these metrics: H_∞ is defined as *min-entropy*, a worst-case security metric for human-chosen passwords, i.e., when a user chooses the password with highest probability. G is defined as *guesswork*, representing the expected number of sequential guesses to find a password of an account if attackers proceed in an optimal order, i.e., trying passwords in descending order of the password probability. μ_α is called *marginal guesswork* or *α-work-factor*, which measures the expected number of guesses needed to succeed with probability α. *Marginal success rate* or *β-success rate*, λ_β, represents the probability that attackers can correctly guess the password of an account given β guesses. G_α, the *α-guesswork*, reflects the expected number of guesses per account to achieve a success rate α.

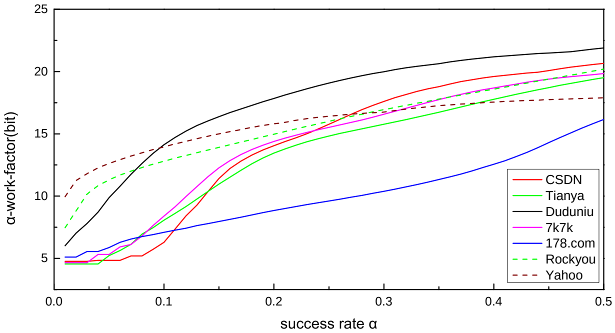


Fig. 3. The expected number of guesses needed to succeed with a success rate α (α-work-factors, $\tilde{\mu}_\alpha$) of all seven websites. The dash lines represent English websites and solid lines map to Chinese websites.

To be more intuitive to programmers and cryptographers, we can convert these metrics into units of bits by taking the logarithmic value. We use a tilde over each letter to denote the values that are converted into bits: $\tilde{G}$, $\tilde{\mu}_\alpha$ and $\tilde{G}_\alpha$.

In this section, we follow the same assumption proposed by Bonneau [2][30] that attackers know the exact distribution of the target password set and calculate the password strength, i.e., the attackers utilize the distribution of passwords to crack passwords in the same website. We call it *intra-site guessing*. In the next section, we relax the assumption, and we examine the guessing efficiency if the attackers are only aware of password distribution of other websites.

2.3.2 Resistance to Intra-Site Guessing

We summarize the calculated metrics for each website in Table 16 and Figure 3, and we draw the following observations:

- In Table 16, we observe that the *β-success-rates* (λ_5, λ_{10}) of RockYou and Yahoo are much lower than

TABLE 15
Metrics [2][30] list used in our analysis. $\mathcal{X}$ refers to the probability distribution of passwords; N refers to the number of distinct passwords in a password set; p_i refers to the probability of the i-th password in $\mathcal{X}$ where $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_N$.

Metric	**Formula**	**Term**	**Description**
$H_\infty(\mathcal{X})$	$-\log_2(p_1)$		Worst-case security metric
$G(\mathcal{X})$	$\sum_{i=1}^{N} p_i \cdot i$	*guesswork*	The expected number of sequential guesses to find the password of an account if attackers proceed in optimal order
$\tilde{G}(\mathcal{X})$	$\log_2(2 \cdot G(\mathcal{X}) - 1)$		Bit representation of $G(\mathcal{X})$
$\mu_\alpha(\mathcal{X})$	$\min\left\{j \in [1,N] \mid \sum_{i=1}^{j} p_i \geq \alpha\right\}$	*α-work-factor*	The expected number of guesses needed to succeed with probability α
$\tilde{\mu}_\alpha(\mathcal{X})$	$\log_2\left(\frac{\mu_\alpha(\mathcal{X})}{\lambda_{\mu_\alpha}}\right)$		Bit representation of $\mu_\alpha(\mathcal{X})$
$\lambda_\beta(\mathcal{X})$	$\sum_{i=1}^{\beta} p_i$	*β-success rate*	The probability that attackers can correctly guess the password of an account given β guesses
$G_\alpha(\mathcal{X})$	$(1-\lambda_{\mu_\alpha}) \cdot \mu_\alpha + \sum_{i=1}^{\mu_\alpha} p_i \cdot i$	*α-guesswork*	The expected number of guesses per account to achieve a success rate α
$\tilde{G}_\alpha(\mathcal{X})$	$\log_2\left(\frac{2 \cdot G_\alpha(\mathcal{X})}{\lambda_{\mu_\alpha}} - 1\right) + \log_2\left(\frac{1}{2-\lambda_{\mu_\alpha}}\right)$		Bit representation of $\tilde{G}_\alpha(\mathcal{X})$

TABLE 16
Resistance to guessing. H_∞ is the *min-entropy* for the most likely passwords. For $\tilde{G}$, H_∞, and $\tilde{G}_\alpha$, a larger value maps to stronger security. For λ_β, a smaller value indicates a lower possibility of successful password cracking. Overall, the table shows that a small portion of Chinese passwords are repeated and weak, but guessing a majority of Chinese passwords is as hard as guessing English ones.

	$\tilde{G}$	H_∞	λ_5	λ_{10}	$\tilde{G}_{0.25}$	$\tilde{G}_{0.5}$
CSDN	21.29	4.77	9.41%	10.44%	15.60	20.30
Tianya	21.49	4.55	7.15%	8.11%	14.67	19.11
Duduniu	22.55	6.02	2.74%	3.51%	18.94	21.59
7k7k	21.01	4.65	7.19%	8.20%	15.38	19.49
178.com	20.40	5.11	6.40%	8.74%	9.50	15.67
RockYou	22.65	6.81	1.71%	2.05%	15.88	19.80
Yahoo	18.03	8.05	0.78%	1.01%	16.31	17.68

those of Chinese websites, i.e., given $\beta(e.g., 5, 10)$ guesses, the probability of guessing Chinese passwords correctly is higher. This phenomenon shows that Chinese websites have a lot of repeated passwords, but the $G_{0.25}$ and $G_{0.5}$ are similar (less than 3) for both Chinese and English websites (except `178.com`). Thus, it may be easier to guess a small proportion of Chinese passwords, but for a majority of Chinese passwords, guessing them becomes as hard as guessing English ones.

- In Figure 3, the values of *α-work-factor* of `CSDN`, `Tianya` and `7k7k` are small if the expected success rate α is small, but it increases quickly along with α. This phenomenon indicates that although part of Chinese users use weak passwords that are easy to guess, a considerable number of users still carefully select passwords to protect their accounts. In addition, the users of `Duduniu` tend to choose better passwords. One possible explanation is that `Duduniu` involves monetary transaction and users tend to choose more secure passwords.

3 Cross-Region Guessing

In this section, we would like to answer the following questions.

- Given that attackers only have the password distribution of English websites, how well can they guess the passwords of Chinese websites?

TABLE 17
β-success-rates of cross-site guessing. The data in columns 2 and 3 map to the scenarios that we used each Chinese datasets to guess `RockYou` passwords, and the data in columns 4 and 5 map to the ones that we used `RockYou` passwords to guess the ones of each Chinese website. These data show that the cross-site guessing between Chinese and English users is hard.

	Chinese Websites		**RockYou**	
	→ RockYou		**→ Chinese Websites**	
	λ_5	λ_{10}	λ_5	λ_{10}
CSDN	0.31%	0.35%	3.79%	7.11%
Tianya	1.24%	1.34%	4.78%	5.16%
Duduniu	1.18%	1.50%	2.11%	2.27%
7k7k	1.19%	1.28%	4.69%	4.97%
178.com	0.93%	1.00%	3.19%	3.33%

- Given the knowledge of the differences between Chinese and English passwords, can attackers improve the efficiency of guessing the passwords of Chinese websites?

The following two subsections answer these two questions.

3.1 Cross-Site Password Guessing

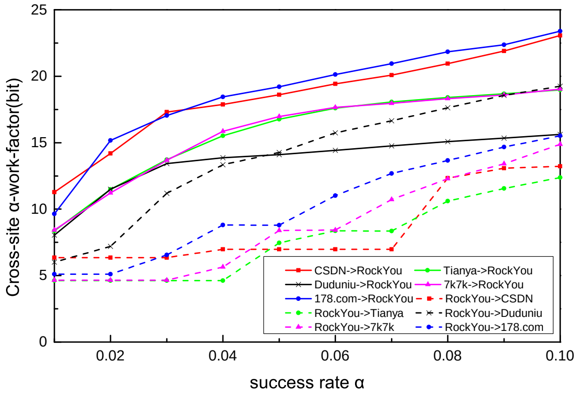


Fig. 4. α-work-factors ($\tilde{\mu}_\alpha$) of cross-site guessing, i.e., the expected number of guesses needed to succeed with a success rate α. "`X->Y`" means using the `X`'s optimal order to guess `Y`'s passwords. For example, "`CSDN->RockYou`" means using the `CSDN`'s optimal order to guess `RockYou`'s passwords.

In this section, we examine how well attackers can guess passwords from a website when they only possess a password set of another website, and we call such scenarios as cross-site password guessing. This represents the situation where attackers want to crack passwords of a website whose passwords have never been leaked. We modify the metrics that are modeled for the intra-website password guessing (listed in Table 15) to evaluate cross-site password guessing. We use two metrics, *α-work-factors* and *β-success-rates*, to evaluate the resistance to cross-site guessing. We denote these two metrics by adding a check symbol:

$$\check{\mu}_\alpha(\mathcal{X}) = \min\left\{ j \in [1, N_{other}] | \sum_{i=1}^{j} p_{(other)i} \geq \alpha \right\} \quad (1)$$

$$\check{\tilde{\mu}}_\alpha(\mathcal{X}) = \log_2\left(\frac{\check{\mu}_\alpha(\mathcal{X})}{\check{\lambda}_{\check{\mu}_\alpha}}\right) \quad (2)$$

$$\check{\lambda}_\beta(\mathcal{X}) = \sum_{i=1}^{\beta} p_{(other)i} \quad (3)$$

In the above metrics, $p_{(other)i}$ refers to the probability of the other websites' i-th password in $\mathcal{X}$. For example, we utilize the `CSDN`'s optimal password order to estimate the strength of `Tianya`'s passwords, and $\mathcal{X}$ is the probability distribution of `Tianya`. In the `CSDN`'s optimal order, "123456" is the first password. Given that in `Tianya`'s passwords "123456" accounts for 0.52%, $p_{(CSDN)1}$ is 0.52%. Different from the similar metrics used in the literature [2], we assume the attackers have no knowledge of the target password set. Thus, we use $p_{(other)}$ rather than p in Equation 1 and Equation 3.

Using the methods mentioned above, we examine two scenarios: (1) given the passwords from the five Chinese websites as a prior knowledge, how well can we guess the passwords of `RockYou`; (2) given the passwords of `RockYou`, how well can we guess the passwords of the five Chinese websites. Note that we did not take `Yahoo` into consideration because of its small data size. The results of *α-work-factors* and *β-success-rates* of cross-site guessing are shown in Figure 4 and Table 17, where we can conclude that cross-site guessing is much harder than intra-site guessing (shown in Figure 3 and Table 16).

A lower *β-success-rates* means that the probability of correct guesses given β guesses are lower. In the case of using the information of Chinese passwords to guess the `RockYou` passwords, the β-success rates ($\check{\lambda}_5$ to $\check{\lambda}_{10}$) (listed in the 2nd and 3rd columns of Table 17) are lower than the intra-site guessing ones, i.e., $\lambda_5 = 1.71\%$ and $\lambda_{10} = 2.05\%$ for `RockYou`. In the case of using the information of the `RockYou` passwords to guess Chinese passwords, the β-success rates ($\check{\lambda}_5$ to $\check{\lambda}_{10}$) (listed in the 4th and 5th columns of Table 17) are also lower than the corresponding intra-site guessing listed in Table 16. A higher *α-work-factors* means that it takes a larger number of guesses to hit the right passwords. Compared with intra-site guessing (shown in Figure 3), for the same α value, the *α-work-factors* of the cross-site guessing (shown in Figure 4) is larger. Thus, cross-site guessing is much harder. In general, attackers do not have the optimal orders of the target datasets. However, provided with the knowledge of the target dataset, how can we make the cross-site guessing easier and more efficient? Specifically, can we guess Chinese passwords easier with English passwords? We will answer this question in Section 3.2.

3.2 Guessing with Knowledge of the Differences

Several password models can be used for guessing. We chose the PCFG-based guessing method and Markov-based guessing method, proved to be efficient in password guessing [31][32][11], as representatives of these password guessing models to examine whether the aforementioned rules are useful for guessing Chinese passwords. In the following, we design experiments with the two popular methods to make cross-site guessing against passwords in `CSDN`, `Tianya` and `7k7k` based on the knowledge of the difference between Chinese and English passwords.

3.2.1 Methodology

We are interested in two questions: (1) How important are Pinyins and date formats for guessing Chinese passwords? (2) Given that attackers are only aware of the English password distribution, can they synthesize a password distribution utilizing the differences that we have observed to improve the efficiency of cracking Chinese passwords?

To answer these questions, we generated the following password training sets common for both methods.

- **RockyouTS**: This training set contains passwords that are randomly chosen from `RockYou`. This represents a training set that only contains English password information.
- **MRockyouTS**: This training set also contains passwords from `RockYou`. However, the passwords are carefully selected so that its distribution follows the Chinese password distribution: 50% of the passwords are digit-only, and 10% are letter-only. This data set helps to examine whether the structure of passwords is enough to assist password guessing.
- **RockyouDuduTS**: Half of the passwords of this training set are randomly chosen from `Duduniu`, and the other half are randomly chosen from `RockYou`. This dataset helps to examine whether combined samples of Chinese and English passwords can assist password guessing.
- **DuduTS**: This training set contains passwords randomly chosen from `Duduniu` only. This represents the scenario that attackers manage to obtain Chinese password sets.

Note that all training sets contain 2,000,000 passwords, respectively. Besides the four training sets, we also create other types of sources for the PCFG-based method and Markov-based method, which will be depicted in the following subsections. Note that we only use *RockYou* to represent the passwords of English users here. Thus, our results should be biased towards the characteristics of *RockYou*.

Setup of Probabilistic Context-Free Grammar. The PCFG-based guessing method [14] increases the efficiency of password cracking process by trying passwords according to a decreasing order of password probability. The key of PCFG is the generation of password rules (or structures) which may be generated using multiple variations [32], [33]. The rules can be constructed either from passwords themselves or word-mangling templates that can be filled in with dictionary words. In our experiments, we built rules from three sources: (1) password sets, (2) dictionaries, and optionally (3) dates.

With the password sets listed above, we construct two dictionaries to examine the effect of Pinyins in password guessing:

- **EDict**: This dictionary is a combination of the *Dic-0294* and *English-Lower*. *Dic-0294* is obtained from a password guessing website [19] and *English-lower* is obtained from John the Ripper's public website [34]. *EDict* has 869,310 unique entries in total.
- **CDict**: To form this dictionary, in addition to *EDict*, we add 20,000 most frequently used Pinyins extracted from the five Chinese websites' passwords. The size of *CDict* is larger than *EDict* by about 2.3%.

Besides Pinyins, dates also play an important role in password guessing. Since dates are digits, we modify the rules generated by the PCFG directly. We add 20,000 six-digit dates and 20,000 eight-digit dates that are most frequently used in the Chinese websites to the rules. These dates are assigned with the highest probabilities in the observed rules of six-digit numbers and eight-digit numbers, respectively. In total, these rules increase the number of six-digit and eight-digit rules by about 15% for *MRockyouTS* and about 31% for *RockyouTS*. We do not apply these rules to training sets *RockyouDuduTS* and *DuduTS*, because they already contain enough Chinese dates.

Setup of Markov model based guessing. The downside of the PCFG model is that we will never guess out passwords whose template does not occur in the training dataset. An alternative is using Markov model. Different orders, normalization methods and smoothing methods have been proposed to improve the Markov model. Here we give an example of order-1 Markov chain combined with end-symbol based normalization method [11]. As a result, the probability assigned to a password "$c_1c_2...c_l$" is computed as follows:

$$P(c_1c_2\cdots c_l) = P(c_1|c_0)P(c_2|c_1)\cdots P(c_l|c_{l-1})P(c_e|c_l)$$

where c_0 denotes the start symbol and c_e denotes the end symbol. The probability of $P(c_i|c_{i-1})$ and $P(c_e|c_l)$ are learned from the training sets. In our guessing experiments, we used an order-4 Markov model.

In contrast to PCFG, the only source for Markov model is the training set. To show the influence of Pinyins and dates, we created another three training sets, each of which contains 2,000,000 passwords. Firstly, we prepare two small password sets for composing our training sets.

- **Pinyins**: This set contains 200,000 most popular Pinyins used in the five Chinese websites' passwords.
- **Dates**: This set contains 100,000 six-digit dates and 100,000 eight-digit ones which put the year at the beginning. They are randomly selected from five Chinese websites.

Then, the extra training sets are created as follows.

- **RockyouTS_pinyin**: The training set is obtained by randomly replacing 200,000 letter-only passwords in `RockYouTS` with the whole set `Pinyins`.
- **RockyouTS_date**: The training set is obtained by randomly replacing 200,000 digit-only passwords in `RockyouTS` with the whole set `Dates`.
- **RockyouTS_both**: Both `Pinyins` and `Dates` are added into `RockYouTS` to form this training set. 200,000 letter-only passwords and 200,000 digit-only passwords are randomly removed from `RockYouTS` to keep the same size as the other training sets. This set indicates a situation where Pinyins and dates work together for password guessing.

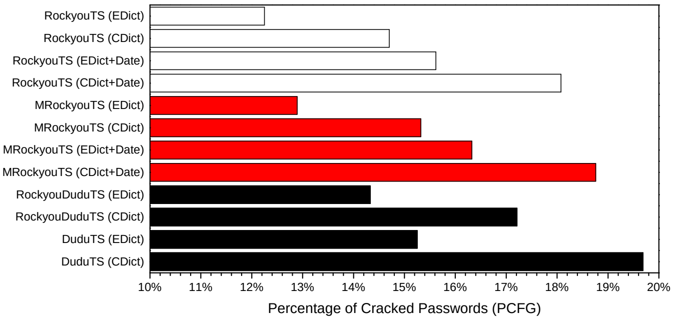


Fig. 5. PCFG-based Method: Chinese Passwords guessed within 10B guesses.

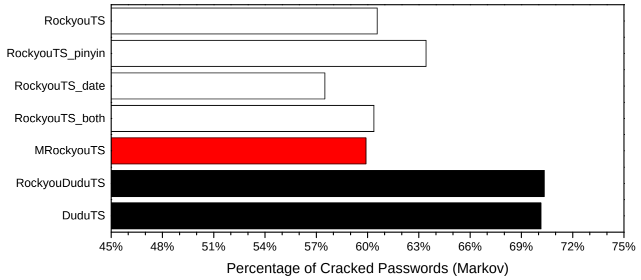


Fig. 6. Markov-based Method: Chinese Passwords guessed within 10B guesses.

In total, these newly added sets increase the percentage of Pinyins and Chinese preferred dates in *RockyouTS*. The result can help to understand how well Pinyins and dates in Chinese conventional format work in guessing the passwords of Chinese websites.

3.2.2 Results of the PCFG and the Markov based Guessing

We used the above sources to guess the passwords of `CSDN`, `Tianya` and `7k7k`, and tried 10 billion guesses per experiment. Figure 5 shows the result of the PCFG based guessing and Figure 6 shows the result of the Markov based guessing.

As shown in Figure 5, the name of the training set is labeled on the left. In the parentheses, *EDict* and *CDict* represents which dictionary the guessing is based on and *Date* means that we added (rather than replaced) the dates to the rules generated by PCFG. According to Figure 5, we have the following conclusions when we use the PCFG-based method.

- Chinese Pinyins and dates play an important role in guessing Chinese passwords. By adding 20,000 Pinyins into the dictionary, we managed to increase the percentage of password guessing. For *RockyouTS*, compared with *EDict*, the guessing efficiency increases by 20% while using *CDict*. This percentage is about 16% after we added date rules into both *EDict* and *CDict*. Furthermore, according to Section 2.2.3, more than half of Chinese passwords are digit-only. For *RockyouTS*, the guessing efficiency increases by 27% after adding dates into *EDict* and it increases by 23% after adding dates into *CDict*. We can observe that, with the same dictionary, adding dates to *RockyouTS* makes its percentage of guessed passwords more than that of *RockyouDuduTS*.

- If we use the training set *RockyouTS* and *MRockyouTS*, the differences between the percentages of guessed passwords are small (less than 1% in all scenarios). This means that the distribution of password categories (e.g., letter-only, digit-only) does not play an important role in password guessing. It is the string patterns that make difference, since Chinese and English users prefer to use different patterns of digits and letters. Thus, using *RockyouDuduTS*, which consists of both English password and Chinese password patterns can help the password guessing.

In total, compared with the guessing efficiency of *EDict*, we increase the guessing efficiency by about 48% using *CDict+Date* for *RockyouTS*.

As shown in Figure 6, however, the conclusions are as follows, when we leverage the Markov-based method to guess the passwords of `CSDN`, `Tianya` and `7k7k`.

- Overall, Markov-based method performs much better than PCFG-based method does. This result is expected. Literature [11][33][35] shows that under appropriate training, Markov-based method could guess more passwords. As [35] said, Markov-method is good at guessing strong passwords and PCFG-based method, a dictionary guessing method indeed, can guess weak passwords efficiently.
- Compared with the results of PCFG-based method, `RockyouTS` and `MRockyouTS` still guess similar number of passwords, but `RockyouDuduTS` guesses a little more passwords than those `DuduTS` does. The reason might be that English passwords have more letter patterns so that Markov Model is trained better to guess letter relevant passwords.
- Adding Pinyins, we managed to increase the efficiency of guessing. `RockyouTS_pinyin` guesses 4.7% more passwords than `RockyouTS` does. However, to our surprise, when we added dates, the result decreases. This might be because we replaced too many digit-only passwords and decreased the number of digit patterns.

In summary, the Markov-based guessing experiments imply that Pinyins play a critical role in guessing Chinese passwords, but when we use dates, we should be more careful.

4 Related Work

Although graphical passwords, biometrics and other alternatives to text-based passwords have been proposed, text-based passwords still predominate today's Internet due to its simplicity. A huge amount of research has shown the characteristics of user-created passwords [36][35][37][21][38][16][17][39]. We discuss related work in three categories: the ones studying password metrics, the ones focusing on empirical study, and the ones investigating regional features.

Password metrics: In terms of measuring the strength of passwords, NIST standards [40] propose to use Shannon's entropy to estimate the strength of a single password. Unfortunately, this method is proved to be inefficient [15]. Guesswork or guessing entropy [2][30][41][42] can also be used to measure the strength of passwords. In addition, guess numbers, which measure how many guesses it would take a guessing algorithm to reach a specific password, are widely used by researchers to measure password strength [11][31][32]. Both guesswork and guess numbers are independent of what the passwords are, but depend on the distribution of the passwords. We modified these metrics to estimate the strength of passwords across websites.

Empirical study: Analyzing passwords empirically is a widely accepted methodology. Morris *et al.* [43] described the history of the password-security-scheme design and studied the password habits of 3,289 Unix users. Howe *et al.* [44] studied the behavior of home computer users because home computer users are more likely to suffer from various attacks, *e.g.*, phishing [45], dictionary attacks [46], heuristic password guessing [14], or brute force attacks. Florencio *et al.* [47] reported a large-scale study of Web passwords habits. The study involved half a million users over a three-month period. They found that on average, each user has 6.5 passwords and about 25 websites accounts. Kelly *et al.* [31] studied 12,000 actual passwords from several perspectives. They found that certain passwords policies that can improve the strength of user-created passwords are underestimated. Mazurek *et al.* investigated the password guessability for an entire university [32]. Our study also belongs to this category and we analyzed a larger password dataset.

Regional features: Languages, regional culture, and many other factors may affect passwords, and thus several work has focused in this area. Abbott *et al.* [48] used a small dataset of native English and Spanish speakers to find distinct usage patterns between the two groups. They focused their research on how language and culture could impact the decisions that users make when creating a password. Bonneau *et al.* [49] investigated the lingering effects of character encoding on the password ecosystem based on password datasets from Chinese, English, Hebrew and Spanish speakers. Sometimes the conclusions drawn by various researchers do not agree with each other. For instance, Bonneau [2] analyzed the language dependency of password guessing and showed that among all `Yahoo` passwords, passwords created by Chinese are almost the hardest to guess. In contrary, Yang *et al.* [50] examined similarities and differences of password construction among four companies and investigated how cultural factors shape user password construction in China. They asserted that on average users in China have weaker senses of security than those in Western countries. We believe the disagreement is the result of dataset selection. That's why we conducted our study using datasets with two groups of websites: five Chinese websites, and two English websites. Such a dataset represents a larger and more diverse corpus of passwords than prior work, and our corpus includes passwords from users that only speak Chinese and know little English. Our large-scale analysis shows that (1) the passwords of both English and Chinese users are similar in strength as shown in Figure 3 and Table 16; (2) if attackers are aware of the fundamental differences between two languages (as pointed out in this paper), they can guess Chinese passwords efficiently.

5 Conclusion and Future Work

To the best of our knowledge, this paper is the first large-scale empirical study on Chinese Web passwords, leveraging a corpus of 100 million publicly available passwords. By comparing Chinese and English passwords, we find that Chinese users prefer digits in their passwords, but the security concerns are common in both Chinese and English users based on the strength measure results for password sets. Moreover, Pinyins and dates also appear often in their passwords. Leveraging these observations, we can improve the guessing efficiency of cracking Chinese passwords by up to 48% based on PCFG using Pinyins and dates, and up to 4.7% based on Markov model using Pinyins. Our research sheds light on the severe threats that passwords are faced with from regional patterns.

In summary, we suggest future researchers to consider the following factors when studying the security of localised passwords:

- Regional patterns, including language patterns, are very popular in localized passwords. For example, different user groups will leverage different rules of the composition and structures [35][50][48][51]. In addition, different languages have their separate characteristics in both writing and colloquialism. Thus, there are some language-related factors that researchers could consider. These factors include variable input methods, possible encodings, keyboard patterns, etc. For example, as investigated in this paper, there are both Pinyin and Shuangpin input methods in Chinese. Users can also use different encodings for usability or some other reasons [49]. Various expressions in different languages may lead to different keyboard patterns in passwords.
- Culture-related factors, such as preference for some digit or date formats and some localised dictionaries, could be considered too. Many researchers have highlighted the influence of cultural factors on users' choosing of passwords. For example, the research in [29] mentioned users' preferences for date formats, such as holidays. Yang *et al.* [50] mentioned that number four is the least frequently used number because of its meaning about unlucky things in Chinese. For Chinese language, localised dictionaries might include Chinese idioms, ancient poetry and Chinese location names, etc. They may exist in other forms for other languages. These factors can help researchers to further understand the semantic meaning of passwords.

In future, it is worth trying to gather additional password datasets from various websites, especially the ones for English users, so as to increase the coverage and variety of data sets. In addtion, with an increasing number of password creation policies being enforced by websites, a direction for future study is to investigate the *status quo* of the password creation policies in Chinese websites and to study the impact of these policies on password statistics. Also, it is very important for web masters in different regions to develop a region-aware password strength meter. The meter may reduce the ratio of real weak passwords which could be strong in current English users oriented password meters.

Acknowledgments

This paper is supported by NSFC (Grant No. 61572136, 61472358), National Key Science and Technology Program, Key Lab of Information Network Security, Ministry of Public Security. We thank Mr. Ding Wang and all anonymous reviewers for their comments.

Appendix A
Method to Remove the Duplicate Passwords from Tianya and 7k7k

Users usually reuse their passwords across websites. When analyzing this behavior, we found that `Tianya` and `7k7k` have an unusually large number of the same accounts (identified by email) with the same passwords. Normally, the password reuse rate between two Chinese websites (*e.g.*, `CSDN` and `Duduniu`) is about 30% to 40%. However, more than 90% common users between `7k7k` and `Tianya` reuse their passwords, *i.e.* they use same passwords in both websites.

Since the statistic features of these duplicate accounts are different from the ones between any other websites, we suspected that the attackers have copied accounts from `7k7k` to `tianya` or the opposite way. In order to eliminate the influence of these duplicate passwords, we remove them from both datasets. Having such a huge number of Chinese passwords, we believe removing these duplicate passwords have negligible influence on the overall analysis results.

References

[1] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *Proceedings of IEEE S&P 2012*. IEEE, 2012, pp. 553–567.

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

[3] CNNIC, "The 35th survey report on Chinese Internet development," http://www.cnnic.cn/hlwfzyj/hlwxzbg/201502/P020150203551802054676.pdf, Jan 2015.

[4] CSDN, http://www.csdn.net/company/about.html.

[5] Tianya, http://help.tianya.cn/about/history/2011/06/02/166666.shtml.

[6] Duduniu, http://baike.baidu.com/view/1557125.htm.

[7] 7k7k, http://www.7k7k.com/html/about.htm.

[8] 178.com, http://www.178.com/s/information/about.html.

[9] Rockyou, http://rockyou.com/ry/about-us.

[10] Yahoo, http://info.yahoo.com/.

[11] J. Ma, W. Yang, M. Luo, and N. Li, "A study of probabilistic password models," in *Proceeding of IEEE S&P 2014*, 2014.

[12] J. O. Pliam, "On the incomparability of entropy and marginal guesswork in brute-force attacks," in *INDOCRYPT*, 2000, pp. 67–79.

[13] S. Boztas, "Entropies, guessing, and cryptography," Department of Mathematics, Royal Melbourne Institute of Technology, Tech. Rep., 1999.

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

[15] M. Weir, S. Aggarwal, M. P. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS 2010, Chicago, Illinois, USA, October 4-8, 2010*, 2010, pp. 162–175. [Online]. Available: http://doi.acm.org/10.1145/1866307.1866327

[16] J. T. R. Veras, C. Collins, "On the Semantic Patterns of Passwords and their Security Impact," in *Proceedings of NDSS 2014*, 2014.

[17] X. de Carn de Carnavalet and M. Mannan, "From very weak to very strong: Analyzing password-strength meters," in *Proceedings of NDSS 2014*, 2014.

[18] D. Schweitzer, J. Boleng, C. Hughes, and L. Murphy, "Visualizing keyboard pattern passwords," in *Proceedings of VizSec 2009*. IEEE, 2009, pp. 69–73.

[19] "Wordlist chinese.zip from outpost9," http://www.outpost9.com/files/WordLists.html.

[20] "The concise Oxford dictionray of current English," http://archive.org/stream/conciseoxforddic00fowlrich/conciseoxforddic00fowlrich_djvu.txt.

[21] D. Malone and K. Maher, "Investigating the distribution of password choices," in *Proceedings of WWW 2012*. ACM, 2012, pp. 301–310.

[22] Computerbits, http://www.computerbits.ie.

[23] Microsoft, http://www.microsoft.com/china/pinyin/.

[24] ZiRanMa, http://www.zrm.com.cn/.

[25] ZiGuang, http://www.unispim.com/.

[26] P. JJ, http://dir.jjol.cn/Pyjj/.

[27] znabc, http://www.znabc.com/.

[28] Sougou, http://pinyin.sogou.com/dict/.

[29] R. Veras, J. Thorpe, and C. Collins, "Visualizing semantics in passwords: The role of dates," in *Proceedings of the Ninth International Symposium on Visualization for Cyber Security*, ser. VizSec '12. New York, NY, USA: ACM, 2012, pp. 88–95. [Online]. Available: http://doi.acm.org/10.1145/2379690.2379702

[30] J. Bonneau, S. Preibusch, and R. Anderson, "A birthday present every eleven wallets? The security of customer-chosen banking pins," in *Proceedings of FC 2012*, 2012.

[31] P. Kelley, S. Komanduri, M. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. Cranor, and J. Lopez, "Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms," in *Proceedings of IEEE S&P 2012*, 2012, pp. 523–537.

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

[33] M. Dürmuth, F. Angelstorf, C. Castelluccia, D. Perito, and C. Abdelberi, "OMEN: Faster Password Guessing Using an Ordered Markov Enumerator," in *Engineering Secure Software and Systems - 7th International Symposium, ESSoS 2015, Milan, Italy, March 4-6, 2015. Proceedings*, 2015, pp. 119–132. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-15618-7_10

[34] "Wordlist from John the Ripper," http://download.openwall.net/pub/passwords/wordlists/.

[35] M. Dell'Amico, P. Michiardi, and Y. Roudier, "Password strength: An empirical analysis," in *Proceedings IEEE INFOCOM 2010*. IEEE, 2010, pp. 1–9.

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

[37] C. Kuo, S. Romanosky, and L. F. Cranor, "Human selection of mnemonic phrase-based passwords," in *Proceedings of the Second Symposium on Usable privacy and security*. ACM, 2006, pp. 67–78.

[38] D. A. Sawyer, "The characteristics of user-generated passwords," DTIC Document, Tech. Rep., 1990.

[39] J. Bonneau and S. E. Schechter, "Towards reliable storage of 56-bit secrets in human memory," in *Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20-22, 2014.*, 2014, pp. 607–623. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/bonneau

[40] W. E. Burr, D. F. Dodson, E. M. Newton, R. A. Perlner, W. T. Polk, S. Gupta, and E. A. Nabbus, "NIST special publication 800-63-1 electronic authentication guideline," 2006.

[41] J. L. Massey, "Guessing and entropy," in *Proceedings of 1994 IEEE International Symposium on Information Theory*. IEEE, 1994, p. 204.

[42] C. Cachin, "Entropy measures and unconditional security in cryptography," Ph.D. dissertation, Swiss Federal Institute of Technology Zurich, 1997.

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

[44] A. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne, "The psychology of security for the home computer user," in *Proceedings of IEEE S&P 2012*, 2012, pp. 209–223.

[45] G. Xiang and J. I. Hong, "A hybrid phish detection approach by identity discovery and keywords retrieval," in *Proceedings of WWW'09*, 2009, pp. 561–570.

[46] B. Pinkas and T. Sander, "Securing passwords against dictionary attacks," in *Proceedings of ACM CCS 2002*, pp. 161–170.

[47] D. Florencio and C. Herley, "A large-scale study of web password habits," in *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, 2007, pp. 657–666.

[48] J. Abbott and V. M. Garcia, "Password differences based on language and testing of memory recall," *NNGT Int. J. on Information Security*, vol. 2, 2015.

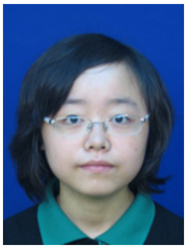
[49] J. Bonneau and R. Xu, "Of contraseñas, sysmawt, and mìmǎ: Character encoding issues for web passwords," in *Web 2.0 Security & Privacy*, May 2012. [Online]. Available: http://www.jbonneau.com/doc/BX12-W2SP-passwords_character_encoding.pdf

[50] C. Yang, J. long Hung, and Z. Lin, "An analysis view on password patterns of Chinese Internet users," *Nankai Business Review International*, vol. 4, pp. 66–77, 2013.

[51] A. G. Voyiatzis, C. Fidas, D. N. Serpanos, and N. M. Avouris, "An Empirical Study on the Web Password Strength in Greece." in *Panhellenic Conference on Informatics*, P. Angelidis and A. Michalas, Eds. IEEE Computer Society, 2011, pp. 212–216. [Online]. Available: http://dblp.uni-trier.de/db/conf/pci/pci2011.html#VoyiatzisFSA11

Weili Han (M'08) is an associate professor at Fudan University. His research interests are mainly in the fields of Access Control, Digital Identity, IoT security. He is now the members of the ACM, SIGSAC, IEEE, and CCF. He received his Ph.D. at Zhejiang University in 2003. Then, he joined the faculty of Software School at Fudan University. From 2008 to 2009, he visited Purdue University as a visiting professor funded by China Scholarship Council and Purdue Uinversity. He serves in several leading conferences and journals as PC members, reviewers, and an associate editor.

Zhigong Li is a graduate student at Software School, Fudan University. His research interests focus on passwords security and system security.



Lang Yuan is a graduate student at Software School, Fudan University. Her research interests focus on password security and system security.

Wenyuan Xu received her Ph.D. degree in electrical and computer engineering from Rutgers University in 2007. She is a professor at Zhejiang University. Her research interests include wireless networking, network security and privacy. Dr. Xu received the NSF Career Award in 2009. She is a member of the ACM and IEEE, and has served on the technical program committees for several IEEE/ACM conferences on wireless networking and security.