A Large-Scale Empirical Analysis of Chinese Web Passwords

Zhigong Li and Weili Han, ***Fudan University;*** **Wenyuan Xu,** ***Zhejiang University***

https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/li_zhigong

This paper is included in the Proceedings of the 23rd USENIX Security Symposium.

August 20–22, 2014 • San Diego, CA

ISBN 978-1-931971-15-7

Open access to the Proceedings of the 23rd USENIX Security Symposium is sponsored by USENIX

A Large-Scale Empirical Analysis of Chinese Web Passwords

Zhigong Li, Weili Han
Software School, Fudan University
Shanghai Key Laboratory of Data Science, Fudan University

Wenyuan Xu
Department of Electronic Engineering, Zhejiang University

Abstract

Users speaking different languages may prefer different patterns in creating their passwords, and thus knowledge on English passwords cannot help to guess passwords from other languages well. Research has already shown Chinese passwords are one of the most difficult ones to guess. We believe that the conclusion is biased because, to the best of our knowledge, little empirical study has examined regional differences of passwords on a large scale, especially on Chinese passwords. In this paper, we study the differences between passwords from Chinese and English speaking users, leveraging over 100 million leaked and publicly available passwords from Chinese and international websites in recent years. We found that Chinese prefer digits when composing their passwords while English users prefer letters, especially lowercase letters. However, their strength against password guessing is similar. Second, we observe that both users prefer to use the patterns that they are familiar with, *e.g.*, Chinese Pinyins for Chinese and English words for English users. Third, we observe that both Chinese and English users prefer their conventional format when they use dates to construct passwords. Based on these observations, we improve a PCFG (Probabilistic Context-Free Grammar) based password guessing method by inserting Pinyins (about 2.3% more entries) into the attack dictionary and insert our observed composition rules into the guessing rule set. As a result, our experiments show that the efficiency of password guessing increases by 34%.

1 Introduction

Passwords are the most widely used credentials for authenticating Web users around the world, including the users that do not speak English. Text-based passwords are likely to remain the dominant mechanism for authenticating users for the foreseeable future [7][19]. Meanwhile, researchers are still in the process of understanding the security strength of passwords and exploiting methods to improve password guessing. Although insightful, most existing work focuses on passwords of English users. Little work has studied the impact of regional convention and languages on password selection utilizing a large dataset of passwords. One exception is Bonneau [6], 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 his dataset (*i.e.*, `Yahoo` users are familiar with English). In this paper, we analyze passwords of non-English speakers, specifically, Chinese users, which represent *618* million Internet users as of the end of 2013 [12], and compare them with passwords of English users.

To understand the differences between Chinese and English passwords, this paper leverages over 100 million leaked and publicly available passwords from several popular Chinese websites (`CSDN` [13], `Tianya` [33], `Duduniu` [17], `7k7k` [5], and `178.com` [4]) and English websites (`RockYou` [30] and `yahoo` [37]). These Chinese websites only provide Chinese webpages, and we consider their users as *Chinese users*. 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*. Note that, these websites (except `Duduniu`, which is an e-commerce website) provide similar services, i.e., non-monetary ones 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.

The unfortunate leakages of the large volume of passwords provides us an opportunity to understand password differences between the two groups of users in depth. Such analysis is important, because it enables

	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	15,940,099
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**	**105,836,958**

Table 1: Basic information of leaked passwords of the websites that are analyzed in this paper. We removed the duplicated accounts between `Tianya` and `7k7k` from the `Tianya` dataset. See details in Appendix A.

better password guessing evaluation and can guide web masters to protect the accounts.

We designed analysis tools and leveraged the guessing resistance indicators (such as α-*work-factors* [28] and β-*success-rates* [10]) to find the differences among accounts of multiple websites, and found the preference of the two groups of users. Then, we improved the efficiency of the Probabilistic Context-Free Grammar (PCFG) based password guessing method [35] by adding regionally preferred patterns (*i.e.* Pinyins) into the dictionary and modifying the generated guessing rules. We summarize our findings and main contributions as follows:

- **Different Characters Sets**: Chinese users prefer digits in their passwords, while English users prefer letters, especially lowercase letters. 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 Dates**: Both Chinese and English users prefer to use language-related patterns as passwords. That is, Chinese users prefer Chinese Pinyins and English users prefer English words. As for dates, both groups prefer their conventional formats. That is, Chinese prefer dates with the year at the beginning and English users prefer dates with the year at the end.
- **Improvement of 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 34% in guessing Chinese passwords using a PCFG based guessing method. This confirms that the Pinyins and date's rules are important in guessing Chinese passwords.

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 [6] and PCFG based methods [35]. 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.

At the end of 2010, an incident that is 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` [13] is one of the most popular Chinese IT professional communities, similar to `MSDN`. `Tianya` [33] is the largest online forums and blogs in China. `7k7k` [5] and `178.com` [4] are two websites providing game infor-

	Chinese	English
1	123456 (2.17%)	123456 (0.88%)
2	123456789 (0.65%)	12345 (0.24%)
3	111111 (0.59%)	123456789 (0.23%)
4	12345678 (0.39%)	password (0.18%)
5	000000 (0.34%)	iloveyou (0.15%)

Table 2: The most popular passwords and their occurrence percentages.

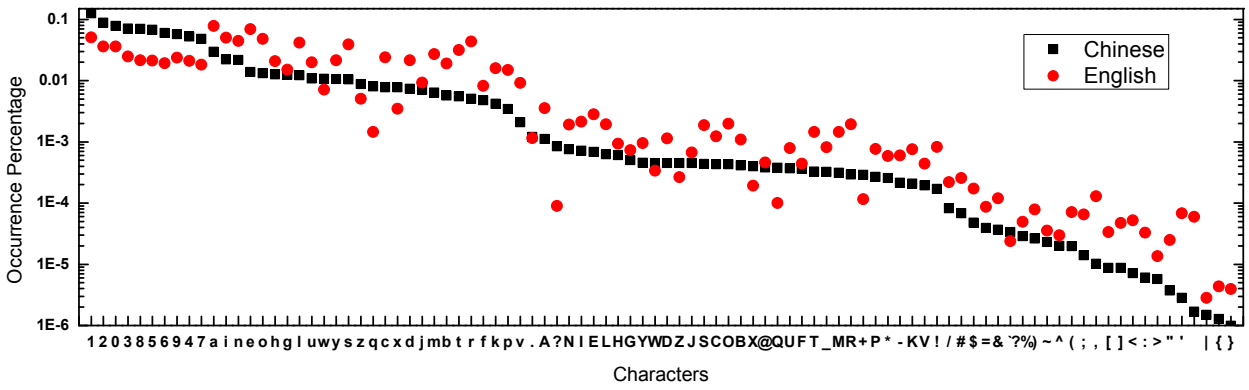

Figure 1: Character distribution, i.e., the occurrence percentage of each character for Chinese and English passwords. The characters are arranged in a descending order according to the percentages in Chinese passwords.

mation and online flash games. Duduniu [17] 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 also happened to English websites as well. In 2009, attackers broke into the database of RockYou and released the 32 million passwords (in plaintext) to the public. In 2012, Yahoo's accounts were leaked. A hacking group 'DD3Ds Company' utilized a union-based SQL injection to obtain login details of about 450 thousand user accounts.

The raw files contain duplication and blank passwords that can affect the analysis. For instance, we detected that attackers copied a portion of 7k7k passwords to Tianya, because the password duplication rate between Tianya and 7k7k is much more than the rate between any other two websites (i.e., about 90% between Tianya and 7k7k and about 30% between any other two websites). We thus removed these duplicate passwords in Tianya using the method described in Appendix A. After removing the accounts with blank passwords and filtering out duplicated accounts, we obtained 105,836,958 accounts, as detailed in Table 1. Finally, we imported them into MySQL for further analysis.

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 4.14% of all Chinese passwords and 1.69% of all English passwords, which shows that Chinese passwords are more congregated.
- 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 top 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

	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	60.77%	11.13% (10.75%)	26.41% (23.03%)	0.14% (0.12%)	0.32%	1.14% (0.49%)
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 3: Compositions of passwords. 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. 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.

	# 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	635	DDDDDD	19.51%
178.com	459	DDDDDD	15.48%
RockYou	803	LLLLLL	5.40%
Yahoo	1165	LLLLLL	9.19%

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 passwords in both Chinese websites and English ones. D represents a digit, and L represents a lowercase letter.

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 because of the word patterns in either languages. 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 used less 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 aspects. (1) we divided passwords according to their compositions and calculated the percentages in seven category (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 [35]. 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 Table 3 and Table 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. Chinese characters cannot be entered directly as a password, and digits 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

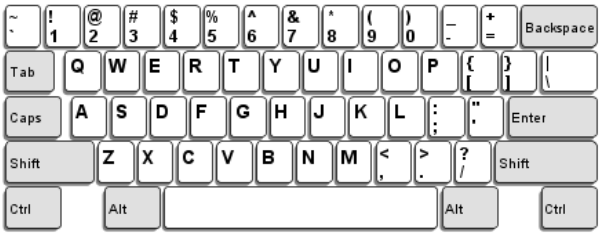

Figure 2: A typical layout of a keyboard (103P) used in China, which is the same as an English keyboard layout.

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 2.2.6.

- 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 bother with symbols.

2.2.4 Keyboard Patterns

Sometimes, users prefer to create their passwords according to keyboard patterns [32]. Thus, we analyzed the percentages of three primary keyboard patterns. Note that Chinese users utilize standard English keyboards (shown in Figure 2), 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.*, asdfhj.
- **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* or *Zig Zag*, *e.g.*, zxcfgh.

	Chinese	English
Same Row	8.31% (0.55%)	2.42% (0.25%)
Zig Zag	0.26%	0.06%
Snake	0.27%	0.08%

Table 5: Percentage 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.

Algorithm to 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 coordinates of 1 (and !) are (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 using the algorithm illustrated in Algorithm 1, where $isAdjacent(pos1, pos2)$ determines whether two letters located in the coordinates $pos1$ and $pos2$ are adjacent in the same row or column.

Result. The statistics analyzed by Algorithm 1 is shown in Table 5, from 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 0.2% more *Zig Zag* passwords for Chinese than English users. This could be because 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. Trained with Pinyin since primary school, Chinese computer users are familiar with it. Pinyin is the most popular method to input Chinese characters to a computer because it requires almost no extra training for Chinese. Typically, a Chinese character is entered by multiple keystrokes. Although other input methods, such as Wubi, exist, these methods are not as popular due to their steep learning curves.

Since websites do not support passwords composed of Chinese characters directly, unsurprisingly, just like the words in English passwords, Pinyins are widely used in passwords of Chinese users. Ignoring the tones, typically, 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 [3]. However, users may use various compositions of multiple Pinyins in their passwords. For example, the password *nihao*, is composed of Pinyins *ni* and *hao*.

	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	44.70% (4.97%)	10.04% (1.12%)	21.09% (5.84%)	7.02% (1.94%)
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 6: Percentage 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 percentage 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.

	Top Chinese Pinyins	Top English Words
1	woaini (1.47%)	password (1.28%)
2	li (1.06%)	iloveyou (0.98%)
3	wang (0.97%)	love (0.76%)
4	tianya (0.89%)	angel (0.59%)
5	zhang (0.84%)	monkey (0.45%)

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.

Algorithm to 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 *helloworld* is composed of English words *hello* and *world*. For English words, we chose the Oxford English Dictionary [1] and extracted more than 20,000 commonly used English words.

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 Tries, we can identify if a password is composed of Chinese Pinyins or English words. The algorithm to insert entries into the Trie is shown in Algorithm 2. In our experiments, we constructed two Tries: one is constructed out of Chinese Pinyins, and the other is built based on the more than 20,000 commonly used English words. The procedure to identify if a password is composed of Chinese Pinyins or English words is shown in Algorithm 3. The structure *node* has two properties. The first is named as *child*, which is an array of *node* and represents the child nodes. The second is a boolean, *isValue*, which represents if the string from the root to the current node is a valid value. The algorithm will try to match the password with the known strings from Trie recursively.

Note that because it is hard to determine the semantic meaning, a password may be semantically meaningless even if it is a composition of Chinese Pinyins or English words. Furthermore, some passwords can be interpreted as compositions of Pinyins and English words at the same time. We removed the passwords with both Pinyins and English words in our analysis.

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.*, the ones contain 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 Table 6 and 7, we draw the following conclusions:

- Out of the letter-only passwords, Pinyins are the dominant patterns for Chinese users in composing their passwords, and English words dominate the English passwords. Even when we consider all categories of 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.

- Interestingly, it seems that love is always the main theme of human beings. As shown in Table 7, *love*

	# Consecutive Exactly Eight Digits	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	2,470,204	32.41%	0.18%	0.37%
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 8: 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*.

	# Consecutive Exactly Six Digits	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	3,999,958	24.34%	2.63%	0.88%
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%

Table 9: 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.

and *iloveyou* are ranked at the second and the third in English passwords. Meanwhile, *woaini* is the top ranked Pinyin, which means *I love you* in Chinese.

- The Pinyins of names are widely used in Chinese passwords. The Pinyins *li*, *wang* and *zhang*, listed as the top used Pinyins for passwords in Table 7, are among the most popular surnames in China. Note that it is difficult to identify first names in Chinese, because 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 the website name, is ranked at the fourth in Chinese Pinyins.

- 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 expression, 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 because either *yao* or *wei* can be a surname.

The influence of Chinese Pinyins in password guessing is discussed in Section 3.2.

2.2.6 Dates

Given that digits are commonly used in passwords, we try to understand the meaning of these digits. Since dates are typically represented as a string of digits, in this subsection we analyze the usage of dates in passwords.

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 classify eight-digit dates into *YYYYMMDD*, *MMDDYYYY* and *DDMMYYYY*. The results are shown in Table 8 and 9. 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 80% of them as *DDMMYY*. Then, we assigned 20% of the ambiguous passwords to *MMDDYY* and 80% to *DDMMYY*.

Furthermore, there may be false positive where a general six-digit number is considered as a date. For example, *123123* could be considered as *December 31, 1923*,

	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	73.90%	25.51% (24.61%)	0.18%	0.41% (0.37%)
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 10: 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.

but most likely it is just two consecutive *123*. Thus, we selected 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 most 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.

Table 8 and Table 9 show the results. For example, the 29.24% in the first row in Table 8 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 and digit, symbol and digit, letter and digit and symbol passwords out of all passwords that contain dates (both six-digit and eight-digit dates). As shown in Table 10, 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.

[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.*

	Beginning	Middle	End
CSDN	21.68%	4.32%	74.00%
Tianya	27.33%	4.75%	67.07%
Duduniu	24.76%	1.36%	73.88%
7k7k	32.17%	2.70%	65.13%
178.com	22.30%	1.03%	76.67%
RockYou	27.40%	3.91%	68.69%
Yahoo	22.66%	5.00%	72.34%

Table 11: Positions of dates. The percentages of passwords that contains dates at the beginning, the middle, or the end.

Date Position. To understand the position of dates in passwords, we analyzed those passwords that contain dates (digit-only passwords are not included).

We categorize the position of the dates as *beginning*, *middle*, and *end*, and summarize the results in Table 11. For both Chinese and English users, they prefer to have dates appear at the end of passwords and rarely place them in the middle.

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 [6][8], which are designed to evaluate the password strength in different regions.

As shown in Table 12, we briefly introduce these metrics: H_∞ is defined as *min-entropy*, a worst-case secu-

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

Table 12: Metrics [6][8] list used in our analysis. $\mathscr{X}$ 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 $\mathscr{X}$ where $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_N$.

rity metric for human-chosen passwords, i.e., when a user chooses the mostly likely password. G is defined as *guesswork*, representing the expected number of sequential guesses to find a password of an account if an attacker proceeds in an optimal order, i.e., trying passwords in a 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 an attacker 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.

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 as proposed by Bonneau [6][8], *i.e.*, attackers know the exact distributions 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 13 and Figure 3, and we draw the following observations:

- In Table 13, we observe that the β-*success-rates* (λ_5, λ_{10}) of `RockYou` and `Yahoo` are much lower than 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) between 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 value of α-*work-factors* of `CSDN`, `Tianya` and `7k7k` are small if the expected success rate α is small, but it grows quickly with the increase of α. This phenomenon indicates that although part of Chinese users use the 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 secure passwords.

	$\tilde{G}$	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	21.25	4.75	6.53%	7.61%	15.22	19.63
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

Table 13: 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 possiblity of successful password cracking. Overall, the table shows that a small portion of Chinese passwords are repreaded and weak, but guessing a majority of Chinese passwords is as hard as guessing English ones.

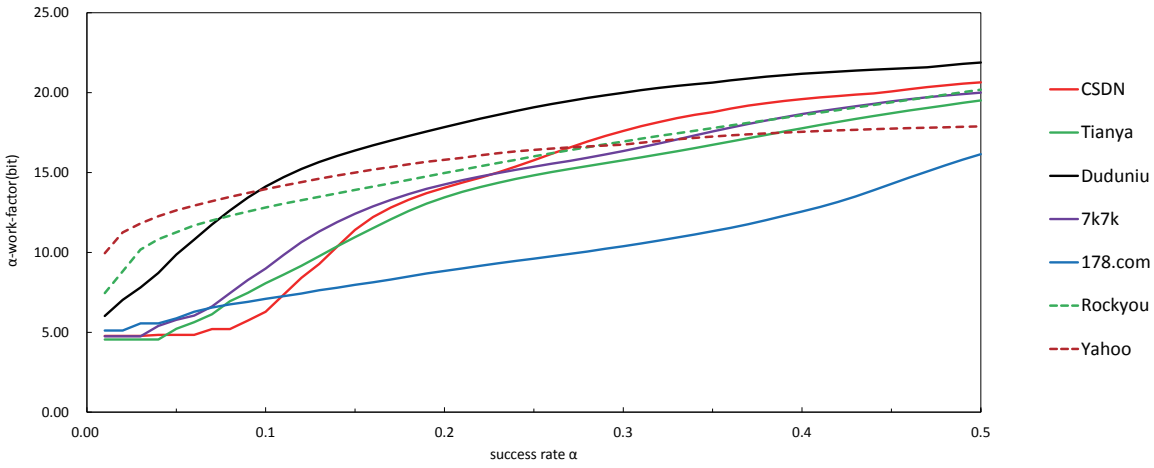

Figure 3: The expected number of guesses needed to succeed with a success rate α (α-work-factors, $\tilde{\mu}_\alpha$) of all seven websites. The dash lines represent English websites and solid lines map to Chinese websites.

3 Cross-Region Guessing

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

- Given that an attacker only has the password distribution of English websites, how well can she guess the passwords of Chinese websites?
- Given the knowledge of the differences between Chinese and English passwords, can an attacker improve the efficiency of guessing the passwords of Chinese websites?

The following two subsections answer these two questions.

3.1 Cross-Site Password Guessing

In this section, we examine how well an attacker can guess passwords from a website when she only possesses a password set of another website, and we call such scenarios as cross-site password guessing. This represents the situation when an attacker 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 12) 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:

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

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

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

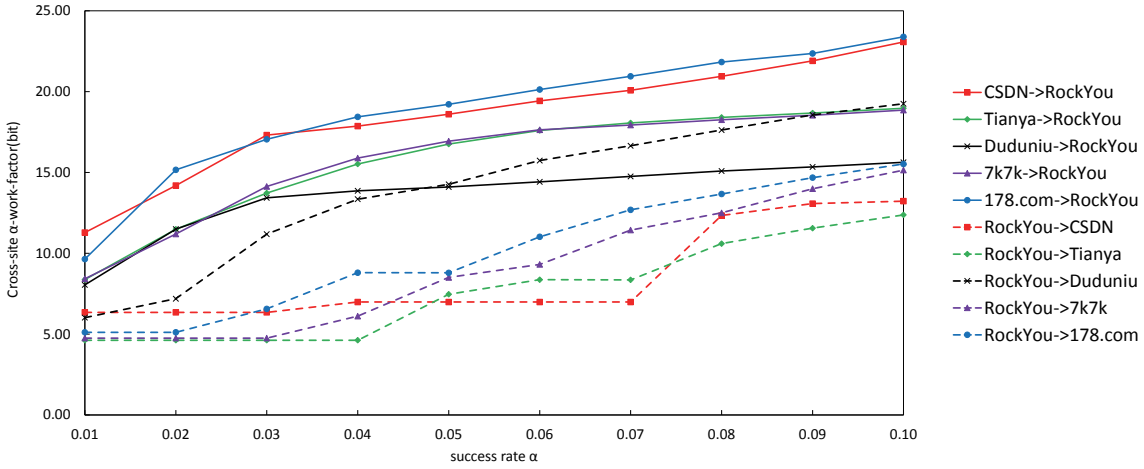

Figure 4: α-work-factors ($\tilde{\mu}_\alpha$) of cross-site guessing, i.e., the expected number of guesses needed to succeed with a success rate α. "X->Y" means using the X's optimal order to guess Y's passwords. For example, "CSDN->RockYou" means using the CSDN's optimal order to guess RockYou's passwords.

	Chinese Websites → RockYou		**RockYou → Chinese Websites**	
	$\check{\lambda}_5$	$\check{\lambda}_{10}$	$\check{\lambda}_5$	$\check{\lambda}_{10}$
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	1.20%	1.28%	4.39%	4.66%
178.com	0.93%	1.00%	3.19%	3.33%

Table 14: β-success-rates of cross-site guessing. The data in columns 2 and 3 maps to the scenarios that we used each Chinese datesets to guess Rockyou passwords, and the data in columns 4 and 5 maps to the ones that we used Rockyou passwords to guess the ones of each Chinese website. These data shows that the cross-site guessing between Chinese and English users is hard.

In the above metrics, $p_{(other)i}$ refers to the probability of the other websites' i-th password in $\mathscr{X}$. For example, we utilize the CSDN's optimal password order to estimate the strength of Tianya's passwords, and $\mathscr{X}$ 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%.

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 14, where we can conclude that cross-site guessing is much harder than intra-site guessing (shown in Figure 3 and Table 13).

A lower β-*success-rates* means that the probability of correct guesses given β guesses are lower. In cases 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 14) are lower than the intra-site guessing ones, i.e., $\lambda_5 = 1.71\%$ and $\lambda_{10} = 2.05\%$ for RockYou. In cases 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 14) are also lower than the corresponding intra-site guessing listed in Table 13. 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 harder.

Algorithm 1 Identify Keyboard Patterns

Input: S: a string
Output: the keyboard pattern of S

```
if S.length < 4 then
    return NO_PATTERN
end if
letters[] ← S.toCharArray()
samerow ← TRUE
zigzag ← TRUE
for i = 1; i < letters.length(); i++ do
    pos1 ← letters[i − 1]
    pos2 ← letters[i]
    if isAdjacent(pos1, pos2) then
        samerow ← samerow&isSamerow(pos1, pos2)
        zigzag ← zigzag&!isSamerow(pos1, pos2)
    else
        return NO_PATTERN
    end if
end for
if samerow then
    return SAME_ROW
end if
if zigzag then
    return ZIG_ZAG
end if
return SNAKE
```

Algorithm 2 Insert into the Trie

Input:
S: a string (a Chinese Pinyin or English word) that needs to be inserted into the Trie
Root: the root of the Trie

```
S ← S.toLowercase()
letters[] ← S.toCharArray()
node ← Root
for i = 0; i < letters.length(); i++ do
    pos ← letters[i] − 'a'
    node.child[pos].val ← letters[i]
    node ← node.child[pos]
end for
node.isValue ← TRUE
```

3.2 Guessing with Probabilistic Context-Free Grammar

The PCFG-based guessing method [35] increases the efficiency of password cracking process by trying passwords according to a decreasing order of password probability. The key of PCFG is to generate password rules (or structures). The rules can be constructed either from passwords themselves or word-mangling templates that can be filled in with dictionary words, for example. In our experiments, we built rules from three sources: (1) password sets, (2) dictionaries, and optionally (3) dates. We chose to use PCFG to examine whether the aforementioned rules are useful for guessing Chinese passwords, because it has been shown to be efficient in password guessing [21][24].

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 an attacker is only aware of the English password distribution, can she synthesize a password distribution utilizing the differences that we have observed to improve the efficiency of cracking Chinese passwords?

To answer those questions, we created rules out of three types of sources for the PCFG-based guessing method: password training sets, dictionaries, and dates. For password training sets, we generated the following ones. Note that all training sets contain 2,000,000 passwords, respectively.

- **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 an attacker manages to obtain Chinese password sets.

In order to examine the effect of Pinyins in password guessing, we construct two dictionaries:

- **EDict**: This dictionary is a combination of the *Dic-0294* and *English-Lower*. *Dic-0294* is obtained

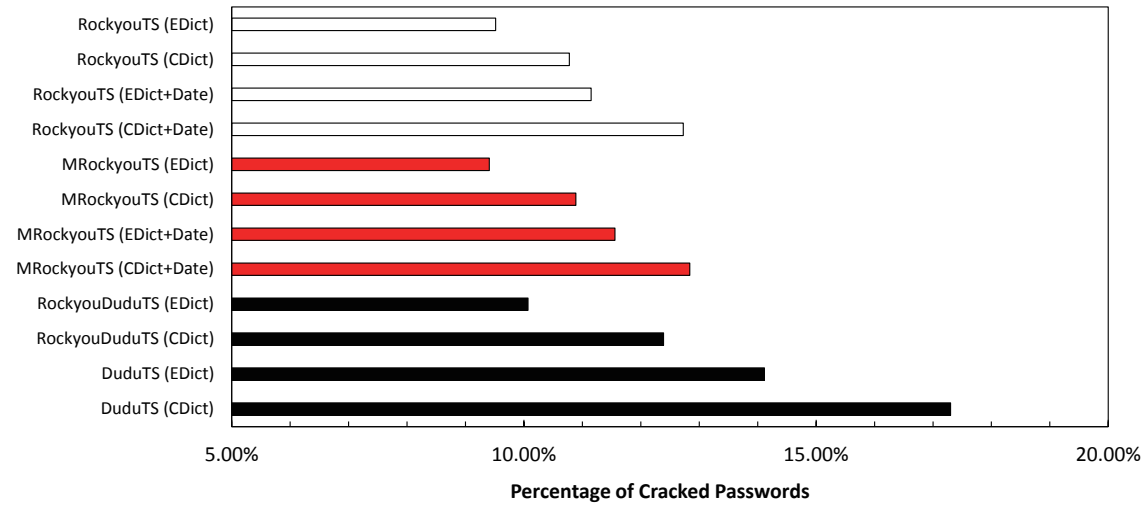

Figure 5: Passwords guessed within 10B guesses. Terminologies are explained in Section 3.2.1 in detail.

from a password guessing website [3] and *English-lower* is obtained from John the Ripper's public website [2]. *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 from the five Chinese websites. As a result, 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.

We used the above dictionaries and the modified rule set to guess the passwords of CSDN, and try 10 billion guesses per experiment.

3.2.2 Results of the PCFG 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 the dates to the rules generated by PCFG. According to Figure 5, we have the following conclusion.

- 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*, from *EDict* to *CDict*, the guessing efficiency increases by 13% and from *EDict+Date* to *CDict+Date*, the guessing efficiency increases by 14%.

  Furthermore, according to Section 2.2.3, more than half Chinese passwords are digit-only. For *RockyouTS*, the guessing efficiency increases by 17% after adding dates into *EDict* and it increases by 18% after adding dates into *CDict*. Last but not least, under the same dictionary, adding dates changes the percentage of guessed passwords of *RockyouTS* 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 0.45% in all scenarios). This means that the distribution of password categories (e.g., letter-only, digit-only, etc.) 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 both English password and Chinese password patterns can help the password guessing.

In total, from *EDict* to *CDict+Date*, we increase the guessing efficiency by 34% for *RockyouTS*. This guessing experiment imply that Pinyin and date's rules should be considered in password protection in websites. *E.g.*,

Algorithm 3 IdentifyComposition

Input:
S: a string that needs to match elements of Trie
Root: the root of the Trie

Output:
Whether the string S is composed of the element (s) in the Trie.

```
1: if S is NULL or S.length() is 0 then
2:    return FALSE
3: end if
4: letters[] ← S.toCharArray()
5: node ← Root
6: for i = 0; i < letters.length; i++ do
7:    pos ← letters[i]−'a'
8:    if node.child[pos] is NULL then
9:       if i is 0 then
10:         return FALSE
11:      end if
12:      if node.isValue is FALSE then
13:         return FALSE
14:      end if
15:      return IdentifyComposition(S.substring(i))
16:   else
17:      node ← node.child[pos]
18:      if IdentifyComposition(S.substring(i + 1)) is
         TRUE and node.isValue is TRUE then
19:         return TRUE
20:      end if
21:   end if
22: end for
23: return node.isValue
```

Web masters should tell Chinese users to reduce the usage of Pinyin or dates in composing their passwords.

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 ease of implementation. A large body of research has shown the characteristics of user-created passwords [14][16][22][23][31][29][15].

Morris *et al.* [25] described the history of the design of the password security scheme and studied the password habits of 3,289 Unix users. Yan *et al.* [38] studied the password memorability and security. They found that users rarely choose passwords that are both hard to guess and easy to remember. Howe *et al.* [20] studied the behavior of home computer users because home computer users are more likely to suffer from various attacks, *e.g.*, phishing [36], dictionary attacks [27], heuristic password guessing [35], or brute force attacks. Florencio *et al.* [18] 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.* [21] studied 12,000 actual passwords from several perspectives. They found that certain passwords policies which can improve the strength of user-created passwords are underestimated. In addition, a blacklist of weak passwords improves the security of passwords greatly. However, the aforementioned literature rarely mentioned the password difference between different regions, especially between Chinese and English users.

Bonneau [6] analyzed the language dependency of password guessing. The results show that among all `Yahoo` passwords, passwords created by Chinese are almost the hardest to guess. However, our experiments show that (1) the passwords of both English and Chinese users are similar in strength as shown in Figure 3 and Table 13; (2) if an attacker is aware of the fundamental differences between two languages (as pointed out in this paper), she or he can guess Chinese passwords efficiently. Moreover, our empirical study is based on two groups of websites: five Chinese websites, and two English websites, which represents a larger and more diverse corpus of passwords than `Yahoo` data set in Bonneau's work, and our corpus include passwords from users that only speak Chinese, unlike the Chinese users in Bonneau's work who should be familiar with English. Bonneau *et al*. [9] also investigated the lingering effects of character encoding on the password ecosystem based on password datasets from Chinese, English, Hebrew and Spanish speakers. Comparing with the results in [9], our large-scale empirical analysis in this paper also shows that the strength of the passwords of Chinese and English users is similar. Moreover, we firstly quantitatively measure how an attacker can leverage the lingering effects to crack more Chinese passwords.

In terms of measuring the strength of passwords, NIST standards [11] propose to use Shannon's entropy to estimate the strength of a single password. Unfortunately, this method does not work well. Bonneau [6][8] proposed a set of metrics to measure the strength of passwords. These metrics are independent of what the passwords are, but depend on the distribution of the passwords. We modify these metrics to estimate the strength of passwords across websites. In addition, Kelly *et al.* [21] used guess numbers to measure the strength of passwords.

Guessing passwords has attracted much attention. Narayanan *et al.* [26] discussed a password-guessing algorithm based on Markov model. In this model, guessing passwords is based on the frequency of each character. Weir *et al.* [35] proposed a PCFG based password guess-

ing method. The PCFG generates password structures in the highest probability order based on a training set of passwords. Then, it generates word-mangling rules and guesses passwords from these rules. This approach provides us with an opportunity to examine the differences between Chinese and English passwords. In addition, Veras *et al*. [34] employed Natural Language Processing techniques to understand the semantic patterns in passwords, then cracked more passwords than a state-of-the-art approach did.

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. Moreover, Pinyins and dates also appear often in their passwords. Leveraging these observations, we show that by adding rules and Pinyins into the dictionary for guessing passwords, we can improve the guessing efficiency of cracking Chinese passwords by 34%.

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 worthy exploring the semantic meanings of the Chinese passwords.

6 Acknowledgement

This paper is supported by Key Lab of Information Network Security, Ministry of Public Security (C13612), CNNIC DNSLab, Natural Science Foundation of Shanghai (12ZR1402600), 12th Five-Year National Development Foundation for Cryptography (MMJJ201301008), 1000 Young Talent plan from the China Central Organization, supported by the Fundamental Research Funds for the Central Universities (2013QNA4019). We also would like to thank Ari Juels for his suggestion and anonymous reviewers for their comments. Weili Han is the corresponding author.

References

[1] The concise oxford dictionray of current english. http://archive.org/stream/conciseoxforddic00fowlrich/conciseoxforddic00fowlrich_djvu.txt.

[2] Wordlist from john the ripper. http://download.openwall.net/pub/passwords/wordlists/.

[3] Wordlist from outpost9. http://www.outpost9.com/files/WordLists.html.

[4] 178.COM. http://www.178.com/s/information/about.html.

[5] 7K7K. http://www.7k7k.com/html/about.htm.

[6] BONNEAU, J. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *Proceedings of 2012 IEEE Symposium on Security and Privacy (SP)* (2012), pp. 538–552.

[7] BONNEAU, J., HERLEY, C., VAN OORSCHOT, P. C., AND STAJANO, F. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proceedings of 2012 IEEE Symposium on Security and Privacy (SP)* (2012), IEEE, pp. 553–567.

[8] BONNEAU, J., PREIBUSCH, S., AND ANDERSON, R. A birthday present every eleven wallets? the security of customer-chosen banking pins. In *Proceedings of the 16th International Conference on Financial Cryptography (FC '12)* (2012).

[9] BONNEAU, J., AND XU, R. "of contraseñas, sysmawt, and mìmă: Character encoding issues for web passwords". In *Web 2.0 Security & Privacy* (May 2012).

[10] BOZTAS, S. Entropies, guessing, and cryptography. Tech. rep., Department of Mathematics, Royal Melbourne Institute of Technology, 1999.

[11] BURR, W. E., DODSON, D. F., NEWTON, E. M., PERLNER, R. A., POLK, W. T., GUPTA, S., AND NABBUS, E. A. Nist special publication 800-63-1 electronic authentication guideline, 2006.

[12] CNNIC. The 33rd survey report on chinese internet development. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201401/t20140116_43820.htm, Jan 2014.

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

[14] DAS, A., BONNEAU, J., CAESAR, M., BORISOV, N., AND WANG, X. The tangled web of password reuse. In *Proceedings of NDSS 2014* (2014).

[15] DE CARN DE CARNAVALET, X., AND MANNAN, M. From very weak to very strong: Analyzing password-strength meters. In *Proceedings of NDSS 2014* (2014).

[16] DELL'AMICO, M., MICHIARDI, P., AND ROUDIER, Y. Password strength: An empirical analysis. In *Proceedings IEEE INFOCOM 2010* (2010), IEEE, pp. 1–9.

[17] DUDUNIU. http://baike.baidu.com/view/1557125.htm.

[18] FLORENCIO, D., AND HERLEY, C. 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.

[19] HERLEY, C., AND VAN OORSCHOT, P. A research agenda acknowledging the persistence of passwords. *IEEE Security & Privacy 10*, 1 (2012), 28–36.

[20] HOWE, A., RAY, I., ROBERTS, M., URBANSKA, M., AND BYRNE, Z. The psychology of security for the home computer user. In *Proceedings of 2012 IEEE Symposium on Security and Privacy (SP)* (2012), pp. 209–223.

[21] KELLEY, P., KOMANDURI, S., MAZUREK, M., SHAY, R., VIDAS, T., BAUER, L., CHRISTIN, N., CRANOR, L., AND LOPEZ, J. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proceedings of 2012 IEEE Symposium on Security and Privacy (SP)* (2012), pp. 523–537.

[22] KUO, C., ROMANOSKY, S., AND CRANOR, L. F. Human selection of mnemonic phrase-based passwords. In *Proceedings of the Second Symposium on Usable privacy and security* (2006), ACM, pp. 67–78.

[23] MALONE, D., AND MAHER, K. Investigating the distribution of password choices. In *Proceedings of the 21st International Conference on World Wide Web* (2012), ACM, pp. 301–310.

[24] MAZUREK, M. L., KOMANDURI, S., VIDAS, T., BAUER, L., CHRISTIN, N., CRANOR, L. F., KELLEY, P. G., SHAY, R., AND UR, B. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM Conference on Computer and Communications Security* (2013), ACM, pp. 173–186.

[25] MORRIS, R., AND THOMPSON, K. Password security: A case history. *Communications of the ACM 22*, 11 (1979), 594–597.

[26] NARAYANAN, A., AND SHMATIKOV, V. Fast dictionary attacks on passwords using time-space tradeoff. In *Proceedings of the 12th ACM Conference on Computer and Communications Security* (2005), ACM, pp. 364–372.

[27] PINKAS, B., AND SANDER, T. Securing passwords against dictionary attacks. In *Proceedings of the 9th ACM Conference on Computer and Communications Security* (2002), ACM, pp. 161–170.

[28] PLIAM, J. O. On the incomparability of entropy and marginal guesswork in brute-force attacks. In *INDOCRYPT* (2000), pp. 67–79.

[29] R. VERAS, C. COLLINS, J. T. On the semantic patterns of passwords and their security impact. In *Proceedings of NDSS 2014* (2014).

[30] ROCKYOU. `http://rockyou.com/ry/about-us`.

[31] SAWYER, D. A. The characteristics of user-generated passwords. Tech. rep., DTIC Document, 1990.

[32] SCHWEITZER, D., BOLENG, J., HUGHES, C., AND MURPHY, L. Visualizing keyboard pattern passwords. In *Proceedings of 6th International Workshop on Visualization for Cyber Security (VizSec 2009)* (2009), IEEE, pp. 69–73.

[33] TIANYA. `http://help.tianya.cn/about/history/2011/06/02/166666.shtml`.

[34] VERAS, R., COLLINS, C., AND THORPE, J. On the semantic patterns of passwords and their security impact. In *Proceedings of NDSS 2014* (2014).

[35] WEIR, M., AGGARWAL, S., DE MEDEIROS, B., AND GLODEK, B. Password cracking using probabilistic context-free grammars. In *Proceedings of the 30th IEEE Symposium on Security and Privacy* (2009), IEEE, pp. 391–405.

[36] XIANG, G., AND HONG, J. I. A hybrid phish detection approach by identity discovery and keywords retrieval. In *Proceedings of the 18th Internation Conference on World Wide Web (WWW'09)* (2009), pp. 561–570.

[37] YAHOO. `http://info.yahoo.com/`.

[38] YAN, J., BLACKWELL, A., ANDERSON, R., AND GRANT, A. Password memorability and security: empirical results. *IEEE Security Privacy 2*, 5 (2004), 25–31.

A Method to Remove the Copied Passwords in Tianya

`Tianya` and `7k7k` have an unusually large number of the same accounts (identified by email) with the same passwords. Since the statistic features of these duplicated accounts are different from the ones between any other websites, thus we suspected that the attackers have copied accounts from `Tianya` to `7k7k` or vice versa.

To investigate whether the accounts are copied from `Tianya` to `7k7k` or vice versa, we performed the following analysis. We first divided all accounts from `Tianya` and `7k7k` into two groups: One group contains the users who have the same accounts and passwords both at `Tianya` and `7k7k`, and the other contains the users who do not. We call the passwords of the two groups *reused passwords* and *not-reused passwords*.

After analyzing the compositions (e.g., digit-only passwords) of the *reused passwords* and *not-reused passwords*, we found that the proportions of various compositions are similar between the *reused passwords* and the `7k7k`'s *not-reused passwords*, but different with *Tianya*'s *not-reused passwords*. As a result, we believe that it is likely that accounts have been copied from `7k7k` to `Tianya` and we deleted the *reused passwords* from `Tianya`.