How does personal information reside in human-chosen passwords? –A quantitative study

Yue Li
College of William & Mary
yli@cs.wm.edu

Haining Wang
University of Delaware
hnw@udel.edu

Kun Sun
College of William & Mary
ksun@cs.wm.com

ABSTRACT

Left blank.

Categories and Subject Descriptors

[**Security and privacy**]: *Human and societal aspects of security and privacy*; [**General and reference**]: *Metrics*

General Terms

Security

Keywords

passwords, password cracking, data processing, password protection

1. INTRODUCTION

Left blank

2. PERSONAL INFO IN PASSWORDS

Human-generated passwords are long criticized to be weak. Numerous works have shown that due to memorability requirement, users are more likely to use meaningful strings as their passwords. Therefore, passwords are usually very different from real random strings. For example, "password" is more likely a password than "ziorqpe". As a result, most passwords are within only a small portion of the large password space, making password guessing a lot easier. A natural question is: how do users choose their passwords so that they are different from random strings? The answer to the question has great significance for it has strong security implication to both users as well as systems. If an attacker knows exactly how users construct their passwords, cracking their passwords will become an easy task. On the other hand, if a user knows how other users construct their passwords, the user can easily improve his/her password strength by avoid using these password construction methods.

To this end, researchers have done much to unveil the composition of passwords. Traditional dictionary attacks on passwords have shown that users tend to use dictionary words to construct their passwords. [][] claims that the distribution of characters in passwords is very similar to that in their native languages and people are prone to use words in their languages. [] Shows password words distribution is not the same as word distribution in the language. [Markov] shows that passwords are phonetically memorable. [PCFG] shows that using dictionary words to guess passwords is effective. [][][] indicate that users use keyboard strings such as "qwerty" and "qweasdzxc", trivial strings such as "password", "123456", and date strings such as "19951225" in their passwords.

As far as we see, most studies are done at a macro level. We now study user passwords in an individual base. We would like to show that user personal experience plays an important role when users create their passwords. Intuitively, people tend to choose their passwords based on their personal information because human beings are limited by their memory – totally unrelated passwords are much less memorable.

2.1 12306 Dataset

In recent years, many password datasets are exposed to the public. Recent works on password measurement or password cracking are usually based on these datasets. Some of these datasets, such as Rockyou, are very large such that they even constitute millions of passwords. Now we are going to use a dataset which we call 12306 dataset to illustrate how personal information is used in user passwords.

2.1.1 Introduction to dataset

At the end of year 2014, a Chinese dataset is exposed to the public by anonymous attackers. It is said that the dataset is obtained using social engineering[], in which attackers use datasets at hand to try other websites. We call this dataset 12306 dataset because all passwords are from a website www.12306.com. The website is the official website for online railway ticket booking for Chinese users.

12306 dataset contains over 130,000 Chinese passwords. Having witnessed so many large datasets been leaked out, the size of 12306 dataset is just medium. What makes it special is that together with plain text passwords, the dataset also carries several types of user personal information. For example, user's name, ID number, etc. As the website needs real ID number to register and people need to provide real information to book a ticket, information in the dataset is considered reliable.

Table 1: Most Frequent Passwords

Rank	Password	Amount	Percentage
1	123456	389	0.296%
2	a123456	280	0.213%
3	123456a	165	0.125%
4	5201314	160	0.121%
5	111111	156	0.118%
6	woaini1314	134	0.101%
7	qq123456	98	0.074%
8	123123	97	0.073%
9	000000	96	0.073%
10	1qaz2wsx	92	0.070%

2.1.2 Basic Measurement

We do fundamental measurement to reveal some characteristics of 12306 dataset. After appropriate cleansing, we remove a minor part of passwords (0.2%), with 131,389 good passowrds left for analysis. Note that websites may have different password creation policy. With strict password policy, users may apply mangling rules (For example, $abc- > @bc$ or $abc1$)to their passwords to fulfill the policy requirement. As 12306 website has changed its password policy after the password leakage, we do not know exactly the password policy at the time the dataset is leaked. However, from the dataset, we infer the password policy is quite simple – all passwords need to be no shorter than 6 symbols. There is no restriction on what type of symbols are used. Therefore users are not forced to apply much mangling to their passwords.

The average length of passwords in 12306 dataset is 8.44. Then we show the most common passwords in 12306 dataset. They are listed in Table 1.

From Table 1 we can see that the dominating passwords are trivial passwords (123456, a123456, etc), keyboard passwords (1qaz2wsx and 1q2w3e4r), and "I love you" passwords. Both "5101314" and "woaini1314" means "I love you forever" in Chinese. The most commonly used Chinese passwords are similar to previous studies [fudan]. However, 12306 dataset is much less congregated. The most popular password "123456" accounts less than 0.3% of all passwords while the number is 2.17% in [Fudan]. We believe that the sparsity is due to the importance of the website so that users are less prone to use trivial passwords like "123456", etc.

Then, we show the basic structure of passwords. The most popular password structures are shown in Table 2. Our result again shows that Chinese users prefer to use digits in their passwords instead of letters as in English-speaking users. The 5 top structures all have significant portion of digits, in which at most 2 or 3 letters are appended in front.

We reckon that the reason behind may be Chinese users lack vocabulary because Chinese use non-ASCII character set. Digits seem to be the best choice when creating a password.

In conclusion, 12306 dataset is a Chinese password dataset that has general Chinese password characteristics. However, its passwords are more sparse than previously studied datasets.

2.2 Personal Information

As we have mentioned, 12306 dataset not only contains user passwords, it also carries multiple types of personal information. They are:

Table 2: Most Frequent Password Structures

Rank	Structure	Amount	Percentage
1	D_7	10893	8.290%
2	D_8	9442	7.186%
3	D_6	9084	6.913%
4	L_2D_7	5065	3.854%
5	L_3D_6	4820	3.668%
6	L_1D_7	4770	3.630%
7	L_2D_6	4261	3.243%
8	L_3D_7	3883	2.955%
9	D_9	3590	2.732%
10	L_2D_8	3362	2.558%

"D" represents digits and "L" represents English letters. The number indicates the segment length. For example, D_7 means the password contains 7 digits in a row.

```
1. Name: User's Chinese name
2. Email address: User's registered email address
3. Cellphone number: User's registered cellphone number
4. Account name: the account used to log on the system,
may contain digits and letters. For example, "myacct123".
5. ID number: Government issued ID number.
```

Note that the government issued ID number is an 18-digit powerful number. These digits actually show personal information as well. Digit 1-6 represents the birth place of the owner, Digit 7-14 represents the birthday of the owner, and digit 17 represents the gender of the owner – odd number means male and even number means female. We take out the 8-digit birthday information and treat it separately because birthday information is very important in a password. Therefore, we finally have 6 types personal information - 1)Name, 2)Birthday, 3)Email, 4)Cellphone 5)Account name, and 6) ID number (birthday not included).

2.2.1 New Password Representation

To better illustrate how personal information correlates to user passwords, we develop a new representation of password which add more semantic symbols beside the conventional "D", "L" and "S" symbols, which means digit, letter, and special symbol accordingly. We try to match password to the 6 types of user personal information, and express the passwords with these personal information. For example, a password "alice1987abc" may be represented as $[Name][Birthday]L_3$ instead of $L_3D_4L_3$ in a traditional measurement. We substitute personal information with corresponding tag ([Name] and [Birthday] in this case). For the segments that are not matched, we still use "D","L", and "S" to describe the types of characters.

We believe representation like $[Name][Birthday]L_3$ is better than $L_5D_4L_3$ since it more accurately describe the composition of user passwords. We apply the matching to the whole 12306 dataset to see how these personal information tag appear in such password representations.

2.2.2 Matching Method

In order to make personal information password representations, an essential question will be: How do we match the personal information to user passwords? To answer this question, we show the algorithm we used in Algorithm 1.

The high level idea is that we find all substrings of the password and sort them in descending length order. Then we try to match the substrings from longest to shortest to all types of personal information. If one match is found, the leftover password segments are recursively applied the match function until no further match is found. Segments that are not matched to personal information will be processed using the traditional "LDS" method.

Algorithm 1 Match personal information with password

```
1: procedure MATCH(pwd,infolist)
2:     newform ← empty_string
3:     if l thenen(pwd) == 0
4:         return empty_string
5:     end if
6:     substring ← get_all_substring(pwd)
7:     reverse_length_sort(substring)
8:     for eachstring ∈ substring do
9:         if len(eachstring) ≥ 2 then
10:            if matchbd(eachstring,infolist) then
11:                tag ← "[BD]"
12:                leftover ← pwd.split(eachstring)
13:                break
14:            end if
15:            ...
16:            if matchID(eachstring,infolist) then
17:                tag ← "[ID]"
18:                leftover ← pwd.split(eachstring)
19:                break
20:            end if
21:        else
22:            break
23:        end if
24:    end for
25:    if leftover.size() ≥ 2 then
26:        for i ← 0 to leftover.size()-2 do
27:            newform ← MATCH(leftover[i],infolist) +
   tag
28:        end for
29:        newform ← MATCH(leftover[leftover.size()−
   1])+newform
30:    else
31:        newform ← seg(pwd)
32:    end if
33:    return newform
34: end procedure
```

Note that we did not show specific matching algorithm to each type of the personal information (line 10 and line 16). To keep Algorithm 1 clean and simple, we describe the matching methods as follows.

First we make sure the password segments are at least length of 2 for matching. For segment of length 1, we directly map it to digit, letter, or special character. We try to match segments with length 2 or more to each kind of the information. For name information, we first convert Chinese names into Pinyin form, which is alphabetic representation of Chinese. Then we compare password segments to 10 possible permutations of the names, which include $lastname + firstname$, $last\_initial + firstname$, etc. If the segment is exactly same as any of the permutations, we consider a match is found. We list all the 10 permutations in the Appendices. For birthday information, we list 17 possible permutations and compare password segments to each of the permutation, if the segment is same as any permutations, we consider a match is found. We list all the birthday permutations in the Appendices. For account name, cellphone number, and ID number, we further restrain the length of segment to be at least 4 to avoid coincidence. We believe a match of length 4 is very likely to be an actual match. If the segment is a substring of any of the 3 personal information, we regard it a match to the corresponding personal information.

Table 3: Most Frequent Password Structures

Rank	Structure	Amount	Percentage
1	D7	7105	5.407%
2	[ACCT]	6103	4.644%
3	[NAME][BD]	5410	4.117%
4	D6	4873	3.708%
5	[BD]	4470	3.402%
6	D8	4233	3.221%
7	L1D7	3286	2.500%
8	[NAME]D7	2941	2.238%
9	[NAME]D3	2363	1.798%
10	[NAME]D6	2241	1.705%

Table 4: Most Popular Personal Information

Rank	Information Type	Amount	Percentage
1	[BD]	30674	23.34%
2	[NAME]	29653	22.56%
3	[ACCT]	17065	12.98%
4	[EMAIL]	4229	3.218%
5	[ID]	2918	2.220%
6	[CELL]	529	0.402%

2.2.3 Matching Result

After applying Algorithm 1 to 12306 dataset. We found that 71,037 out of 131,389 (54.1%) of the passwords contain at least one of the 6 types of personal information. Apparently, personal information is an essential part of user passwords and most users put certain personal information in their passwords. We believe the rate could be higher if we have more personal information at hand. However, this percentage has served its purpose properly. Then We present the top 10 password structures in Table 3 and most commonly used personal information in Table 4. Based on Table 3 and Table 4, we have the following observations

1. The second and third structures are perfectly matched to personal information. 3 out of the top 10 structures are composed by pure personal information and 6 out of the top 10 structures have personal information segment. The dominating structures D_7, D_6, and D_8 in Table 2 still rank fairly high.

2. Birthday, name, and account name are most popular personal information in user passwords. Over 20% passwords in our dataset contain birthday or name information. On the other hand, much less people use email and ID number in their passwords. Further more, only few people include their cellphone number in their passwords.

3. Set aside personal information, digits are still dominating user passwords. Only one structure from the top 10 structures has one letter segment with minimum length (1). The result confirms that Chinese users prefer to use digits in their passwords.

4. An interesting observation is that although account name has merely half percentage as birthday and name information, the structure [ACCT] ranks highest among all structures that contain personal information. The reason behind may be that users tend to use their account names as their passwords instead of using them as part of their passwords.

2.2.4 Gender difference

We are also interested in the difference of password composition between males and females. Note that although the dataset does not have a gender column, user ID number actually has gender information (The second last digit in ID number represents gender). We realized that the dataset is biased in gender, with 9,856 females and 121,533 males in it. To balance the number, we randomly select 9,856 males from the male pool and compare them with females. The average length of passwords for males and females are 8.41 and 8.51, which are quite similar. It shows that males and females do not differ much in the length of their passwords. We then apply the matching method to each of the genders. We found that 55% of male passwords contain personal information while 45% of female passwords contain personal information. Therefore we conclude that generally males put more personal information than females in their passwords. It indicates females have wider thought when it comes to password. It also implies that females have more complex passwords, and therefore maybe more secure. We list the top 10 structures for each gender in Table 5 and personal information usage in Table 6. From the tables we have the following observations:

1. For males, 6 out of the top 10 structures contain personal information. Yet for females, only 3 out of top 10 structures contain personal information. It further implies that males are more likely to consider personal information when creating passwords.

2. For both males and females, [ACCT], [NAME][BD], and [BD] are three most frequent structures with personal information. However, The percentage of males are much higher than that of females. Averagely 47.3% more males are constructing their passwords following the 3 patterns than females.

3. From Table 6 we can see the percentage of each type of personal information in the passwords. Interestingly males and females are very different in the usage of name information. Males use their names as frequent as their birthday (23.43% passwords of males contain their names) while only 13.03% passwords of females contain names. We also notice that the name usage mostly contribute the 10% difference in personal information usage between males and females.

4. Users seem not like applying mangling rules on their passwords. We notice that the several most frequent password structures are either pure personal information (such as $[BD]$) or strings (such as D_6) that do not relate to any of the personal information. Structures like $[NAME]D_7$ are less likely to appear in user password.

Table 5: Most Frequent Structures in Different Gender

Rank	Male		Female	
	Structure	Percentage	Structure	Percentage
1	D_7	5.732%	D_6	4.870%
2	[ACCT]	4.799%	D_7	4.220%
3	[NAME][BD]	4.190%	[ACCT]	4.017%
4	[BD]	3.591%	D_8	3.246%
5	D_6	3.520%	[NAME][BD]	2.607%
6	D_8	2.861%	[BD]	2.221%
7	L_1D_7	2.840%	L_2D_6	2.079%
8	[NAME]D_7	2.333%	L_2D_7	1.704%
9	[NAME]D_6	1.968%	L_1D_7	1.684%
10	[NAME]D_3	1.785%	L_3D_6	1.663%

Table 6: Most Frequent Personal Information in Different Gender

Rank	Male		Female	
	Information Type	Percentage	Information Type	Percentage
1	[BD]	23.75%	[BD]	20.30%
2	[NAME]	23.43%	[ACCT]	13.24%
3	[ACCT]	13.06%	[NAME]	13.03%
4	[EMAIL]	2.881%	[EMAIL]	4.403%
5	[ID]	2.343%	[ID]	1.461%
6	[CELL]	0.263%	[CELL]	0.385%

2.3 Service Information

2.4 Ethical Consideration

We do realize that studying leaked datasets involves much ethical concern. Like many other works, we only use this dataset for researching purpose. We will not expose any user personal information of in this dataset.

3. CORRELATION QUANTIFICATION

3.1 Coverage

4. PERSONAL-PCFG, AN INDIVIDUAL-ORIENTED PASSWORD CRACKER

4.1 Attack Scenarios

5. PASSWORD PROTECTION

6. RELATED WORKD

Left blank

7. REF

8. ACKNOWLEDGMENTS

9. REFERENCES

APPENDIX

A. FULL LIST OF BIRTHDAY AND NAME IN MATCHING ALGORITHM