A Large-Scale Study of Web Password Habits of Chinese Network Users

Zhipeng Liu
1. College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Yudao Street 29, Nanjing, Jiangsu, 210016, P.R. China
2. Department of Software Engineering Nanjing University of Posts and Communications Nanjing, Jiangsu Province, China
Email: liuzhipengcs@139.com

Yefan Hong
Department of Software Engineering Nanjing University of Posts and Communications Nanjing, Jiangsu Province, China

Dechang Pi
College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Yudao Street 29, Nanjing, Jiangsu Province, P.R. China
Email: dc.pi@nuaa.edu.cn

***Abstract*—nowadays, more and more people in China are connecting to Internet; the network environment becomes less secure due to large amount of network intruders. However, many of the network users tend to set easy passwords for their convenience. These vulnerable passwords increase the risk of information leakage. So it is necessary for us to analyze the habits and strength of the passwords set by the Chinese network users. In this paper, we collect over 20 million pieces of data published on the Internet by network intruders and analyze the features of passwords through statistical methods. We find some interesting patterns in order to quantify password strength through comprehensive analysis of password length, type, and other variables. Finally, we propose some suggestions for setting secure passwords.**

***Index Terms*—password habits, password strength, Chinese network users, password security; web security**

I. INTRODUCTION

Passwords are words made up of a group of characters or numbers that are used for authentication, to prove identity or gain access to a resource [1]. Passwords have been widely utilized in web environments, and they are near universal means for identification authentication [2]. Users should keep their own passwords secure and safe from being cracked by network intruders. Otherwise they will risk almost a certain loss.

While the majority of users know that we will have more risk with simple passwords, a convenient login is extremely attractive; so users often ignore the risks and set simple passwords. For instance, if a user has 20 network accounts, it is not practical for him to remember 20 different passwords. The general approach is to apply 3 to 5 passwords for all accounts. Because of this, in recent years, a large amount of phishing sites have appeared on the Internet. These sites try to obtain the users' passwords by attracting users to login in the disguised system. Some sites try to entice users to download and install malicious programs in their operating systems. These programs are designed to capture detailed information of the owners of the infected systems. The network intruders utilize this information to try other sites in order to obtained valuable information. At the end of 2011, the databases of a number of Chinese Internet corporations have been attacked by network intruders. From December 21st to 31st, tens of millions of user account information, including user names, passwords, registered mail addresses and other information, have been openly published on the Internet. The Chinese Software Development Network (CSDN), the largest online community of developers in China, was also involved in this event and the information of 6 million users was compromised [3]. Renren Network, the largest Social Network Sites (SNS) in China has also been intruded, and the information of 2 million users has been released. Later, these sites have put up notices or sent e-mails to each user to ask them to change their passwords for security reasons. In this paper, we have gathered hundreds of millions of data to study the web password habits of Chinese network users and measure the password strength.

We introduce the sources of our data and the methods of processing them in Section 2. We present our analysis of the findings and results in Section 3. Finally, we propose some suggestions for setting secure passwords.

II. RELATED WORK

The study of the habit of passwords, and password strength for authentication has been analyzed including length, types, and its comparative security [4-7]. In 1979, Morris and Thompson analyzed the password strength on the UNIX operating system [8] by attempting to crack hashed passwords. An experiment carried by Morris and Thompson [9] has examined widely-used weak

doi:10.4304/jsw.9.2.293-297

passwords on a number of machines. They found that 86% of the passwords were too short, with digits or lowercase only, which was easily found in dictionaries. Later, Klein etc. [4, 10] have utilized certain rule-sets that aimed at excluding weak passwords with a system called pro-active password checkers or password strength meters (PSM). Florencio and Herley [2] have reported the results of a large scale study of password use and password reuse habits, which involved half a million users over a three month period. Castelluccia, Duermuth and Perito [11] have shown how to build secure adaptive password strength meters and proposed the first adaptive password strength meter (APSM) in 2012. Gaw and Felten [12] have collected the usage of password habits of 49 participants and presented a lot of password management strategies. Dinei, Cormac and Baris [13] have concluded that forcing users to choose strong passwords appears no effect to the security of accounts in web environments. Matteo Dell'Amico, Pietro Michiardi and Yves Roudier [14] have carried out an elaborate analysis of the password strength of Internet applications on 3 datasets. They have found a “diminishing returns” principle which means a success of password guess decreases by orders of magnitude for attacks. Other work related to web security include [15-17].

From description above we can understand that there have been many researches on user password habits. However, most of the data have been obtained by survey methods, and the scale of data was limited. To the best of our knowledge, no research has been carried out to analyze the web password habits of Chinese network users. In this paper, we use the actual data collected on the Internet which were published by network intruders. It is unprecedented that we have collected over 20 million pieces of data, more than most of the existing surveys about the information about the information of network users in China.

TABLE I

DATA FORMATS OF 4 SITES

Sites	Package(s)	Data Format
CSDN	1	Username # Password # E-mail
Renren	1	E-mail Password
178	11	ID Username Password
7k7k	4	E-mail Password

III. METHODS

In this paper, we analyze data from four sites, including CSDN.net (Hereinafter referred to as "CSDN"), Renren Network (Hereinafter referred to as "Renren"), 178.com (Hereinafter referred to as "178"), and 7k7k.com (Hereinafter referred to as "7k7k"). More than 90 percent of the best programmers in China have accounts on CSDN; we infer that the passwords features of CSDN members are representatives of Chinese computer professionals.

We obtained the number of packages and the format of the data by reading the first 10 lines of 4 sites using head command in Linux operating systems.

Table 1 shows the data of 178 and 7k7k has been stored in different packages; with 11 and 4 packages for 178 and 7k7k, respectively. We merge them into one file before data processing.

We observe that there are totally 3 data formats among the 4 sites. So we write 3 scripts in python languages in order to standardize and preprocess the data. First, we remove the invalid information, such as some items in Renren package are divided into three messages. Second, we extract passwords from each package for passwords pattern extraction. Third, as we need to match the username or E-mail address with passwords in Section 3, we also separately extract usernames and E-mail addresses into another file.

Table 2 shows the data items of 4 sites. The data of 7k7k accounts for 37% of the total data items. 178 accounts for 34%, CSDN 24% and Renren 5% respectively.

We use python language with regular expressions to analyze these data sets.

IV. EXPERIMENTAL RESULTS

A. Password Length Analysis

The most common method to obtain stronger passwords is to enhance password length.

We assume $S = m^n$ as password strength where m denotes the number of candidate characters, and n denotes the length of passwords. There is no doubt that increasing password length improves the password strength exponentially. We present the results of password length analysis in Figure 1.

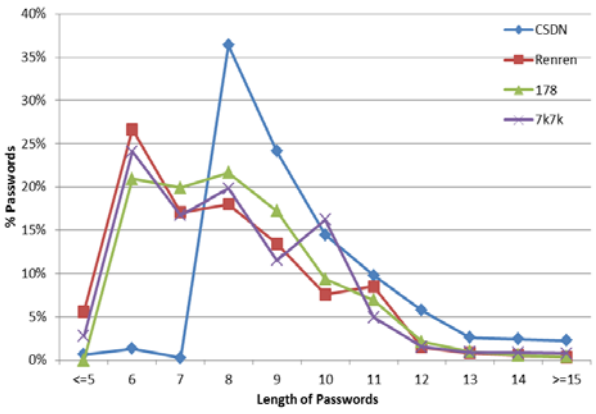

Figure 1. Passwords length of 4 sites

We can observe the distribution and average value of password length for 4 sites. The average password length is ranging from 7.7477 to 9.4571 for the 4 sites; we can see that the average password length of CSDN users is the longest among 4 sites. The majority of CSDN password lengths are in the range from 8 to 11 (blue flags in Figure 1). This result reflects the computer professionals in China are very conscious about information security. They also know that increasing the length of passwords can improve security levels.

However, the situations of password lengths in other 3 sites are not optimistic. Renren, mainly used by young students, tends to set short passwords. It reflects that most young people, especially students, often ignore the requirements of the system to set secure passwords. According to our offline survey, 22 students of 30 feel that it is troublesome to set a long password. They tend to persuade themselves that the contents behind their accounts are of little value to network intruders. 178 and 7k7k serves online players, the average password lengths of them are similar to Renren. We can observe that many people in China tend to set short passwords for convenience. This is not conducive to the security of user passwords and is likely to cause information leakages. Klein has reported that it is possible to obtain about 25% of passwords on a UNIX system by brute force attack [4], and he suggests that it is dangerous for passwords which are 8 characters long. From this we can infer that at least 25% passwords of web users in China are inherently vulnerable.

B. Types of Passwords Analysis

Adding complexity to passwords is another way to achieve stronger passwords. An experiment carried by Grampp and Morris found that weak passwords, such as names followed by a single digit, were in widespread use on a number of machines they examined in a corporate network [8]. The result of P. Leong and C. Tham [7] also support these conclusions. In summary, we shouldn't set simple passwords. However, although most users know this as common sense, they often tend to ignore it. One reason for this is that complex passwords are hard to remember, and the processes to retrieve the passwords are complicated in most network situations.

The alphabetical size is the sum of the sizes of the different types of characters. These types and sizes are lowercase (26), uppercase (26), digits (10) and special (22) [2]. We define a as the number of candidate digits, b as the number of candidate lowercase, c as the number of candidate uppercase and d as the number of candidate special characters. x, y, and z, are defined as length of digits, lowercase and uppercase respectively. Then we have 4 cases:

1） Digits only, $S = a^n = 10^n$;

2） Lowercase only, $S = b^n = 26^n$

3） Letters only but including uppercase, $S = C_n^y b^y \cdot C_{n-y}^{n-y} c^{n-y} = C_n^y 26^y \times 26^{n-y} = C_n^y 26^n$

4） Include digits, lowercase, uppercase and special,

$$S = C_n^x a^x \cdot C_{n-x}^y b^y \cdot C_{n-x-y}^z c^z \cdot C_{n-x-y-z}^{n-x-y-z} d^{n-x-y-z}$$
$$= C_n^x 10^x \cdot C_{n-x}^y 26^y \cdot C_{n-x-y}^z 26^z \cdot 32^{n-x-y-z}$$
$$= C_n^x C_{n-x}^y C_{n-x-y}^z 10^x \cdot 26^{y+z} \cdot 32^{n-x-y-z}$$

It is obvious that for the same password length, more complex type of password leads to the stronger password strength.

Figure 2 depicts the format of the passwords and the number of character types they contains.

We can observe the percentage of passwords which belongs to different password types. The password types are linked with the perceived importance of the password owners. The proportion of the digits-only password types for both Renren and 7k7k are very high. The percentage for Renren is nearly 70% and 7k7k is nearly 60% of all passwords. The proportions of passwords including special characters for these two sites are relatively low (both less than 30%). This means that passwords of these two sites are not of a strong type. However, passwords belong to CSDN and 178 are relatively safer in password strength. This is due to the lower percentage of digits-only passwords and the higher percentage of including special characters.

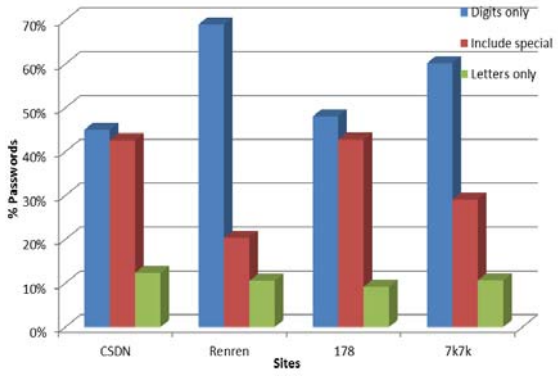

Figure 2. Types of passwords for 4 sites

In general, users of CSDN are better in the combination of passwords. However, the percentage of digits-only passwords is still higher than passwords including special characters. We can infer that the percentage of digits-only passwords of network users in China exceed 50%, while the percentage of passwords including special characters is less than 30%. This means that it is relatively easier for network intruders to decipher passwords of Chinese network users.

There are many methods to facilitate people to remember passwords more easily. Here, we list 4 methods: passwords with birthdays, mobile numbers, user names or E-mail addresses. Here we discuss the first 2 methods and the other 2 will be discussed in Section 4.3.

Many users prefer to use birthdays as their email account passwords. In recent years, with the popularity of mobile phones, it is also common for people to use mobile numbers as common elements of passwords. However, in our daily lives, the chances of personal information leakage are much higher. Our birthdays and mobile numbers can even be obtained by Google search engine.

If network intruders have been informed of such information, they usually use brute-force methods to attack accounts via dictionary. The dictionary contains passwords which a number of network users often use. It is often a text file [8]. As a result, there are a lot of risks despite of convenience. The distribution of passwords with the information of birthdays or mobile numbers is depicted in and Figure 3.

We can observe the percentage of passwords which contain birthdays and mobile phone numbers in Figure 3. It is clear that in CSDN, more than 12% network users use this method to set their passwords, and this will bring a huge security risk. Users in the remaining 3 sites, however, have a lower proportion(less than 10%).

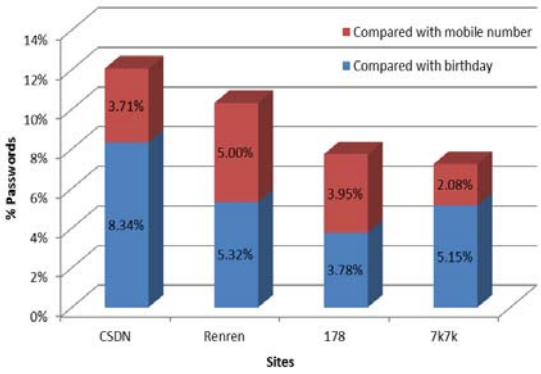

Figure 3. Information about passwords for 4 sites

C. Password with User Names or E-mail Addresses

It is extremely dangerous to set passwords with user names or E-mail addresses because these information are easy to be obtained. However, there are still many network users who are in favor of this method.

Table 2 shows the distribution of passwords with user names or E-mail addresses. The percentage of using user names or E-mail addresses of CSDN is relatively high. From Section 4.1 and Section 4.2 We know that users of CSDN tend to choose long and complex passwords. However, most of their passwords contain user names or E-mail addresses, which is extremely vulnerable.

TABLE III
DISTRIBUTION OF PASSWORDS WITH USER NAMES OR E-MAIL ADDRESSES

Password Contains	CSDN		Renren		178		7k7k	
Neither	5,688,694	88.49%	1,289,795	94.30%	8,742,390	96.36%	9,942,268	97.57%
Username	482,183	7.50%			330,576	3.64%		
E-mail	257,754	4.01%	77,990	5.70%			247,928	2.43%
Both	138,070	2.15%						

D. Password Strength Analysis

Password strength is a measurement of the effectiveness of preventing a password from brute-force attacks. It estimates how many attempts an attacker would need to "guess" password correctly. The password strength is a function of length, complexity, and unpredictability [5]. Password strength can also be described as "Strong" or "Weak". Thus, passwords which are longer and contain more types of characters are considered as "Strong" passwords. However, if they are matched with user names or E-mail addresses, these kinds of passwords is considered to be weak.

We investigate the percentage of all password features according to password length, which include digits, lowercase, uppercase and special characters. We analyze and summarize password strength for our 4 network sites. To facilitate our analysis, we transform primary 7 levels (Very safe, Safe, Very strong, Strong, Moderate, Weak, Very Weak) into 3 levels (Strong, Moderate, Week). Which means, Very Safe, Safe and very strong are merged into "Strong", Strong and Moderate are merged into "Moderate", Week and Very Week are merged into "Weak".

TABLE IV
DISTRIBUTION OF PASSWORD STRENGTH OF CSDN

Password Features					Password strength		
Length	Digits	Lower-case	Upper-case	Special Characters	Strong	Moderate	Weak
<=5	96.13%	3.75%	0.09%	0.03%	0.00%	0.00%	100.00%
6	90.70%	9.07%	0.16%	0.07%	0.00%	0.06%	99.93%
7	61.44%	37.45%	0.76%	0.35%	0.04%	0.58%	99.38%
8	75.34%	23.10%	1.15%	0.41%	0.29%	2.12%	97.59%
9	71.67%	26.28%	1.60%	0.46%	0.44%	2.61%	96.96%
10	60.07%	37.06%	2.26%	0.61%	0.65%	4.17%	95.18%
11	69.08%	28.36%	1.92%	0.63%	0.99%	4.47%	94.54%
12	58.37%	37.98%	2.78%	0.86%	1.61%	6.06%	92.33%
13	47.96%	47.20%	3.57%	1.26%	2.85%	9.57%	87.58%
14	52.82%	42.74%	3.36%	1.08%	2.99%	7.96%	89.05%
>=15	44.11%	49.38%	4.39%	2.13%	7.13%	12.55%	80.32%

TABLE II
DISTRIBUTION OF PASSWORD STRENGTH FOR RENREN, 178 AND 7K7K

Password Features					Password strength		
Length	Digits	Lower-case	Upper-case	Special Characters	Strong	Moderate	Weak
<=5	88.15%	10.93%	0.54%	0.39%	0.00%	0.06%	99.94%
6	85.64%	13.90%	0.35%	0.12%	0.00%	0.12%	99.89%
7	81.19%	18.25%	0.45%	0.11%	0.00%	0.12%	99.87%
8	74.34%	25.01%	0.47%	0.18%	0.08%	0.86%	99.05%
9	69.08%	30.13%	0.61%	0.18%	0.14%	1.07%	98.79%
10	61.26%	33.04%	5.45%	0.25%	0.46%	18.22%	81.32%
11	77.52%	20.33%	0.93%	1.22%	5.78%	1.44%	92.78%
12	47.35%	50.23%	1.24%	1.18%	0.81%	3.61%	95.58%
13	41.91%	55.96%	1.33%	0.81%	1.27%	5.75%	92.98%
14	47.88%	47.88%	1.38%	2.86%	1.49%	16.57%	81.93%
>=15	41.50%	49.49%	2.00%	7.01%	3.40%	7.73%	88.88%

We then examine the password strength. The results of password strength of CSDN are listed in Table 3, and the results of other 3 sites are listed in Table 4.

V. SUGGESTIONS OF SETTING SECURE PASSWORDS

Humans are the center of information activities [10], and information need protection. Nowadays, the most widely used form of authentication is to require a user to provide the specified password. People should pay attention to the password management in case of information leakage. Setting secure passwords can prevent losses from network intruders via password dictionary or professional decryption software to crack passwords [11]. The most important thing is that people should set secure passwords. We present 3 suggestions below:

1. Avoid simple passwords

Simple passwords are easy to remember, which facilitates authentication process. However, according to our analysis, simple passwords are detrimental for

protection purpose. The network users in China have to reconsider password setting procedures.

2. Substituting digits with special characters

To add special characters to our passwords is another good way for setting good passwords. For example, we substitute "@" for "2", substitute "*" for "8". These special characters can enhance password strength for they are not belonging to letters and numbers.

3. Using a domain suffix

We can set a "basic" password of 6-8 characters. Then we add a "variable" password which denotes the domain names. So the "final" password consists of the "basic" password with "variable" password. For instance, if we specify the "basic" password "abc123", then for Gmail.com, we assign the "variable" password as "gmail" and the final password would be "abc123gmail". This password is much better than it "basic" partners.

VI. CONCLUSION

The large-scale data of passwords allows us to measure the password strength of Chinese people. Many of the network users tend to set easy passwords for their convenience. These vulnerable passwords increase the risk of information leakage. We have collected over 20 million pieces of data published on the Internet by network intruders and analyzed the features of passwords through statistical methods. We have found some patterns to quantify password strength through comprehensive analysis of password length, type, and other variables. We have shown that the password strength of Chinese users is weak, and the Chinese web users have a low sense of information security. We have proposed some suggestions for setting secure passwords.

ACKNOWLEDGMENT

This research is sponsored by Qing Lan Project, the 333 project of Jiangsu Province， the Aeronautical Science Foundation of China (No. 20111052010) and 2012 Jiangsu Graduates Innovation Project (CXZZ12_0163).

REFERENCES

[1] Wikipedia.Password. http://en.wikipedia.org/wiki/Password.

[2] Dinei Florencio, Cormac Herley. *A Large-Scale Study of web password habits*. International World Wide Web Conference Committee (IW3C2). Banff, Alberta, Canada, pp.657:665, 2007.

[3] Jinghua Times. *600 million users information has been compromised on CSDN*. http://tech.sina.com.cn/i/2011-12-22/02326546996.shtml. 2011.12.22.

[4] Daniel V. Klein. *"Foiling the Cracker" : A Survey of, and Improvements to, Password Security*. In Proceedings of the 2nd USENIX Security Workshop. Pages 5-14, August 1990.

[5] Wikipedia. Password strength. http://en.wikipedia.org/wiki/Password_strength.

[6] *How to make a password strength meter like Google* v2.0[EB/OL] http://www.codeandcoffee.com/2007/07/16/how-to-make-a-password-strength-meter-like-google-v20/.

[7] P. Leong and C. Tham, *UNIX password encryption considered insecure*. In: USENIX Winter Conference Proceedings (Jan. 1991)

[8] F.T.Grampp and R. H. Morris. *UNIX Operating System Security*. Bell System Tech. Journal, 1984.

[9] R. Morris and K. Thompson. *Password Security: A Case History*. Comm. ACM, 1979.

[10] M. Bishop and D. V. Klein. *Improving system security via proactive password checking*. Computers & Security, 14(3), 1995.

[11] Claude Castelluccia, Markus Duermuth, Daniele Perito. *Adaptive Password-Strength". Meters from Markov Models.* Network & Distributed System Security Symposium, NDSS 2012.

[12] Shirley Gaw and Edward W. Felten. *Password management strategies for online accounts*. In Proceedings of the 2nd symposium on Usable privacy and security, Page 44-55, July 2006.

[13] Dinei Florêncio, Cormac Herley and Baris Coskun. *Do strong web passwords accomplish anything.* Usenix Hot Topics in Security, June 2007.

[14] Matteo Dell'Amico, Pietro Michiardi and Yves Roudier. *Password strength: an empirical analysis*. In INFOCOM 2010 Proceedings. March 2010.

[15] R Padmavathy and Chakravarthy Bhagvati. *A Small Subgroup Attack for Recovering Ephemeral Keys in Chang and Chang Password Key Exchange Protocol*. Journal of Software, Vol 6, No 4. 2011.

[16] Zuowen Tan. *An Authentication and Key Agreement Scheme with Key Confirmation and Privacy-preservation for Multi-server Environments*. Journal of Software, Vol 6, No 11. 2011.

[17] Xiangguo Cheng, Chen Yang and Jia Yu. *A New Approach to Group Signature Schemes*. Journal of Software, Vol 6, No 4. 2011.

Zhipeng Liu was born in 1980. He received the B.S. and M.S. degree in computer science and technology from Nanjing University of Posts and Telecommunications, in 2002, 2005, respectively. Now, he is a Ph.D. candidate in Nanjing University of Aeronautics and Astronautics. His research interests are data mining and distributed systems.

Yefan Hong was born in 1988. He is now a graduate student of Nanjing University of Posts and Telecommunications. His research interests are database systems and data mining.

Dechang Pi was born in 1971. He was a Ph.D. of Nanjing University of Aeronautics and Astronautics (NUAA) of China and now he is a professor and Ph.D. supervisor in NUAA. His research interests are data mining and database systems etc.