Analysis of Software Security level based on system clustering and correlation

Shizhuang YIN, Quan SHI, Yadong WANG, Shuai WANG, Weiyi WU
Shijiazhuang Campus of Army Engineering University, Shijiazhuang, Hebei Province, 050003
E-mail: 18525742489@163.com, 18525742489@oec.mtn

Abstract. With the wide use of electronic devices, passwords are used increasingly frequently. How to effectively evaluate password security needs our further research. In this paper, the known leaked passwords obtained from the network are analyzed by cluster analysis, frequency analysis, and correlation analysis by using system clustering and correlation, the strength grade model of the password is established, and the relationship among the factors of the password is analyzed. It shows that most people's personal information plays an important role in building passwords.

1. Introduction
With the continuous development of science and technology, electronic products are widely used. Passwords[1] are used increasing frequently. Passwords are the most popular authentication methods, mainly because they are easy to implement, do not require special hardware or software, and are familiar to users and developers. Unfortunately, multiple password database leaks show that users tend to choose passwords that are easy to guess, mainly made up of ordinary strings (for example, password "123456", "I love you") and their variants. Not long ago, it was reported that the credit company Equifax announced that the personal information of 143 million people in its system had been maliciously leaked by hackers[2], but worryingly, if a hacker wants to access your online data by directly cracking your password, it may take less than hour to do so. Now the worst news is that scientists have used AI to create an application that combines existing tools to deduce more than 1/4 account passwords from a set of more than 43 million LinkedIn profiles. The researchers say the consequences would be unthinkable if hackers also mastered the technology.

The goal of the researchers is to test the security of passwords[3], but the technology is originally a "double-edged sword". The new research aims to create a set of generative countermeasure network PassGAN[4], by applying deep learning, in which the artificial neural network includes a "generator" and a "discriminator". The "generator", which is responsible for producing manual output (such as images) similar to actual examples (actual photos), "discriminators", attempts to detect real examples from fake examples. They help each other until the generator becomes a skilled counterfeiting machine. Giuseppe Ateniese, one of the researchers, compared generators and discriminators to police crime profilers and witnesses accordingly.

Prior to using the most powerful password guessing programs John Ripper and hashCat used many techniques[5], John Ripper is directly brute force cracking, randomly trying a combination of many characters until you get the correct answer, hashCat infers characters from previously set passwords. In
some websites, these two sets of programs have successfully cracked more than 90% of the passwords, but because cracking requires manual coding and building attack plans, it will actually take a whole year.

2. password strength model

2.1. the main contents of this article

the data are preprocessed according to Annex 1.

the processed data are analyzed by cluster analysis[6], which is systematic clustering and fast clustering, respectively.

A password strength model is established based on clustering results and correlation analysis[7]. Combining the conclusions and the results of frequency analysis[8], we return to the rationality of the results obtained by the codebook demonstration.

2.2. data preprocessing

For more than 3000 passwords collected from the network, we extract four main eigenvalues[9]:

- Password length in Excel, use \( c_2 = \text{len}(b2) \); to get the length of the password.
- For the first letter of the password, use \( d_2 = \text{left}(b2) \) in Excel; get the first letter of the password, and then use the code command to convert the English letters and symbols into ASCII codes.
- Whether the password is mixed. In Excel, use \( e_2 = \text{IF}(\text{OR}(\text{ISNUMBER}(\text{FIND}(["A";"B";"C";"D";"E";"F";"G";"H";"I";"J";"K";"L";"M";"N";"O";"P";"Q";"R";"S";"T";"U";"V";"W";"X";"Y";"Z";"a";"b";"c";"d";"e";"f";"g";"h";"i";"j";"k";"l";"m";"n";"o";"p";"q";"r";"s";"t";"u";"v";"w";"x";"y";"z"]),)) \) to determine whether there are uppercase letters, return 1 for existence, and 0 for non-existence. Then use the same method to determine whether lowercase letters exist. Finally, the two values are manipulated to get whether the password itself is mixed or not.
- Whether the password is mixed with English numbers. In Excel, use \( f_2 = \text{IF}(\text{OR}(\text{ISNUMBER}(\text{FIND}(["1";"2";"3";"4";"5";"6";"7";"8";"9";"0"], B4)),"1","0"); to judge whether there is a number, if there is a number, it returns 1, and if it does not exist, it is 0. Then use the same method to determine the existence of letters. Finally, the two values are operated to get whether the password itself is mixed with English numbers.
- Add a disturbance to the data, that is, Article 22 adds an empty password with all its eigenvalues represented by 0.

2.3. Cluster analysis

The results obtained by systematic clustering and fast clustering methods are as follows: when the data are imported into SPSS[10], the results are as follows:

As the figure is longer, see the experimental results in Annex 3 for details. After systematic clustering, we set up a fast clustering with class 4, and the results are roughly the same. There are small changes in some classifications, which do not affect the overall results.

According to the results, we can see that users' passwords can be roughly divided into four categories, while the fourth type of data is relatively small. After reviewing the original data, we find that the added disturbance data is divided into the fourth category, and its specific values are shown in Table 1:

| Password | Length | Initials | case | Digital English |
|----------|--------|----------|------|-----------------|
| me       | 2      | 109      | 0    | 0               |
| a        | 1      | 97       | 0    | 0               |
| 22       | 2      | 2        | 0    | 0               |
| e        | 1      | 101      | 0    | 0               |
| m        | 1      | 109      | 0    | 0               |
| y        | 1      | 121      | 0    | 0               |
This kind of password we think is a stupid password, that is, there is little difference between setting a password and not setting it.

According to the results of clustering, we can divide the password into four levels. Because there are many contents that need password protection and are familiar with computer knowledge, the complexity of the password used is high, so it is difficult to decipher it. According to the complexity of the password, we divide each segment of the password into four levels, as shown in Table 2.

| Password level | Length | Initials | case | Digital |
|----------------|--------|----------|------|---------|
| 1              | 1      | 0        | 0    | 0       |
| 2              | 1      | 1        | 0    | 0       |
| 3              | 1      | 1        | 1    | 0       |
| 4              | 1      | 1        | 1    | 1       |

According to the password composition ratio, a password complexity value is assigned to each age group (in the case of password1, the password complexity is 1: 0, 1: 0, 0: 2), and the hierarchical correlation coefficient is used as shown in Formula 1[11].

\[
γ_0 = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2
\] (1)

We can calculate the correlation between password holders and the complexity of their use of passwords.

3. Frequency analysis and correlation analysis

Firstly we analyze frequency of password firstly and its concrete result is shown in Figure 1. Through analysis we find that the maximum number of passwords is 6 which coincides with our usual passwords such as bank cards mobile phones etc.

According to the distribution data of cipher length we draw strip graph 1. Figure 1 shows that length distribution of cipher length conforms to normal distribution and \( P (\mu > 6) > P (\mu < 6) \) indicating that people’s safety consciousness increases and more people use longer length codes.

![Fig. 1 password length distribution](image-url)
We analyze frequency analysis of converted ASCII codes with detailed results. Its distribution is shown in figure 2. It shows that people tend to use lowercase letters and numbers when setting passwords, which shows less use for capital letters and special characters, which accords with users' password setting for convenience memory when setting passwords.

Fig 2 Distribution map of password acronym

Tables 3 reflect whether people use numbers and English letters simultaneously when setting passwords. Tables 4 reflect whether people use uppercase letters and lowercase English letters simultaneously when setting passwords.

Table 3 whether the numbers are mixed in English

| Frequency | percentage | effective percentage | cumulative percentage |
|-----------|------------|----------------------|-----------------------|
|           |            |                      |                       |
| Effective | 1          | 296                  | 8.3                   | 8.3 100.0 |
| Total     | 3546       | 3250                 | 91.7                  | 91.7 91.7 |

Table 4 whether to mix size

| Frequency | percentage | effective percentage | cumulative percentage |
|-----------|------------|----------------------|-----------------------|
|           |            |                      |                       |
| Effective | 1          | 145                  | 4.1                   | 4.1 100.0 |
| Total     | 3546       | 3401                 | 95.9                  | 95.9 95.9 |

By observing we find that people who use passwords more than 90% don't write uppercase letters and English or Chinese words simply use simple numbers and lowercase letters indicating that people need to improve complex passwords.

The above four bar graphs show the frequency of the four eigenvalues in the sample, and the corresponding value of the first letter is its ASCII code. Through these pictures, we can see that certain letters and passwords of certain length appear more frequently, while people who use mixed passwords are very few. It can be seen that people use less passwords that are difficult to remember, such as mixed passwords. To analyze the correlation between the various factors of the password, we have done the correlation analysis. We analyze four features by using Pearson correlation and saliency correlation.
As shown in the table 5, the password length and the other three quantities are significantly correlated at 0.01 level (both sides), while the first letter is related to the case mix, and the number length mixing is only related to the password length, from which we can analyze that as the password length increases, people surely have a stronger sense of secrecy, and it is easier to take other measures to strengthen the password strength. Capital letters generally appear at the beginning of a password or in words with special meaning, and numbers are generally personal information such as birthdays, phone numbers, special dates, and so on. Specific examples are shown in Table 6:

As shown in the table 5, the password length and the other three quantities are significantly correlated at 0.01 level (both sides), while the first letter is related to the case mix, and the number length mixing is only related to the password length, from which we can analyze that as the password length increases, people surely have a stronger sense of secrecy, and it is easier to take other measures to strengthen the password strength. Capital letters generally appear at the beginning of a password or in words with special meaning, and numbers are generally personal information such as birthdays, phone numbers, special dates, and so on. Specific examples are shown in Table 6:

### Table 5 correlation analysis

| Password length | Initials of passwords | Whether size mixing or not | Whether digital English mixing |
|-----------------|-----------------------|-----------------------------|-------------------------------|
| Pearson correlation. | 1 | .118** | .044** | .114** |
| Saliency (bilateral) |  |  |  |  |
| N | 3546 | 3546 | 3546 | 3546 |
| Pearson correlation. | .118** | 1 | -.220** | -.025 |
| Saliency (bilateral) |  |  |  |  |
| N | 3546 | 3546 | 3546 | 3546 |
| Pearson correlation. | .044** | -.220** | 1 | -.037* |
| Saliency (bilateral) |  |  |  |  |
| N | 3546 | 3546 | 3546 | 3546 |
| Pearson correlation. | .008 | .000 | .137 | .029 |
| Saliency (bilateral) |  |  |  |  |
| N | 3546 | 3546 | 3546 | 3546 |
| Pearson correlation. | .114** | -.025 | -.037* | 1 |
| Saliency (bilateral) |  |  |  |  |
| N | 3546 | 3546 | 3546 | 3546 |

**. Significant correlations were observed at 0.01 levels (bilaterally).

*. Significant correlations were observed at 0.05 levels (bilaterally).

### Table 6 correlation analysis

| Password. | Password length | The initials of a password. | Whether the size is mixed | whether the number is mixed in English. |
|-----------|----------------|-----------------------------|--------------------------|----------------------------------------|
| ABC123    | 6              | 65                          | 1                        | 1                                      |
| Bond007   | 7              | 66                          | 1                        | 1                                      |
| NCC1701   | 7              | 78                          | 1                        | 1                                      |
| OU812     | 5              | 79                          | 1                        | 1                                      |
| Front242  | 8              | 70                          | 1                        | 1                                      |
| password1 | 9              | 112                         | 0                        | 1                                      |
| abc123    | 6              | 97                          | 0                        | 1                                      |
| a1b2c3    | 6              | 97                          | 0                        | 1                                      |
| a12345    | 6              | 97                          | 0                        | 1                                      |
| a1b2c3d4  | 8              | 97                          | 0                        | 1                                      |
| bond007   | 7              | 98                          | 0                        | 1                                      |
| david1    | 6              | 100                         | 0                        | 1                                      |
| happy1    | 6              | 104                         | 0                        | 1                                      |

4. Conclusions

Through a large-scale statistical survey, the relationship between passwords themselves is analyzed in detail to provide support for the generation of password dictionaries. The statistical survey found that the number of passwords and the composition of passwords used by password holders are closely related to their age and education, so when generating password dictionaries, passwords should be composed
mainly of numbers and letters, supplemented by some common other characters; the number of passwords is mainly less than 12 digits[12]. How to effectively make a password dictionary is one of the hot issues in the field of deciphering, and it is an important work related to whether the password can be successfully deciphered. In this paper, we analyze the relationship between the password itself in detail through 3546 known leaked passwords. According to the statistical survey, it is found that the number of passwords and the composition of passwords used by password holders are closely related to their age and education level. Based on the results of the correlation calculation of the data, it is suggested that the corresponding classification option of the password should be added when writing the software, to improve the efficiency of deciphering the password. Therefore, in the preparatory work of the password dictionary, we should focus on collecting all kinds of information (ID card number, birthday, mobile phone, family, telephone, and other digital information) of the password holder and his related persons (husband, wife, parents, children, male and female friends, etc.).

In addition, we find that the first letter of the password is related to the size of the mix, and the length of the number is only related to the length of the password, from which we can analyze that with the increase of the length of the password, people have a stronger sense of secrecy. It is easier to take other measures to strengthen the strength of the password. Capital letters generally appear at the beginning of a password or in words with special meaning, and numbers are generally personal information such as birthdays, phone numbers, special dates, and so on. In the length of the password, the frequency of the 6-digit password is larger and roughly shows a normal distribution, while the frequency of "b", "c" and "n" in the first letter is higher.

Through argumentation this paper shows that some websites require password must reach certain length simultaneously and contain numbers uppercase letters lowercase letters special characters etc. which are necessary to improve user account security. But this also puts forward higher memory requirements for users how to make passwords safe while convenient memory will be our next step research direction. Meanwhile, this paper collected passwords randomly, compared with overall sample size small, there are some analysis deviations.

Reference
[1] Qin Zhiguang. Research on current situation and Development of cryptographic algorithms [J]. Computer applications, 2004, 24 (002): 1-4.
[2] Zhang Hong, Zhang Shi Ruimin, Ding Mn. Statistical Analysis of relationship between password Holder Information and password used, 2013 (1): 86-86.
[3] Chen Ming. Analysis and Research on Security of White Box Code [D]. 2015.
[4] Xu Yifeng. Overview of theoretical models and applications of generating counter Networks [J]. Journal of Jinhua vocational Technology College, 2017, 017 (003): 23-29.
[5] Weng Jie, Wu Qiang, Yang Canqun. OpenCL-based algorithm for solving MD5 problem [J]. Computer Engineering, 2011.
[6] Fang Kaitai. Cluster analysis (i) [J]. Practice and Cognition of Mathematics, 1978 (1): 23-29.
[7] Li Youmei, Liang Xun. Overview of Canonical correlation Analysis [J]. Journal of China Metrology University, 2017, 28 (001): 23-29.
[8] elected. Statistical Analysis of frequency of Common function invocation [J]. Computer Engineering and Science, 1995, 000 (001): 27-31.
[9] Fang Yuying, Xu Hong. Automatic search for integral feature of lightweight packet cipher GRANULE [J]. Journal of Information Engineering University, 2019, 020 (003): 23-29.
[10] Ma Liping. Standardization of Statistical data -- dimensionless method of Statistical data -- Research and application of Modern Statistical Analysis method of — City.
[11] Wang Xiaoyan, Wang Xiaoyan. Correlation coefficient between grade correlation coefficient and Spelman rank [J]. Journal OF Guangdong LIGHT INDUSTRY INSTITUTE OF LIGHT INDUSTRY, 2006, 5 (004): 26-27.
[12] Lu Zhixu, Qiu Weidong,. Dictionary generation method based on data Mining Technology [J]. Information Security and Communication secrecy 2011 (11): 63-65.