Poster: Analysis and Verification of Practical Password Strength Meters

Taku Sugai
Toho University
5513054s@nc.toho-u.ac.jp

Akira Kanaoka
Toho University
akira.kanaoka@is.sci.toho-u.ac.jp

I. Intruduction

Pursuit journey of moving to a stronger authentication method than a password is hampered by the strong deployability that the password has [1]. Although various authentication methods have been proposed one after another, passwords are still widely used due to difficulty in deploying and implementation.

There are several approaches to make password authentication itself stronger by adopting a mechanism to encourage users to set strong passwords. One is a password composition policy. A service side does not accept a password unless it exceeds a certain limit. Typical policies include ”Passwords are 8 characters or more” and ”At least one letter of the alphabet is used”. The other is a password strength meter. It calculates and displays the score how strong the password entered by the user has. And It has the effect of allowing the user to set a stronger password.

Several studies about password composition policy and password strength meter has been conducted from several directions. Such research results have not led to a successful development yet.

Dell’Amico, et. al focused on the password strength meter and pointed out that the entropy of the password itself is not reflected in the score by the meter. Then they proposed a method that can make the approximation fast did[2]. Although the indications by Dell’Amico et. al were well accepted, on the other hand, it has not been investigated how the strength score is calculated and how much the score deviates from the entropy.

In this poster, we extract sites that adopt password strength meter from Alexa Top 100 ’s website and analyze calculation method of each score. How the meter of each service behaves and how to calculate the score by the method is clarified, and classification and comparison are done. Also, using the data set of the password that was actually used, investigate the distribution of scores by each meter and clarify the difference.

II. Survey on Practical Password Strength Meters

In this research, the actual condition of the strength meter used at the Alexa Top 100 sites is investigated.

Although the password strength meter is used not only for Web sites but also on smartphones and PC client applications, we conducted a survey limited to Web sites in this poster.

A. Survey on Password Input Request in User Registration Phase

Many of the Top 100 sites have a mechanism for user registration, where a password has been entered. When entering the password, there were some differences depending on the site, such as those with password composition policy and those with password strength meter. Also, since Alexa’s sites are ranked by domain, there are cases where there are several top-level sites within the Top 100 site that have the same service but different service domains for each country. The most common one is Google’s site, 18 of Top 100 ranked as Google’s service domain of each country. Google centralizes user registration of related services along with accounts.google.com as well as user registration in each service domain. YouTube.com is also included. Similar service deployment was seen by Amazon and Microsoft.

As a result of the investigation, when removing the duplication of the aggregated amount in each service and enumerating the adoption situation of the password strength meter, the use of a total of 13 password strength meters was confirmed. In addition, some of the Alexa Top 100 sites contain services of China, and they had a screen for user registration, but most of them first let us enter the number of the mobile phone, and in this survey we could not investigate until the previous registration. There is a possibility that the initial code is input through the SMS, then the password setting is made there, and the composition policy and the strength meter are there.

B. Timing for composition policy confirmation and score calculation

The timing when the composition policy is confirmed and the password strength is calculated is roughly classified into those to confirm and calculate them locally (on the browser) using JavaScript and those to send the input password to the server to do it remotely. There are Twitter, Yandex, eBay, Reddit, Tumblr, and Apple that can remotely implement configuration policy confirmation. In addition, there are Google, eBay, tumblr, NAVER, which perform score calculation remotely.

C. Operations used for score calculation

The score calculation method was analyzed for nine services that calculate scores locally. How the information used will affect the score depends on the service. The table II shows the information that we are using, out of the nine services for the eight services excluding Dropbox. While each of the eight

Information used	Service
Password Length	Twitter, Yahoo!Japan, VK, Yandex, Reddit, Mail.ru, Apple, Rakuten,
Use of the same character	Twitter,
Continuous use of characters	Yahoo!Japan, Mail.ru, Rakuten, Dropbox
Continuous use of phrases	Yahoo!Japan, Mail.ru
Number of types of characters used	Twitter, VK, Mail.ru, Rakuten,
Presence of symbols	Twitter, Yahoo!Japan, Apple
Presence of numbers	Twitter, Yahoo!Japan, Reddit, Mail.ru, Rakuten,
Whether capitalization is used	Yahoo!Japan, Reddit, Rakuten,
Match with registered words	VK, Rakuten,

TABLE I. INFORMATION USED FOR SCORE CALCULATION

Service	Score range	Avg.	Var.	N-Avg.	N-Var.
Mail.ru	0-3	1.00	0.43	33.07	469.25
Apple	0-4	1.01	0.02	30.40	14.70
Rakuten	0-100	33.84	382.03	33.84	382.03
Reddit	0-100	20.02	166.57	20.02	166.57
Twitter	0-100	27.54	228.72	27.54	228.72
VK	0-4	2.07	0.70	51.84	435.20
Yahoo!Japan	0-4	1.81	0.18	45.29	113.09
Yandex	0-100	38.55	14.69	38.55	14.69
tumblr	0-5	0.52	0.64	10.44	254.27
NAVER	0-4	1.34	0.97	33.49	603.70
Google	0-4	1.91	1.63	47.81	1015.79

TABLE II. AVERAGE SCORE AND VARIANCE OF EACH METER

services uses its own calculation method, Dropbox calculates using the library zxcvbn [3].

Zxcvbn performs a large different score calculation from these eight services .

III. FEATURE ANALYSIS OF EACH STRENGTH METER

As a result of the analysis in Section 3, it turned out that the meter of each service greatly differs for each service such as score calculation method. In this section, therefore, we analyze the score of each meter using RockYou password data set.

For each meter that calculates scores locally, since each calculation method is known, a program that simulates computation was created on a PC for analysis and the score of each password was calculated. For the meter that calculates the score remotely, since the calculation method is not known, a crawler that accesses the URL to be calculated is created and the score with each password is calculated. For eBay, the maximum number of accesses per day from the same IP address is set, so analysis using a crawler could not be performed.

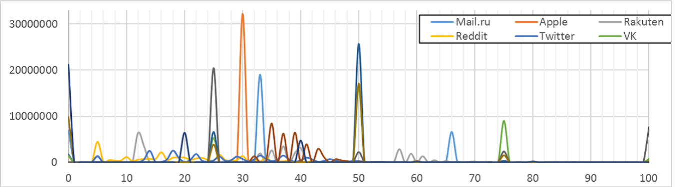

Fig. 1. Score distribution by each meter (After Nomalization)

Since the score range varies depending on the meter of each service, direct comparison is difficult. Therefore, comparison is made by normalizing the range of the score. Normalization uses the appearance of each meter to make the score range of all meters 0-100. For example, when the meter appearance is displayed at equal intervals in each score in the score width of 0 - 4, each score is multiplied by 25.

The average score and variance of each meter with its normalized value are shown in Table II. In the table ”N-Avg.” means ”Average by normalized value”. It can be seen that there is a big difference in the average score for each service. The score distribution in the normalized state is shown in Fig 1. The characteristics of each meter are shown well and it can be seen that Reddit spreads evenly on each score except for its distribution to other scores other than having a high peak, while the peak height is low. In addition, it is understood that each peak position also differs depending on the meter, and there is no unified or similar score distribution in each service.

IV. CONCLUSION

In this poster, we analyzed the behavior of password strength meter used at Alexa Top 100 site, and investigated and analyzed the strength score distribution using actual password data set. As a result of the survey, it turned out that the strength meter was used in many sites. On the other hand, the score calculation method was not a unified or similar method, but each was calculating by a unique method. Even when using the same password data set, it was found that the score average, variance, distribution are greatly different. This can be said that Dell’Amico and others point out that the score calculation method adopted by many sites does not reflect the original entropy of the password.

In the future, further investigation will be done as to how much the meter deviates from the actual entropy and it is expected that correction will be required. As a secondary result, several services were also observed to transmit password data remotely. It is believed that the possibility of future threats being discovered by this will also be discussed in the future.

REFERENCES

[1] J. Bonneau, C. Herley, P. C. v. Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In Proceedings of the 2012 IEEE Symposium on Security and Privacy, SP ’12, pages 553-567, Washington, DC, USA, 2012. IEEE Computer Society.

[2] M. Dell’Amico and M. Filippone. Monte carlo strength evaluation: Fast and reliable password checking. In Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security, CCS ’15, pages 158-169, New York, NY, USA, 2015. ACM.

[3] Dropbox, ”dropbox/zxcvbn: A realistic password strength estimaor”, github, https://github.com/dropbox/zxcvbn