Effectiveness and Soundness of Commercial Password Strength Meters

Shukun Yang[1], Shouling Ji[1], Xin Hu[2], and Raheem Beyah[1]

1. Georgia Institute of Technology — *{syang87, sji}@gatech.edu*
rbeyah@ece.gatech.edu

2. IBM T. J. Watson Research Center — *huxin@us.ibm.com*

Introduction

➢Hundreds of millions of passwords including those from Google, Yahoo! and LinkedIn were leaked over the past decade. No need to worry because your password is not among those leaked? Wrong! The leaked passwords can be used to significantly facilitate attacks on your passwords!

➢Are we safe with the existing commercial password strength meters? When you register your account on a website, should you relax when it shows you a "Strong" rating for your password? In this work as follows, we show how much you can rely on these password strength meters.

Methodology

Previous work has implied inconsistencies of the feedback from various password meters [1]. We further employ the leaked datasets and the existing state-of-the-art cracking algorithms to test the effectiveness and accuracy of different meters. We leverage the ground truth that passwords which can be cracked in a reasonable amount of time are labeled "weak". Then we test the commercial meters on both the original password datasets and the cracked password datasets to evaluate their feedback. From the results we can see the accuracy and effectiveness of the tested meters.

Password Datasets

- We consider 15 large-scale real world password datasets which contain ~200M passwords. They are summarized in Table 1.

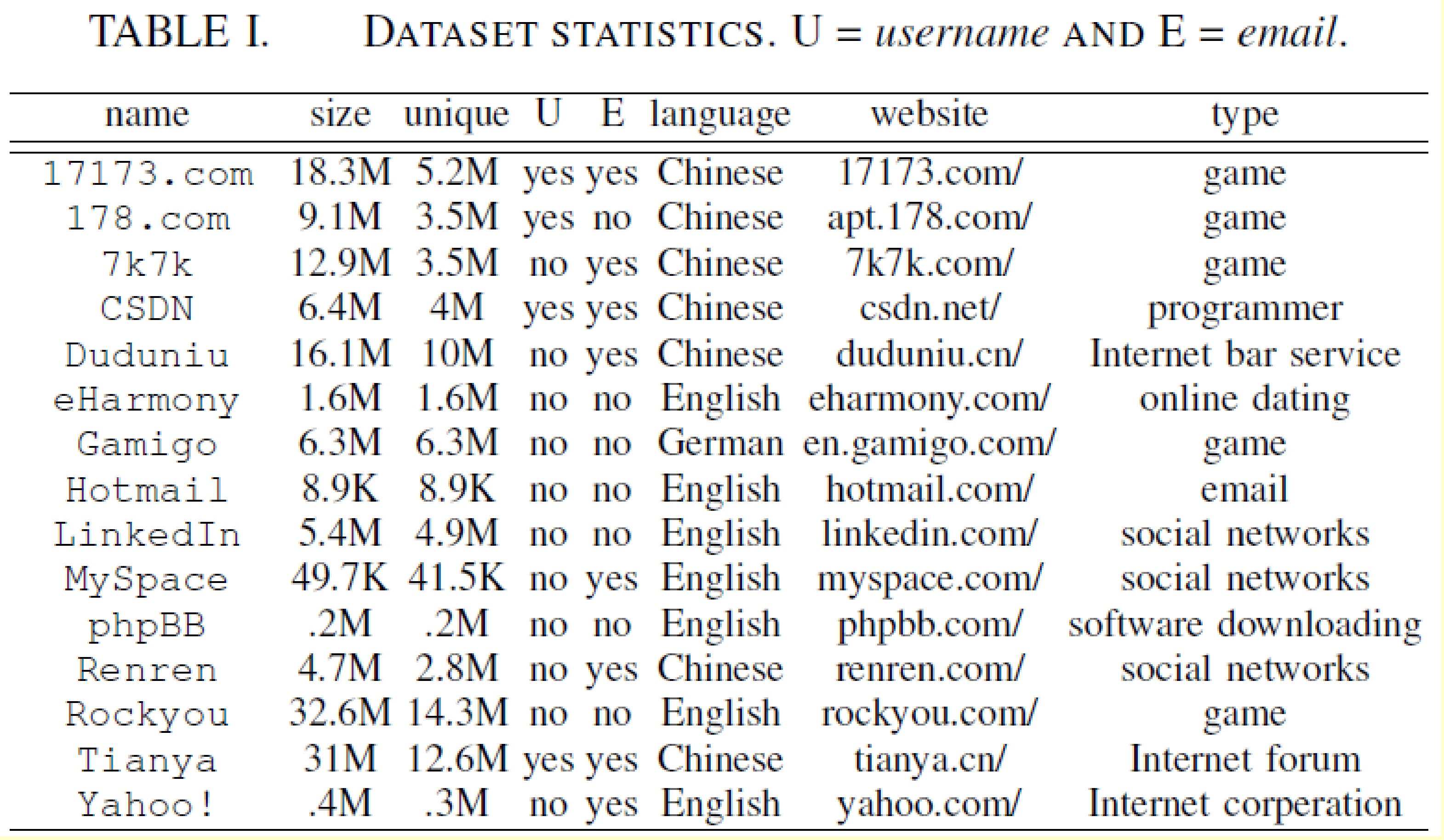


TABLE I. DATASET STATISTICS. U = *username* AND E = *email*.

name	size	unique	U	E	language	website	type
17173.com	18.3M	5.2M	yes	yes	Chinese	17173.com/	game
178.com	9.1M	3.5M	yes	no	Chinese	apt.178.com/	game
7k7k	12.9M	3.5M	no	yes	Chinese	7k7k.com/	game
CSDN	6.4M	4M	yes	yes	Chinese	csdn.net/	programmer
Duduniu	16.1M	10M	no	yes	Chinese	duduniu.cn/	Internet bar service
eHarmony	1.6M	1.6M	no	no	English	eharmony.com/	online dating
Gamigo	6.3M	6.3M	no	no	German	en.gamigo.com/	game
Hotmail	8.9K	8.9K	no	no	English	hotmail.com/	email
LinkedIn	5.4M	4.9M	no	no	English	linkedin.com/	social networks
MySpace	49.7K	41.5K	no	yes	English	myspace.com/	social networks
phpBB	.2M	.2M	no	no	English	phpbb.com/	software downloading
Renren	4.7M	2.8M	no	yes	Chinese	renren.com/	social networks
Rockyou	32.6M	14.3M	no	no	English	rockyou.com/	game
Tianya	31M	12.6M	yes	yes	Chinese	tianya.cn/	Internet forum
Yahoo!	.4M	.3M	no	yes	English	yahoo.com/	Internet corperation

- **Ethical Consideration**: *The 15 datasets are publicly available. We use the datasets only for research purposes.*

Password Cracking Algorithms

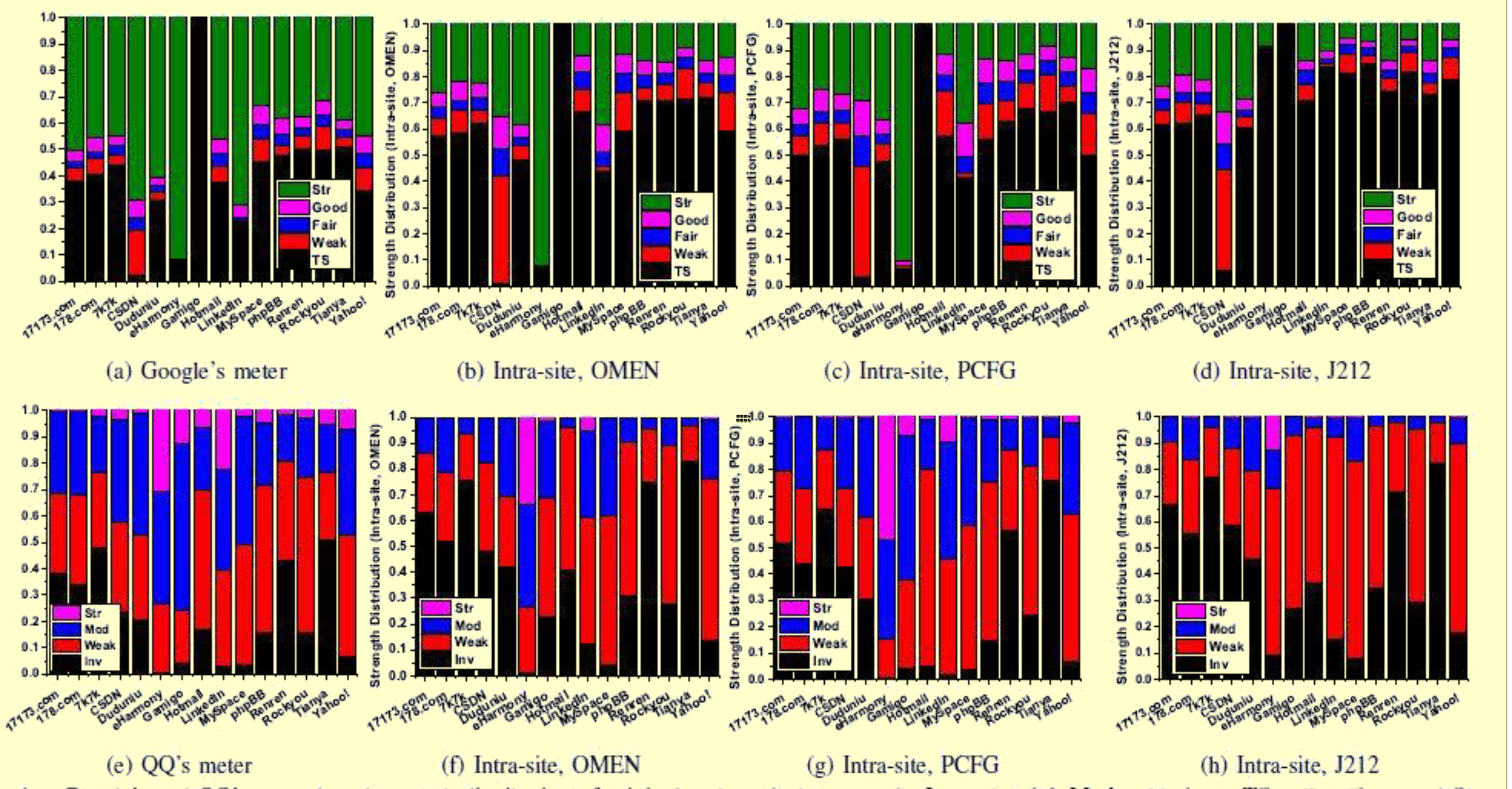
- **Password cracking tool.** We employ the latest versions of *John the Ripper* (JtR) [2]. We evaluate its *Incremental mode* (JtR-Inc) which is an intelligent brute-force algorithm, *Single mode (JtR-S)* which leverages user's social profile information, and the very powerful *Markov mode (JtR-M)* which is based on Markov Model.
- **Probabilistic Context Free Grammar (PCFG) based scheme.** The idea of PCFG-based schemes is to generate password guesses in terms of password structures and password frequency/probability (from high to low) [3].
- **Semantics based PCFG scheme.** We employ the *Semantic Guesser* which is an enhanced version of PCFG that takes into account the semantic significance in the passwords [4].
- **Markov model based scheme.** Markov model based schemes generate password guesses based on a trained Markov model. Except for the JtR-M above, we employ the *Ordered Markov ENumerator* (OMEN) [5].

Evaluation and Results

- **Meters:** In our evaluation, we pick two meters that are representative, Google's meter (English), and QQ's meter (Chinese). Google is well-known for its services including Gmail, search engine and etc. QQ is popular in China for its instant messaging service, mailbox and so on.

- **Intra-site:** Within a password dataset, we randomly select 30% of the passwords for training, and the rest are used for testing.

- **Setup:** We first test the15 original datasets and observe the strength feedback from meters of Google and QQ (shown in Fig.1 (a) and (e)). Then we try to crack all datasets using 3 state-of-the-art password cracking algorithms: ① ***OMEN*** (with 10^9 guesses), ② ***PCFG*** (with 10^9 guesses), ③ ***JtR-M212*** (*level* = 212, with ~10^9 guesses). Finally, we test the two meters on only the cracked passwords in the 15 datasets and observe the results in Fig. 1.



(a) Google's meter (b) Intra-site, OMEN (c) Intra-site, PCFG (d) Intra-site, J212

(e) QQ's meter (f) Intra-site, OMEN (g) Intra-site, PCFG (h) Intra-site, J212

Fig. 1. Google's and QQ's meter-based strength distribution of original and cracked passwords. **Inv** = *Invalid*, **Mod** = *Moderate*,**TS** = *Too Short*, and **Str** = *Strong*.

Conclusion: QQ's meter is more conservative and stringent compared to Google's meter. Similar observations of inconsistent and inaccurate feedback apply to other meters as well. The commercial meters need to be improved! Users need better guidance to aid in choosing strong passwords.

References

[1] X. C. Carnavalet and M. Mannan, *From Very Weak to Very Strong: Analyzing Password-Strength Meters*, NDSS 2014.

[2] John the Ripper-bleeding-jumbo, *https://github.com/magnumripper/JohnTheRipper*.

[3] M. Weir, S. Aggarwal, B. Medeiros, and B. Glodek, *Password Cracking Using Probabilistic Context-Free Grammars*, S&P 2009.

[4] R. Veras, C. Collins, and J. Thorpe, *On the Semantic Patterns of Passwords and their Security Impact*, NDSS 2014.

[5] M. Dürmuth, A. Chaabane, D. Perito, and C. Castelluccia, *When Privacy meets Security: Leveraging Personal Information for Password Cracking*, CoRR, 2013.