Effectiveness and Soundness of Commercial Password Strength Meters
Shukun Yang\(^1\), Shouling Ji\(^3\), Xin Hu\(^2\), and Raheem Beyah\(^1\)

1. Georgia Institute of Technology \{syang87, sj\}@gatech.edu
2. IBM T. J. Watson Research Center rbeyah@ece.gatech.edu
3. GaTech 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!

- 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.

Evaluation and Results

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Unique</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>17173.com</td>
<td>18.3M</td>
<td>5.2M</td>
<td>Chinese</td>
</tr>
<tr>
<td>178.com</td>
<td>9.1M</td>
<td>3.5M</td>
<td>Chinese</td>
</tr>
<tr>
<td>7kk</td>
<td>12.9M</td>
<td>3.5M</td>
<td>Chinese</td>
</tr>
<tr>
<td>CSdn</td>
<td>6.4M</td>
<td>4M</td>
<td>English</td>
</tr>
<tr>
<td>Duoduni</td>
<td>161M</td>
<td>10M</td>
<td>Chinese</td>
</tr>
<tr>
<td>eHarmony</td>
<td>1.6M</td>
<td>1.6M</td>
<td>English</td>
</tr>
<tr>
<td>Gamigo</td>
<td>6.3M</td>
<td>6.3M</td>
<td>English</td>
</tr>
<tr>
<td>Hotmail</td>
<td>8.9K</td>
<td>8.9K</td>
<td>English</td>
</tr>
<tr>
<td>Linkedln</td>
<td>5.4M</td>
<td>4.9M</td>
<td>English</td>
</tr>
<tr>
<td>MySpace</td>
<td>49.7K</td>
<td>41.5K</td>
<td>English</td>
</tr>
<tr>
<td>P@ssw0rd</td>
<td>2M</td>
<td>2M</td>
<td>English</td>
</tr>
<tr>
<td>Renren</td>
<td>4.7M</td>
<td>2.8M</td>
<td>Chinese</td>
</tr>
<tr>
<td>Rockyou</td>
<td>32.6M</td>
<td>14.3M</td>
<td>English</td>
</tr>
<tr>
<td>Tianya</td>
<td>31M</td>
<td>12.6M</td>
<td>Chinese</td>
</tr>
</tbody>
</table>

- **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 E Numerator (OMEN) [5].

- **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 the 15 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.

**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