From Very Weak to Very Strong: Analyzing Password-Strength Meters

Xavier de Carné de Carnavalet  Mohammad Mannan

Concordia University, Montreal, Canada
Password-strength meter/checker

Password: •••••••

Good
What is this work about?

We analyzed why is this:

- Very weak
- Weak
- So-so
- Good
- Great!
What is this work about?

And why is that (same password):

- Dropbox: Password: ●●●●●●●●
  Strength: Very weak

- Google: Password: ●●●●●●●●
  Password strength: Weak
  Use at least 8 characters. Don’t use a password from another site, or something too obvious like your pet’s name. Why?

- Microsoft: Password: ●●●●●●●●
  Strength: Medium

- Twitter: Password: ●●●●●●●●
  Password is okay.

- FedEx: Password: ●●●●●●●●
  You entered a strong password. Longer passwords are even more secure and should include a mixture of mixed-case letters, numbers and special characters.
Our motivations

1. Recent studies: meters really guide users to choose better passwords [Ur et al., USENIX Security’12] and [Egelman et al., CHI’13]

2. Deployed meters impact hundreds of millions of users

3. Built by up-to-billion-dollar IT companies

4. They don’t seem reliable...
Tested 11 web services/applications
11 dictionaries: 3,895,247 unique passwords
Analysis setup (1/3)

1. 11 dictionaries: 3,895,247 unique passwords
2. Top500, cracking tools (e.g., JtR) worm dictionaries, database leaks (e.g., RockYou)
11 dictionaries: 3,895,247 unique passwords

Top500, cracking tools (e.g., JtR) worm dictionaries, database leaks (e.g., RockYou)

Mangling & leet transformations

password → Password1+ or p@5$w0rd
Analysis setup (2/3)

1. Understanding of functionalities (involve some RE)
2. JavaScript (whitebox) and/or server-side (blackbox)
3. 52+ million tests
Analysis setup (3/3)

1. Analyze results
2. Understand checkers profile
3. Find common weaknesses
In theory

Designing PSMs is non-trivial:
- No straightforward academic literature to follow
- Failure of NIST recommendations
- How to deal with password leaks, cultural references?
In practice

- Custom “entropy” based on:
  - Perceived complexity
  - Password length
  - Number of charsets used
  - Known patterns
  - Comparison with dictionary of common passwords (blacklist)

- More entropy $\simeq$ more secure password
- Everyone invents their own algorithm
Meters heterogeneity

1. Each meter reacts differently to our dictionaries
2. Strength results vary widely from one to another

Example: *Password1*
- Obvious, Very weak, Weak (x3), Poor, Moderate (blacklisted), Medium (x2), Strong (x3), Very strong
- By Microsoft itself (3 versions): strong, weak and medium!

3. Some simple dictionaries score significantly higher than others
Simple mangling rules/leet transformations allow bypassing password requirements

Example: Consider \{Top500, C&A, Cfkr and JtR\}

How many passwords are medium or better?

<table>
<thead>
<tr>
<th>Web service</th>
<th>Regular</th>
<th>Mangled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skype</td>
<td>10.5%</td>
<td>78%</td>
</tr>
<tr>
<td>Google</td>
<td>0.002%</td>
<td>26.8%</td>
</tr>
</tbody>
</table>
Password policies

1. Password policies not often explicitly stated
2. Rules for measuring strength unexplained to users
3. Differences in policies:
   - Very stringent: assign strengths only for 3+ charsets (FedEx)
   - Promotion of single-charset passphrases (Dropbox)
4. Google and Yahoo!, lots of personal info, but lenient policy...
Google checker: some results

Password strength distribution:

Inconsistencies:

1. testtest is weak
2. testtest0 is strong
3. testtest1 is fair
4. testtest2 is good
5. testtest3 is strong...
6. Strength is time-dependent
One checker to rule them all

Password Multi-Checker

<table>
<thead>
<tr>
<th>Services</th>
<th>Strength scores</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>Moderate (Blacklisted)</td>
<td>2/3</td>
</tr>
<tr>
<td>Dropbox</td>
<td>Very Weak</td>
<td>1/5</td>
</tr>
<tr>
<td>Drupal</td>
<td>Strong</td>
<td>4/4</td>
</tr>
<tr>
<td>eBay</td>
<td>Medium</td>
<td>4/5</td>
</tr>
<tr>
<td>FedEx</td>
<td>Strong</td>
<td>4/5</td>
</tr>
<tr>
<td>Google</td>
<td>Weak</td>
<td>2/5</td>
</tr>
<tr>
<td>Microsoft (v1)</td>
<td>Strong</td>
<td>3/4</td>
</tr>
<tr>
<td>Microsoft (v2)</td>
<td>Weak</td>
<td>1/4</td>
</tr>
<tr>
<td>Microsoft (v3)</td>
<td>Medium</td>
<td>2/4</td>
</tr>
<tr>
<td>PayPal</td>
<td>Weak</td>
<td>2/4</td>
</tr>
<tr>
<td>Skype</td>
<td>Poor</td>
<td>1/3</td>
</tr>
<tr>
<td>Twitter</td>
<td>Obvious</td>
<td>2/6</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>Very Strong</td>
<td>4/4</td>
</tr>
</tbody>
</table>
Summary (1/2)

Facts:
- Passwords are not going to disappear anytime soon
- Users will continue to choose weak passwords

Current solutions:
- Stringent policies (user resentment?)
- Influence users in choosing better passwords, \textit{willingly}
  - Provide feedback on the quality of chosen passwords
  - Should be consistent and avoid confusion
Reality:

1. Commonly-used meters are highly inconsistent
2. Fail to provide coherent feedback, sometimes blatantly misleading
3. Often have very ad-hoc design
4. Simple transformations not taken into account
What can be done?

1. Common API to reduce confusion (e.g., Dropbox with zxcvbn)
2. Real-time cracking with state-of-the art techniques to assess passwords?
3. Passphrases (be careful at simple structures)
4. Password popularity, Markov models, PCFG, semantic?
Thanks

To recap:

1. Meters less robust than expected from such large companies
2. Companies should stop misleading users
3. Opportunities for academic research

Contact: x_decarn@ciise.concordia.ca
Project URL: http://goo.gl/0E5Ieu
Additional slides
Percentage of dic. assigned “good” or +

Base dictionaries:

```
<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Drupal</th>
<th>Yahoo!</th>
<th>Dropbox</th>
<th>Microsoft</th>
<th>PayPal</th>
<th>FedEx</th>
<th>Twitter</th>
<th>Skype</th>
<th>eBay</th>
<th>Apple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cfkr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JtR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C&amp;A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RY5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phpBB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

“Advanced” dictionaries:

```
<table>
<thead>
<tr>
<th></th>
<th>Drupal</th>
<th>Yahoo!</th>
<th>Google</th>
<th>PayPal</th>
<th>FedEx</th>
<th>eBay</th>
<th>Twitter</th>
<th>Dropbox</th>
<th>Skype</th>
<th>Apple</th>
<th>Microsoft</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top500+M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cfkr+M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JtR+M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C&amp;A+M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RY5+M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
FedEx: Password strength distribution

X. de Carné de Carnavalet
NDSS’14: Analyzing Password-Strength Meters
FedEx: Password strength distribution

Very weak? Fine…
FedEx: Targeted dictionary

Refined mangling rules:

1. capitalize, append a digit and a symbol
2. capitalize, append a symbol and a digit
3. capitalize, append a symbol and two digits
4. capitalize, append a symbol and a digit, and prefix with a digit

Gives 121,792 words from \{Top500, JtR, Cfkr\}

1. 60.9\% is now very strong
2. 9.0\% is strong
3. 29.7\% is medium
4. 0.4\% is very weak