Birthday, Name and Bifacial-security: Understanding Passwords of Chinese Web Users

Ding Wang and Ping Wang, Peking University; Debiao He, Wuhan University; Yuan Tian, University of Virginia

August 16, 2019 • Santa Clara, CA, USA
Password is everywhere
Passwords are influenced by native languages

[Bonneau 2012; Li 2014]

Research Questions

• What are the **structural** or **semantic** characteristics of Chinese passwords comparing to English passwords?
• Given the differences in the passwords, how to measure the **strength** of Chinese passwords?
Outline

1. Our Datasets
2. Characteristics of Chinese pwds
3. Strength of Chinese pwds
4. Implications and Insights
Password datasets

- Datasets from nine high-profile websites (from data breach)
- Six from Chinese websites, three from English websites
- A total of 106 million

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Web Service</th>
<th>Language</th>
<th>Leaked Time</th>
<th># of PWDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tianya</td>
<td>Social forum</td>
<td>Chinese</td>
<td>Dec. 2011</td>
<td>31,761,424</td>
</tr>
<tr>
<td>7k7k</td>
<td>Gaming</td>
<td>Chinese</td>
<td>Dec. 2011</td>
<td>19,138,452</td>
</tr>
<tr>
<td>Dodonew</td>
<td>E-commerce</td>
<td>Chinese</td>
<td>Dec. 2011</td>
<td>16,283,140</td>
</tr>
<tr>
<td>178</td>
<td>Gaming</td>
<td>Chinese</td>
<td>Dec. 2011</td>
<td>9,072,966</td>
</tr>
<tr>
<td>CSDN</td>
<td>Programmer forum</td>
<td>Chinese</td>
<td>Dec. 2011</td>
<td>6,428,632</td>
</tr>
<tr>
<td>Duowan</td>
<td>Gaming</td>
<td>Chinese</td>
<td>Dec. 2011</td>
<td>5,024,764</td>
</tr>
<tr>
<td>Rockyou</td>
<td>Social forum</td>
<td>English</td>
<td>Dec. 2009</td>
<td>32,603,387</td>
</tr>
<tr>
<td>Yahoo</td>
<td>Portal (e.g., E-commerce)</td>
<td>English</td>
<td>July 2012</td>
<td>453,491</td>
</tr>
<tr>
<td>Phpbb</td>
<td>Programmer forum</td>
<td>English</td>
<td>Jan. 2009</td>
<td>255,421</td>
</tr>
<tr>
<td>Tianya</td>
<td>7k7k</td>
<td>Dodonew</td>
<td>178</td>
<td>CSDN</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>----------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>123456</td>
<td>123456</td>
<td>123456</td>
<td>123456</td>
<td>123456789</td>
</tr>
<tr>
<td>111111</td>
<td>0</td>
<td>A123456</td>
<td>111111</td>
<td>12345678</td>
</tr>
<tr>
<td>000000</td>
<td>111111</td>
<td>123456789</td>
<td>Zz12369</td>
<td>11111111</td>
</tr>
<tr>
<td>123456789</td>
<td>123456789</td>
<td>111111</td>
<td>Qiulaobai</td>
<td>123123</td>
</tr>
<tr>
<td>123123</td>
<td>123123</td>
<td>5201314</td>
<td>123456aa</td>
<td>00000000</td>
</tr>
<tr>
<td>123321</td>
<td>5201314</td>
<td>123123</td>
<td>wmsxie12</td>
<td>123123123</td>
</tr>
<tr>
<td>5201314</td>
<td>123</td>
<td>a321654</td>
<td>123123</td>
<td>1234567890</td>
</tr>
<tr>
<td>12345678</td>
<td>12345678</td>
<td>12345</td>
<td>000000</td>
<td>88888888</td>
</tr>
<tr>
<td>666666</td>
<td>1234567</td>
<td>000000</td>
<td>qq66666</td>
<td>11111111</td>
</tr>
<tr>
<td>111222tianya</td>
<td>wangyut2</td>
<td>123456a</td>
<td>w2w2w2</td>
<td>147258369</td>
</tr>
</tbody>
</table>

Top-10 Passwords
Comparisons

**Different:**
- Letter distribution
- Semantic patterns

**Similar:**
- Length distribution
- Frequency distribution
Comparisons

Different:
• Letter distribution
• Semantic patterns

Similar:
• Length distribution
• Frequency distribution
Top-3 structural patterns

Chinese Passwords

- D(e.g.,123456)
- LD(e.g.,a12345)
- DL(e.g.,12345a)
- Sum of top-3

English Passwords

- L(e.g.,abcdef)
- LD(e.g.,abc123)
- D(e.g.,123456)
- Sum of top-3
Semantic patterns in passwords

- We construct 22 dictionaries of different semantic categories
Length distribution of passwords

[Graph showing the percentage distribution of password lengths for different services, with peaks at 6 and 8 characters.]
Contradictory results on strength

• Bonneau 2012: Chinese passwords are the hardest to guess
• Li 2014: The strength of Chinese passwords and English passwords is similar
Experimental settings

- Two state-of-the-art password cracking algorithms: **PCFG** and **Markov**
  - Markov: *n*-gram [Narayanan-Shmatikov ACM CCS’05; Ma et al. IEEE S&P’14]

- Training set
  - Chinese: Duowan 4 million; English: Rockyou 4 million;

- Pairing passwords by service type
  - Users create password based on importance of the account
    - [Egelman et al. CHI’11, Stobert-Biddle ACM TOPS’18 ]
    - Tianya vs. Rockyou,
    - Dodonew vs. Yahoo,
    - CSDN vs. Phpbb
Limitation of PCFG

- **Basic idea** of PCFG algorithm: parse pwds into Letter-segment, Digital-segment, Symbol-segment.

  E.g., password123 → L₈D₃  \[\text{count}(L₈D₃) + 1\]
  \[L₈ \to \text{password}\]  \[\text{count}(L₈ \to \text{password}) + 1\]
  \[D₃ \to 123\]  \[\text{count}(D₃ \to 123) + 1\]

- We find that Chinese passwords include many interleaving patterns:
  e.g., “1n1d1g”

- PCFG algorithm **overlooks** the relationship between L and D segment, therefore **underestimates** the probability of them
Improvement on PCFG

• Interleaving conversion: transform interleaving passwords into normal passwords
• Input improvement: adding Pinyin names
• Input improvement: adding dates
Improved PCFG cracking results

- At $10^7$ guesses, our improved PCFG gains 1.70%~4.29% improvements
  - Interleaving conversion
  - Adding Pinyin names
  - Adding dates

![Graph showing improved PCFG cracking results](image-url)
PCFG cracking results

- When guess number is small, Chinese pwds are weaker than English pwds
  - Guess number =100, about 10% Chinese pwds cracked, English 3.5%;
  - Guess number =10^7, >32% Chinese pwds cracked, English >43%;
Markov cracking results

- When guess number is small, Chinese passwords are weaker than English passwords
  - Guess number $=100$, about 9.7% Chinese passwords are cracked, English 3.1%;
  - Guess number $=10^7$, >30% Chinese passwords are cracked, English >36%;
Bifacial-security of Chinese passwords

- When the guess number is small, Chinese passwords are weaker than English passwords;
- When guess number is large (e.g., >10^4), Chinese passwords are stronger than their English counterparts.

<table>
<thead>
<tr>
<th>Algorithm*</th>
<th>Test set</th>
<th>Online guessing</th>
<th>Offline guessing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10^1 10^2 10^3</td>
<td>10^4 10^5 10^6 10^7</td>
</tr>
<tr>
<td>PCFG</td>
<td>Dodonew</td>
<td>0.027 0.044 0.068</td>
<td>0.103 0.150 0.225 0.288</td>
</tr>
<tr>
<td></td>
<td>Yahoo</td>
<td>0.008 0.022 0.063</td>
<td><strong>0.136 0.212 0.316 0.390</strong></td>
</tr>
<tr>
<td></td>
<td>Tianya</td>
<td>0.073 0.105 0.138</td>
<td>0.213 0.295 0.355 0.376</td>
</tr>
<tr>
<td></td>
<td>Rockyou_rest</td>
<td>0.020 0.044 0.110</td>
<td>0.214 0.320 0.438 0.497</td>
</tr>
<tr>
<td></td>
<td>CSDN</td>
<td><strong>0.070 0.105 0.136</strong></td>
<td>0.189 0.229 0.272 0.300</td>
</tr>
<tr>
<td></td>
<td>Phpbb</td>
<td>0.021 0.038 0.087</td>
<td><strong>0.183 0.274 0.369 0.415</strong></td>
</tr>
<tr>
<td>Markov</td>
<td>Dodonew</td>
<td>0.024 0.040 0.060</td>
<td>0.085 0.145 0.212 0.305</td>
</tr>
<tr>
<td></td>
<td>Yahoo</td>
<td>0.007 0.016 0.043</td>
<td><strong>0.097 0.165 0.261 0.361</strong></td>
</tr>
<tr>
<td></td>
<td>Tianya</td>
<td><strong>0.062 0.087 0.118</strong></td>
<td>0.154 0.269 0.386 0.516</td>
</tr>
<tr>
<td></td>
<td>Rockyou_rest</td>
<td>0.018 0.035 0.081</td>
<td><strong>0.159 0.259 0.392 0.503</strong></td>
</tr>
<tr>
<td></td>
<td>CSDN</td>
<td><strong>0.037 0.098 0.116</strong></td>
<td>0.144 0.211 0.260 0.316</td>
</tr>
<tr>
<td></td>
<td>Phpbb</td>
<td>0.019 0.034 0.071</td>
<td><strong>0.146 0.230 0.333 0.436</strong></td>
</tr>
</tbody>
</table>
Outline

1. Our Datasets
2. Characteristics of Chinese pwds
3. Strength of Chinese pwds
4. Implications and insights
For password creation policies

- The finding of “bifacial-security nature” suggests that Chinese passwords are more vulnerable to online guessing attacks.

Suggestions for password policy:
1) A moderate size of blacklist;
2) Password strength meters.
For password strength meters (PSM)

- When designing a PSM, the confounding factors underlying a password distribution can be considered for training-set selection:
  1) Language > Site service type > Password policy;
  2) The closer the training set to the target password, the better

This suggests that there is no single training set that can fit all PSMs.
For password cracking

- Criteria for training-set selection:
  
  Language > Site service type > Password policy;

- PCFG-based attacks are
  
  1) **Lighter weight** (31% less computation and 70% less memory),
  2) **Higher success rates** when the guess number is **small** (e.g. <$10^3$)

- Markov-based attacks achieve **higher** success rates when the guess number is **large** (e.g. >$10^6$):
Takeaways

• We should study passwords from other cultures, because the passwords from different languages are so different
• Future work: investigate Chinese pwd behaviors through user surveys. (e.g, password creation and management behavior)
Thank You!

wangdingg@pku.edu.cn  pwang@pku.edu.cn  hedebiao@whu.edu.cn  yuant@virginia.edu
Post Credits: Interesting Passwords on CSDN

Guess what Chinese poem developers use to create these passwords?

- Willow _0f0=sprintf(“2_Bird_ff0/a”)
- 两个黄鹂鸣翠柳 (Two yellow birds are signing on a green willow)
- while(1)Ape1Cry&&Ape2Cry
- 两岸猿声啼不住 (Apes keep crying along both banks of the river)
- 1dcypsz1/2#0081/2j#ff0
- 一道残阳铺水中，半江瑟瑟半江红 (When sun sets, half of the river is green, and the other half of the river is red)
- Hint: the passwords might include color codes