Introduction

Background:
1. Password
   a. Dominant authentication method [1].
   b. Including meaning segments.
2. Probabilistic context-free grammars (PCFGs)
   a. Model password distributions.
   b. Used for password strength meters and password guessing attacks.

Challenge: How to segment passwords?
1. Existing segmentation methods:
   a. Simple segmentation based on char types (PCFGp [2], PCFGc [3]).
   b. Improved segmentation with external dictionaries (e.g., PCFGc [4]).
   c. “password” is identified as an English word.
   d. “password123” is identified as a keyboard pattern.
   e. But external dictionaries cannot fully and accurately cover the individual segments in passwords, because passwords are different from other types of texts.
2. Inaccurate segmentation leads to misestimation of password probability.
   a. Example: “jordan23” consists of Michael Jordan’s name and his jersey number. Current PCFG models divide it to two independent segments and underestimate its probability.

Contribution:
1. A word extraction method for passwords, extracting individual segments (called words) from passwords.

Word extraction for passwords

Extraction is based on cohesion and freedom, inspired by a method for Chinese words [5].

1. Cohesion is the evaluation of a string’s internal association.
   \[ \text{Coh}(s) = \min_{x \in C^s} \text{PMI}(s_x \ | \ s) \]
   where
   \[ \text{PMI}(s_x \ | \ s) = \log \frac{P(s_x \ | \ s)}{P(s_x) \cdot P(s)} \]

2. Freedom is the evaluation of a string’s independence from its context.
   \[ \text{Fre}(s) = -\sum_{x \in C^s} \log P(x|s) \cdot \log P(x|s) \cdot \log P(x|s) \]
   \[ \text{Fre}(s) = -\sum_{x \in C^s} \log P(x|s) \cdot \log P(x|s) \cdot \log P(x|s) \]
   \[ \text{Fre}(s) = \min_{x \in C^s} \text{Fre}(s) \]

We extract a substring \( s \) in passwords as a word if \( \text{Coh}(s) \geq T_c \) and \( \text{Fre}(s) \geq T_f \), where \( T_c \) and \( T_f \) are empirically set to 0.01 and 1.08, respectively.

WordPCFG

1. Extract words from passwords.
2. Segment passwords using the dictionary of words.
3. Train the probabilities of segments and templates.

Datasets

Passwords leaked from online services.

Table 1. Password dataset information

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Unique</th>
<th>Total</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rockyou</td>
<td>14,526,790</td>
<td>32,581,870</td>
<td>Social Network</td>
</tr>
<tr>
<td>000Webhost</td>
<td>10,585,709</td>
<td>15,251,073</td>
<td>Web Hosting</td>
</tr>
<tr>
<td>Clixsense</td>
<td>1,628,471</td>
<td>2,222,046</td>
<td>Online Surveys</td>
</tr>
<tr>
<td>CSDN</td>
<td>4,037,605</td>
<td>6,428,277</td>
<td>IT Community</td>
</tr>
<tr>
<td>Dodonew</td>
<td>10,135,260</td>
<td>16,258,891</td>
<td>Online Gaming</td>
</tr>
<tr>
<td>Duowan</td>
<td>3,119,060</td>
<td>4,982,730</td>
<td>Gaming Portal</td>
</tr>
</tbody>
</table>

Results

Table 2. Extracted words from passwords via our method

<table>
<thead>
<tr>
<th>Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyboard pattern</td>
<td>qwerasdf1q2w3e xcxcvbn1qaz123456</td>
</tr>
<tr>
<td>English word</td>
<td>superstar skate board lucky dragon</td>
</tr>
<tr>
<td>Chinese pinyin</td>
<td>woini woshi mima baobei haha</td>
</tr>
<tr>
<td>Name</td>
<td>steven wangming</td>
</tr>
<tr>
<td>Phrase</td>
<td>iloveu teamo byebye mylife howareyou</td>
</tr>
<tr>
<td>Hybrid</td>
<td>ilove24 jordan23 welcome24ever</td>
</tr>
</tbody>
</table>

Conclusion

1. Our word extraction method can automatically extract individual segments from passwords.
2. Using this method can precisely segment passwords.
3. Thus, our WordPCFG achieves a significant improvement on password guessing.

References