Improved Probabilistic Context-free Grammars for Passwords Using Word Extraction

PEKING UNIVERSITY

TU Delft — Delft University of Technology

Haibo Cheng [1] Wenting Li [1] Ping Wang [1] Kaitai Liang[2]

[1]Peking University [2]Delft University of Technology

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 (PCFG_W [2], PCFG_M [3]).
      I "password123" → "password/123".
      II "1qa2ws3ed" → "1/qa/2/ws/3/ed".
      This is *inaccurate*.
   b. Improved segmentation with external dictionaries (e.g., PCFG_C [4]).
      I "password" is identified as an English word;
      II "1qa2ws3ed" is identified as a keyboard pattern.
      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.*
   - 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.
2. A new password model—WordPCFG, achieving better performance on guessing attacks.

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.

$$\mathrm{Coh}(s) = \min_{s_1 \| s_2 = s} \mathrm{PMI}(s_1; s_2),$$

where

$$\mathrm{PMI}(s_1; s_2) = \log \frac{p(s_1 \| s_2)}{p(s_1) \cdot p(s_2)}.$$

2. Freedom is the evaluation of a string's independence from its context.

$$\mathrm{Fdm}_l(s) = -\sum_{c \in \Sigma} \Pr(c \| s) \cdot \log \Pr(c \| s),$$

$$\mathrm{Fdm}_r(s) = -\sum_{c \in \Sigma} \Pr(s \| c) \cdot \log \Pr(s \| c),$$

$$\mathrm{Fdm}(s) = \min_{x \in \{r,l\}} \mathrm{Fdm}_x(s).$$

We extract a substring s in passwords as a word if $\mathrm{Coh}(s) \geq T_c$ and $\mathrm{Fdm}(s) \geq T_f$, where T_c and T_f are empirically set to 0.01 and 1.0, respectively.

WordPCFG

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

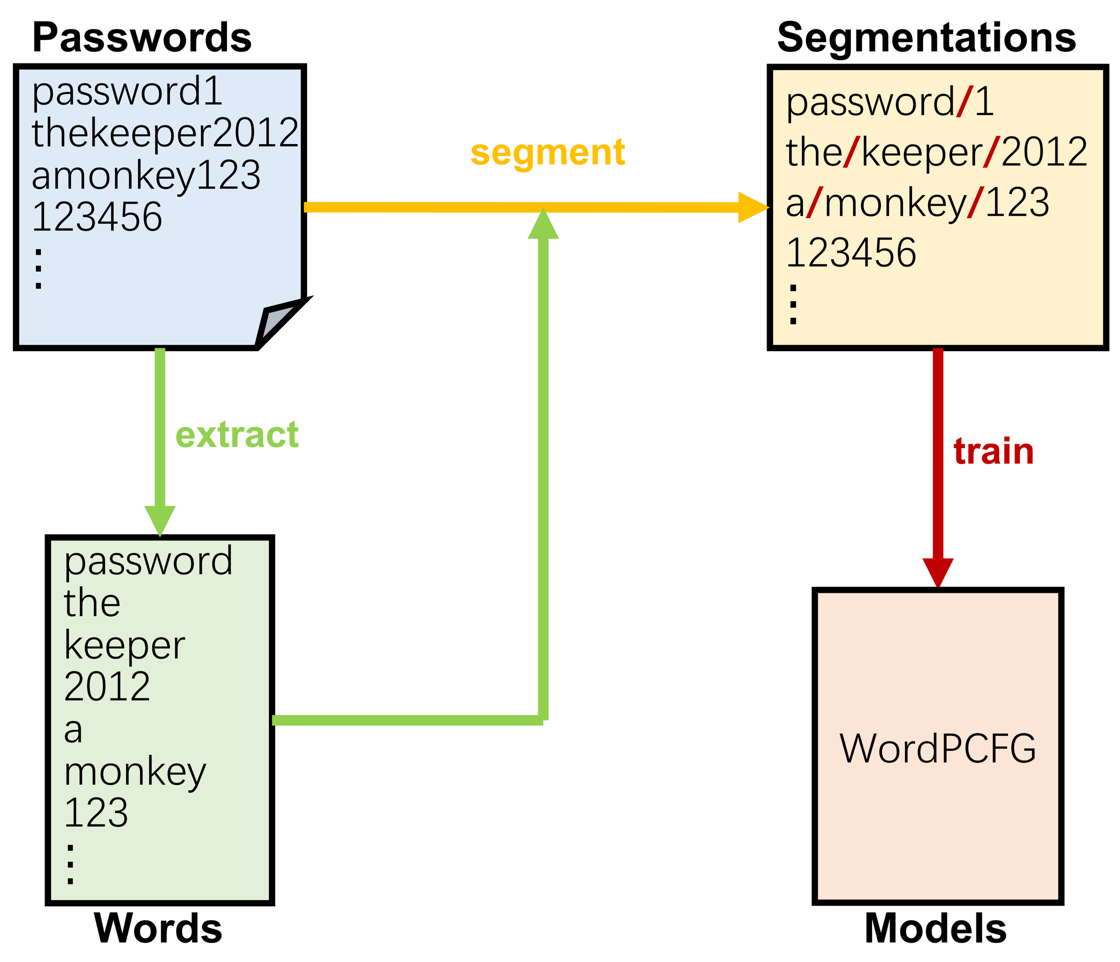

Figure 1. The training process for WordPCFG

Datasets

Passwords leaked from online services.

Table 1. Password dataset information

Dataset	Unique	Total	Service
Rockyou	14,326,970	32,581,870	Social Network
000Webhost	10,583,709	15,251,073	Web Hosting
Clixsense	1,628,471	2,222,046	Online Surveys
CSDN	4,037,605	6,428,277	IT Community
Dodonew	10,135,260	16,258,891	Online Gaming
Duowan	3,119,060	4,982,730	Gaming Portal

Results

Table 2. Extracted words from passwords via our method

Type	Examples
Keyboard pattern	qwerasdf 1q2w3e zxcvbn 1qaz 123456
English word	superstar skateboard lucky dragoon
Chinese pinyin	woaini woshi mima baobei haha
Name	steven wangming
Phrase	iloveu teamo byebye mylife howareyou
Hybrid	kobe24 jordan23 welcome2 4ever

Results

To show the accuracy of WordPCFG, we leverage it for guessing attacks.

1. Attack: Crack passwords in descending order of probabilities.
2. Experimental setting: randomly shuffle the dataset, and use one half for training and the rest for testing.
3. Performance:
   a. WordPCFG achieves a significant improvement, when the guessing number climbs to 10^{10}.
   b. WordPCFG can crack 83.04%–95.47% passwords, achieving a 12.96%–71.84% improvement.

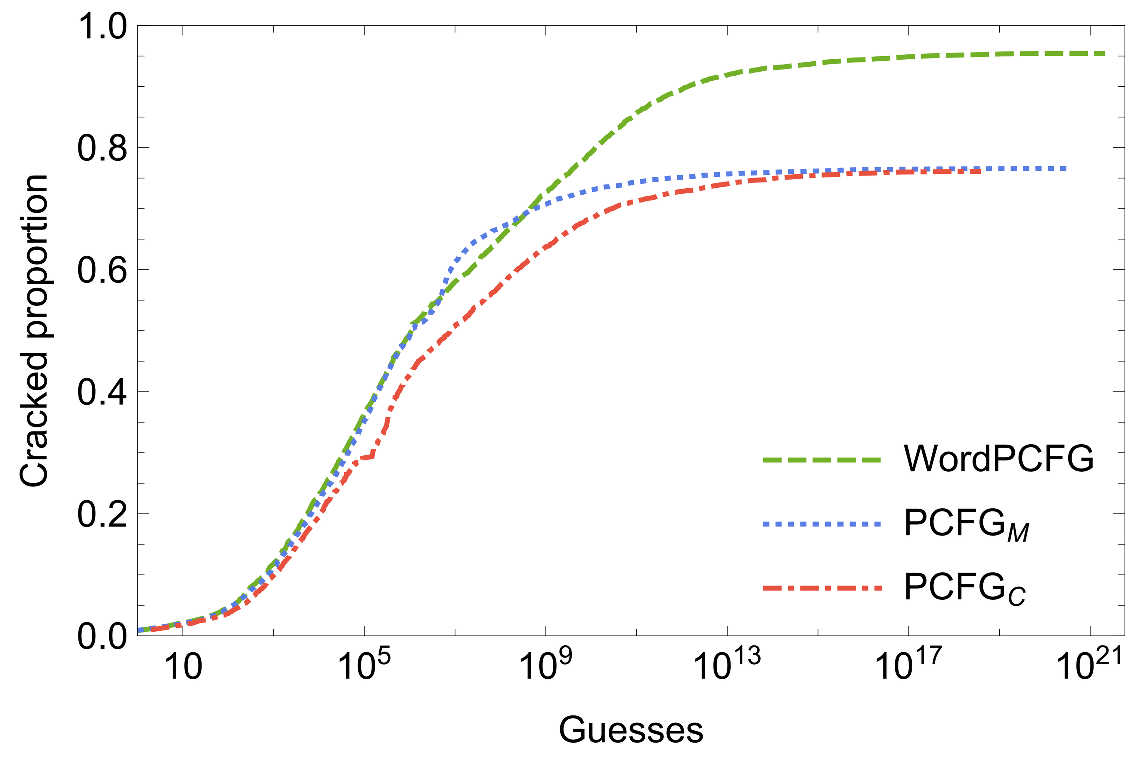

Figure 2. Rockyou

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

[1] Joseph Bonneau et al. "The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes". In: *Proc. IEEE S&P 2012*, pp. 553–567.

[2] Matt Weir et al. "Password Cracking Using Probabilistic Context-Free Grammars". In: *Proc. IEEE S&P 2009*, pp. 391–405.

[3] Jerry Ma et al. "A Study of Probabilistic Password Models". In: *Proc. IEEE S&P 2014*, pp. 689–704.

[4] Rahul Chatterjee et al. "Cracking-resistant password vaults using natural language encoders". In: *Proc. IEEE S&P 2015*, pp. 481–498.

[5] Shan He and Jie Zhu. "Bootstrap method for Chinese new words extraction". In: *Proc. IEEE ICASSP 2001*. Vol. 1, pp. 581–584.