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

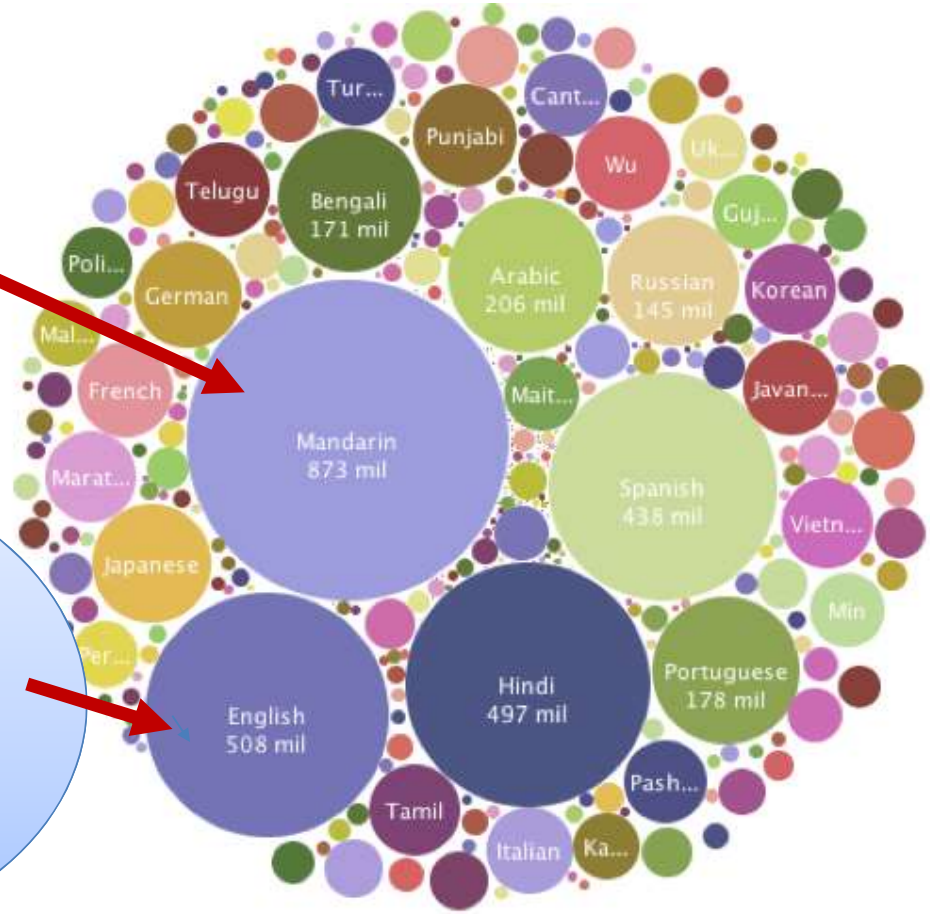
August 16, 2019 • Santa Clara, CA, USA

Password is everywhere

Passwords are influenced by native languages

[Bonneau 2012; Li 2014]

[Morris-Thompson 1979; Klein 1990; Yan et al. 2004; Bonneau 2012; Weir et al. 2010 ; Veras et al. 2012; Mazurek et al. 2013; Pearman et al. 2017; Becker et al. 2018]



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

Dataset	Web Service	Language	Leaked Time	# of PWDs
Tianya	Social forum	Chinese	Dec.2011	31,761,424
7k7k	Gaming	Chinese	Dec. 2011	19,138,452
Dodonew	E-commerce	Chinese	Dec. 2011	16,283,140
178	Gaming	Chinese	Dec. 2011	9,072,966
CSDN	Programmer forum	Chinese	Dec. 2011	6,428,632
Duowan	Gaming	Chinese	Dec. 2011	5,024,764
Rockyou	Social forum	English	Dec. 2009	32,603,387
Yahoo	Portal (e.g., E-commerce)	English	July 2012	453,491
Phpbb	Programmer forum	English	Jan. 2009	255,421

Top-10 Passwords

Tianya	7k7k	Dodonew	178	CSDN	Duowan	Rockyou	Yahoo	Phpbb
123456	123456	123456	123456	123456789	123456	123456	123456	12346
111111	0	A123456	111111	12345678	111111	12345	password	passsword
000000	111111	123456789	Zz12369	11111111	123456789	123456789	welcome	phpbb
123456789	123456789	111111	Qiulaobai	dearbook	123123	Password	ninja	qwerty
123123	123123	5201314	123456aa	00000000	000000	lloveyou	abc123	12345
123321	5201314	123123	wmsxie12	123123123	5201314	princess	123456789	12345678
5201314	123	a321654	123123	1234567890	123321	123321	12345678	letmein
12345678	12345678	12345	000000	88888888	a123456	rockyou	sunshine	111111
666666	1234567	000000	qq66666	11111111	suibian	12345678	princess	1234
111222tianya	wangyut2	123456a	w2w2w2	147258369	12345678	abc123	qwerty	123456789

Outline

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

Comparisons

Different:

- Letter distribution
- Semantic patterns

Similar:

- Length distribution
- Frequency distribution

Comparisons

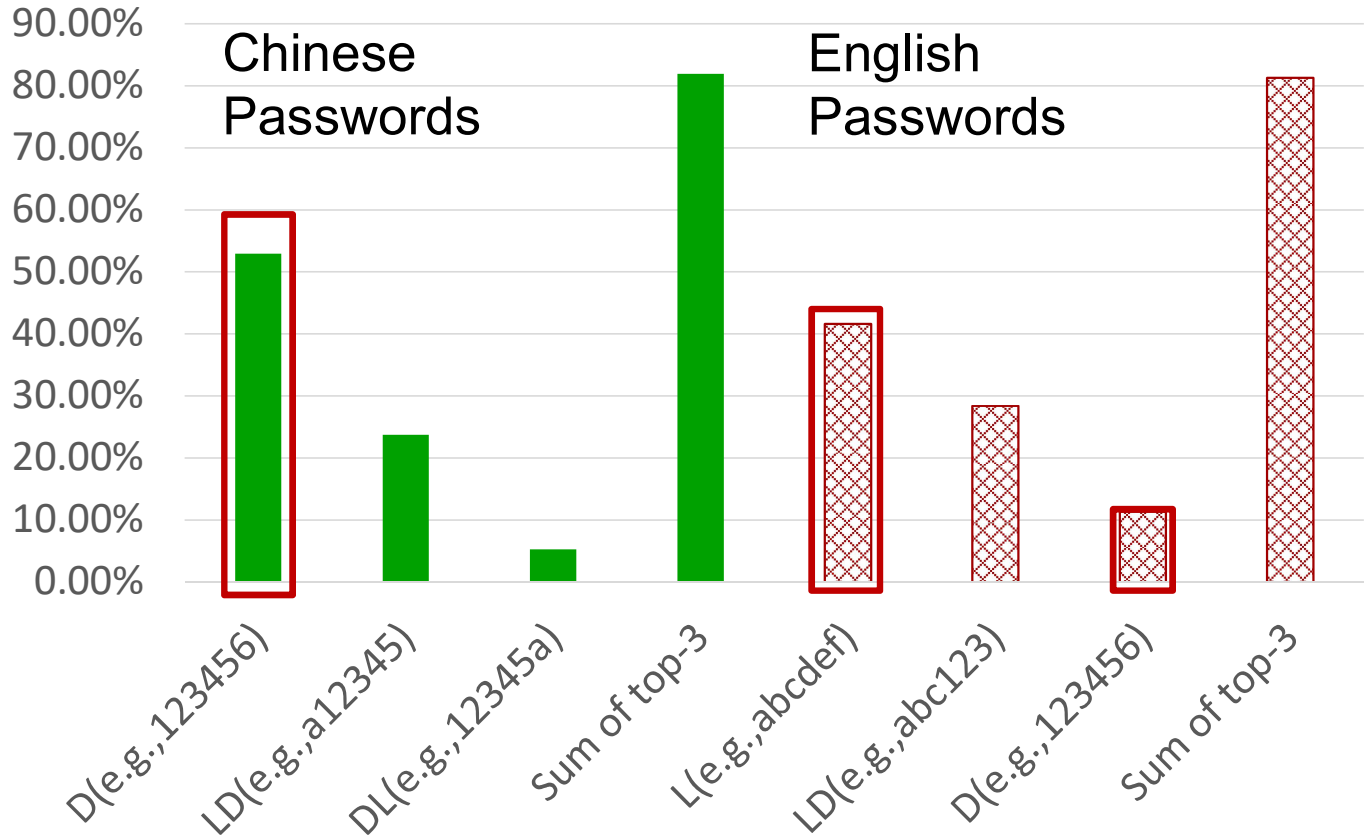
Different:

- Letter distribution
- Semantic patterns

Similar:

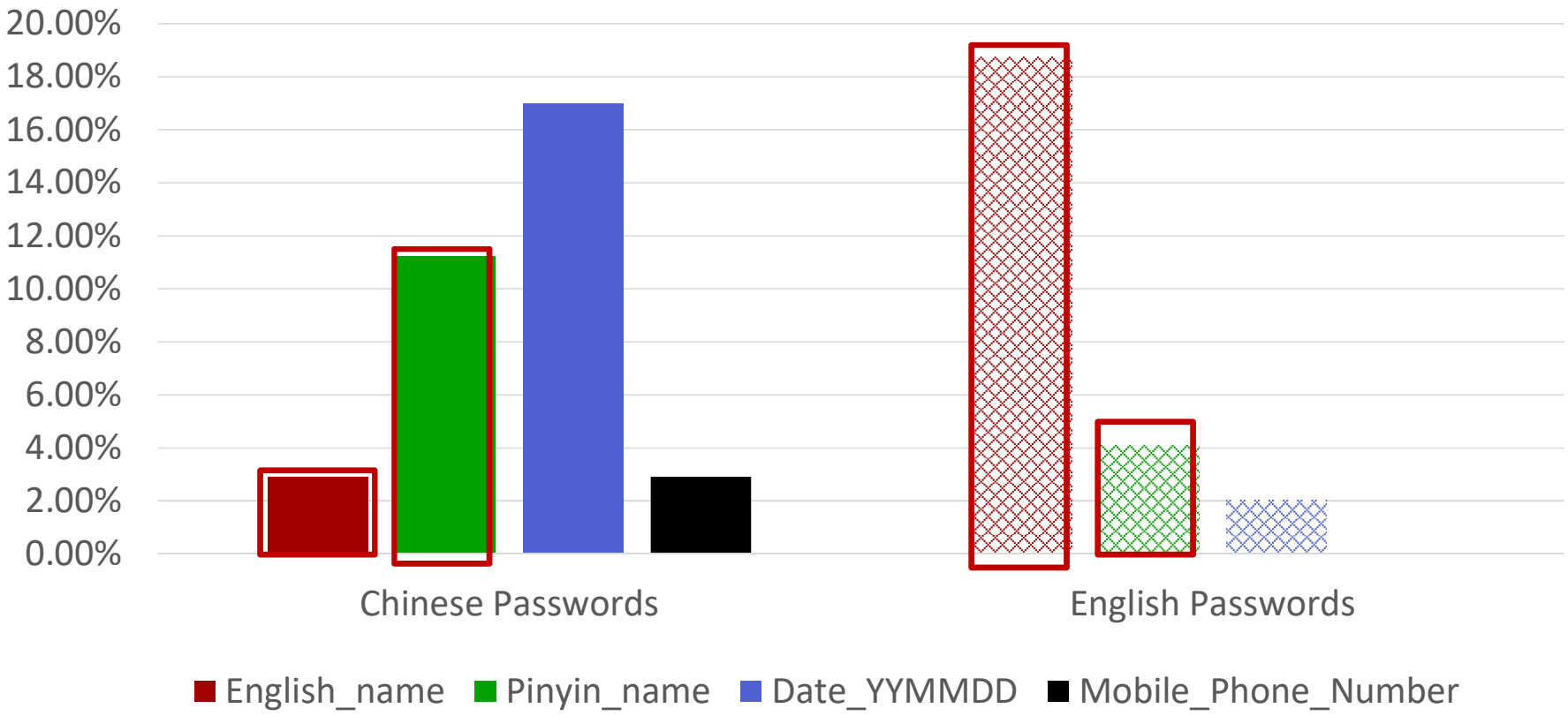
- Length distribution
- Frequency distribution

Top-3 structural patterns

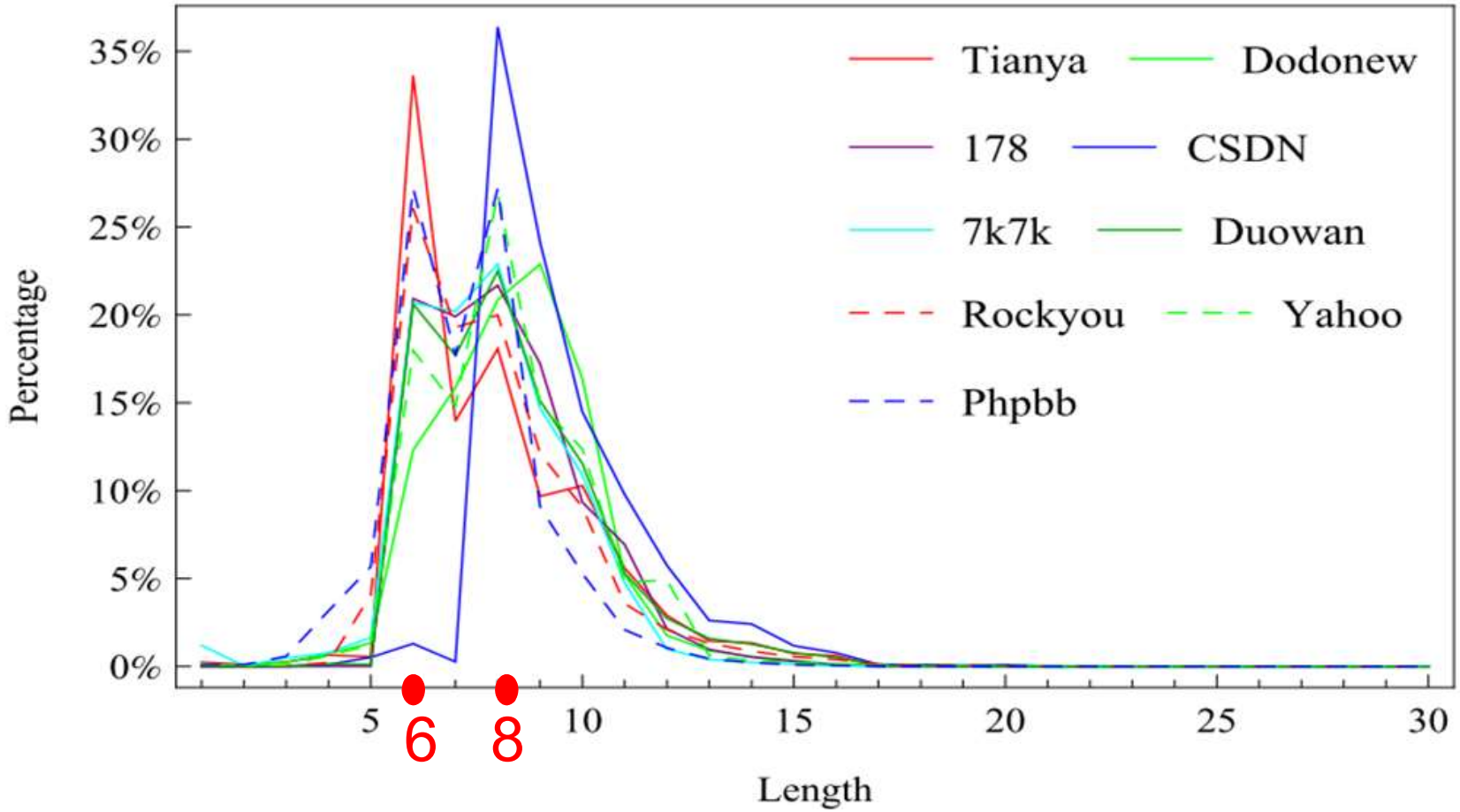


Semantic patterns in passwords

- We construct 22 dictionaries of different semantic categories



Length distribution of passwords



Outline

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

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

  PCFG: Probabilistic Context-Free Grammars [Weir et al . IEEE S&P'09; Ma et al. IEEE S&P'14] 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 algorihtm: parse pwds in to Letter-segmnet, Digital-segmnet, Symbol-segmnets.

E.g.,	password123 → L_8D_3	count(L_8D_3) +1
	L_8 → password	count(L_8 → password) +1
	D_3 → 123	count(D_3 → 123) +1

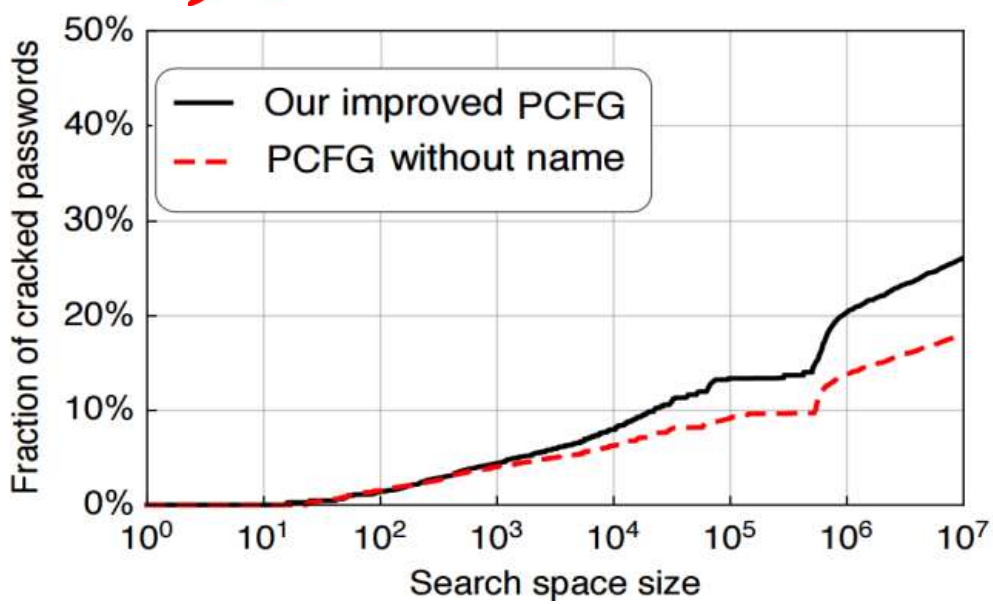
- We find that Chinese passwords include many interleaving patterns: e.g., “1n1d1g”
- PCFG algorithm overlooks the relationship between L and D segment, therefore underestimates the probability of them

Improvement on PCFG

- Interleaving conversion： transform interleaving passwords into normal passwords
- Input improvement: adding Pinyin names
- Input improvement: adding dates

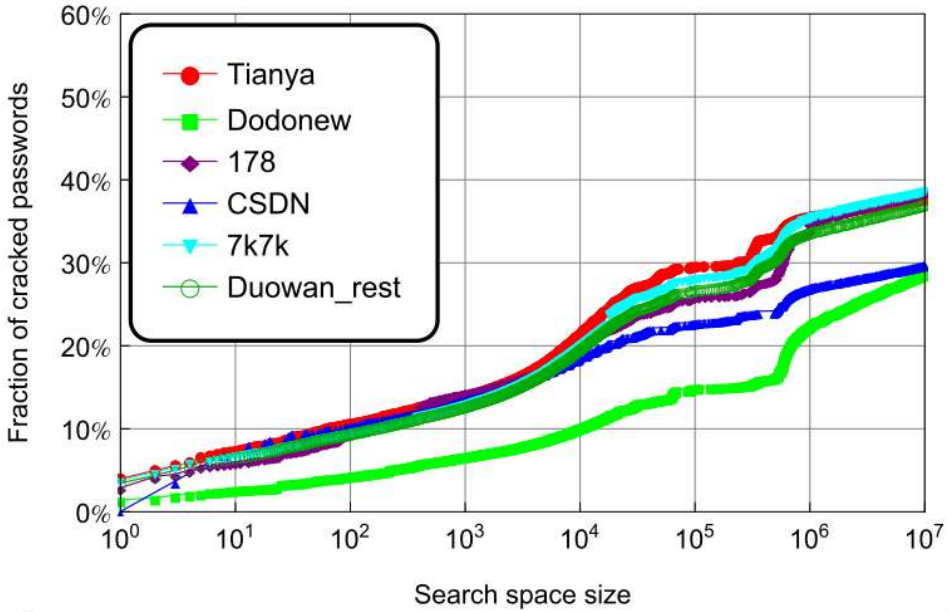
Improved PCFG cracking results

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

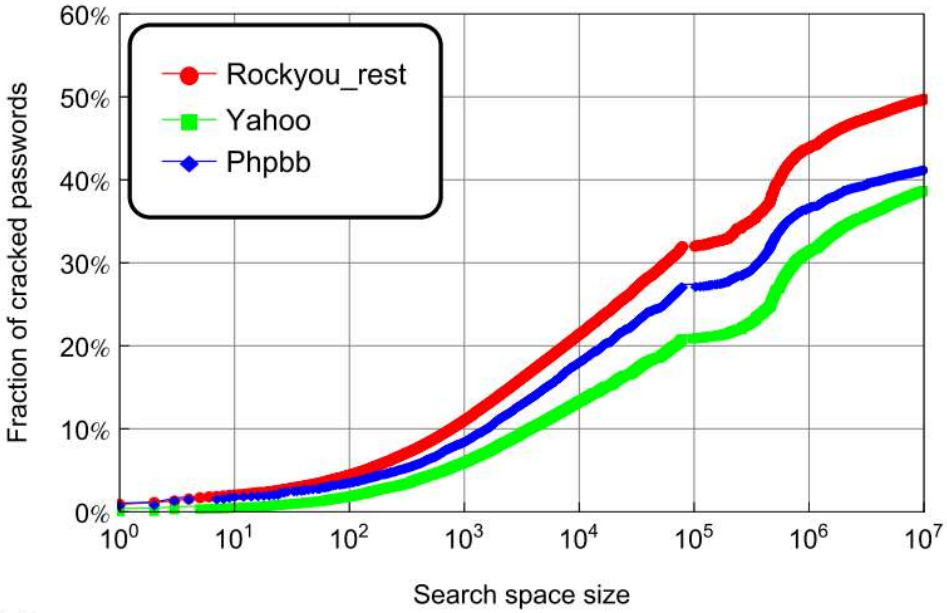


PCFG cracking results

- 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%；



(a) PCFG-based attacks on six Chinese datasets



(b) PCFG-based attacks on three English datasets

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.

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

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 desinging a PSM, the con-founding 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),

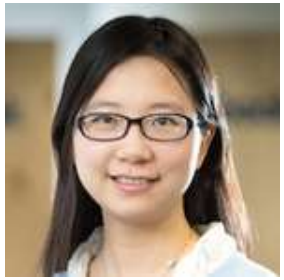
  2) Higher success rates when the guess number is small (e.g. $<10^3$)

- 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.educn 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