Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Characterizing the semantics of passwords: The role of Pinyin for Chinese Netizens

Gang Han[a,b], Yu Yu[c,*], Xiangxue Li[d,e,f,**], Kefei Chen[b,g], Hui Li[a]

[a] School of Electronics and Information, Northwestern Ploytechnical University, China
[b] Science and Technology on Communication Security Laboratory, China
[c] Department of Computer Science and Engineering, Shanghai Jiaotong University, China
[d] Department of Computer Science and Technology, East China Normal University, China
[e] Westone Cryptologic Research Center, China
[f] National Engineering Laboratory for Wireless Security, Xi'an University of Posts and Telecommunications, China
[g] School of Science, Hangzhou Normal University, China

ARTICLE INFO

Keywords:
Passwords
Semantics
Pattern
Human factors

ABSTRACT

Password-based authentication is the current dominant technology for online service providers to confirm the (claimed) identities of legitimate users. Semantic patterns reflect how people choose their passwords, and understanding the patterns is useful in developing policies, guidelines and good practices to secure the password-based mechanism. Semantic patterns are hard to recognize in general and they may vary for people of different spoken languages, cultures, and ethnicity groups, etc. However, it is possible to investigate them in a specific context. In this paper, we manage to characterize the Pinyin semantics of passwords from the Chinese Netizens (up to 591 million), thanks to the well-defined structures of the Pinyin phonetic system.

We perform a comprehensive analysis on the (publicly available) compromised password datasets from several leading Chinese sites for social networking, (micro)blogging, Internet forums, gaming, dating, and various other online service providers in China. The number of passwords in total sums to over 141 million, of which the largest site leaks more than 30 million on its own. Our findings show that over 4% of passwords from our datasets represent Pinyin (including names), another nearly 5% of passwords represent concatenations of Pinyin and date (i.e., Pinyin with a date prefix/suffix), and the next 17% of passwords are combinations of Pinyin and numeric (non-date) prefix/suffix. A majority (over 93%) of pure Pinyin passwords are transcribed from only 2–4 Chinese characters. The pure numeric pattern and the pattern containing special symbols are also studied. Over 76% of the passwords can be covered by the patterns of pure numeric and concatenation of Pinyin and digits. Special symbols appear in only 2.66% of the passwords, and they are most likely (with a percentage of 82.85%) in the middle. To the best of our knowledge, this is the first large scale study of its kind, and might yield other interesting insights into the semantic role Pinyin plays (either as good practice guidance on strengthening password security, or for improving password guessing attack).

1. Introduction

In 2014, a collection of private pictures of various celebrities were posted and disseminated on websites and social networks. The hackers could have taken advantage of a security issue in the iCloud API which allowed them to guess victims' passwords. And this prompts the question on the security of sensitive data stored in the cloud. As modern mobile devices, including phones, generally upload pictures and other media to the cloud provider, access to the cloud services will provide the attackers access to such sensitive data. The cloud and SaaS (Software-as-a-Service) applications are great targets for these attacks because those applications have to deal with password-based user identities and because they are accessible from anywhere in the world.

User authentication is a central component of currently deployed security infrastructures. Three main techniques are used for user authentication: knowledge-based systems (what the user knows), token-based systems (what the user possesses), and biometrics-based systems (what the user is). Of them, knowledge-based (typically, password-based) schemes have a long history and are the current predominant authentication method for online services. In this paper

[*] Corresponding author.
[**] Corresponding author at: Department of Computer Science and Technology, East China Normal University, China.
E-mail addresses: yuyu@yuyu.hk (Y. Yu), xxli@cs.ecnu.edu.cn (X. Li).

http://dx.doi.org/10.1016/j.csi.2016.10.006
Received 13 March 2016; Received in revised form 17 September 2016; Accepted 10 October 2016
Available online xxxx

we focus on textual passwords.

We have seen considerable efforts studying the usage and characteristics of passwords [10,11,14,17,19,24,28,32,41]. For example, the authors of [41] explored password vulnerabilities and threats in a university context, including best practices for password syntax, security, and policy. Using password lists from four online sources (hotmail, flirtlife, computerbits, rockyou), Malone and Maher [17] investigated whether Zipf's law is a good candidate for describing the frequency with which passwords are chosen.

Despite decades of password research, there is consistent difficulty in collecting realistic data to analyze. This explains why existing password studies suffer from one or more of the following drawbacks [18]: limited-scale datasets, data from experimental studies rather than from deployed authentication systems, no access to plaintext passwords, etc.

Although we know that patterns (e.g., similarity to dictionary words and the types/positions of characters used) exist in user chosen passwords, we still do not have a good grasp of how people choose them [4,5,11,13,21] and the nature and presence of semantic patterns in user-chosen passwords remains somewhat of a mystery. Semantic patterns are useful mnemonics that help people remember their passwords; they also have the potential to heavily impact security if the pattern defines a small number of passwords that an attacker can use in a guessing attack.

Understanding the semantic patterns behind the passwords that people choose is not an easy task, some researchers thus focus on dates in passwords. Bonneau [4] indicated that numbers appear to be commonly used in passwords across language groups, nations, and other population groups. Bonneau and Preibusch [5] observed that dates are common amongst 4-digit sequences, but their findings do not generalize to what a date pattern from a password looks like[1] and its connections to other texts within the passwords. Veras et al. [39] found that nearly 5% of passwords in the RockYou dataset represent pure dates (either purely numerical or mixed alphanumeric representations).

Semantic patterns are hard to recognize in general and they may vary for people of different spoken languages, cultures, and ethnicity groups, etc. However, it is possible to investigate them in a specific context. We focus in this paper on the passwords for Chinese Netizens (up to 591 million as of 2013 [8]) and manage to investigate the Pinyin semantics of passwords thanks to the well-defined structure of Pinyin. Pinyin [27], formally Hanyu Pinyin, is the official phonetic system of China and Singapore for transcribing the Mandarin pronunciations of Chinese characters into the Latin alphabet. It is often used to teach Standard Chinese and spell Chinese names in foreign publications and may be used by many Chinese IME (Input Method Editor) systems (such as Google Pinyin and Microsoft Pinyin) for entering Chinese characters into computers.

Our analysis of password patterns from large-scale, real-world datasets is fueled by the leaks of hundreds of millions of passwords from popular websites during the last few years in China [43,2,7,15]. We examine large scale datasets of over 141 million passwords. Our findings show that over 4% of passwords in our datasets represent pure Pinyin (including Chinese names), the next nearly 5% of passwords represent concatenations of Pinyin and date (representations of Pinyin with date prefix/suffix), and another 17% of passwords are concatenations of Pinyin and other (non-date) digits (representations of either Pinyin followed by digits or digits followed by Pinyin). A majority (over 93%) of pure Pinyin passwords are transcribed from only 2 to 4 Chinese characters. The pure numeric pattern and the pattern containing special symbols are also discussed. Over 76% passwords can be covered by the patterns of pure numeric and concatenation of Pinyin and digits. Special symbols appear in only 2.66% of the passwords, and they are most likely (with a percentage of 82.85%) in the middle. This is the first large scale study of its kind, and might yield other interesting insights into the semantic role Pinyin plays (either as good practice guidance on strengthening password security, or for improving password guessing attack).

2. Related work

There is extensive literature on password guessing and distribution [1,4,11,14,17,18,32,40], and that user-chosen passwords fall into predictable patterns has been well documented. Some checkers [9] can detect weak patterns such as common words, repetitions, easy keyboard sequences, common semantic patterns (e.g., dates and years) and natural character sequences (e.g., `123` and `gfedcba`). Morris and Thompson found that a large fraction of passwords on a Unix system are easily guessable [20]. Three decades later, Florencio and Herley [11] showed that web users gravitate toward the weakest passwords allowed and reported the results of a large scale study of password use and password re-use habits by getting extremely detailed data on password strength, the types and lengths of passwords chosen, and how they vary by site.

Bonneau and Preibusch [5] provided the first published estimates of the difficulty of guessing a human-chosen 4-digit PIN. They used a set of patterns, including five different date patterns (e.g., MMDD) and found that guessing PINs based on the victims' birthday will enable a competent thief to gain use of an ATM card once for every 11–18 stolen wallets, depending on whether banks prohibit weak PINs such as `1234`.

Veras et al. [39] focused on passwords characterized by sequences of 5–8 digits and found that in the RockYou dataset, which contains over 32 million passwords, over 15% of passwords contain sequences of 5–8 consecutive digits, 38% of which could be classified as a date. This represents significantly more dates than one would expect to parse from a randomly generated set of numbers of the same length.

Uchida [36] used a password pattern methodology to generate strong and memorable passwords. Their trick is that if a password pattern is chosen that is easily referenced by a physical cue or tool, the password itself can be strong but also memorable.

Veras et al. [38] leveraged Natural Language Processing to analyze semantic patterns in leaked passwords. They found that most passwords in the RockYou dataset are semantically meaningful, containing terminologies related to love, sex, profanity, animals, alcohol and money.

3. Data preparation

In this section, we discuss our data sources which provide the realistic textual passwords from deployed authentication systems. There are many sets of passwords (e.g., 40 million from the OpenWall Mangled Wordlist [25], 32 million from the website RockYou [37], and 47,000 from MySpace [33], etc.) belonging to sites which were hacked and the lists of passwords were leaked to the public domain subsequently.

During the past few years many security breaches in leading websites in China led to the disclosure of passwords of hundreds of millions of users [2,7,15]. This is the first time such huge volume passwords were leaked to the public in China although we have seen similar leakage many times in Western world [3,6,12,31]. These leaked password lists provide the largest samples of real-world passwords to date, offering an enormous opportunity for empirically grounded research. An attacker might be able to obtain a very accurate distribution for a given site by correlating user statistics.

3.1. Datasets

Our password datasets belong to a number of major online sites

[1] There is a variety of formats for dates both numerically (e.g., 31052014, 05312014, and 20140513) and literally (e.g., may312014 and wuyue2013).

Table 1
The number of passwords from each site.

Site	#pass
163	1,846,207
126	9,356,887
CSDN	6,427,428
7K7K	19,103,896
178	9,049,078
766	5,808,460
17173	18,545,683
ispeak	9,645,862
mop	2,660,850
renren	4,735,089
tianya	30,895,992
weibo	4,732,385
duowan	8,028,166
pconline	5,442,614
ys168	328,576
zhenai	5,247,667

that were hacked, causing the disclosure of huge volume of passwords. The largest one contributes over 30 million passwords to our analysis. The authenticity of the data sources is either confirmed by the owners themselves or validated by the victim users [2,7]. After filtering out the dirty data, we have 141,854,840 passwords in total. We mention that the sites where our data were taken from include many popular online service providers (Table 1) in China and we list a few examples as below.

- Type 1: Email services

  Two of our data sources are mail.163.com and www.126.com. Both are email services provided by the NetEase Inc. [22], a popular web portal (ranked 27 by Alexa as of April 2014 [29]) and NASDAQ listed company. The former is allegedly the biggest Chinese provider for personal email services. We extract 11,203,094 passwords from the password datasets for the above email systems.
- Type 2: CSDN

  Another data source is the Chinese Software Developer Network (CSDN, www.csdn.net), which is one of the biggest networks of software developers in China and provides Web forums, blog hosting, IT news, and other services. CSDN has about 10 million registered users and is the largest developer community in China [23]. The CSDN users are mainly IT professionals, computer engineers and scientists, who have better security awareness than other normal people. We get 6,427,428 passwords from the CSDN data source.
- Type 3: Others (Internet forums, gaming, dating, etc.)

  The next example is Sina Weibo (www.weibo.com the most popular Twitter-like microblogging site in China used by well over 30% of Chinese Netizens [42]. Sina Weibo was owned by the biggest Chinese web portal Sina corp (NASDAQ:SINA), and recently it became a spinoff (NASDAQ:WB) separated from Sina corp.

  The other 12 sites of our data sources include Internet forums, gaming, dating service providers such as Tianya, 7k7k, 17173, zhenai. For example, as a popular Internet forum in China, Tianya Club is the 33rd most visited site in China and 241st overall by Alexa as of May 2014 [30]. We get 124,224,318 passwords from these data sources.

With the many ways we use the Internet, it is easy to consider some passwords less important than others. On the other hand, however, all passwords are important because wrongdoers can piece together the information users store online and use it for their benefit. By learning users' current password structure, attackers can increase their chances of guessing passwords for critical websites such as users' bank account or users' company's email account.

Table 2
Pinyin (1/3).

final sound	zh	ch	sh	r	
a	zha	cha	sha		a
o					o
e	zhe	che	she	re	e
ai	zhai	chai	shai		ai
ei	zhei		shei		ei
ao	zhao	chao	shao	rao	ao
ou	zhou	chou	shou	rou	ou
an	zhan	chan	shan	ran	an
ang	zhang	chang	shang	rang	ang
en	zhen	chen	shen	ren	en
eng	zheng	cheng	sheng	reng	eng
ong	zhong	chong		rong	
u	zhu	chu	shu	ru	wu
ua	zhua	chua	shua	rua	wa
uo	zhuo	chuo	shuo	ruo	wo
uai	zhuai	chuai	shuai		wai
ui	zhui	chui	shui	rui	wei
uan	zhuan	chuan	shuan	ruan	wan
uang	zhuang	chuang	shuang		wang
un	zhun	chun	shun	run	wen
ueng					weng

3.2. Parsing

Pinyin, is the official phonetic system for transcribing the Mandarin pronunciations of Chinese characters into the Latin alphabet in China. It is used to spell Chinese names in foreign publications and may be used as an input method to enter Chinese characters into computers. The spelling of Chinese geographical or personal names in Pinyin has become the most common way to transcribe them in English. Pinyin has also become the dominant method for entering Chinese text into computers in China.

Unlike European languages, clusters of letters – initials and finals – and not consonant and vowel letters, form the fundamental elements in Pinyin. Tables 2–4 summarize the syllables of Mandarin Chinese as shown in the combinations of initial and final sounds and as spelled in Hanyu Pinyin.

Thanks to the well-defined structure of Chinese language, we can use these tables to parse a given password that contains letters purely. For example, the common password `woaini` (meaning: i love you) can be parsed as `wo`, `ai`, and `ni`.

It is known that some examples of the most commonly used passwords include `password, password1, PASSWORD`, etc. As well as these common passwords, users also commonly use family or pets names such as charlie, thomas, or fluffy, sports teams or sports players names. Some users use common names and just add a number after it, e.g., `charlie1`. Here we also address the question: what does a Chinese name (the first name and family name) mean for passwords?

As a part of Pinyin, Chinese names also represent semantics in passwords. We mention the book "Hundred Family Surnames" [35] composed in the early Song Dynasty (969–1127) that contained 411 Chinese surnames and was later expanded to 504 (including 444 single-character surnames and 60 double-character surnames). Table 5 illustrates the first 128 family names in China (the order is not according to the actual arrangement of surname of population, but easy to read, learn and remember). We remark that different surnames of Chinese may be represented as the same spelling of Pinyin. Among the family names, however, most surnames are rare, and most people have populous surnames. For example, 21% of the population of Chinese have the most populous 0.595% of surnames: there are 95 million, 93 million, and 90 million people with the surnames wang, li and zhang respectively. The first names of Chinese are of one character or two characters basically. Combining the first names with the family names, we can see that Chinese names are often composed of two

Table 3
Pinyin (2/3).

final sound	b	p	m	f	d	t	n	l	g	k	h	z	c	s
a	ba	pa	ma	fa	da	ta	na	la	ga	ka	ha	za	ca	sa
o	bo	po	mo	fo										
e			me		de	te	ne	le	ge	ke	he	ze	ce	se
ai	bai	pai	mai		dai	tai	nai	lai	gai	kai	hai	zai	cai	sai
ei	bei	pei	mei	fei	dei	tei	nei	lei	gei	kei	hei	zei		
ao	bao	pao	mao		dao	tao	nao	lao	gao	kao	hao	zao	cao	sao
ou		pou	mou	fou	dou	tou	nou	lou	gou	kou	hou	zou	cou	sou
an	ban	pan	man	fan	dan	tan	nan	lan	gan	kan	han	zan	can	san
ang	bang	pang	mang	fang	dang	tang	nang	lang	gang	kang	hang	zang	cang	sang
en	ben	pen	men	fen	den		nen		gen	ken	hen	zen	cen	sen
eng	beng	peng	meng	feng	deng	teng	neng	leng	geng	keng	heng	zeng	ceng	seng
ong					dong	tong	nong	long	gong	kong	hong	zong	cong	song
u	bu	pu	mu	fu	du	tu	nu	lu	gu	ku	hu	zu	cu	su
ua									gua	kua	hua			
uo					duo	tuo	nuo	luo	guo	kuo	huo	zuo	cuo	suo
uai									guai	kuai	huai			
ui					dui	tui			gui	kui	hui	zui	cui	sui
uan					duan	tuan	nuan	luan	guan	kuan	huan	zuan	cuan	suan
uang									guang	kuang	huang			
un					dun	tun	nun	lun	gun	kun	hun	zun	cun	sun
ueng														

characters or three characters (e.g., Yu Yu and Xiangxue Li). With well-defined structure, these names can contribute to password analysis.

4. Semantic patterns discovered

Besides the semantic pattern Pinyin in passwords, we also notice some other patterns. This section discusses our findings.

4.1. Length distribution

Fig. 1 shows the length distribution of our data sources. One can see that the dominant password length ranges from 6 to 11 letters. The average lengths of type 1, type 2, and type 3 are about 8.086814, 9.433813, and 8.3156 respectively.

Since the users registered at type 2 have better security awareness, we expect that the passwords of type 2 should be stronger than those of type 3, which is confirmed by the length distributions of the passwords (see Fig. 1).

As emails may communicate much sensitive information, it is a natural hypothesis that the email passwords should be much stronger than those of type 3. However, the resulting length distributions do not coincide with this. This may be due to the fact that the two email systems are built at the early stage of Chinese Internet development and thus the systems had loose requirements on choosing the passwords and the users did not know much about the password strength at that time.

4.2. Letter characteristics

We check the letter frequency of the passwords, where the frequency of letter α ($\alpha = a, \ldots, z$) is computed by

$$\frac{\text{the number of } \alpha \text{ in the datasets}}{\text{the sum of the numbers of a to z in the datasets}}.$$

We first quantify the usage of upper- and lower-case letters in all 141,854,840 passwords. Table 6 illustrates the number of each uppercase letter in the passwords. And Table 7 shows the number of each lowercase letter in the passwords.

It is clear to see from Tables 6 and 7 that users are far more likely to use lowercase letters when setting their passwords. This reflects the fact that people do not like to press two keys at the same time or switch between uppercase and lowercase. Another result from the comparison may tell us that the sites may not require the users to take uppercase letters when choosing passwords.

We then do not distinguish uppercase or lowercase to compute letter frequency. Fig. 2 shows the comparisons for the passwords in type 1, type 2 and type 3.

One can see that the three curves have similar shape. This implies that in general Chinese users follow the same characteristics to choose their passwords. We can thus view our data sources as a whole and

Table 4
Pinyin (3/3).

final sound	b	p	m	f	d	t	n	l	z	c	s	zh	ch	sh	r	j	q	x	
i	bi	pi	mi		di	ti	ni	li	zi	ci	si	zhi	chi	shi	ri	ji	qi	xi	yi
ia					dia			lia								jia	qia	xia	ya
ie	bie	pie	mie		die	tie	nie	lie								jie	qie	xie	ye
iao	biao	piao	miao		diao	tiao	niao	liao									qiao	xiao	yao
iu			miu		diu		niu	liu								jiu	qiu	xiu	you
ian	bian	pian	mian		dian	tian	nian	lian								jian	qian	xian	yan
iang							niang	liang								jiang	qiang	xiang	yang
in	bin	pin	min				nin	lin								jin	qin	xin	yin
ing	bing	ping	ming		ding	ting	ning	ling								jing	qing	xing	ying
iong																jiong	qiong	xiong	yong
v							nv	lv								ju	qu	xu	yu
ve							nue	lue								jue	que	xue	yue
van																juan	quan	xuan	yuan
vn																jun	qun	xun	yun

Table 5
Family names.

Zhao	Qian	Sun	Li	Zhou	Wu	Zheng	Wang	Feng	Chen	Chu	Wei	Jiang	Shen	Han	Yang
Zhu	Qin	You	Xu	He	Lu	Shi	Zhang	Kong	Cao	Yan	Hua	Jin	Wei	Tao	Jiang
Qi	Xie	Zou	Yu	Bai	Shui	Dou	Zhang	Yun	Su	Pan	Ge	Xi	Fan	Peng	Lang
Lu	Wei	Chang	Ma	Miao	Feng	Hua	Fang	Yu	Ren	Yuan	Liu	Feng	Bao	Shi	Tang
Fei	Lian	Cen	Xue	Lei	He	Ni	Tang	Teng	Yin	Luo	Bi	Hao	Wu	an	Chang
Le	Yu	Shi	Fu	Pi	Bian	Qi	Kang	Wu	Yu	Yuan	Bu	Gu	Meng	Ping	Huang
He	Mu	Xiao	Yin	Yao	Shao	Zhan	Wang	Qi	Mao	Yu	Di	Mi	Bei	Ming	Zang
Ji	Fu	Cheng	Dai	Tan	Song	Mao	Pang	Xiong	Ji	Shu	Qu	Xiang	Zhu	Dong	Liang

Table 8 shows the letter frequency of the whole datasets. The high frequency letters are (in order of frequency) aA, iI, nN, eE, oO, hH, lL, gG, and bB, fF, kK, pP, vV are on the other side. Comparatively, taking from Pavel Micka's website [26], which cites Robert Leward's Cryptological Mathematics [16], we know that the high frequency letters in English are E, T, A, O, I, N, S, H, and K, J, X, Q, Z are on the other side.

Note that for the 52 letters in the datasets, Shannon entropy is 4.74231, which is far less than 5.70044(=log $\frac{1}{52}$, the entropy of a single random letter).

4.3. Special symbols

Now we move on to passwords that have one or more special symbols (i.e., non-alphanumeric). Table 9 shows the numbers of passwords that contain special symbols.

As expected, type 1 and type 2 have more percentages than type 3, since email accounts may convey sensitive information and the users of CSDN have better security-awareness.

In all 141,854,840 passwords, there are 3,776,287 ones that have special symbols. Interestingly, special symbols are much more likely to occur in the middle (82.8506%) than at the beginning (a special symbol followed by an alphanumeric sequence, 4.0593%) or at the end (an alphanumeric sequence with a special symbol suffix, 13.9011%), as shown in Fig. 3.

In all these 3,776,287 passwords that have special symbols, special symbols occur 6,320,821 times. The distribution of special symbols used in the passwords is shown in Fig. 4.

As anticipated, symbols “. @_ + * - !” are the most frequently used special ones in user-chosen password. For example, separators (e.g., “-” and “.”) are typically used to delimit the elements of a date (year, month, and day). However, people tend to avoid the use of special characters such as “: ><’ } | ” {” that have low percentages.

Considering all the 26 uppercase letters, the 26 lowercase letters, 10 digits, and 32 special symbols in all the 141,854,840 passwords, we have Shannon entropy 4.715885, which is far less than 6.554589(=log $\frac{1}{94}$). This implies that the sampling of password is far from a uniform selection from the alphabet.

4.4. Pinyin

Now we check the Pinyin characteristics in the passwords. We classify all our passwords into four groups (Fig. 5).

- Group 1: Numeric
  In this group, all the passwords are purely digit sequences.
- Group 2: Pinyin
  We check the cases that users use Pinyin or their names as their passwords and focus on 6 classes.

  (1) Pure Pinyin class, i.e., a password is a full spelling for Chinese characters. For example, the password `woaini` is transcribed from the 3 Chinese characters meaning `iloveyou`. When parsing the passwords that are purely with letters, we use the Chinese language characteristics.

  (2) Pinyin and date class, i.e., the passwords with the form of the concatenation of Pinyin and dates, e.g., `woaini20140101` and `20140101woaini`.

  (3) Pinyin and other non-date digits class, i.e., the passwords with the form of the concatenation of Pinyin and numbers, e.g., `woaini0` or `0woaini`.

  (4) Pure name class. As mentioned above, we suppose that the first name is composed of one or two Chinese characters since the first names longer than 3 characters are pretty rare. Moreover, we only consider the single-character surnames and omit double-character surnames since the later is so rare that can be negligible.

  With all these in mind, we consider four cases: full spelling of surname followed by the first letter(s) of the first name, e.g., `lixx` and `yuy`; the first letter(s) of the first name followed by full spelling of surname, e.g., `xxli` and `yyu`; full spelling of first name followed by the first letter of surname, e.g., `xiangxuel` and `qiz`; the first letter of surname followed by full spelling of the first name, e.g., `lxiangxue` and `zqi`. Note that full spelling of name (e.g., `lixiangxue`) is already covered by the class (1).

  (5) Name and date class, i.e., the passwords with the form of the concatenation of name (as in the class (4)) and dates, e.g., `lixx20140101` and `20140101lixx`.

  (6) Name and other number class, i.e., the passwords with the

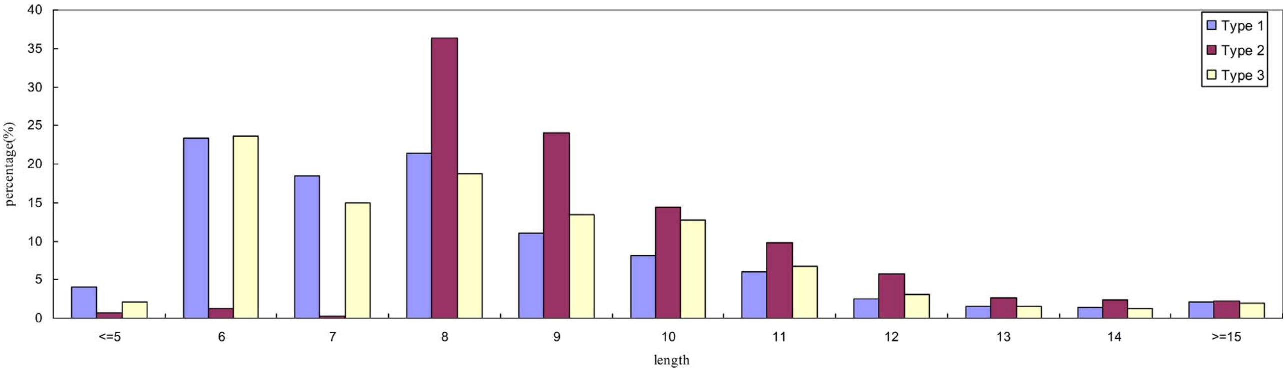

Fig. 1. Length distribution.

Table 6
The number of each uppercase letter in the datasets.

A	B	C	D	E	F	G	H	I	J	K	L	M
1,217,497	567,259	633,253	602,231	921,879	496,591	611,067	716,812	824,093	564,111	292,639	768,333	432,497
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
868,857	476,807	368,990	416,934	408,319	596,193	425,964	453,926	272,433	483,438	454,124	511,445	490,687

Table 7
The number of each lowercase letter in the datasets.

a	b	c	d	e	f	g	h	i	j
39,196,859	8,464,096	12,157,187	11,320,925	21,222,803	7,220,631	16,827,585	17,601,168	30,228,811	9,871,620
k	l	m	n	o	p	q	r	s	t
7,155,053	17,155,900	10,691,674	29,803,454	20,308,013	5,849,266	10,336,563	9,194,101	16,047,483	9,430,935
u	v	w	x	y	z				
15,464,585	3,704,925	13,546,378	10,744,993	14,871,757	11,088,826				

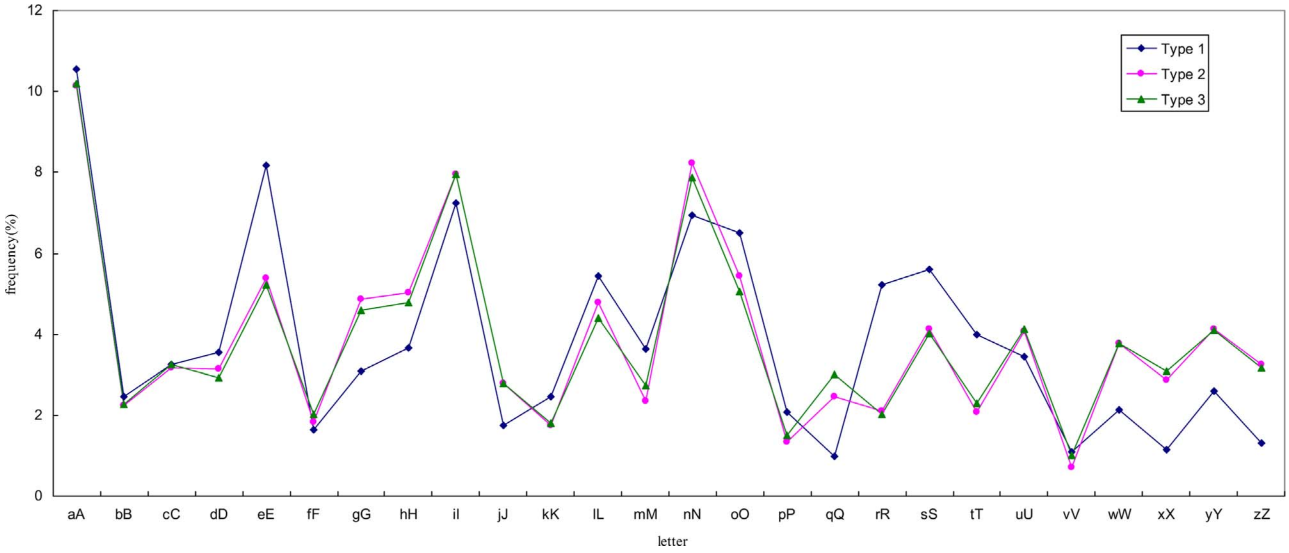

Fig. 2. Comparison of letter frequency in three types.

Table 8
Letter frequency (%).

aA	bB	cC	dD	eE	fF	gG	hH	iI	jJ	kK	lL	mM
10.2071	2.2651	3.2451	2.9323	5.2307	2.0155	4.6035	4.7751	7.9679	2.7754	1.8089	4.3903	2.7229
nN	oO	pP	qQ	rR	sS	tT	uU	vV	wW	xX	yY	zZ
7.8796	5.0684	1.5117	3.0133	2.0209	4.0114	2.2919	4.1265	1.0137	3.7652	3.0993	4.0877	3.1706

Table 9
The numbers of passwords that contain special symbols.

types	#pass[a]	#pass(s)[b]	Percentage (%)
Type 1	11,203,094	387,698	3.4606
Type 2	6,427,428	229,365	3.5685
Type 3	124,224,318	3,159,224	2.5432

[a] The numbers of passwords in the type.
[b] The numbers of passwords that contain special symbols.

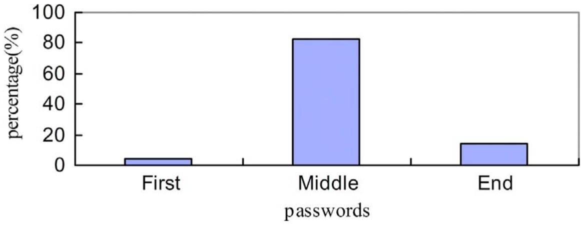

Fig. 3. Position of special symbol.

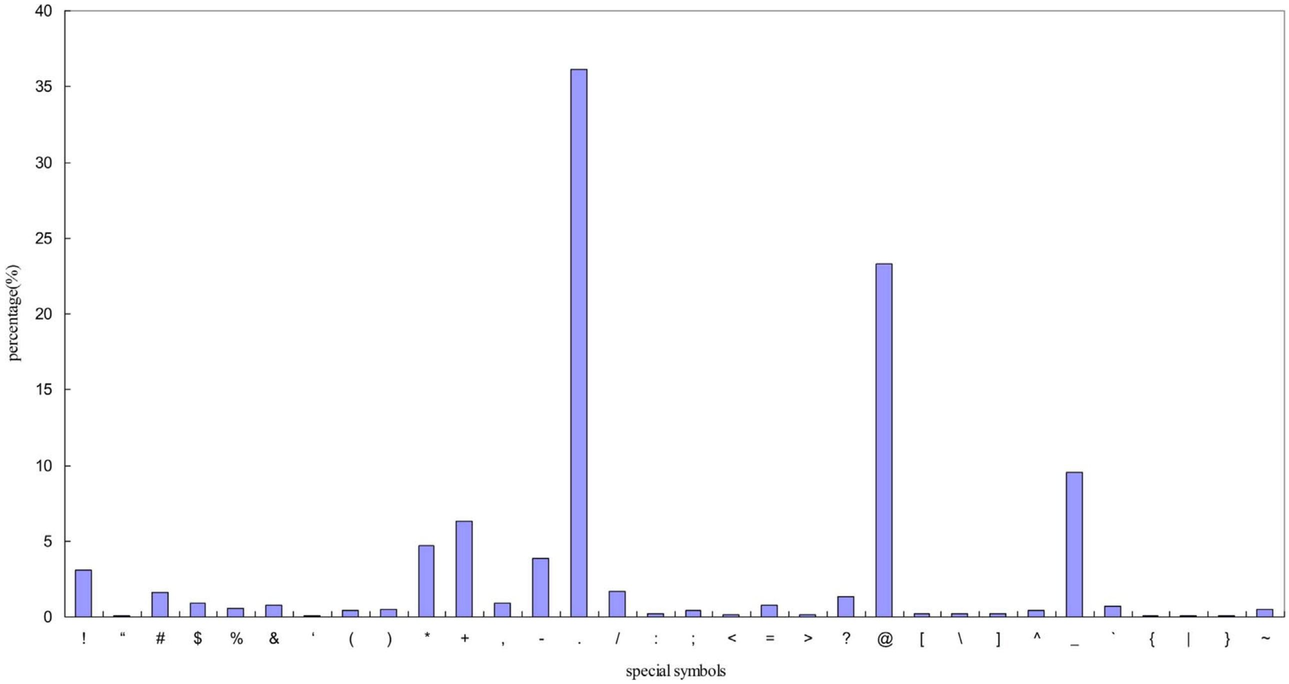

Fig. 4. Special symbol distribution.

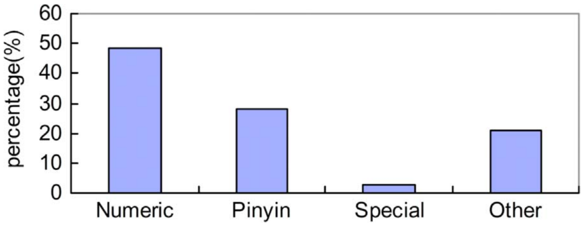

Fig. 5. Four groups.

form of the concatenation of name (as in the class (4)) and other digits, e.g., lixx0 and 0lixx.

- Group 3: passwords that require special symbols ! ” # $ % & ’ () * +, -. /:; < = > ? @ [\] ^ _ ‘ {|}˜.
- Group 4: other passwords, i.e., not in above three groups.

For the large scale datasets of over 141 million passwords, we find that over 4% of passwords represent pure Pinyin (the class (1)) and pure names (the class (4)), nearly 5% of passwords represent concatenations of “Pinyin or name” and dates (the class (2) and the class (5)), and those of ‘Pinyin and number’ (the class (3)) and ‘name and number’ (the class (6)) constitute the next 17.1852%. Fig. 6 shows the details. Combining pure numeric pattern (group 1) and the Pinyin pattern (group 2), we see that over 76% passwords can be covered.

We remark that in the group 4, Pinyin may occur with the form of a mixture. Namely, Pinyin is followed by other symbols, e.g., woaini-darling. This pattern may be further discussed in the future.

For the passwords in the pure Pinyin class, we analyze their length distribution and the number of Chinese characters each password is transcribed from, and Figs. 7 and 8 show the results. One can see from Fig. 7 that 6–11 letters are dominant. And according to Fig. 8 over 93% of the users choose the full spelling of Pinyin that correspond to two, three or four Chinese characters as their passwords, which is for the convenience and mnemonics of users.

4.5. Common passwords

As most users know English as a second language, we compare some commonly used passwords. These passwords are Pinyin and English words of the same meaning. And the passwords of English words are extracted from most popular password lists released on the Internet. The purpose is to find out the users' preference of native language to their second language when choosing passwords.

We address that in China, education in English, rises at an unprecedented speed and scale in recent decades. For students of different levels (from primary school pupils to doctoral students of universities), a course in a foreign language is compulsory; and English is compulsory for about 90% of those students. The education of English as a foreign language is affecting the life of nearly one billion Chinese people.

Table 10 tells us that Chinese tend to use their native language as passwords, even though they know English well. To validate this, we can further extract the most popular passwords in our data sources. Table 11 illustrates these passwords.

Several recent leaks of large password datasets have revealed that certain popular choices, such as 123456, are exceedingly common. The

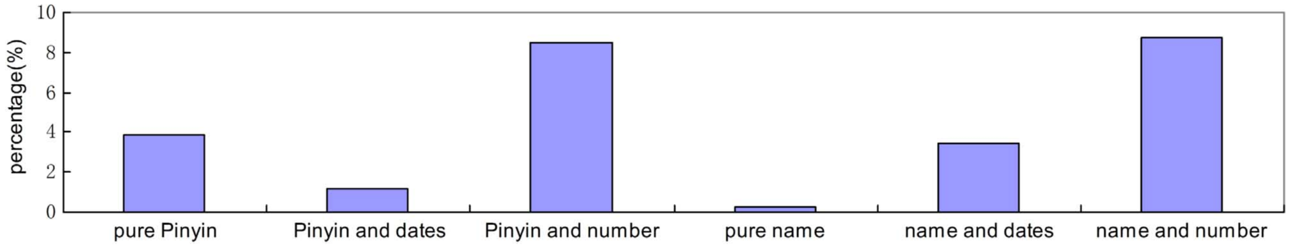

Fig. 6. Group 2-Pinyin.

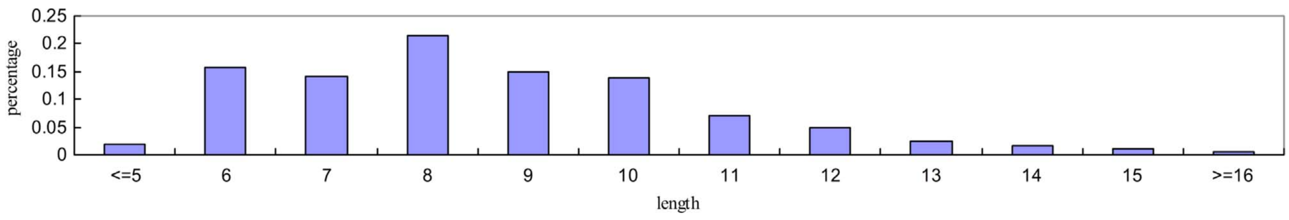

Fig. 7. Length distribution in the class of pure Pinyin.

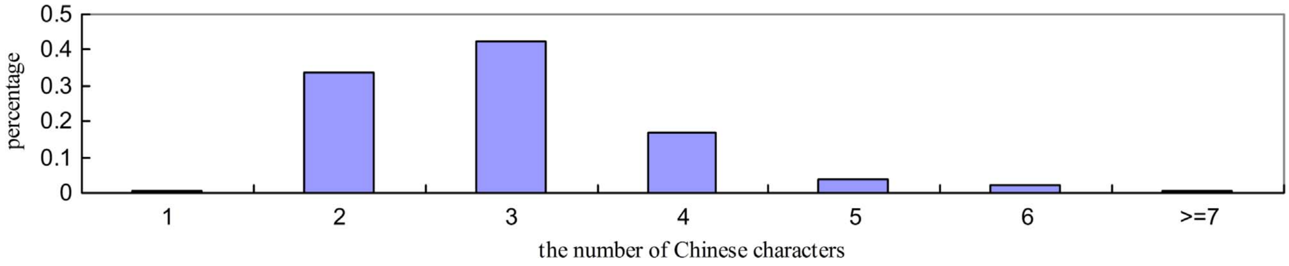

Fig. 8. The number of Chinese characters for passwords in the class of pure Pinyin.

Table 10
Comparison of some common passwords.

String	Count of passwords that contain the string
`woaini` vs. `iloveu` + `iloveyou`	372,575 vs. 74,533
`mima` vs. `password`	56,892 vs. 72,255
`nihao` vs. `hello`	46,251 vs. 28,219
`suibian` vs. `whatever`	3690 vs. 2391
`nicai` vs. `guess`	2479 vs. 1427

Table 11
Top 25 most commonly used and worst passwords.

No.	Password	Count	Percentage (%)	Password[a]
1	123456	4,013,525	2.82932	123456
2	123456789	967,961	0.68236	password
3	111111	939,597	0.66237	12345678
4	123123	459,618	0.32401	qwerty
5	12345678	390,417	0.27522	abc123
6	000000	389,161	0.27434	123456789
7	5201314	275,253	0.19404	111111
8	0	252,437	0.17795	1234567
9	123321	178,683	0.12596	iloveyou
10	11111111	153,118	0.10794	adobe123
11	a123456	149,510	0.10540	123123
12	111222tianya	148,414	0.10462	admin
13	666666	132,828	0.09364	1234567890
14	1234567	122,760	0.08654	letmein
15	888888	115,846	0.08167	photoshop
16	123	110,343	0.07778	1234
17	1314520	106,125	0.07481	monkey
18	7758521	104,825	0.073896	shadow
19	654321	97,071	0.068430	sunshine
20	woaini	96,459	0.067998	12345
21	1234567890	95,972	0.067655	password1
22	121212	86,230	0.060787	priness
23	112233	83,319	0.058735	azerty
24	88888888	82,208	0.05795	trustno1
25	123123123	72,479	0.05109	000000

[a] From top 25 most commonly used and worst passwords of 2013 [34].

last column of Table 11 lists top 25 most commonly used and worst passwords of 2013, released by SplashData, a company that makes password management and productivity apps [34]. The Adobe hackers and their 38 million victims explain why the list includes `adobe123` and `photoshop`. The same happens to the password `111222tianya` and `dearbook` (the 12th and 46th on our list, respectively) that show features specific to the website or service. This offers a good reminder not to base the password on the name of the website or application being accessing. Some sites, for example Twitter, have noticed this and implement banned password lists, which includes many of the more common passwords, including the name of the site.

For space saving, we only list top 25 passwords in Table 11. The comparison in Table 11 shows much difference between the common lists of Chinese users and Western users. There are many English word(s) on the Western list. For our list, however, there is no English word(s). On the contrary, `woaini` is Pinyin, and there are another 3 passwords related to Chinese language: `5201314`, `1314520` and `7758521` have the same sound with some Chinese characters when read in Chinese (the first two mean “i love you forever”, and the last means “kiss me and i love you”).

When looking at the top 50 most commonly used passwords in our data sources, we only find one password that is composed of English word(s), i.e., `password` (count: 58,071).

5. Conclusion

5.1. Our outputs

Understanding semantic patterns in passwords can help us to better understand how people choose their passwords, which can help inform usable password policies and password creation guidelines. If we discover semantic patterns that do not lead to security vulnerabilities, they can be used as the building blocks of successful and usable new password guidelines and policies. On the other hand, if any of these semantic patterns do lead to security vulnerabilities, they are still useful as they can be used to help us build stronger and more appropriate password blacklists and proactive checks.

Although semantic patterns are generally difficult to recognize computationally, we make an interesting step towards aiding discovery of interesting semantic patterns in user choice by analyzing Pinyin structures in passwords. These semantic patterns have security implications – most notably, they enable the creation of language-independent password guessing dictionaries, which require no a priori knowledge of the users. We remark that given a dictionary of Pinyin corresponding to one and two Chinese characters, for which the maximal length of the Pinyin can be up to 12, and the file size of the dictionary is only 1395 KB. These dictionaries could be successful in an offline attack or against systems that do not implement account lockout policies. The Pinyin analysis of passwords helps uncover patterns that can improve techniques for cracking passwords.

5.2. End user's selection of password

End users are often the weakest link in the chain of information security, and there is no surprise that end users are often the first point of call to be exploited by attackers. As showed by our outputs in the work, end users should select strong password to secure their sensitive information. The easier a password is for the owner to remember generally means it will be easier for an attacker to guess. Passwords based on thinking of a phrase and taking the first letter of each word are believed to be as memorable as naively selected passwords, and as hard to crack as randomly generated passwords. Combining two or more unrelated words, or having a personally designed algorithm for generating obscure passwords, are also candidate good methods.

For the authentication service, many organizations specify a password policy that sets requirements for the composition and usage of passwords, typically dictating minimum length, required categories (e.g. upper and lower cases, numbers, and special characters), and prohibited elements (e.g., own name, date of birth, address, and telephone number). This would be helpful for the end users in selecting strong passwords.

Acknowledgments

The work was supported by the National Natural Science Foundation of China Grant (Nos. 61472249, 61572192, 61572149, and 61571191), Science and Technology on Communication Security Laboratory (9140C110203140C11049), and International Science & Technology Cooperation & Exchange Projects of Shaanxi Province (2016KW-038).

References

[1] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, Jonathan M. Smith. Smudge attacks on smartphone touch screens, in: Proceedings of the 4th USENIX Conference on Offensive Technologies, WOOT'10, ACM, New York, 2010, pp. 1–7.

[2] Baidu ⟨http://baike.baidu.com/view/7167245.htm⟩.

[3] J. Bonneau, The Gawker Hack: How a Million Passwords Were Lost, Light Blue Touchpaper Blog ⟨http://www.lightbluetouchpaper.org/2010/12/15/the-gawker-hack-how-a-million-passwordswere-lost/⟩, December 2010.

[4] J. Bonneau, The science of guessing: analyzing an anonymized corpus of 70 million passwords, in: Proceedings of IEEE Symposium on Security and Privacy, IEEE Security and Privacy'12, 2012, pp. 538–552.

[5] J. Bonneau, S. Preibusch, A birthday present every eleven wallets? The security of customer-chosen banking pins, in: Proceedings of the International Conference on Financial Cryptography, FC'12, Springer, Berlin, 2012, pp 25–40.

[6] P. Bright, Sony Hacked Yet Again, Plaintext Passwords, E-mails, DOB posted, Ars Technica ⟨http://arstechnica.com/techpolicy/2011/06/sonyhacked-yet-again-plaintext-passwords-posted/⟩, June 2011.

[7] CNCERT ⟨http://www.cert.org.cn/publish/main/11/2012/20120330183913861664205/20120330183913861664205.html⟩, 2011.

[8] CNNIC ⟨http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201307/P020130717505343100851.pdf⟩.

[9] Xavier de Carne de Carnavalet, Mohammad Manna, From very weak to very strong: analyzing password-strength meters, in: Proceedings of 2014 Network and Distributed System Security (NDSS) Symposium, NDSS'14, ACM, New York, 2014, pp. 1–16.

[10] Solar Designer. John the Ripper, 1996–Present ⟨http://www.openwall.com/john/⟩.

[11] D. Florencio, C. Herley, A large-scale study of web password habits, in: Proceedings of the 16th International Conference on World Wide Web, WWW'07, ACM, New York, 2007, pp. 657–666.

[12] D. Goodin, Hackers Expose 453,000 Credentials Allegedly Taken From Yahoo Service ⟨http://arstechnica.com/security/2012/07/yahoo-servicehacked/⟩.

[13] C. Herley, P. Van Oorschot, A research agenda acknowledging the persistence of passwords, IEEE Secur. Priv. 10 (2012) 28–36.

[14] M. Jakobsson, M. Dhiman, The benefits of understanding passwords, in: Proceedings of 7th USENIX Workshop on Hot Topics in Security, HotSec'12, USENIX, Washington, 2012, pp. 1–6

[15] M. Kumar, China Software Developer Network (CSDN) 6 Million User Data Leaked, The Hacker News ⟨http://thehackernews.com/2011/12/chinasoftware-developer-network-csdn-6.html⟩, December 2011.

[16] Robert Edward Lewand, Cryptological Mathematics, The Mathematical Association of America, 2000

[17] David Malone, Kevin Maher, Investigating the distribution of password choices, in: WWW'12, Proceedings of the 21st International Conference on World Wide Web, ACM, New York,2012, pp. 301–310.

[18] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, Blase Ur, Measuring password guessability for an entire university, in: Proceedings of 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, ACM, New York, NY, USA, 2013, pp. 173–186.

[19] B.D. Medlin, J.A. Cazier, An empirical investigation: health care employee passwords and their crack times in relationship to hipaa security standards, Int. J. Healthc. Inf. Syst. Inform. 2 (2007) 39–48.

[20] R. Morris, K. Thompson, Password security: a case history, Commun. ACM 22 (1979) 594–597.

[21] M. Weir, S. Aggarwal, M. Collins, H. Stern, Testing metrics for password creation policies by attacking large sets of revealed passwords, in: Proceedings of the ACM Conference on Computer and Communications Security, CCS'10, ACM, New York, 2010, pp. 162–175

[22] NetEase Inc. ⟨http://en.wikipedia.org/wiki/NetEase⟩.

[23] The Chinese Software Developer Network ⟨http://en.wikipedia.org/wiki/CSDN⟩.

[24] Schneier on Security, Information Leakage from Keypads ⟨https://www.schneier.com/blog/archives/2009/07/information_lea_1.html⟩.

[25] Openwall ⟨http://www.openwall.com/wordlists/⟩.

[26] Micka Pavel, Letter Frequency (English) ⟨http://en.algoritmy.net/article/40379/letter-frequency-english⟩.

[27] Pinyin ⟨http://en.wikipedia.org/wiki/pinyin⟩.

[28] S. Ragan, Report: Analysis of the Stratfor Password List ⟨http://www.thetechherald.com/articles/report-analysis-of-the-stratforpassword-list⟩, 2012.

[29] Site Ranking for NetEase (163.com) by Alexa ⟨http://www.alexa.com/siteinfo/163.com⟩.

[30] Site Ranking for Tianya Club (tianya.cn) by Alexa ⟨http://www.alexa.com/siteinfo/tianya.cn⟩.

[31] Rapid7, Linkedin Passwords Lifted ⟨http://www.rapid7.com/resources/infographics/linkedinpasswords-lifted.html⟩.

[32] D.A. Sawyer, The characteristics of user-generated passwords (Ph.D. thesis), 1990.

[33] B. Schneier, Myspace Passwords Aren't So Dumb ⟨http://www.wired.com/politics/security/commentary/securitymatters/2006/12/72300⟩, December 2006.

[34] SplashData, "Password" Unseated by "123456" on Splashdata's Annual "Worst Passwords" List ⟨http://splashdata.com/press/worstpasswords2013.htm⟩.

[35] Hundred Family Surnames ⟨http://en.wikipedia.org/wiki/Hundred_Family_Surnames⟩.

[36] Katsuya Uchida, Password Pattern Selection Methodology: A Simple Security Technique ⟨www.uchidak.com/eng/20111225uchidarandpw.pdf⟩.

[37] A. Vance, If Your Password Is 123456, Just Make It Hackme. New York Times (New York edition), January 21, 2010.

[38] Rafael Veras, Christopher Collins, Julie Thorpe, On the semantic patterns of passwords and their security impact, in: Proceedings of 2014 Network and Distributed System Security (NDSS) Symposium, NDSS'14, ACM, New York, 2014, pp. 1–16

[39] Rafael Veras, Julie Thorpe, Christopher Collins, Visualizing semantics in passwords: the role of dates, in: Proceedings of the Ninth International Symposium on Visualization for Cyber Security, VizSec'12, ACM, New York, 2012, pp. 88–95

[40] Emanuel von Zezschwitz, Anton Koslow, Alexander De Luca, Heinrich Hussmann, Making graphic-based authentication secure against smudge attacks, in: IUI'13, March 19–22, 2013, Santa Monica, CA, USA, ACM, New York, 2013, pp. 277–286

[41] Melissa Walters, Tampa Erika Matulich, Assessing password threats: implications for formulating university password policies, J. Technol. Res. 02 (2010) 1–9.

[42] Sina Weibo ⟨http://en.wikipedia.org/wiki/Sina_Weibo⟩.

[43] ZDNet, Chinese Hacker Arrested for Leaking 6 Million Logins ⟨http://www.zdnet.com/blog/security/chinese-hacker-arrested-for-leaking-6-million-logins/11064⟩.