The Emperor's New Password Creation Policies:

An Evaluation of Leading Web Services and the Effect of Role in Resisting Against Online Guessing*

Ding Wang and Ping Wang

School of EECS, Peking University, Beijing 100871, China
wangdingg@mail.nankai.edu.cn; pwang@pku.edu.cn

Abstract. While much has changed in Internet security over the past decades, textual passwords remain as the dominant method to secure user web accounts and they are proliferating in nearly every new web services. Nearly every web services, no matter new or aged, now enforce some form of password creation policy. In this work, we conduct an extensive empirical study of 50 password creation policies that are currently imposed on high-profile web services, including 20 policies mainly from US and 30 ones from mainland China. We observe that no two sites enforce the same password creation policy, there is little rationale under their choices of policies when changing policies, and Chinese sites generally enforce more lenient policies than their English counterparts.

We proceed to investigate the effectiveness of these 50 policies in resisting against the primary threat to password accounts (i.e. online guessing) by testing each policy against two types of weak passwords which represent two types of online guessing. Our results show that among the total 800 test instances, 541 ones are accepted: 218 ones come from trawling online guessing attempts and 323 ones come from targeted online guessing attempts. This implies that, currently, the policies enforced in leading sites largely fail to serve their purposes, especially vulnerable to targeted online guessing attacks.

Keywords: User authentication, Password creation policy, Password cracking, Online trawling guessing, Online targeted guessing.

1 Introduction

Textual passwords are perhaps the most prevalent mechanism for access control in a broad spectrum of today's web services, ranging from low value news portals and ftp transfers, moderate value social communities, gaming forums and emails to extremely sensitive financial transactions and genomic data protection [27]. Though its weaknesses (e.g., vulnerable to online and offline guessing [42]) have been articulated as early as about forty years ago and various alternative authentication schemes (e.g., multi-factor authentication protocols [26, 52] and graphical passwords [56]) have been successively suggested, password-based authentication firmly stays as the dominant form of user authentication over the

* This is the full version of our paper that is to be appeared in Proc. of 20th European Symposium on Research in Computer Security (ESORICS 2015), Vienna, Austria, Sept. 21-25, 2015.

Internet. Due to both economical and technical reasons [25], it will probably still take the lead on web authentication in the foreseeable future.

It has long been recognised that system-assigned passwords are hardly usable [1, 5], yet when users are allowed to select passwords by themselves, they tend to prefer passwords that are easily memorable, short strings but not arbitrarily long, random character sequences, rendering the accounts protected by user-generated passwords at high risk of compromise [6, 17, 54]. It is a rare bit of good news from recent password studies [16, 47, 50] that, if properly designed, password creation policies do help user select memorable yet secure passwords, alleviating this usability-security tension. Unsurprisingly, nearly every web service, no matter new or aged, follows the fashion and now enforces some form of password creation policy. Generally, a password creation policy[1] is composed of *some password composition rules* and *a password strength meter* (see Fig. 1). The former requires user-generated passwords to be satisfied with some complexity (e.g., a combination of both letters and numbers) and nudges users towards selecting strong passwords [10, 39], while the latter provides users with a visual (or verbal) feedback [16, 50] about the password strength during registration.

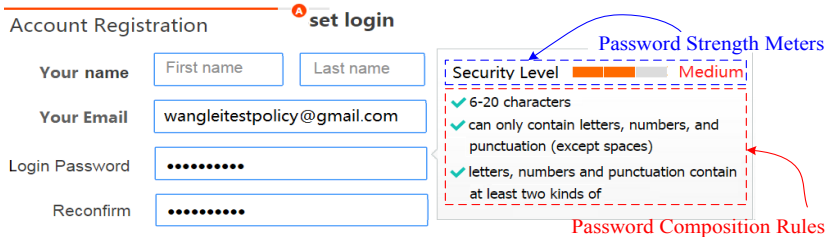

Fig. 1. A typical example of password creation policy

However, to what extent can the widely-deployed password creation policies on the Internet be relied upon has long been an open issue. In 2007, Furnell [19] initiated an investigation into the password practices on 10 popular websites and found that, password rules and meters are vastly variable among the examined sites and none of them can perform ideally across all of the evaluated criteria.

In 2010, Bonneau and Preibush [8] conducted the first large-scale empirical study of password policy implementation issues in practice. By examining 150 different websites, they observed that bad password practices were commonplace and particularly, highly inconsistent policies were adopted by individual sites, which suggests that there is a lack of widely accepted industry standards for password implementations. At the meantime, Florêncio and Herley [18] investigated the rationale underlying the choices of password policies among 75 high-profile websites and found that, greater security demands (e.g., the site scale, the value protected and the level of severity of security threats) generally do not constitute the dominant factor for selecting more stringent password rules. Instead, these Internet-scale, high value web services (e.g., e-commerce sites like Paypal and online banking sites like Citibank) accept relatively weak passwords

[1] We use "password policy" and "password creation policy" interchangeably, and don't consider other password policies like storage [4], expiration [12] and recovery [45].

and these sites bearing no consequences from poor usability (e.g., government and university sites) usually implement restrictive password rules.

To figure out whether leading websites are improving their password management policies as time goes on, in 2011 Furnell [20] made an investigation into 10 worldwide top-ranking sites and compared the results with those of the study [19] he performed in 2007. Disappointingly, he reported that, during the four-year intervening period there has been hardly any improvement in password practices while the number of web services and security breaches has increased greatly. In 2014, Carnavalet and Mannan [11] investigated the problem of to what extent the currently deployed password strength meters are lack of sound design choices and consistent strength outcomes. They systematically evaluated 13 meters from 11 high-profile web services by testing about 4 million passwords that are leaked from popular online services as well as specifically composed passwords. It is found that most meters in their study are "quite simplistic in nature and apparently designed in an ad-hoc manner, and bear no indication of any serious efforts from these service providers" [11]. Fortunately, most meters can correctly assign sensible scores to highly weak popular passwords, e.g., at least 98.4% of the top 500 passwords [9], such as `password`, `123456`, `iloveyou` and `qwerty`, are considered "weak" or "very weak" by every meter.

Motivations. However, most of the existing works [8, 18–20] were conducted five years ago, while the online world has evolved rapidly during the intervening period. In early 2010, Twitter had 26 million monthly active Users, now this figure has increased tenfold;[2] In Nov. 2010, Gmail had 193 million active users, now this figure reaches 500 million;[3] In April 2010, Xiaomi, a privately owned smartphone company headquartered in Beijing, China, just started up, now it has become the world's 3rd largest smartphone maker (ranked after Apple and Samsung) and there are 100 million Xiaomi users worldwide who rely on its cloud service.[4] All these three sites have recently been the victims of hacking and leaked large amounts of user credentials [37, 40, 43]. As we will demonstrate, they all (as well eight other sites examined in this work) have changed their policies at least once during the past five years. Moreover, at that time how to accurately measure password strength was an open problem and there were few real-life password datasets publicly available, and thus the methodologies used in these earlier works are far from systematic (mature) and satisfactory.

The sole recent work by Carnavalet and Mannan [11] mainly focuses on examining password meters from 13 sites, paying little attention to the other part of password policies (i.e., password composition rules). Due to the fact that a password (e.g., `Wanglei123`) measured "strong" by the password meter of a site (e.g., AOL) may violate the password rule of this site, finally it is still rejected by the site. In addition, many sites (e.g., Edas, AOL and Sohu) enforce *mandatory* password rules but *suggestive* meters, a password metered "weak" might pass the password rule of these sites, and finally this "weak" password is still accepted.

[2] `http://www.statista.com/statistics/282087/`
[3] `http://thefusejoplin.com/2015/01/choose-google-gmail-yahoo-mail/`
[4] `https://www.techinasia.com/xiaomi-miui-100-million-users/`

Consequently, the question of how well these sites *actually* reject weak passwords and withstand online guessing remains unanswered.

Another limitation of existing works is that little attention has been given to non-English web services. As typical hieroglyphics, Chinese has been the main language used in a total of over 3.64 million web services until 2014 and about 0.95 million new web services that started up in 2014 (which means 0.95M new password policies come out and impact on common users.) [24]. What's more, Chinese web users, who have reached 649 million by the end of 2014 [13], have been the largest Internet population in the world and account for a quarter of the world's total netizens. Therefore, it is important (and interesting) to investigate what's the strengths and weaknesses of the current password policies in Chinese web services as compared to their English counterparts.

Our contributions. The main contributions of this work are as follows:

(1) First, we propose a systematic, evidence-grounded methodology for measuring password creation policies and investigate the status quo of policies enforced by 50 leading web services (with special emphasis on Chinese web services) with a total of ten application domains. We find that, generally, gaming sites, email sites, e-commerce sites and non-profit organizations manage with the least restrictive password rules, while the sites of IT manufacturers impose the most stringent ones; Web portals, email sites, e-commerce sites and technical forums tend to provide explicit feedbacks of the password strength to users, while sites of security companies, IT manufacturers and academic services, ironically, often do not bother to provide users with any piece of information about password strength.

(2) Second, we explore the differences in password policy choices between English sites and Chinese sites. Compared to their English counterparts, Chinese sites, in general, are more undaunted (audacious) in their password rule choices, while there is no significant difference between these two groups of sites with regard to the password meter choices.

(3) Third, we employ state-of-the-art password cracking techniques (including the probabilistic-context-free-grammar (PCFG) based and Markov-Chain-based) to measure the strength of the 16 testing passwords that are used to represent two primary types of online password guessing attempts. This provides a reliable benchmark (ordering) of the actual strength of these testing passwords beyond intuitive (heuristic) estimates as opposed to previous works like [11, 20]. We observe that most of the meters overestimate the strength of at least some of these 16 passwords, rendering the corresponding web services vulnerable to online guessing.

The structure of this paper is as follows: Our methodology is elaborated in Sec. 2; Our results are presented in Sec. 3. The conclusion is drawn in Sec. 4.

2 Our methodology

As there is little research on studying password practices and the approaches used in the few pioneering works [8, 11, 18, 20] are far from systematic and may

be demoded over the past five years, in the following we take advantage of state-of-the-art techniques and elaborate on a systematic methodology for measuring password policies. As far as we know, for the first time several new approaches (e.g., the use of large-scale real-life passwords as corroborative evidence, the use of targeted online guessing to measure password strength, and the classification and selection of testing passwords) are introduced into this domain.

2.1 Selecting representative sites

To investigate the status quo of password creation policies deployed in today's Internet (with special emphasis on Chinese web services), first of all we selected ten themes of web services that we are most interested in and that are also highly relevant to our daily online lives: web portal, IT corporation, email, security corporation, e-commerce, gaming, technical forum, social forum, academic service and non-profit organization. Then, for each theme we choose its top 5 sites according to the Alexa Global Top 500 sites list based on their traffic ranking (`http://www.alexa.com/topsites`). Some companies (e.g., Microsoft and Google) may offer various services (e.g., email, search, news, product support) and have a few affiliated sites, fortunately they generally rely on the same authentication system (e.g., Windows Live and Google Account) to manage all consumer credentials and we can consider all the affiliated sites as one. Similarly, for each theme we also choose its top 10 sites that are among the Alexa Top 500 Chinese sites rank list. In this way, there are 15 leading sites selected for each theme: 5 from English sites and 10 from Chinese sites. Further, we randomly selected 5 sites out of these 15 sites for each theme, resulting in 50 sites used in this work (see Table 5): 20 from English sites and 30 from Chinese sites.

We note that though our selected websites have a wide coverage, yet many other themes are still left unexplored, such as e-banking, e-health and e-government. The primary reason why we does not include them is that, they rely heavily on multi-factor authentication techniques in which passwords play a much less critical role. In addition, the number of sites allocated for each theme is also limited. Nonetheless, our sample characterizes the current most recognised and leading portion of the online web services, which attract the majority of the visit traffic [28, 31]. Therefore, the password practices used by these sites will impact on the major fraction of end-users and may also became a model for other less leading sites (which generally are with less technical, capital and human resources). Further considering the amount of work incurred for one site, an inspection of 50 sites is really not an easy task, let alone an initial study like ours (as there is no sophisticated procedure to follow, we have to carry out an iterative process of data collection). In the future work, we are considering to increase the number of sites for each theme to 10 or possibly 20, and the investigation results as well as a set of evidence-supported, practicable policy recommendations will be made available at the companion site `http://wangdingg.weebly.com/password-policy.html`.

2.2 Measuring password policy strength

The task of measuring strength of *a policy* is generally accomplished by evaluating strength of *the password dataset* generated under this policy, and a number

of methods for tackling the latter issue have been proposed, including statistical-based ones (e.g., guessing entropy and α-guesswork [6]) and cracking-based ones (e.g., [34, 53]). However, these methods all require access to a real password dataset with sufficient size. Fortunately, we note that Florêncio and Herley [18]'s simple metric —N_{min}·$\log_2 C_{min}$— is not subject to this restriction and sufficient for our purpose, where N_{min} is the minimum length allowed and C_{min} is the cardinality of the minimum charset imposed.[5] For instance, the strength of a policy that requires a user's password to be no short than 6 and must contain a letter and a number is 31.02($=6 \cdot \log_2 36$) bits. This metric well characterizes the minimum strength of passwords allowed by the policy, providing a lower bound of the policy strength. We adopt this metric in our work.

Table 1. Basic information about the seven password datasets used in this work

Dataset	Services	Location	Language	When leaked	How leaked	Total passwords
Rockyou	Social	USA	Englsih	Dec. 14, 2009	SQL injection	32,603,387
Tianya	Social	China	Chinese	Dec. 4, 2011	Hacker breached	30,233,633
7k7k	Gaming	China	Chinese	Dec. 2, 2011	Hacker breached	19,138,452
Dodonew	Ecommerce	China	Chinese	Dec. 3, 2011	Hacker breached	16,231,271
CSDN	Programming	China	Chinese	Dec. 2, 2011	Hacker breached	6,428,287
Duowan	Gaming	China	Chinese	Dec. 1, 2011	Insider disclosed	4,982,740
Yahoo	Portal	USA	English	July 12, 2012	SQL injection	453,491

2.3 Exploiting real-life password datasets

Our work relies on seven password datasets, a total of 124.9 million real-life passwords (see Table 1), to train the cracking algorithms and learn some basic statistics about user password behaviors in practice. Five datasets of Chinese web passwords, namely Tianya (31.7 million), 7k7k (19.1 million), Dodonew (16.3 million), Duowan (8.3 million) and CSDN (6.4 million), were all leaked during Dec. 2011 in a series of security breaches [36]. Tianya is the largest social forum in China, 7k7k, Dodonew and Duowan are all popular gaming forums in China, and CSDN is a well-known technical forum for Chinese programmers.

Two datasets of English web passwords, namely Rockyou (32.6 million) and Yahoo (0.5 million), were among the most famous datasets in password research [35, 53]. Rockyou is one of the world's largest in-game video and platform for premium brands located in US, and its passwords were disclosed by a hacker using a SQL injection in Dec. 2009 [3]. This dataset is the first source of large-scale real-life passwords that are publicly available. Yahoo is one of the most popular sites in the world known for its Web portal, search engine and related services like Yahoo Mail, Yahoo News and Yahoo Finance. It attracts "more than half a billion consumers every month in more than 30 languages". Its passwords were hacked by the hacker group named D33Ds in July 2012 [55]. We will pay special attention to this site because it has changed its password policy, as far as we can confirm, at least three times during the past five years.

2.4 Measuring password strength

Essentially, the strength of a password is its guessing resistance against the *assumed* attacker. This equals the uncertainty this attacker has to get rid of, and naturally the idea of shannon entropy was suggested to measure password

[5] This implicitly assumes that users are least-effort ones.

strength, called NIST entropy [10]. Later, NIST entropy was found to correlates poorly with guess resistance and can at best serve as a "rough rule of thumb" [34, 53]. In contrast, the *guess-number* metric, which is based on password cracking algorithms (e.g., PCFG-based and Markov-based [35]), was shown to be much more effective, and it has been used in a number of following works like [38, 47].

However, we note that the traditional use of guess-number metric generally implicitly assumes that the attacker is a random, trawling attacker $\mathcal{A}_{tra}$ (i.e., not targeting a selected user). In many cases this is apparently not realistic. For a targeted attacker $\mathcal{A}_{tar}$, with the knowledge of the name of the target user, she can drastically reduce the guess number required to find the right password. In this work, we consider these two kinds of attacker and suppose that the targeted attacker know of the user's name. This assumption is reasonable because, for $\mathcal{A}_{tar}$ to launch a targeted attack, he must know some specific information about the victim user $\mathcal{U}_v$, and $\mathcal{U}_v$'s name is no-doubt the most publicly available data.

To take advantage of name information in cracking, we slightly modify the PCFG-based and Markov-based algorithms by specially increasing the probability of the name-related letter segments. This can be easily achieved in PCFG-based attacks [35]. For instance, assuming the victim's name is "wanglei", after the PCFG-based training phase, one can increase the probability of the item "$L_4 \rightarrow wang$" in the PCFG grammars to that of the most popular L_4 segment and similarly, the item "$L_7 \rightarrow wanglei$" to that of the most popular L_7 segment.

Algorithm 1: Our Markov-Chain-based generation of targeted guesses

Input: A training set $\mathcal{TS}$; A name list *nameList*; The victim user's name *victimName*; The size k of the guess list to be generated (e.g., $k = 10^7$)
Output: A guess list L with the k highest ranked items

```
Pre-Training:
    for name ∈ nameList do
        trieTree.insert(name)
    for password ∈ TS do
        for letterSegment ∈ splitToLetterSegments(password) do
            if InTrieTree(letterSegment) then
                if isFullName(letterSegment) then
                    password.replace(letterSegment, victimName.fullName)
                if isSurName(letterSegment) then
                    password.replace(letterSegment, victimName.surName)
                if isFirstName(letterSegment) then
                    password.replace(letterSegment, victimName.firstName)
Ordinary Markov-Chain-based training on the pre-trained set TS using
Good-Turing smoothing and End-Symbol normalization (see [51]);
Produce a list L with top-k guesses in decreasing order of probability.
```

However, for Markov-based attacks since there is no concrete instantiation of "letter segments" during training, we substitute all the name segments (including full, sur- and first names) in training passwords (we use 2M Duowan passwords and 2M CSDN passwords together as training sets) with the victim's corresponding name segments before training. For instance, "zhangwei0327"

Table 2. Two types of passwords modeling two kinds of guessing attacks ('Guess rank' is the order in which the corresponding attacker will try that guess; '–' = not exist)

User	Password	Guess rank in trawling PCFG	Guess rank in trawling Markov	Guess rank in targeted PCFG	Guess rank in targeted Markov
Type A (Hotspot)	123456	**1**	**1**	**3**	**2**
	123456789	**3**	**2**	**1**	**3**
	5201314	**6**	**8**	**9**	**10**
	woaini	**12**	**19**	**30**	**423**
	iloveyou	**43**	**347**	**24**	**359**
	password	**84**	**164**	**34**	**194**
	woaini1314	737	116	1501	32736
	password123	17002	36834	6572	36679
Type B (Name-based)	wanglei	281	595	**64**	**1**
	wanglei123	13929	35852	**324**	**7**
	wanglei1	42627	86999	**3450**	**16**
	wanglei12	169546	235971	11205	**58**
	Wanglei123	3020809	6222672	**323**	**392**
	wang.lei	301547	7856239	2287915	379205
	wanglei@123	5291970	–	1927185	5109
	Wanglei@123	–	–	1927186	206144

Table 3. Popularity of Type A passwords in real-life password datasets

Hotspot	Tianya (31.7M,2011)		Dodonew (16.3M,2011)		7k7k (19.1M,2011)		Duowan (8.3M,2011)		Rockyou (32.6M,2009)		Yahoo (0.5M,2012)	
Password	Rank	Freq.	Rank	Freq.	Rank	Freq.	Rank	Freq.	Rank	Freq.	Rank	Freq.
123456	1	3.98%	1	1.45%	1	3.79%	1	3.43%	1	0.89%	1	0.38%
123456789	4	0.59%	3	0.32%	4	0.63%	3	0.62%	3	0.24%	6	0.05%
5201314	7	0.19%	5	0.19%	6	0.34%	6	0.28%	415	0.01%	5090	0.00%
woaini	17	0.09%	26	0.04%	15	0.09%	18	0.07%	3626	0.00%	–	0.00%
iloveyou	49	0.04%	106	0.01%	53	0.03%	45	0.03%	5	0.15%	16	0.03%
password	86	0.02%	23	0.04%	98	0.02%	87	0.02%	4	0.18%	2	0.18%
woaini1314	295	0.01%	18	0.05%	72	0.02%	57	0.03%	87348	0.00%	–	0.00%
password123	20045	0.00%	8004	0.00%	22462	0.00%	14382	0.00%	1384	0.00%	153	0.01%

is replaced with "wanglei01", "zhao@123" is replaced with "wang@123", and "pingpku@123" is replaced with "leipku@123", where "wang" and "lei" is U_v's surname and first name in Chinese Pinyin, respectively. Our basic idea is that *the popularity of name-based passwords in the training sets largely reflects the probability of the targeted user to use a name-based password, and the clever attacker $\mathcal{A}_{tar}$ will base on this probability to exploit U_v's name.* Our Markov-based algorithm for targeted online guessing is shown as Algorithm 1. One can easily see that, based on our idea, besides Chinese Pinyin names, this algorithm can be readily extended to incorporate names in any other language (e.g., "James Smith" in English), and to incorporate other user-specific data (such as account name and birthdate) to model a more knowledgeable targeted attacker.

To avoid ambiguity, we only consider name segments no shorter than 4. To determine whether a password picked from the training set includes a name or not, we first build a name-based Trie-tree by using the 20 million hotel reservation data leaked in Dec., 2013 [22]. This name dataset consists of 2.73 million unique Chinese full names and thus is adequate for our purpose. We also add 504 Chinese surnames which are officially recognized in China into the Trie-tree. These surnames are adequate for us to identify the first names of Chinese users in the Trie-tree to be used in PCFG-based targeted guess generation.

Table 4. Popularity of Type B passwords in real-life datasets

Name dictionary	Tianya	Dodonew	7k7k	Duowan	Average Chinese	Rockyou	Yahoo	Average English
Pinyin_surname(len≥4)	6.34%	10.04%	7.14%	8.44%	7.99%	1.38%	1.29%	1.34%
Pinyin_fullname(len≥4)	9.87%	15.90%	11.42%	13.42%	12.65%	5.37%	3.61%	4.49%
Pinyin_name_total(len≥4)	10.91%	18.06%	14.81%	14.92%	14.68%	5.36%	4.21%	4.78%

2.5 Selecting testing passwords

As we have mentioned in Section 2.3, we measure how the 50 password policies we are interested in are resistant to two types of guessing attacker, i.e., a trawling attacker $\mathcal{A}_{tra}$ and a targeted attacker $\mathcal{A}_{tar}$ (with the victim's name). The aim of $\mathcal{A}_{tra}$ is to break *as many accounts as possible* with a few password trials [6], while $\mathcal{A}_{tar}$ intends to break *the single account* of the given victim user $\mathcal{U}_v$.

To be effective, $\mathcal{A}_{tra}$ would try the most popular passwords in decreasing order of probability with regard to *the targeting population*, while $\mathcal{A}_{tar}$ would try the most popular passwords in decreasing order of probability with regard to *the specific user*. As shown in Table 2, we use Type A passwords (we call hotspot passwords) to represent the attempts $\mathcal{A}_{tra}$ will try and Type B passwords (we call Chinese-Pinyin-name-based passwords) to represent the attempts $\mathcal{A}_{tar}$ will try, respectively. As revealed in [51], Chinese web users create a new type of passwords, named "Chinese-style passwords", such as `woaini`, `5201314` and `wanglei123` based on their language. Note that, "wanglei" is not a random string of length 7 but a highly popular Chinese name, among the top-20 list of Chinese full names [49]; "520" sounds as "woaini" in Chinese, equivalent to "i love you" in English; "1314" sounds as "for ever and ever" in Chinese. Thus, both "woaini1314" and "5201314" mean "I love you for ever and ever". Such passwords are extremely popular among Chinese users (see Table 3) and thus are as dangerous as internationally bad passwords like `iloveyou` and `password123`.

In the following we show why these two types of passwords are weak and can really serve as representatives of password attempts that the aforementioned two types of attacker would try. Table 3 reveals that, all the eight Type A passwords are among the top-200 rank list in at least one web services. More specifically, all the Type A passwords (except `woaini1314` and `password123`) are among the top-100 rank list in the four Chinese web services, while `woaini1314` is only slightly less popular (i.e., with a rank 295) in Tianya and English services, and `password123` is comparatively much more popular in English services, i.e., with a rank 153 in Yahoo and a rank 1384 in Rockyou, respectively. Besides popularity, these eight Type A passwords are also different in length, culture (language) and composition of charsets. Therefore, they well represent the characteristics of potential passwords that a trawling attacker $\mathcal{A}_{tra}$ would try.

As stated in Section 2.4, to model a targeted guessing attacker $\mathcal{A}_{tar}$, we mainly focus on the case that $\mathcal{A}_{tar}$ knows of the victim's name. Without loss of much generality, we assume the victim is a Chinese web user, named "wanglei". From Table 4 (and see more data in [51]) we can see that Chinese users really love to include their (Pinyin) names into passwords: an average of 14.68% of Chinese users have this habit. That is, given a targeted user, it is confident to predict that

there is a chance of 14.68% that she includes her name into her password, and $\mathcal{A}_{tar}$ would gain great advantage by making use of this fact. We conservatively deal with the ambiguities during the name matching. For instance, there are some English surnames (e.g., Lina) may coincide with a Chinese full name, and we take no account of such names when processing English datasets. Well, how does a user uses her name, which can be seen as a word, to build a password? There are a dozen of mangling rules to accomplish this aim, and the most popular ones [14,30] include appending digits and/or symbols, capitalizing the first letter, leet etc. This results in our eight Type B passwords. One can see that the guess rank in Markov-based targeted attack (see the last column in Table 2) quite accords with the rank of general user behaviors as surveyed in [14]. This implies the effectiveness of our Markov-based targeted attacking algorithm.

2.6 Collecting data from sites

To obtain first-hand data on password policy practices, we create real accounts on each site, read the html/PHP/Javascript source code of the registration page, and test sample passwords to see the reaction of the meter when available. We note that there are many unexpected behaviors of sites. For example, in some sites (e.g., Edas, Easychar and Yahoo) the descriptions of password policies are not explicitly given (or the information explicitly given are not complete), and additional data about policies can only be extracted from the feedbacks of the server after one have actually clicked the "submit" button. Consequently, for all sites and every password testing instance, we press the "submit" button down and take note of the response to avoid missing anything important.

Initially, considering the great amount of manual workload involved, we attempt to automate the collection of data from each site by using PHP/Python scripts or web spiders. However, we have to abandon this idea mainly due to four reasons: 1) A large portion of sites (38%) prevent automated registration by requiring users to solve CAPTCHA puzzles when registration; 2) 18% sites need to input the verification code received by user's mobile phone to accomplish the registration; 3) 8% sites involve a verification code to be received by the user's email before the user can input the password; 4) Information displayed on each site is highly heterogeneous, as demonstrated in Section 3, no two sites share the same password policy, and thus batch processing hardly works. As a result, the whole data collection process is manually handled. To assure accuracy, every process is conducted at least twice (at intervals of more than one week) and the collected data all has been cross validated by the authors.

3 Our results

In this section, we first present the status quo of the password policies employed in the 50 web services studied, and then examine the effectiveness of these policies in resisting against online guessing attacks. All of the data were collected from these services between the months of Jan. to Feb. in 2015.

3.1 Password composition rules in the wild

For each password composition rule, we investigate the following six common requirements: length limits, charset requirement, whether rules are explicitly s-

tated, whether allowing symbols, whether using a blacklist and whether deterring the use of personal data. The results are summarized in Table 5.

Length limits. All sites but one impose a minimum length limit. 60% sites require passwords to be no shorter than 6, 30% sites require passwords to be no shorter than 8, with the remaining 8% sites raging from 5, 7 to 9. It is interesting to see that, all sites from the IT corporation category enforce a minimum-8 length limit. Is this because that these services care the security of user account more than other services examined? We will explore this question later.

At the meantime, 72% sites impose a maximum length limit no larger than 64, as far as they can be identified. Surprisingly, 22% sites do not allow passwords to be longer than 16. As it is cognitively impossible for common users to remember complex non-linguistic strings yet attack vectors are increasing, passphrases have recently received much interest and shown to be more useable than passwords [29,33,48], and actually, they have been used successfully and gain popularity (see `http://correcthorsebatterystaple.net/`). However, passphrases are highly likely to exceed such maximum length limits and thus are prohibited.

Further considering that, increasing the password length is generally more effective in enhancing password security than extending the charsets [23, 44], it is more advisable to set a maximum length limit that is large enough (e.g., 64).

Charset requirement. Among the 50 sites, 23 sites (46%) implement some charset requirements. Once again, all sites from the IT corp. category enforce a charset requirement, while other categories do not show this feature. Remarkably, 3 Chinese sites require that a digit-only password cannot be shorter than some minimum length (e.g., 9). This may be due to their insight into the fact that Chinese users highly love to use digit-only passwords—according to one of our earlier works [51], an average of 52.93% Chinese users use digit-only passwords.

Symbol acceptance. It is perhaps surprising to note that four sites (including both English and Chinese sites) prevent symbols to be included into passwords. Theoretically, among the 95 printable ASCII characters, 33 ones are symbols, excluding which would largely reduce an attacker's search space. It has also been established empirically that passwords with symbol(s) are generally much secure than passwords with no symbol [38, 53]. The only plausible reason for forbidding symbols that we can imagine is to prevent SQL injection attacks, yet such attacks can be well prevented by properly handling the escape characters. It is really beyond comprehension why these four sites forbid symbols.

Using blacklist. As recommended in NIST-800-63 [10], a blacklist of sufficient size (e.g., at least 50,000) is highly desirable in prevent popular passwords which are particularly vulnerable to statistical attacks [46]. US-CERT also suggest the use of blacklist [39]. However, only 16 sites (32%) impose a blacklist and none of their blacklists are adequate. For instance, the blacklist of Twitter only contains 370 bad passwords and ironically, the blacklist of IEEE only consists of the famous "`password`". Also note that, all email sites impose a blacklist; 33% Chinese services impose a blacklist, and this figure for English services is 30%.

Checking user info. As highlighted in both NIST-800-63 [10] and NIST-800-118 [44], uses tend to use their personal data (e.g., account name and personal

Table 5. An overview of the password composition rules in the selected web services ('-' means a length limit of larger than 64; '∅' means no charset requirement; 'Blacklist' means a list of banned popular passwords or structures (e.g., repetition); 'User info' considers two types of a user's personal information, i.e. name and account name.)

Web Services		Len. limits		Charset Requirement	Rules Explicit	Accept Symbol	Using blacklist	Checking user info
		Min	Max					
Web Portal	Sina	6	16	∅	Yes	Yes	No	No
	China.com	6	-	1^+lower,1^+upper,1^+digit	No[a]	Yes	No	No
	Tecent	6	16	Not a number with len<9	Yes	Yes	No	No
	Ifeng	6	20	∅	Yes	Yes	No	Account
	Yahoo	7	30	∅	No	Yes	No	Both[b]
IT Corp.	Microsoft	8	16	Any 2 charsets	Yes	Yes	No	Both
	Intel	8	15	1^+letter,1^+digit,1^+symbol	Yes	Yes	No	Account
	Apple	8	32	1^+lower,1^+upper,1^+digit	Yes	Yes	No	Account
	Lenovo	8	20	Any 2 of letter,digit,symbol	No	Yes	No	No
	Huawei	8	60	1^+letter,1^+digit,1^+symbol	Yes	Yes	No	Account
Email	139	6	16	Not a number with len<8	No[a]	Yes[c]	Yes	No
	163	6	16	∅	Yes	Yes	Yes	Account
	AOL	8	16	∅	Yes	Yes	Yes	Both
	Sohu	6	16	∅	Yes	Yes	Yes	No
	Gmail	8	-	∅	Yes	Yes	Yes	Both
Security Corp.	Rsing	6	-	∅	Yes	Yes	No	NO
	Symantec	8	25	1^+letter,1^+digit	Yes	Yes	No	Account
	Kaspersky	6	16	∅	Yes	Yes	No	NO
	McAfee	8	32	1^+letter,1^+digit,no symbol	Yes	No	No	No
	360	6	20	∅	Yes	Yes	Yes	No
Ecommerce	Taobao	6	20	Any 2 of letter,digit,symbol	Yes	Yes	No	Account
	Jd.com	6	20	∅	Yes	Yes	Yes	Account
	Dangdang	6	20	∅	Yes	Yes	No	No
	Amazon	6	-	∅	Yes	Yes	No	No
	Meituan	6	32	∅	Yes	Yes	No	No
Gaming	17173	6	20	Not digits only	No[a]	Yes	Yes	No
	Duowan	8	20	Not a number with len<9	Yes	Yes	No	No
	4399.com	6	20	∅	Yes	Yes	No	No
	Sdo.com	6	30	Only letter and digit	Yes	No	Yes	No
	Wanmei	6	16	Only letter and digit	Yes	No	No	No
Technical Forum	CSDN	6	20	∅	Yes	Yes	No	No
	51CTO	8	20	∅	Yes	Yes	No	No
	ChinaUnix	6	24	Any 2 of letter,digit,symbol[d]	Yes	Yes	No	Account
	Hack80	9	-	∅	Yes	Yes	No	No
	Pediy.com	5	-	∅	No[e]	Yes	Yes	No
Social Forum	Tianya	6	-	1^+letter,1^+digit	Yes	Yes	Yes	No
	BBS.xiaomi	8	16	Any 2 of letter,digit,symbol	Yes	Yes	No	No
	Renren	6	20	∅	Yes	Yes	No	No
	Twitter	6	-	∅	Yes	Yes	Yes	Account
	Facebook	6	-	∅	Yes	Yes	Yes	No
Academic Service	WoS	8	-	1^+letter,1^+digit,1^+special	Yes	Yes	No	No
	CNKI	6	20	No symbol(except '_')	Yes	No	Yes	No
	Cjc.ac.cn	1	-	∅	Yes	Yes	No	No
	Easychair	6	40	Not digits only	No	Yes	No	No
	Edas	7	-	1^+letter,1^+digit	No	Yes	Yes	No
Non-profit Org.	IEEE	8	64	1^+digit	Yes	Yes	Yes[f]	No
	ACM	6	26	∅	Yes	Yes	No	No
	W3C	8	-	∅	Yes	Yes	No	No
	CCF	6	32	∅	No	Yes	No	No
	CACR	6	-	∅	Yes	Yes	No	No

[a] China.com, 139 and 17173 only explicitly require that password must be no shorter than 6, yet when one submits a password that do not fulfill the charset requirement, they prompt that more type(s) of character(s) is(are) needed.

[b] Yahoo checks whether user's personal name are incorporated in the password yet it is case sensitive, e.g., "wanglei123" will not be blocked if we input the surname 'Wang' instead of 'wang'.

[c] 139 only accepts six kinds of symbols (i.e., `_~@#$^`).

[d] ChinaUnix explicitly states that a password must contain two types of characters, yet it accepts passwords (e.g., "123456789" and "qwertasdfg") that are measured as "medium" or "strong".

[e] There is no explicit rule in Pediy.com, yet when one submits a password shorter than 5, it prompts that an accepted password must be no shorter than 5.

[f] IEEE's blacklist only includes one item (i.e., "password"), which is explicitly stated.

name) to build passwords for better memorization, and accordingly, preventing the use of personal data in a password can raise the min-entropy of this password. However, only 14 sites (28%) disallow account name and/or personal name to be included into passwords. Among these 14 sites, 9 come from English sites.

Explicit rules. Despite the long-standing use of and familiarity with passwords, good password practices have not become "an established part of our security culture", and "even basic provision of guidance can help to deliver a tangible improvement" [21]. Consequently, it is crucial for sites to provide users with explicit advice and guidance, otherwise the implicit rules would only provide users with frustration and fatigue. However, there are still 9 sites (including 3 sites from English sites and 6 from Chinese sites) that do not make the password rules explicit, leaving the users to try their luck to comply with the required rules.

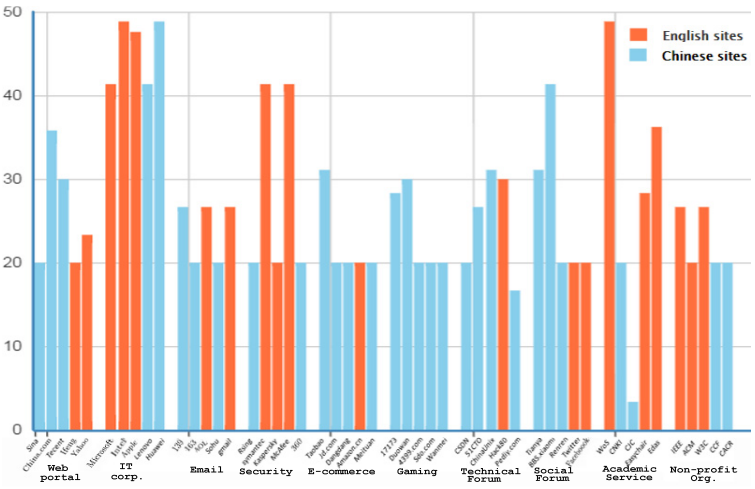

Fig. 2. Strength (in bits, see Sec. 2.2) of the 50 password composition rules

Summary. Despite the long-standing use of passwords and long-recognised importance of the provision of sound password practices, many leading web services seem to lose their lead in enforcing sensible password rules. As no two services examined share the same password rule, there seems to be no generally agreed-upon practice. In 2010, Bonneau and Preibusch [8] found "many aspects of password implementation are not standardised", while our results suggest that after five years of development, basic password practices are still highly diversified. What's worse, policy recommendations from major authorities are also quite different from each other (e.g., US-CERT [39] vs. NIST [44] vs. DISA [15]) and often far from practicable (e.g., "use different passwords on different systems" [39, 44] and "users must not be able to reuse any of their previous 10 passwords." [15]). This greatly impairs their authoritativeness. Unsurprisingly, a large fraction of high-profile sites (e.g., Yahoo, Apple, Microsoft and Kaspersky) each maintains their own, even arcane, unnecessary and illogical rules. Security background or abundant capital, engineering resources do not correlate with

noticeable advantages in policy strength (see Fig. 2). In addition, generally English sites implement more demanding rules than their Chinese counterparts.

3.2 Password strength meters in the tangle

To give users a feedback about the goodness of their selected passwords, password strength meters are employed to accomplish this aim. Recent research has shown that password meters, especially those with a timely [47], easily comprehensible [16] and accurate [50] feedback, can lead to tangible improvements in password security. Table 6 shows that 26 sites (10 English and 16 Chinese) employ a meter.

Among these 26 sites with a meter, 5 sites (including 3 from English sites and 2 from Chinese sites) only verbally show the password strength, and a mere 9 sites (including 5 from English sites and 4 from Chinese sites) impose mandatory strength requirements. Further considering that there are only 20 English sites out of 50 sites investigated, this suggests that, generally, English sites are more stringent in ensuring password security. It is also worth noting that, some sites (i.e., 139, IEEE and Hack80) provides strength feedback to a user only when the user's password meets with their password composition rules first.

According to Furnell's 2007 investigation [19], only two of the 10 sites he examined provide a meter, while his 2011 investigation [20] saw a great advancement: 6 out of 10 sites provide a meter during user registration. However, our results show no advancement in password meter adoption during the past five years.

It is interesting to note that, most sites from the categories of web portal, email, e-commerce and technical forum employ a password meter, while most sites from the categories of IT corp., security corp. and academic service do not provide thus feature. Further considering that the later categories of sites typically employ more restrictive password rules (see Fig. 2), one would really be confused about what's their ultimate purpose of imposing a password creation policy from the user prospective. From the site prospective, as composition rules is highly more easy to be implemented and maintained than a password meter, and thus different choices mean different engineering cost involved. Consequently, one plausible (yet ironical) reason may be that, IT corp. sites, security corp. sites and academic service sites do not provide a meter due to engineering cost. Another reason may be that, due to the "failure of the academic literature to provide approaches that are convincingly better than current practices" [7], these technically-savvy sites are aware of the ineffectiveness of the current password meters, yet there is no adequate, concrete and well-grounded knowledge (e.g., about architectures, frameworks, algorithms, metrics and guidelines) available for them to get things (towards) right, and they are hitting a brick wall.

3.3 Online guessing attackers at large

We proceed to investigate the effectiveness of these 50 policies in resisting against the primary threat to password accounts, i.e. online guessing. As detailed in Section 2.5, we specially select two types of weak passwords to model the two different types online guessing (i.e., trawling and targeted) attacks, respectively. Each type of passwords is composed of 8 testing passwords, and each password is tested against every service, meaning a total of 800(=2*8*50) testing instances.

Table 6. An overview of the password strength meters in the selected web services ('∅' stands for no strength scale; '-' stands for non-existence; 'Monotonicity' stands for whether an additional character contributes to a better score)

Web Services		Strength score scale	Verbal or visual	Monotonicity	Least score enforcement
Web Portal	Sina	Very weak, Weak, Medium, High	Both	Yes	Weak
	China.com	Weak, Medium, Strong	Both	Yes	Medium
	Tecent	Weak,Medium,Strong	Both	Yes	∅
	Ifeng	Low, Medium, High	Both	Yes	∅
	Yahoo	1(Easy to guess),2(weak),3(Medium), 4(Strong), 5(Very Strong)[a]	Visual[b]	Yes	3(Medium)
IT Corp.	Microsoft	∅	None	–	∅
	Intel	∅	None	–	∅
	Apple	Weak, Medium, Strong	Verbal	No	Medium
	Lenovo	∅	None	–	∅
	Huawei	∅	None	–	∅
Email	139	∅	None	No	∅
	163	Weak,Medium,Strong	Both	Yes	∅
	AOL	Weak, Strong, Brilliant	Both	Yes	∅
	Sohu	Weak, Medium, Strong	Both	Yes	∅
	Gmail	Too short, Weak, Fair, Slightly strong, Strong	Both	No	Fair
Security Corp.	Rsing	∅	None	–	∅
	Symantec	∅	None	–	∅
	Kaspersky	∅	None	–	∅
	McAfee	∅	None	–	∅
	360	∅	None	–	∅
Ecommerce	Taobao	Low, Medium, High	Both	Yes	Medium
	Jd.com	Weak, Medium, Strong	Both	Yes	∅
	Dangdang	Weak, Medium, Strong	Both	Yes	∅
	Amazon	∅	None	–	∅
	Meituan	Weak, Medium, Strong	Both	–	∅
Gaming	17173	Weak, Medium, Strong	Verbal	–	∅
	Duowan	∅	None	–	∅
	4399.com	∅	None	–	∅
	Sdo.com	Weak, Medium, Strong	Both	–	∅
	Wanmei	Low, Medium, High	Both	–	∅
Technical Forum	CSDN	Low, Medium, High	Both	Yes	∅
	51CTO	Weak, Medium, Strong	Both	Yes	∅
	ChinaUnix	Weak,Medium,Strong	Both	Yes	Medium
	Hack80	Weak, Medium, Strong	Both	Yes	∅
	Pediy.com	∅	None	–	∅
Social Forum	Tianya	∅	None	–	∅
	BBS.xiaomi	∅	None	–	∅
	Renren	Weak, Fair, Very brilliant	Verbal	Yes	∅
	Twitter	Too obvious/short, NSE, Can be more secure,Ok, Medium, Strong,Perfect	Visual[b]	No	Not secure enough (NSE)
	Facebook	∅	None	–	∅
Academic Service	WoS	∅	None	–	∅
	CNKI	∅	None	–	∅
	Cjc.ac.cn	∅	None	–	∅
	Easychair	∅	None	–	∅
	Edas	Weak, Medium, Strong	Verbal	No	∅
Non-profit Org.	IEEE	Should be stronger, Good, Great	Both	No	∅
	ACM	∅	None	–	∅
	W3C	Very weak, Weak, Sufficient, Strong, Very strong	Verbal	Yes	Sufficient
	CCF	∅	None	–	∅
	CACR	∅	None	–	∅

[a] According to the results obtained in 2013 [11], the password meter of Yahoo was "Weak, Strong, Very strong". Yet, at the time of this writing Yahoo has changed its policy and divides its strength bar into five scales. Although Yahoo's meter only verbally displays the strength score when the password is "1 (Easy to guess)" or "2 (Weak)", and for the other cases, only a visual progress bar is in place, fortunately one can identify the total number of such cases (i.e., three). In line with its scores in 2013, we suppose the three scores corresponding to these three scales that are not verbally displayed are "3(Medium)", "4(Strong)" and "5(Very Strong)", respectively.

[b] The password meters of Yahoo and Twitter only verbally displays the strength score when a password can not be accepted. When a password can be accepted, *only* a visual progress bar is in place. Consequently, their meters is deemed to be visually displayed.

Our results (see Table 7) show that among the 800 testing instances, 541 ones are accepted, where 257 ones are accepted without providing any strength information, 83 ones are accepted while they are metered "weak/low", and each site accepts at least two instances (passwords). Among these 259 rejected instances, 221 ones are rejected by password rules, 17 ones are rejected by password meters, and 21 ones are rejected by both the password rule and meter.

This has at least two important implications. First, considering that at least 2 (and an average of 10.8) weak passwords are allowed by every site and that, ironically, 15 leading sites, including many technically savvy services (e.g., Kaspersky, Rsing and ACM) and financially sound services (e.g., Amazon and Dangdang), accept all the 16 weak passwords like "`123456`", "`woaini`" and "`wanglei`", it is really difficult to refuse the implication that the password policies imposed by the 50 sites largely fail to serve their purpose—resisting online guessing. Second, currently, password rules are overwhelmingly dominant in the filtering of bad passwords, and password meters should have played a more important role.

Perhaps unsurprisingly, password strength scores of the 50 selected sites are highly inconsistent, which accords with previous work [11]. Very often, inherently weak passwords (e.g., "`password123`" and "`wanglei1`") pass the check of password rules and is labeled as strong by password meters, and they are accepted by sites of significant value (e.g., all of the five e-commerce sites); the same password receives highly inconsistent strength outcomes from different password meters and is accepted or rejected for unintelligible reason. For instance, "`Wanglei@123`" is measured as "weak" by Yahoo, "medium" by Sohu and "strong" by Gmail; It is rejected by McAfee (which accepts "`wanglei123`"). These *inaccuracies* provide users with a false sense of security, and what's worse, these *inconsistencies* cause user confusion in selecting a stronger password, both of which would lead the "weakest link" (i.e., common users) in the security chain to be weaker.

Particularly, among the 541 accepted instances, 323 ones (i.e., 59.7%) are used for the test against targeted online guessing, which suggests that web services on today's Internet are comparatively more vulnerable to targeted attacks (at least, against Chinese users). The right part of Table 7, further shows that, most of the meters largely overestimate the strength of Type B passwords and Chinese sites show no better performance, which renders such kind of passwords at large over the Internet and also provides a false sense of security to common users.

Some remarks. To the best of our knowledge, 15 web services studied in this work have been the victims of hacking and leaked large amounts of user credentials (see some shivery news [2, 40, 41, 43]). As far as can be confirmed, among these 15 leaked sites, 9 ones have changed their password policies during the past five years. More specifically, Yahoo has changed its length limits from 6-32 to 7-30 in the last year as compared to the data reported in 2014 by [11]; Apple changed its some lenient charset requirement to the current "1^{+}lower,1^{+}upper,1^{+}digit" in 2012 according to [32]; As compared to the data reported in 2011 by [20], Microsoft has changed its length limits from 6-16 to 8-16 and its meter ratings from {Weak, Medium, Strong} to {$\emptyset$}, Gmail has changed its meter ratings from {Weak, Fair, Good, Strong} to the current {Too short, Weak, Fair, $\cdots$},

Table 7. An overview of the evaluation results of 16 passwords on 50 web services

	123456	123456789	5201314	woaini	iloveyou	password	woaini1314	password123	wanglei	wanglei123	wanglei1	wanglei12	Wanglei123	wang.lei	wanglei@123	Wanglei@123
Sina	Ⓜ,VW	✓,W	Ⓜ,VW	✓,W	Ⓜ,VW	Ⓜ,VW	✓,S	✓,S	✓,W	✓,S	✓,M	✓,M	✓,S	✓,M	✓,S	✓,S
China.com	⊗,W	⊗,W	⊗,W	⊗,W	⊗,W	⊗,W	⊗,M	⊗,M	⊗,W	⊗,M	⊗,M	⊗,M	✓,M	⊗,M	✓,S	✓,S
Tecent	Ⓚ,W	✓,W	Ⓚ,W	✓,W	✓,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,W	✓,M	✓,S	✓,M	✓,S	✓,S
Ifeng	✓,L	✓,L	✓,L	✓,L	✓,L	✓,L	✓,M	✓,M	✓,L	✓,M	✓,M	✓,M	✓,H	✓,M	✓,H	✓,H
Yahoo	⊗,W	✓,M	Ⓜ,W	Ⓜ,W	Ⓜ,W	Ⓜ,W	✓,M	✓,S	Ⓜ,W	✓,M	✓,M	✓,M	⊗,W	✓,M	✓,M	Ⓜ,W
Microsoft	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ
Intel	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓
Apple	Ⓚ,W	Ⓚ,W	⊗,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,M	Ⓚ,M	Ⓚ,W	Ⓚ,M	⊗,W	Ⓚ,W	✓,M	Ⓚ,W	Ⓚ,M	✓,S
Lenovo	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓	Ⓚ	✓	✓	✓	✓	✓	✓	✓
Huawei	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓
139	Ⓚ	Ⓚ	Ⓚ	✓,W	✓,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	Ⓚ	✓,S	✓,S
163	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	✓,M	✓,S	✓,S
AOL	Ⓚ	Ⓚ,W	Ⓚ	Ⓚ	Ⓚ,W	Ⓚ,W	✓,S	✓,S	Ⓚ	Ⓚ,S	Ⓚ,S	Ⓚ,S	Ⓚ,S	Ⓚ,W	Ⓚ,B	Ⓚ,B
Sohu	Ⓚ,W	✓,W	✓,W	✓,W	Ⓚ,W	✓,W	Ⓚ,W	✓,M	✓,W	✓,M	✓,M	✓,M	✓,M	✓,W	✓,M	✓,M
Gmail	Ⓚ,TS	Ⓚ,W	Ⓚ,TS	Ⓚ,TS	⊗,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	✓,F	✓,F	✓,SS	✓,SS	✓,F	✓,S	✓,S
Rsing	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Symantec	Ⓚ	Ⓚ	Ⓚ,	Ⓚ	Ⓚ	Ⓚ	✓	✓	Ⓚ	✓	✓	✓	✓	Ⓚ	✓	✓
kaspersky	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
McAfee	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓	Ⓚ	✓	✓	✓	✓	Ⓚ	Ⓚ	✓
360	Ⓚ	Ⓚ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Taobao	Ⓚ,L	Ⓚ,M	Ⓚ,M	Ⓚ,L	Ⓚ,M	Ⓚ,M	✓,M	✓,M	Ⓚ,L	✓,M	✓,M	✓,M	✓,M	✓,M	✓,M	✓,H
Jd.com	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,M	✓,S	✓,W	✓,M	✓,M	✓,M	✓,M	✓,M	✓,S	✓,S
Dangdang	✓,W	✓,W	✓,W	✓,W	✓,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	✓,M	✓,S	✓,S
Amazon	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Meituan	✓,W	✓,W	✓,W	✓,W	✓,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,W	✓,M	✓,S	✓,W	✓,S	✓,S
17173	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	✓,M	✓,S	✓,S
Duowan	Ⓚ	✓	Ⓚ	✓	✓	✓	✓	✓	Ⓚ	✓	✓	✓	✓	✓	✓	✓
4399.com	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sdo.com	Ⓚ,W	Ⓚ,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	✓,M	✓,M	✓,M	✓,M	✓,M	Ⓚ	Ⓚ	Ⓚ
Wanmei	✓,L	✓,L	✓,L	✓,L	✓,L	✓,L	✓,M	✓,M	✓,L	✓,M	✓,L	✓,M	✓,H	Ⓚ	Ⓚ	Ⓚ
CSDN	✓,L	✓,L	✓,L	✓,L	✓,L	✓,L	✓,H	✓,H	✓,L	✓,H	✓,M	✓,M	✓,H	✓,M	✓,H	✓,H
51CTO	✓,W	✓,W	✓,W	✓,W	✓,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	✓,M	✓,S	✓,S
Chinaunix	Ⓚ,W	✓,M	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	✓,M	✓,M	Ⓚ,W	✓,M	✓,M	✓,M	✓,M	✓,M	✓,S	✓,S
Hack80	Ⓚ	✓,W	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓,M	✓,M	Ⓚ	✓,M	Ⓚ	✓,M	✓,M	Ⓚ	✓,S	✓,S
Pediy.com	Ⓚ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Xiaomi	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓	Ⓚ	✓	✓	✓	✓	✓	✓	✓
Tianya	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓	Ⓚ	✓	✓	✓	✓	✓	✓	✓
Renren	✓,W	✓,W	✓,W	✓,W	✓,W	✓,W	✓,M	✓,M	✓,W	✓,M	✓,M	✓,M	✓,S	✓,M	✓,S	✓,S
Twitter	Ⓚ,TO	Ⓚ,TO	✓,CMS	✓,NSE	Ⓚ	Ⓚ,TO	✓,OK	Ⓚ,TO	✓,NSE	✓,OK	✓,NSE	✓,CMS	✓,M	✓,CMS	✓,S	✓,P
Facebook	Ⓚ	Ⓚ	✓	Ⓚ	Ⓚ	Ⓚ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
WoS	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓	✓
CNKI	Ⓚ	Ⓚ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CJC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Easychair	Ⓚ	Ⓚ	Ⓚ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Edas	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	Ⓚ,W	✓,W	Ⓚ,W	✓,W	✓,W	✓,W	✓,W	✓,M	✓,W	✓,S	✓,S
IEEE	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	Ⓚ	✓,G	Ⓚ	Ⓚ	✓,G	✓,SBS	✓,G	✓,G	Ⓚ	✓,G	✓,G
ACM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
W3C	⊗,VW	Ⓜ,VW	⊗,VW	⊗,VW	Ⓜ,VW	Ⓜ,VW	✓,Suf	✓,Suf	⊗,VW	✓,Suf	Ⓜ,W	Ⓜ,W	✓,S	Ⓜ,W	✓,Suf	✓,VS
CCF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CACR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

1° Notations: ✓: Accepted; Ⓚ: Rejected by password composition rules; Ⓜ: Rejected by password strength meter; ⊗: Rejected by both the password rule and meter.
2° Abbreviations: B = Brilliant; G = Good; H = High; L = Low; M = Medium; P = Perfect; S= Strong; W = Weak; SS = Slightly strong; ; TO = Too obvious; TS = Too short; TW = Too weak; VS = Very strong; VW = Very weak; CMS = Can be more secure; NSE = Not secure enough; SBS= Should be stronger, Suf = Sufficient.
3° The evaluation result for password A vs. site B is in the "X,Y" format, where "X" indicates whether A is accepted or rejected by B, and "Y" indicates A's strength score given by B's password meter. If B is with no meter, "Y" is naturally absent. We note that for three sites with meters (i.e., AOL, Twitter and Hack80), the strength score "Y" is also absent in cases where the password A does not comply with B's password rules.

Twitter has changed its meter ratings from {· · · , Weak, Good, Strong, Very strong} to the current {· · · , Medium, Strong, Perfect}, Facebook has changed its meter ratings from {Weak, Medium, Strong} to {∅}; As victims of the 2011 catastrophic hacking event [36], Duowan changed its length limits from 6-20 to 8-20, CSDN changed its length limits from 8-20 to 6-20, and Tianya added the current charset requirement. In addition, we can identify that AOL has changed its length limits from 6-16 to 8-16 as compared to the data we collected in Mar. 2014, and Taobao changed its length limits from 6-16 to 6-20.

In all, during the past five years, as far as we know, 6 sites have adopted more complex and stringent rules, 1 sites have relaxed its rules, 2 sites have changed their rating scores and 2 sites have chosen not to provide meters at all. While 8 of the 11 changed sites may be towards seemingly more stringent or usable policies, 3 sites are highly going against the trend of good password practices,

bearing no serious efforts from these 3 service providers. In a nutshell, as shown in Table 7, all these 11 "new" password policies still largely fail to serve their purposes and are virtually equivalent to '*the emperor's new password policies*'.

At least, new services or existing ones that wish to establish/change a password policy, should not start the development of yet another policy, but rather consider using or extending the pacemakers' policies to be more consistent with common sense practices and to be more prudent with local, cultural characteristics. Comparatively, among the 50 policies studied, the current policy adopted by Apple is the most effective one against online guessing. However, it is at the cost of usability and leads to great user frustration and fatigue [32]. For example, the "1^+lower,1^+upper,1^+digit" rule highly hinders mobile users. This highlights the imperative needs for more academic efforts to guide the industrial practice.

4 Conclusion

In this work, we have conducted a large-scale empirical analysis of the current state of password creation policies imposed by 50 leading web services by using a systematic, evidence-supported approach. We find that the policies are highly diversified among the studied sites and largely fail to withstand online guessing attacks. Comparatively, password composition rules play a more important role in resisting online guessing than password strength meters, partly because most meters are merely suggestive, and partly because current meters are inaccurate in gauging the strength of passwords. Consistent with previous work [11], highly inconsistent outcomes are given for the same testing password by different meters, which may confuse users and undermine user trust in security advice, defeating the purpose of enforcing password policies in the first place.

As compared to Chinese sites, English sites generally enforce more stringent password polices. We also discuss the factors that may influence a site's choice of password policies. Our results show that, overall, security background or abundant capital, engineering resources do not correlate with noticeable advantages in password practice, as opposed to previous work [8]. A natural future work is to incorporate more sample sites (e.g., medium sites, and sites from other languages and services) and investigate more types of password policies (such as password change, lockout and expiration), gaining a more complete picture of the whole password ecosystem and proposing well-grounded policy recommendations.

Acknowledgment

We are grateful to the anonymous reviewers for their invaluable comments. This research was supported by the NSFC program under Grant No. 61472016.

References

1. Al-Ameen, M.N., Wright, M., Scielzo, S.: Towards making random passwords memorable: Leveraging users' cognitive ability through multiple cues. In: Proc. ACM CHI 2015. pp. 1–10. Seoul, Republic of Korea (April 18-23 2015)
2. Alexander, T.: Leaked passwords (July, 2012), `http://thepasswordproject.com/leaked_password_lists_and_dictionaries`

3. Allan, C.: 32 million Rockyou passwords stolen (Dec 2009), http://www.hardwareheaven.com/news.php?newsid=526
4. Bauman, E., Lu, Y., Lin, Z.: Half a century of practice: Who is still storing plaintext passwords? In: Lopez, J., Wu, Y. (eds.) ISPEC 2015, LNCS, vol. 9065, pp. 253–267. Springer-Verlag (2015)
5. Bishop, M., V Klein, D.: Improving system security via proactive password checking. Computer & Security 14(3), 233–249 (1995)
6. Bonneau, J.: The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In: IEEE S&P 2012. pp. 538–552. San Francisco, USA (May 21-23 2012)
7. Bonneau, J., Herley, C., van Oorschot, P., Stajano, F.: Passwords and the evolution of imperfect authentication. Comm. of the ACM 58(7), 78–87 (2015)
8. Bonneau, J., Preibusch, S.: The password thicket: Technical and market failures in human authentication on the web. In: Proc. WEIS 2010 (June 7-8 2010)
9. Burnett, M.: 10,000 top passwords (June 2011), https://xato.net/passwords/more-top-worst-passwords/
10. Burr, W., Dodson, D., Perlner, R., Polk, W., Gupta, S., Nabbus, E.: NIST SP800-63 – electronic authentication guideline. Tech. rep., NIST, Reston, VA (April 2006)
11. Carnavalet, X., Mannan, M.: From very weak to very strong: Analyzing password-strength meters. In: Proc. NDSS 2014. pp. 1–16. San Diego, CA, USA (2014)
12. Chiasson, S., van Oorschot, P.C.: Quantifying the security advantage of password expiration policies. Designs, Codes and Cryptography (2015), in press, Doi: http://dx.doi.org/10.1007/s10623-015-0071-9
13. CNNIC: CNNIC Released the 35th Statistical Report on Internet Development in China (Feb 2015), http://www.apira.org/news.php?id=1732
14. Das, A., Bonneau, J., Caesar, M., Borisov, N., Wang, X.: The tangled web of password reuse. In: Proc. NDSS 2014. pp. 1–15 (2014)
15. DISA for DoD: Application security and development. Tech. rep., Defense Information Systems Agency (DISA), Reston, VA (July, 2013), doi:http://www.stigviewer.com/stig/application_security_and_development/
16. Egelman, S., Sotirakopoulos, A., Beznosov, K., Herley, C.: Does my password go up to eleven?: the impact of password meters on password selection. In: Proc. CHI 2013. pp. 2379–2388. ACM (2013)
17. Florencio, D., Herley, C.: A large-scale study of web password habits. In: Proc. WWW 2007. pp. 657–666. ACM (2007)
18. Florêncio, D., Herley, C.: Where do security policies come from? In: Proc. ACM SOUPS 2010. pp. 1–14. ACM, Redmond, Washington, USA (July 14-16 2010)
19. Furnell, S.: An assessment of website password practices. Computers & Security 26(7), 445–451 (2007)
20. Furnell, S.: Assessing password guidance and enforcement on leading websites. Computer Fraud & Security 2011(12), 10–18 (2011)
21. Furnell, S., Bär, N.: Essential lessons still not learned? examining the password practices of end-users and service providers. In: Proc. HAS 2013, LNCS, vol. 8030, pp. 217–225. Springer (2013)
22. Goldman, J.: Chinese Hackers Publish 20 Million Hotel Reservations (Dec 2013), http://www.esecurityplanet.com/hackers/chinese-hackers-publish-20-million-hotel-reservations.html
23. Goodin, D.: Anatomy of a hack: How crackers ransack passwords like "qeadzcwrsfxv1331" (May, 2013), http://arstechnica.com/security/2013/05/how-crackers-make-minced-meat-out-of-your-passwords/2/

24. Haikun, C.: Multiply the total to 3.647 million on chinese web sites (Feb, 2015), http://www.changhaikun.com/index.php/2015/04/03/multiply-the-total-to-3-647-million-on-chinese-web-sites/
25. Herley, C., Van Oorschot, P.: A research agenda acknowledging the persistence of passwords. IEEE Security & Privacy 10(1), 28–36 (2012)
26. Huang, X., Xiang, Y., Chonka, A., Zhou, J., Deng, R.H.: A generic framework for three-factor authentication: Preserving security and privacy in distributed systems. IEEE Trans. Parallel Distrib. Syst. 22(8), 1390–1397 (2011)
27. Huang, Z., Ayday, E., Fellay, J., Hubaux, J.P., Juels, A.: Genoguard: Protecting genomic data against brute-force attacks. In: Proc. IEEE S&P 2015. pp. 1–16
28. Ihm, S., Pai, V.S.: Towards understanding modern web traffic. In: Proc. ACM SIGCOMM 2011. pp. 295–312. ACM (2011)
29. Jakobsson, M., Akavipat, R.: Rethinking passwords to adapt to constrained keyboards. Proc. IEEE MoST 2012 pp. 1–11 (2012)
30. Jakobsson, M., Dhiman, M.: The benefits of understanding passwords. In: Proc. HotSec 2012. pp. 1–6. USENIX Association (2012)
31. Jiang, Q., Tan, C.H., Phang, C.W., Sutanto, J., Wei, K.K.: Understanding chinese online users and their visits to websites: Application of zipf's law. International Journal of Information Management 33(5), 752–763 (2013)
32. Johns, R.: Illogical apple id password rules (May, 2012), https://discussions.apple.com/thread/3785494
33. Keith, M., Shao, B., Steinbart, P.: A behavioral analysis of passphrase design and effectiveness. J. of the Assoc. for Inf. Syst. 10(2), 2 (2009)
34. Kelley, P.G., Komanduri, S., Mazurek, M.L., Shay, R., Vidas, T., Bauer, L., Christin, N., Cranor, L.F., Lopez, J.: Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In: Proc. IEEE S&P 2012. pp. 523–537. IEEE (2012)
35. Ma, J., Yang, W., Luo, M., Li, N.: A study of probabilistic password models. In: Proc. IEEE S&P 2014. pp. 538–552. IEEE (2014)
36. Martin, R.: Amid Widespread Data Breaches in China (Dec 2011), http://www.techinasia.com/alipay-hack/
37. Mathew, J.S.: 15,000 twitter credentials stolen and leaked, hacker promises more soon (Aug, 2013), http://www.itcmt.com/2013/08/23/15000-twitter-credentials-stolen-and-leaked-hacker-promises-more-soon/
38. Mazurek, M.L., Komanduri, S., Vidas, T., Cranor, L.F., Kelley, P.G., Shay, R., Ur, B.: Measuring password guessability for an entire university. In: Proc. CCS 2013. pp. 173–186. ACM (Nov 4–8 2013)
39. McDowell, M., Hernan, S., Rafail, J.: Security Tip (ST04-002): Choosing and Protecting Passwords (2013), https://www.us-cert.gov/ncas/tips/ST04-002
40. Millward, S.: Xiaomi now has 100 million users of its android-based mobile os (Feb, 2015), https://www.techinasia.com/xiaomi-miui-100-million-users/
41. Mirante, D., Cappos, J.: Understanding password database compromises. Tech. rep., Polytechnic Institute of NYU, McLean, VA (2013), doi:https://isis.poly.edu/~jcappos/papers/tr-cse-2013-02.pdf
42. Morris, R., Thompson, K.: Password security: A case history. Comm. of the ACM 22(11), 594–597 (1979)
43. Rhodan, M.: Nearly 5 million google passwords leaked on russian site (Sep, 2014), http://time.com/3318853/google-user-logins-bitcoin/
44. Scarfone, K., Souppaya, M.: NIST SP800-118: Guide to enterprise password management. Tech. rep., NIST, Reston, VA (Aug 2013)

45. Schechter, S., Brush, A.B., Egelman, S.: It's no secret. measuring the security and reliability of authentication via secret questions. In: Proc. IEEE S&P 2009. pp. 375–390. IEEE, Oakland, California (May 16-19 2009)
46. Schechter, S., Herley, C., Mitzenmacher, M.: Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks. In: Proc. HotSec 2010. pp. 1–8 (2010)
47. Shay, R., Bauer, L., Christin, N., Cranor, L.F., Forget, A., Komanduri, S., Mazurek, M., Melicher, W., Segreti, S.M., Ur, B.: A spoonful of sugar? the impact of guidance and feedback on password-creation behavior. In: Proc. CHI 2015. pp. 2903–2912
48. Shay, R., Komanduri, S., Durity, A.L., Huh, P.S., Mazurek, M.L., Segreti, S.M., Ur, B., Bauer, L., Christin, N., Cranor, L.F.: Can long passwords be secure and usable? In: Proc. ACM CHI 2014. pp. 2927–2936. ACM (2014)
49. Top 500 chinese pinyin names (Jan, 2015), http://www.data.ac.cn/zrzy/g22.asp
50. Ur, B., Kelley, P.G., Komanduri, S., et al.: How does your password measure up? the effect of strength meters on password creation. In: Proc. USENIX Security 2012. pp. 65–80. Bellevue, WA, USA, (August 8-10 2012)
51. Wang, D., Cheng, H., Wang, P.: Understanding Passwords of Chinese Users: Characteristics, Security and Implications (Jan 2015), http://t.cn/RzSlpDz
52. Wang, D., He, D., Wang, P., Chu, C.H.: Anonymous two-factor authentication in distributed systems: Certain goals are beyond attainment. IEEE Trans. Depend. Secur. Comput. (2014), http://dx.doi.org/10.1109/TDSC.2014.2355850
53. Weir, M., Aggarwal, S., Collins, M., Stern, H.: Testing metrics for password creation policies by attacking large sets of revealed passwords. In: Proc. CCS 2010. pp. 162–175. ACM (October 4-8 2010)
54. Yan, J., Blackwell, A.F., Anderson, R.J., Grant, A.: Password memorability and security: Empirical results. IEEE Security & privacy 2(5), 25–31 (2004)
55. Yap, J.: 450,000 user passwords leaked in Yahoo breach (July 2012), http://www.zdnet.com/article/450000-user-passwords-leaked-in-yahoo-breach/
56. Zhu, B., Yan, J., Bao, G., Mao, M., Xu, N.: Captcha as graphical passwords–a new security primitive based on hard AI problems. IEEE Trans. Inform. Forensics Security 9(6), 891–904 (2014)