Computer Standards & Interfaces xx (xxxx) xxxx–xxxx

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
0920-5489/

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).