Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

User practice in password security: An empirical study of real-life passwords in the wild

Chao Shen [a,*], *Tianwen Yu* [a], *Haodi Xu* [a], *Gengshan Yang* [a], *Xiaohong Guan* [a,b]

[a] *Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an 710049, China*
[b] *Tsinghua University, Haidian District, Beijing 10084, China*

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form 18 March 2016
Accepted 29 May 2016
Available online 31 May 2016

Keywords:
Password characteristics
User practice
Usability
Awareness
Measurement

ABSTRACT

Due to increasing security awareness of password from the public and little attention on the characteristics of real-life passwords, it is thus natural to understand the current state of characteristics of real-life passwords, and to explore how password characteristics change over time and how earlier password practice is understood in current context. In this work, we attempt to present an in-depth and comprehensive understanding of user practice in real-life passwords, and to see whether the previous observations can be confirmed or reversed, based on large-scale measurements rather than anecdotal knowledge or user surveys. Specifically, we measure password characteristics on over 6 million passwords, in terms of password length, password composition, and password selection. We then make informed comparisons of the findings between our investigation and previously reported results. Our general findings include: (1) average password length is at least 12% longer than previous results, and 75% of our passwords have the length between 8 and 10 characters; (2) there is a significant increase of using only numbers as passwords, and easy-to-reach symbols are always the first choice when users added symbols into passwords; (3) there observes a remarkable increase (about 40%) of using combo-meaningful data as passwords, and a striking proportion of using the most common passwords or login names as passwords. Our investigation also includes collecting statistics about the use of symbols, letter-case, and meaningful details, which presents a systematic analysis of password usage. The comparative results indicate that the password characteristics and password practice on this massive password data set are somewhat inconsistent with those from anecdotal knowledge and user surveys, and exhibit a substantial change over time in some ways. Further research needs to build upon this understanding for gaining insight into how password security can be improved.

1. Introduction

Despite a growing number of graphical and biometric authentication mechanisms, passwords remain the dominant method of authentication (Herley and Van Oorschot, 2012; Uellenbeck and Durmuth, 2013). According to NIST specifications, text-based passwords are popular in typical web users' experience since they are conceptually simple, inexpensive to administer, and user-friendly (Burr et al., 2011). Since passwords are

* *Corresponding author.* Tel.: +86 29 82663939.
E-mail addresses: cshen@sei.xjtu.edu.cn (C. Shen), twyu@sei.xjtu.edu.cn (T. Yu), xhguan@sei.xjtu.edu.cn (X. Guan), luanyingjian@stu.xjtu.edu.cn (G. Yang), haodixu@outlook.com (H. Xu).
http://dx.doi.org/10.1016/j.cose.2016.05.007
0167-4048/

commonly used to protect accounts with valuable assets (e.g., Bank or Email accounts), they have increasingly been subjected to several attacks, which mainly exploit users' tendency of choosing simple and poor passwords (e.g., some dictionary words, names, and personal information) User-selected passwords have always been easily guessable and predictable (Adams and Sasse, 1999; Bonneau, 2012; Bunnell et al., 1997; Dell'Amico et al., 2010; Grampp and Morris, 1984; Ji et al., 2015; Mazurek et al., 2013; Morris and Thompson, 1979; Ur et al., 2012), which has been joined with an emerging problem of users unwittingly divulging their passwords to the public [e.g., the recent password-leakage events in Facebook, LinkedIn, and Google (Bogart, 2013; Goodin, 2012; Pagliery, 2014)].

About 37 years ago, Morris and Thompson (1979) found that users had poor practice with their passwords and did not afford sufficient attention to safeguarding their secrets. Since then, we have seen studies on password characteristics (Adams and Sasse, 1999; Bryant and Campbell, 2005, 2006; Campbell and Bryant, 2004; Grampp and Morris, 1984; Zviran and Haga, 1999), which provide some understanding of user practice in password usage and security. But – has this understanding been applied in practice? Has anything changed over time? Moreover, most previous studies favored using passwords from user surveys and anecdotal knowledge, and little attention has been given to real-life passwords and their practical usage, which Dourish et al. (2004) have called "security in the wild".

In this work, we attempt to provide an in-depth and comprehensive understanding of user practice in real-life passwords, and to see whether the previous-observed facts can be confirmed or reversed based on a real-life and large-scale measurement rather than anecdotal knowledge or user surveys. We measure the characteristics of real-life passwords over a large population in terms of password length, password composition, and password selection, and make informed comparisons between our investigation and previous studies. Among our interesting findings are how password characteristics change over time and how earlier password practice is understood in current context: the average password length is 9.46 characters, which is longer than what has been found in the literature, and most of our passwords have the length between 8 and 10 characters; passwords are still dominated by simple structure, and there is a significant increase of using only numbers as passwords, and easy-to-reach symbols are always the first choice when users added symbols into passwords; users prefer to use meaningful data in passwords, and there observes a remarkable increment of selecting combinations of multiple meaningful data as passwords, and a striking proportion of using the most common passwords or the login-names as the passwords. Our investigation also includes collecting statistics about the use of symbols, letter case, and meaningful details, which presents a systematic analysis of password usage. These comparative results indicate that the password characteristics and password practice on this massive password data set are somewhat inconsistent with past anecdotal knowledge and password surveys, and exhibit a substantial change over time in some ways.

This paper is organized as follows. Section 2 describes previous work. Section 3 develops the problem and approach of this study. Section 4 introduces the source of our passwords. Sections 5–7 present our descriptive findings and comparative results between our study and previous studies. Section 8 offers discussions and concludes.

2. Background and related work

The usage of password for authentication has been analyzed at length from the security literature. However, there have been few studies on understanding the characteristics of real-life passwords and the application of this understanding in practice, which need to be examined in the concrete reality of daily usage.

An early study to notice this problem was by Morris and Thompson in 1979 (Morris and Thompson, 1979), in which they examined users' password habit when no constraint was put on their choice. They collected 3289 passwords, and found over 86% of the passwords were extremely weak: being too short, containing only lowercase letters or only digits, and being easily found in dictionaries. Ten years later, Feldmeier and Karn (1990) charted the progress of password security over the previous decade, which shows that most passwords could be easily guessed or cracked. Here we attempt to bring the analysis of the characteristics of real-life passwords in a large-scale measurement, nearly 37 years after the original Morris and Thompson paper.

While some things may have changed since 1979 (e.g., an average password length of more than 6 characters acceptable to users), it is just as true that many users still choose poor and simple passwords, unless forced to do otherwise (Whitten and Tygar, 1998). An earlier experiment by Grampp and Morris (1984) found that weak passwords, e.g., a name followed by a meaningful number, were in a widespread use. Then Riddle et al. (1989) performed a linguistic analysis on 6226 passwords, grouping them into categories such as names, dictionary words, or seemingly random strings, and bore out an opinion that passwords were reasonably short and reflected things close to the user themselves. Our results on password composition partially support these earlier findings (e.g., 75.6% of passwords are composed of meaningful data), and we find that there are more passwords coming from combo-meaningful data other than single-meaningful data.

Later on, Adams and Sasse (1999) conducted a web-based questionnaire about password usage and security. The analysis on 139 responses showed that selecting complex passwords was a difficult task for users, and most of them did not understand the policies of good passwords. Our findings on password selection confirm and extend their observations, and we also observe an increase of about 13% passwords that are composed of random characters.

Zviran and Haga (1999) surveyed 860 users' passwords for presenting a quantitative analysis of password selection and usage. They found the password characteristics since the Internet era had not changed much from those in the pre-personal computer era (Morris and Thompson, 1979). They showed that about 50% of the passwords are with the length of five or fewer characters; 80% of the passwords used only alphabetic characters; 78% of the passwords are based on meaningful data. Our results on the password characteristics reverse their findings to some extent, showing that the average password length is up to 9.46 characters, and 75% of the passwords have a length between 8 and 10 characters. Besides, fewer