Frequency of compromised passwords used by students and staff of Asia Pacific College: an analysis using NIST SP 800-63B and pwned passwords

JV Roig¹, J de la Cuesta², J Castillo¹, J Cabardo², E Casiño³, E Salalima¹ and M Sanchez¹
¹Advanced Research & Consulting, Asia Pacific College, Philippines
²School of Computing and Information Technology, Asia Pacific College, Philippines
³Information Technology Resources Office, Asia Pacific College, Philippines
¹jvr@apc.edu.ph, jvroig@gmail.com

Abstract. The National Institute of Standards and Technology (NIST) released new guidelines in June of 2017 that recommended new standards for managing and accepting user passwords. Among the new guidelines is a requirement that verifiers should check if a user’s supplied password is compromised – that is, already listed in previous breach corpuses. Using a corpus of 320 million breached passwords, the researchers conducted an experiment to gauge what percentage of the population of their home institution, Asia Pacific College, use compromised passwords. The study found that 16.72% overall – or 1 in 6 people – were using passwords that are part of the 320M breach corpus. This paper also provides a methodology that other institutions and companies can use to conduct the same analysis in order to gather data specific to their population that can guide the improvement of their password policies and related IT security services.

1. Introduction
The National Institute of Standards and Technology (NIST) published new recommendations for verifiers (organizations that maintain username-password credentials) for managing user accounts and passwords [1]. Included in their latest publication is a recommendation that user passwords be checked against a list of known passwords obtained from past exploits and breaches [2]. If the checked password is found in such a list, the verifier should alert the user that her password is unsafe and should select a new one.

While this is a very practical advice, the difficulty is in obtaining a corpus of such breached passwords. These breached passwords are actual passwords that have been obtained throughout the years by black hats who have exploited many different services and websites, and eventually publicly dumped into various nooks and crannies of the internet. Fortunately, someone in the security community stepped up to the plate to help institutions get hold of such a breach corpus. Troy Hunt, owner and maintainer of “Have I Been Pwned” (https://haveibeenpwned.com), released his own breach corpus for public download called “Pwned Passwords” – the same breach corpus that powers his online service. This initial release consisted of 320M passwords found in previous data dumps from hacks and breaches of various black hats, and can enable institutions to efficiently comply with NIST’s latest guideline regarding checking of passwords.
With this breach corpus, the IT team of Asia Pacific College (APC; the home institution of the authors of this report) embarked on the implementation of this new guideline. Along the way, the authors collected data about how many faculty, staff, and students used passwords that were found in the list of 320M breached passwords from Troy Hunt, and the results of analysis of that collected data is reported in this paper.

2. Background and Related Literature
It has been long known that the password hygiene of users is generally terrible. Without special guidance, users will generally fall into usage of insecure passwords due to convenience and the properties and limitations of human memory [3][4]. Bryant and Campbell (2006) concluded that a contributing factor in this lack of care for passwords is the belief that they, the users, are not personally at risk [5]. Shay et al (2010) measured user sentiment regarding password policies (based on older NIST password guidelines, circa 2007), and found that respondents felt that the new policies were annoying, and felt generally neutral when asked if it was worth the effort to comply with the new guidelines, which was the same sentiment when asked if they would rather switch back to the old, less stringent policies [6]. Florencio and Herley (2007) found that the average password reuse is 5.67 sites per password per user [7]. They also found that user type an average of 8 passwords per day – no wonder they’d rather reuse passwords (and use very easy to remember and type, i.e., insecure passwords).

Users in general can’t be bothered to keep up or comply with stringent password policies. When forced to do so by systems that strictly enforce complex password rules (say, forcing a combination of upper-case, lower-case, numbers and special characters along with a minimum length, and forcing frequent password changes), the security of individual passwords do increase, but users inevitably cope with negative behavior, such as writing down passwords [8].

There are other purely technology-based solutions that aim to improve the security of passwords, such as the usage of password strength meters. This has been shown to increase the password strength of users [9], but this isn’t a full victory – forcing or nudging users to create a strong, complex password is one thing, making sure users can remember them is another. Forcing a complex password, other than leading to written-down passwords, also lead to forgotten passwords, which increase the workload of IT support people who have to answer an “I-forgot-my-password” call and reset the user’s password. This also means an inconvenience on the part of the user, which motivates her to finally write it down somewhere or make it predictable for her, both of which are password habits that are not supposed to be encouraged.

NIST’s latest guidelines take a more human-oriented approach, instead of the traditional tech-centric policies that has led to the various negative user sentiments and coping mechanisms described above. This report focuses on one specific guideline – checking passwords against a list of known passwords – to measure how many users fail this check within the setting of an educational institution.

3. Experiment Design and Methodology
A list of 320 million breached passwords was obtained through security researcher Troy Hunt of HaveIBeenPwned.com. Hunt publicly released this list after the NIST’s 2017 guidelines was published, specifically to enable institutions to adopt one of the latest guidelines recommended by NIST. This list of passwords does not contain plaintext entries of passwords – instead, every entry is the SHA1 hash of the plaintext passwords. The list also does not contain matched usernames. The list is entirely SHA1 hashes. (https://www.troyhunt.com/introducing-306-million-freely-downloadable-pwned-passwords/)

The research team, enabled by the assistance of APC’s in-house software development team, inserted hooks in APC’s own internal school management system so that the passwords of each user logging in can be compared to the corpus of breached passwords, determining whether the password being used is compromised or not, in accordance with the NIST’s latest guidance.

This special software infrastructure, composed of short PHP codes and an SQLite3 database, was left running and recording results for 3 months, capturing peak periods of system activity – the
midterms and the finals. This experiment period spanned most of the second trimester of school year 2017-2018 (September – December 2017).

Since the corpus of breached passwords contain only SHA1 hashes, the passwords of users are similarly hashed before searching through the corpus for a match. Only the hashes – not the plaintext password – is ever used and recorded by the research team. At no point does the research team ever use or store the actual plaintext password of any user, beyond using them initially in order to produce the SHA1 hash.

Whether a match is found or not, the user is recorded into a special database table created solely for this research. If a match is found, then that user’s record will reflect that his password is compromised. If no match is found, then that user’s record will reflect that his password is not compromised.

Users are not recorded twice. If a previously recorded user logs in again (an expected outcome, since users will log in several times during the experiment period), the software simply updates the user’s existing record. After 3 months, 88.33% of every APC user was recorded by this special infrastructure: 84.62% of employees, 91.52% of College students, and 84.64% of Senior High students.

4. Analysis and Results

4.1. Overall Results
After the experiment period, the researchers recorded a total of 2,249 unique user logins. 376 of these users have passwords that were in the breach corpus from Troy Hunt - meaning 16.72% of APC’s population (employees and students) are using very unsafe passwords. That’s 1 in 6 people using passwords that require minimal effort for black hats to compromise.

4.2. Population Breakdown
The collected data can be broken down into three different user categories – Employee, College Student, and Senior High School. Table 1 shows the breakdown per user category, and compares the recorded users vs the total active users. Overall, almost 90% of the active APC population during the experiment period were successfully recorded by the experiment infrastructure; the unrecorded are those that did not use the school’s integrated information system during the 3-month experiment period (e.g., students that did not bother to check their grades online, or employees that had no need to file their leaves online).

| User Type     | Recorded Users | Actual Population Size | % of Population Recorded |
|---------------|----------------|------------------------|--------------------------|
| Employee      | 220            | 260                    | 84.62%                   |
| College Student | 1,252         | 1,368                  | 91.52%                   |
| Senior High   | 777            | 918                    | 84.64%                   |
| TOTAL         | 2,249          | 2,546                  | 88.33%                   |

4.3. Compromised Passwords Per User Category
Table 2 compares the percentage of accounts per user category that were found to use compromised passwords. As described in Section 3, “compromised passwords” refer to passwords that are in the breach corpus of 320M passwords.

14% of Employee users were found to use compromised passwords, slightly below the combined (all user types) percentage of 16.72%. Employees, however, were also the smallest group of users, comprising only 10% of the recorded data. College students, the largest group in the data set (comprising 55% of the total recorded users), had a slightly higher percentage, at 17.17% found to be using compromised passwords. Senior High School students, at 16.73% with compromised passwords, is almost exactly in line with the overall percentage of 16.72%. They comprise about 35% of the recorded users. The researchers note that while employees aren’t too far behind the overall average of
users with compromised passwords, it is slightly reassuring that employees seem to have better password hygiene than students.

Table 2. Usage of compromised passwords per user category.

| User Type    | Recorded Users | Users with Compromised Passwords | % of Users with Compromised Passwords |
|--------------|----------------|-----------------------------------|---------------------------------------|
| Employee     | 220            | 31                                | 14.09%                                |
| College Student | 1,252          | 215                               | 17.17%                                |
| Senior High  | 777            | 130                               | 16.73%                                |
| TOTAL        | 2,249          | 376                               | 16.72%                                |

4.4. Other Interesting Results

We present in this section some statistics derived from auxiliary data.

4.4.1. Gender. Table 3 presents the percentage of compromised passwords per gender. Male users were found to be slightly worse than female users, with almost 18% of male users (higher than the population average of 16.72%) using compromised passwords. Female users were slightly better than the average, with about 15% using compromised passwords. (The lead author would note here that his male brethren has truly disappointed him.) There are four recorded users classified as “unknown” - this means the experiment apparatus was not able to properly retrieve gender information for four users, as their gender data was missing from APC’s information system. Of the four, two were found to use compromised passwords. Whether these four are all male, all female, or a mix, is mostly irrelevant now. Even if the 2 unknown accounts that did not have compromised passwords were male, and the 2 unknown accounts were female, the percentages will not change much. In that sort of best-case scenario for male users, the percentages would become 17.84% (male compromised) vs 15.32% (female compromised).

Table 3. Usage of compromised passwords per gender.

| Gender | Recorded Users | Users With Compromised Passwords | % of Users with Compromised Passwords |
|--------|----------------|----------------------------------|---------------------------------------|
| Male   | 1,248          | 223                              | 17.87%                                |
| Female | 997            | 151                              | 15.15%                                |
| Unknown| 4              | 2                                | 50.00%                                |
| TOTAL  | 2,249          | 376                              | 16.72%                                |

Do female users just naturally have better password hygiene? That’s an interesting question, but, aside from simply sharing our raw data, outside of the scope of this report.

4.4.2. Password Length. Majority of the APC population use passwords with 8-12 characters. 72% of the population are found within this range, with each length comprising over 10% of the population: 12% have 8 characters, 16% have 9 characters, 17% have 10 characters, 14% have 11, and 13% have passwords that are 12 characters long. The mean password length is 11.2 characters, while the mode is 10 characters. Only 1% of the population use less than 8 characters: 24 users have 7-character passwords, and 1 user has only 4 characters. Over 25% of users have 13 characters or more. 1% of the population have 18-24 characters in their passwords. These facts and figures are shown in Table 4.

5. Conclusion and Recommendations

5.1. Conclusion
This report details the results of an empirical study made in APC to measure the usage of known unsafe passwords found in breach corpuses. A custom-designed infrastructure that hooks into the
existing school information system collected properly anonymized data (non-confidential and non-personally-identifiable) regarding password usage of APC users (employees, college students, and senior high students). 88.33% of the active population was successfully recorded during the experiment period, which spanned most of the second trimester of school year 2017-2018.

Table 4. Password length distribution of APC users.

| Password Length | Number of Users | % of Total Population | Running % of Total Population |
|-----------------|-----------------|-----------------------|------------------------------|
| 4               | 1               | 0.04%                 | 0.04%                        |
| 5               | 0               | 0.00%                 | 0.04%                        |
| 6               | 0               | 0.00%                 | 0.04%                        |
| 7               | 24              | 1.07%                 | 1.11%                        |
| 8               | 275             | 12.23%                | 13.34%                       |
| 9               | 353             | 15.70%                | 29.04%                       |
| 10              | 377             | 16.76%                | 45.80%                       |
| 11              | 321             | 14.27%                | 60.07%                       |
| 12              | 295             | 13.12%                | 73.19%                       |
| 13              | 195             | 8.67%                 | 81.86%                       |
| 14              | 166             | 7.38%                 | 89.24%                       |
| 15              | 99              | 4.40%                 | 93.64%                       |
| 16              | 82              | 3.65%                 | 97.29%                       |
| 17              | 22              | 0.98%                 | 98.27%                       |
| 18              | 15              | 0.67%                 | 98.93%                       |
| 19              | 7               | 0.31%                 | 99.24%                       |
| 20              | 10              | 0.44%                 | 99.69%                       |
| 21              | 4               | 0.18%                 | 99.87%                       |
| 22              | 0               | 0.00%                 | 99.87%                       |
| 23              | 0               | 0.00%                 | 99.87%                       |
| 24              | 3               | 0.13%                 | 100.00%                      |

It was found that 1 in 6 users (16.72%) were using passwords found in the Pwned Passwords breach corpus provided by Troy Hunt. Students fared slightly worse than employees – only 14.09% of employee users used these unsafe passwords, compared to 17.17% for college students and 16.73% for senior high. Categorizing the users per gender also drew an interesting comparison – male users fared slightly worse than female users (17.87% vs 15.15%). Whether this means female users are inherently better is an interesting topic to explore, but outside the scope of this study.

5.2. Limitations of This Study

The current experiment infrastructure used is based solely on Troy Hunt’s trove of 320M breached passwords. It lacks commonsensical checks such as commonly-known passwords used in Filipino (the native language in the research environment of this study). It also lacks checks against patterns that are specific to the organization itself (for example, using “APC” in simple patterns). This means a black hat tasked with cracking passwords will likely have better a success rate than just 1 in 6 passwords. So, while a statistic of 1 in 6 unsafe passwords is already bad, the actual figure is most likely higher. It also bears noting that this study isn’t directly a measure of the prevalence of weak passwords. Unlike most previous password weakness studies, this study only measures inclusion in the Pwned Passwords list. There are certainly weak passwords that won’t necessarily be in that list (such as localized common words and organization-specific terminology or patterns).
5.3. Recommendations and Future Direction

The researchers encourage other organizations to conduct similar experiments and share their results to the community at large. Not only will this enable other organizations to be compliant with the newer NIST guidelines, this can also help shape commonsensical password policies. Creating a repository of unsafe passwords (with which to check user passwords against) based on Troy Hunt’s password list is a good start, but we recommend supplementing these with localized bad passwords (e.g., common Filipino passwords), as well as organization-specific bad passwords (i.e., passwords based on your organization or system name).

Will improved password policies and an aggressive information campaign result in minimizing the use of unsafe passwords found in breach corpuses? Going forward, the researchers plan to conduct another measurement during the following trimester (third trimester, school year 2017-2018) after conducting appropriate information campaigns regarding better password habits and developing newer password policies based on the latest recommendations of NIST. Stay tuned!

References

[1] Grassi P A, Garcia M E and Fenton J L 2017 NIST SP 800-63-3: Digital Identity Guidelines https://doi.org/10.6028/NIST.SP.800-63-3
[2] Grassi P, Fenton J, Newton E, Perlner R, Regenscheid A, Burr W, Richer J, Lefkovitz N, Danker J, Choong Y, Greene K and Theofanos M 2017 NIST SP 800-63B: Digital Identity Guidelines: Authentication and Lifecycle Management https://doi.org/10.6028/NIST.SP.800-63b
[3] Yan J, Blackwell A, Anderson R and Grant A 2004 Password memorability and security: empirical results IEEE Security & Privacy vol 2 no 5 pp 25-31
[4] Adams A, Sasse M A and Lunt P 1997 Making passwords secure and usable People and Computers XII ed H Thimbleby et al (London: Springer)
[5] Bryant K and Campbell J 2006 User behaviours associated with password security and management Australasian J. of Information Syst. vol 14 no 1 pp 81-100
[6] Shay R, Komanduri S, Kelley P, Leon P, Mazurek M, Bauer L, Christin N and Cranor L 2010 Encountering stronger password requirements: user attitudes and behaviors Proc. of the Sixth Symposium on Usable Privacy and Security 2010 p 2
[7] Florencio D and Herley C 2007 A large-scale study of web password habits Proc. of the 16th Int, Conf. on World Wide Web pp 657-66
[8] Stanton J, Stam K, Mastrangelo P and Jolton J 2005 Analysis of end user security behaviors Computers & Security vol 24 no 2 pp 124-33
[9] Egelman S, Sotirakopoulos A, Muslukhov I, Beznosov K and Herley C 2013 Does my password go up to eleven?: the impact of password meters on password selection Proc. of the SIGCHI Conf. on Human Factors in Computing Systems pp 2379-88