';--have i been pwned?

Check if you have an account that has been compromised in a data breach

DATA512 Project - Win Nawat Suvansinpan

pwned?

https://haveibeenpwned.com/ - Troy Hunt

Goals

- An exploration on the patterns of vulnerable passwords.
  - Categorizing them.
  - Investigate changes their “popularity” over the past year.
- An exercise on working with sensitive data without revealing them.

Dataset (~10GB)

- Pwned passwords sorted by count. Collected by Troy Hunt from data dumps. As new dumps happen, new passwords are added, existing passwords have their counts incremented.
- 2 versions
  - July 2019
  - July 2018
- Subset the first 1M rows to meet Github space cap.
- Data is released without license. Written permission is granted.

Hashed passwords	count
hashed_password_1	count_1
...	...
hashed_password_n	count_n

Hypothesis

Hypothesis

- Most vulnerable passwords belong to the following categories:

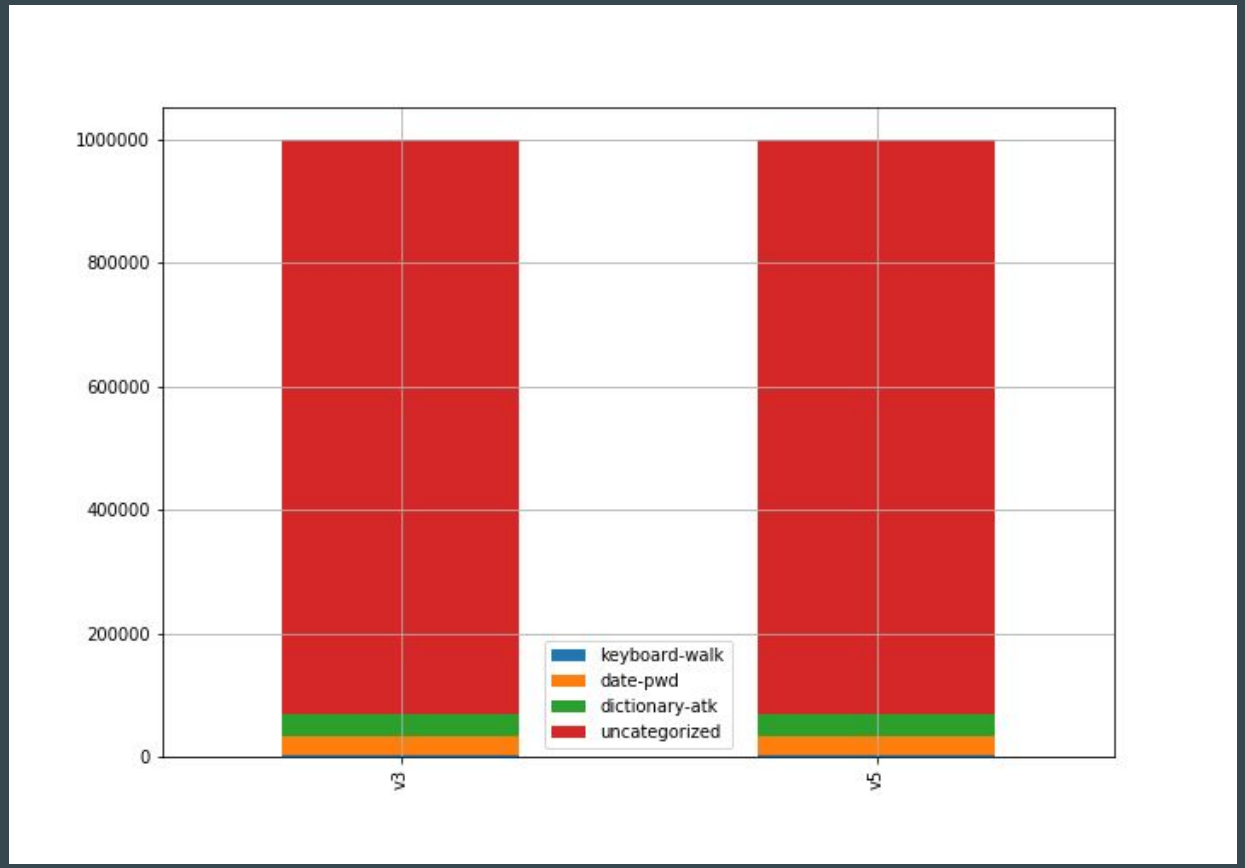
1. Keyboard walk (“asdfghjk”, “qwerty”)
2. Dates (“12052019”)
3. Dictionary attack (wordlist compiled by Openwall.com)
    a. Licensed for use and redistribute.

Questions

- What is the proportion of the passwords that can be classified under the categories above?
- How do the counts of the different categories of compromised passwords change within a year?

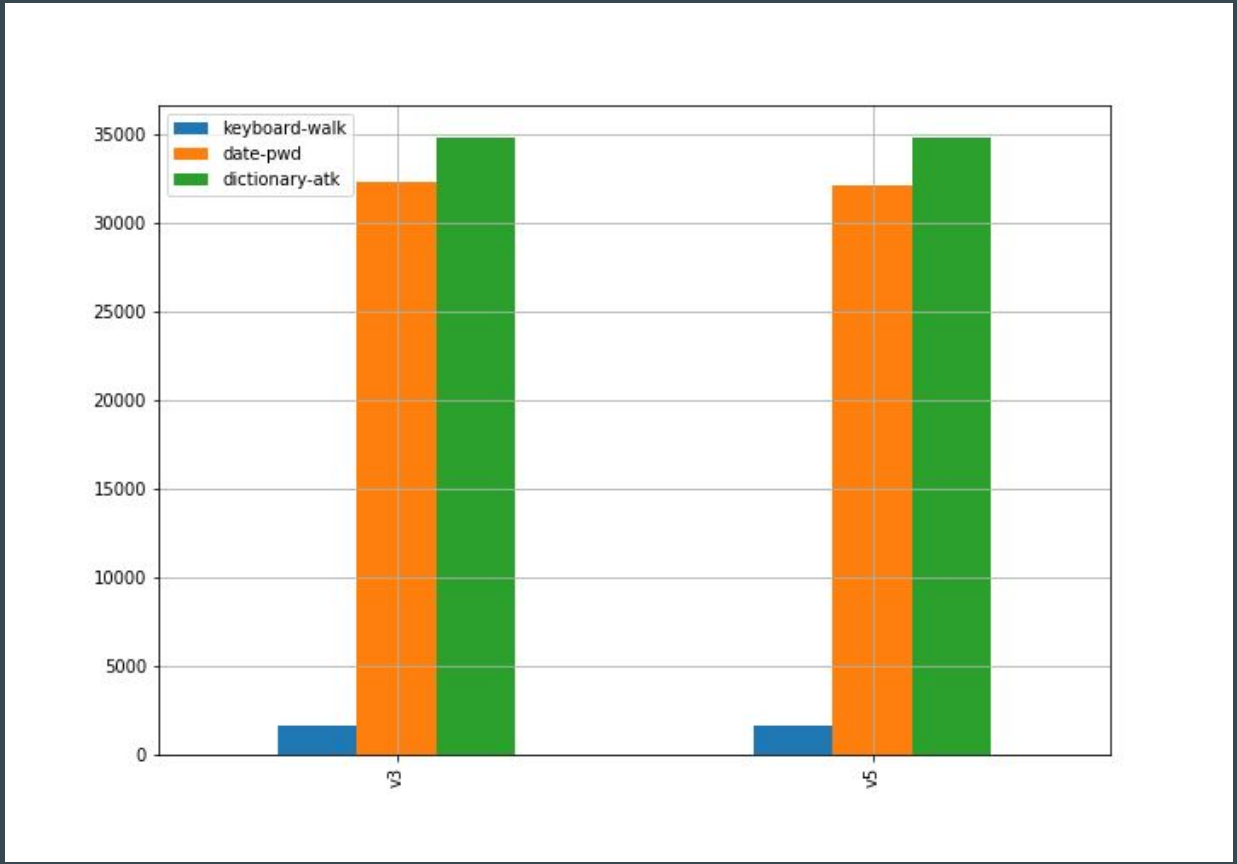
Results

- In progress!



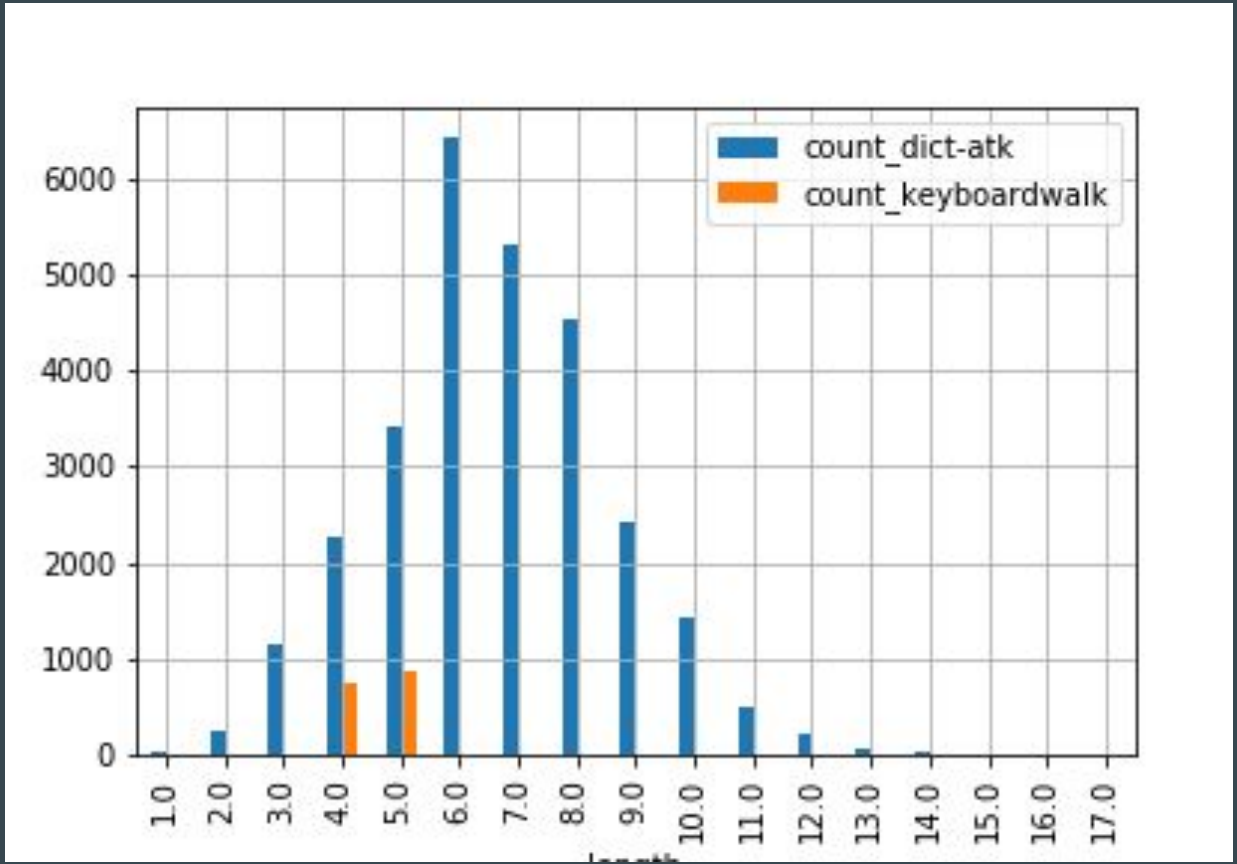
Results

- Just the uncovered passwords
- Some small decrease in count is observed.
- Means compromised passwords are more complicated



Results

- Length histogram
- Random walk passwords:
- Only length 4-5 are compared.



Conclusions

- The leaked passwords are unexpectedly complex.
- Unclear if there is a noticeable movement in each type of passwords.
- Dictionary attack appears to be the most effective so far.
- Most common password length is ~6
- Length of the password increases entropy exponentially!