

# ';--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

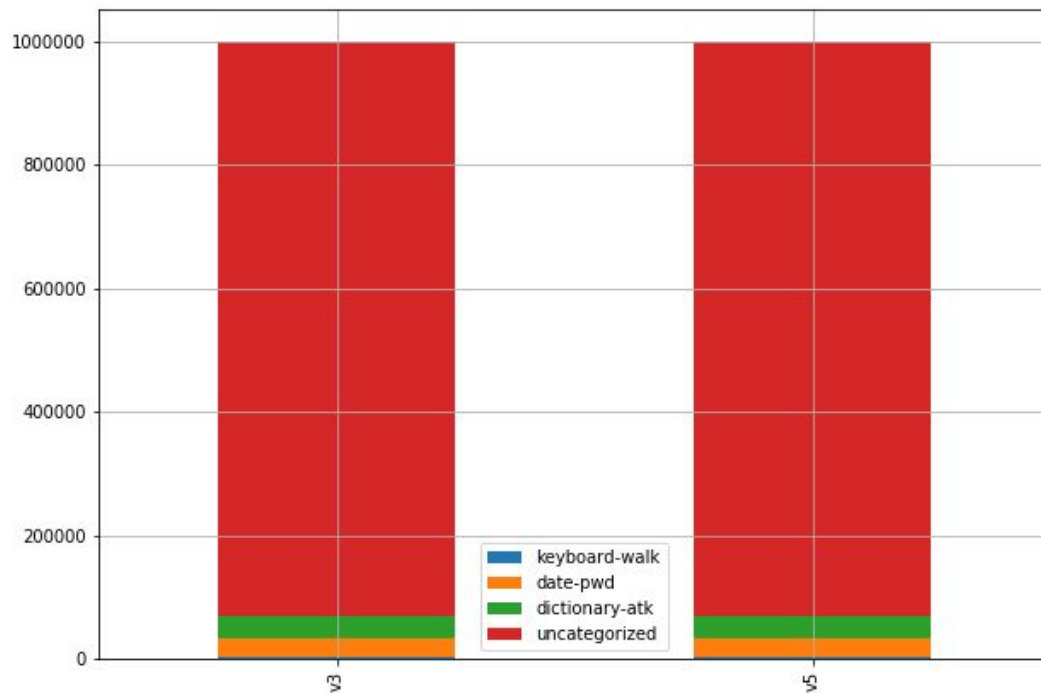
- Most vulnerable passwords belong to the following categories:
  1. Keyboard walk (“asdfghjk”, “qwerty”)
  2. Dates (“12052019”)
  3. Dictionary attack (wordlist compiled by Openwall.com)
    - a. Licensed for use and redistribute.

# 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?

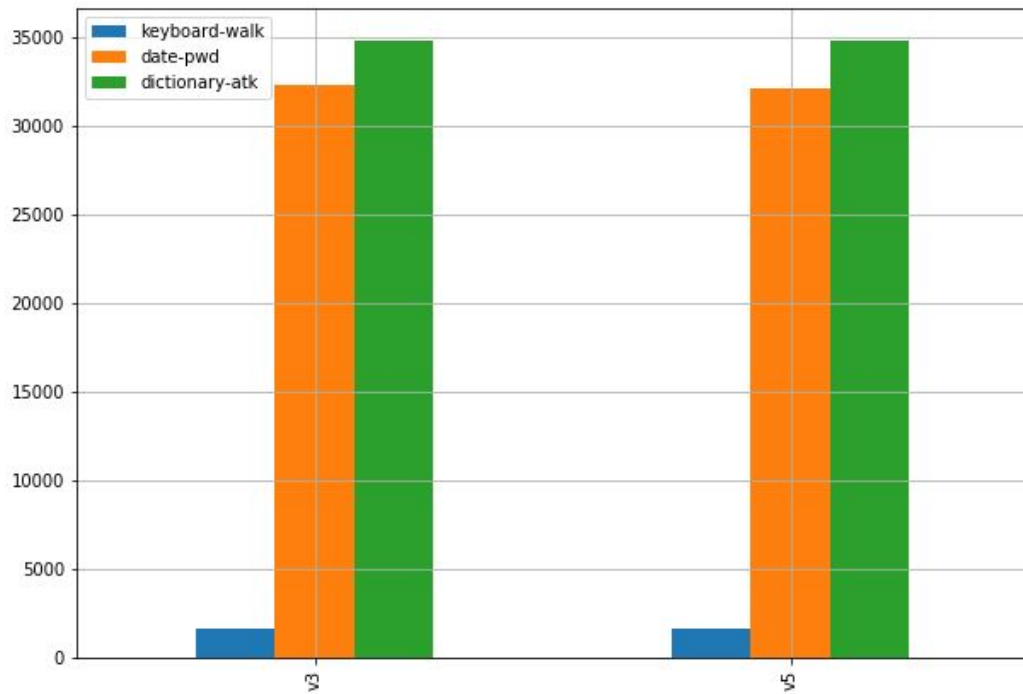
# Results

- In progress!



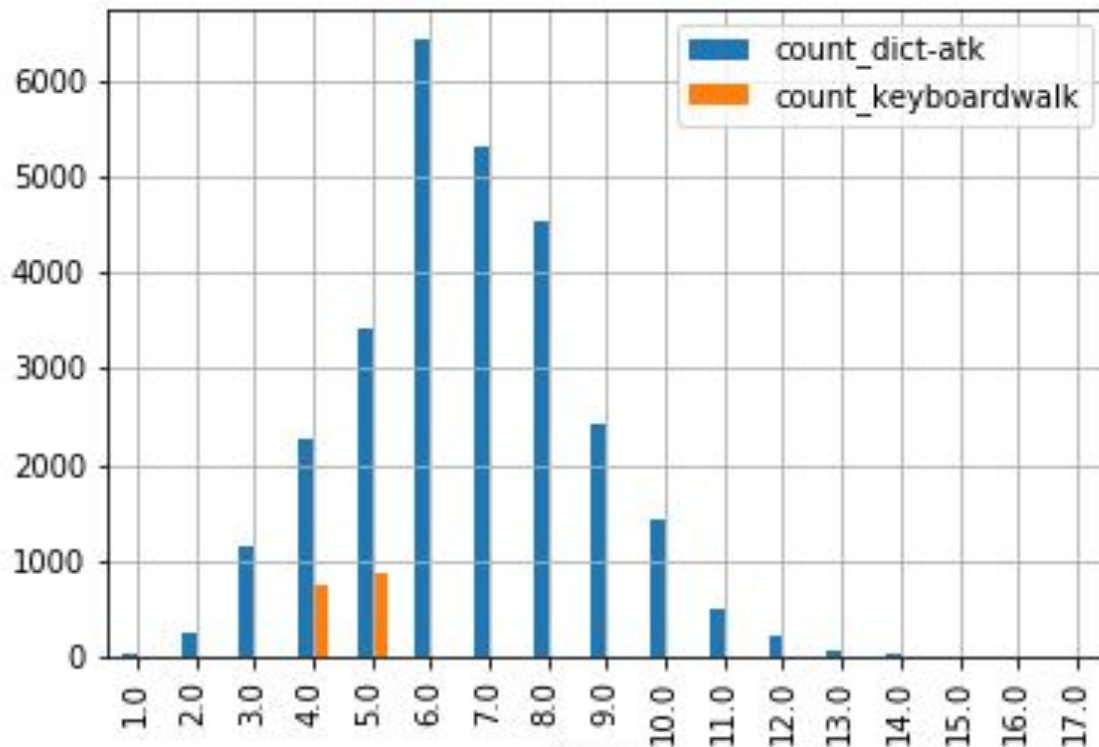
# Results

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



# Results

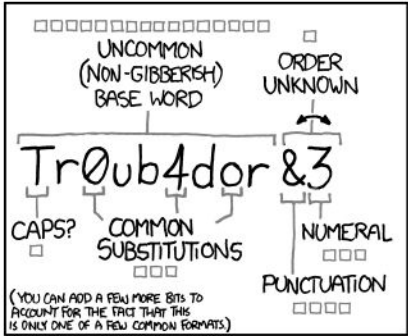

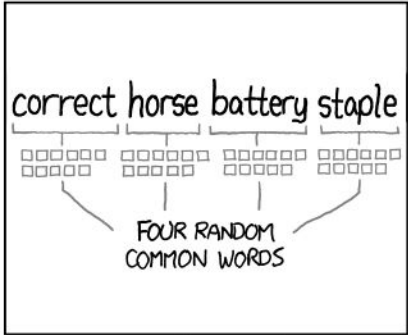
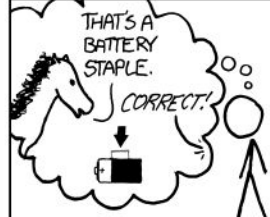
- Length histogram
- Random walk passwords:
- Only length 4-5 are compared.





# 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!

 <p>UNCOMMON (NON-GIBBERISH) BASE WORD</p> <p>ORDER UNKNOWN</p> <p>Tr0ub4dor &amp;3</p> <p>CAPS? COMMON SUBSTITUTIONS NUMERAL PUNCTUATION</p> <p>(YOU CAN ADD A FEW MORE BITS TO ACCOUNT FOR THE FACT THAT THIS IS ONLY ONE OF A FEW COMMON FORMATS)</p>	<p>~28 BITS OF ENTROPY</p> <p><math>2^{28} = 3 \text{ DAYS AT } 1000 \text{ GUESSES/SEC}</math></p> <p>(PLAUSIBLE ATTACK ON A WEAK REMOTE WEB SERVICE: YES, CRACKING A STOLEN HASH IS FASTER, BUT IT'S NOT WHAT THE AVERAGE USER SHOULD WORRY ABOUT.)</p> <p>DIFFICULTY TO GUESS: <b>EASY</b></p>	<p>WAS IT TROMBONE? NO, TROUBADOR. AND ONE OF THE 0s WAS A ZERO?</p> <p>AND THERE WAS SOME SYMBOL...</p>  <p>DIFFICULTY TO REMEMBER: <b>HARD</b></p>
 <p>correct horse battery staple</p> <p>FOUR RANDOM COMMON WORDS</p>	<p>~44 BITS OF ENTROPY</p> <p><math>2^{44} = 530 \text{ YEARS AT } 1000 \text{ GUESSES/SEC}</math></p> <p>DIFFICULTY TO GUESS: <b>HARD</b></p>	<p>THAT'S A BATTERY STAPLE.</p> <p>CORRECT!</p>  <p>DIFFICULTY TO REMEMBER: <b>YOU'VE ALREADY MEMORIZED IT</b></p>

THROUGH 20 YEARS OF EFFORT, WE'VE SUCCESSFULLY TRAINED EVERYONE TO USE PASSWORDS THAT ARE HARD FOR HUMANS TO REMEMBER, BUT EASY FOR COMPUTERS TO GUESS.