

';--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

Dataset - passwords that have been compromised

- 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 10,000 rows for each version to meet Github size cap.
 - 1M rows attempt did not end well.
- Data is released without license. Written permission is granted.

	hash	count
0	7C4A8D09CA3762AF61E59520943DC26494F8941B	23547453
1	F7C3BC1D808E04732ADF679965CCC34CA7AE3441	7799814
2	B1B3773A05C0ED0176787A4F1574FF0075F7521E	3912816
3	5BAA61E4C9B93F3F0682250B6CF8331B7EE68FD8	3730471
4	3D4F2BF07DC1BE38B20CD6E46949A1071F9D0E3D	3120735

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.

Hypothesis

Hypothesis

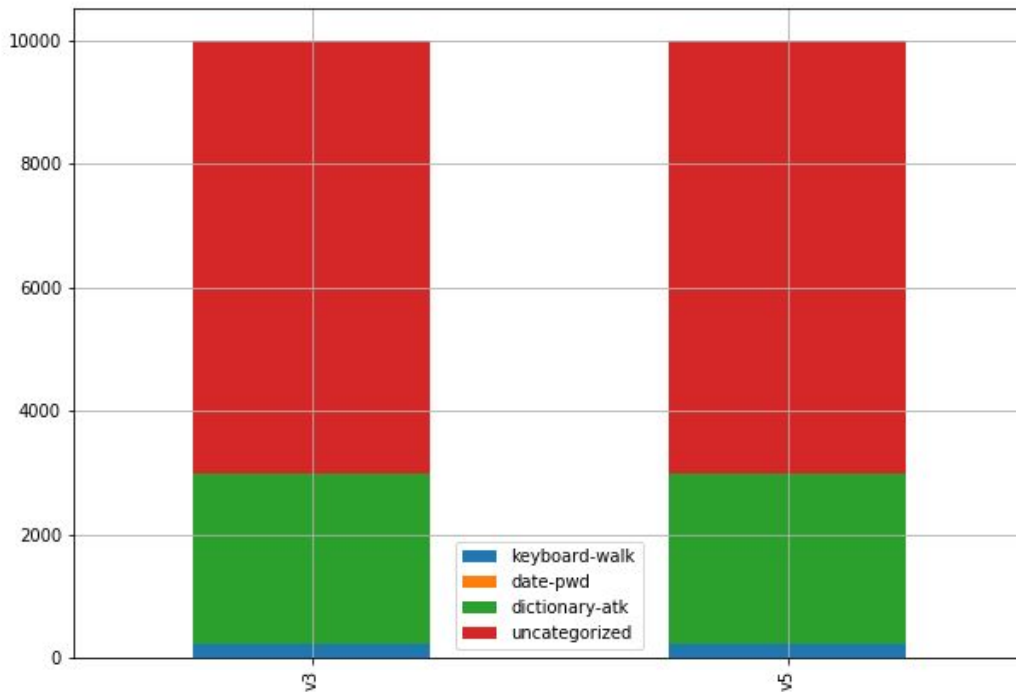
- Majority of vulnerable passwords belong to the following categories:
 1. Keyboard walk (“asdfghjk”, “qwerty”)
 2. Dates (“12052019”)
 3. Dictionary attack (“password” “admin” “welcome”)
 - i. wordlist compiled by Openwall.com
 - a. Licensed for use and redistribution.

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?
(Greater awareness on cybersecurity!)

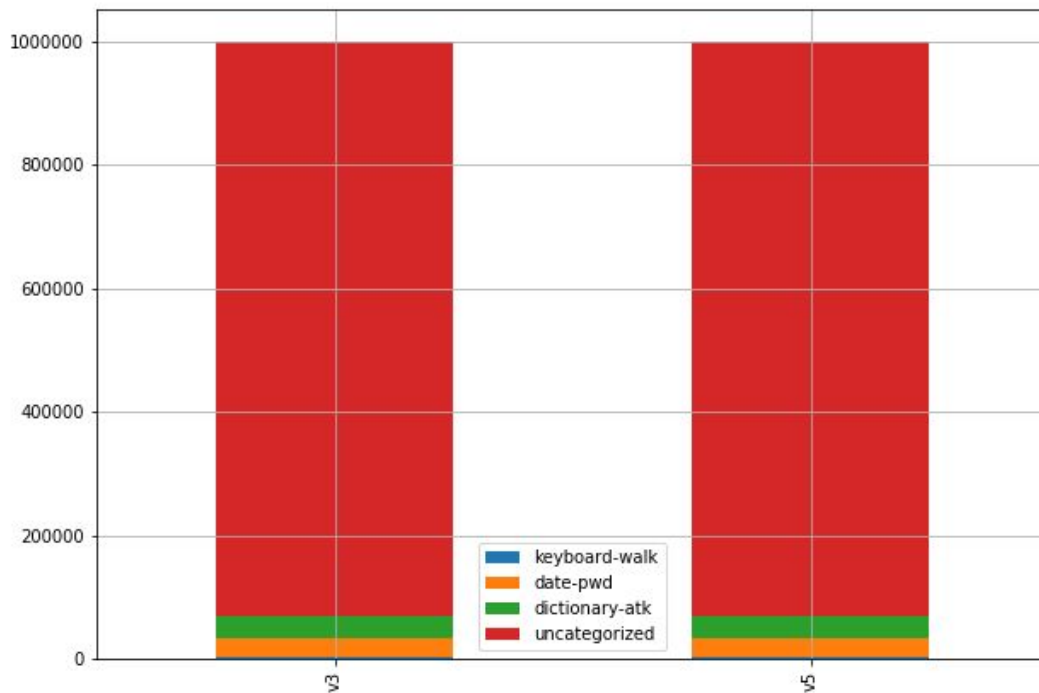
Results

- Underwhelming!
- Dictionary attacks cover about ~30% of the top 10,000.
- Keyboard-walk passwords are still in progress.
- (4-6 characters)



Results

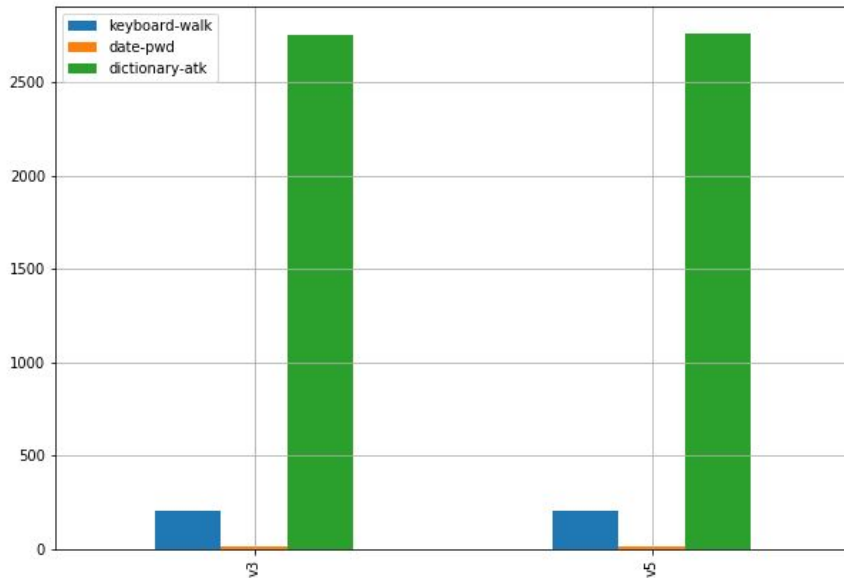
- Underwhelming!
- Keyboard-walk passwords are still in progress.



Results

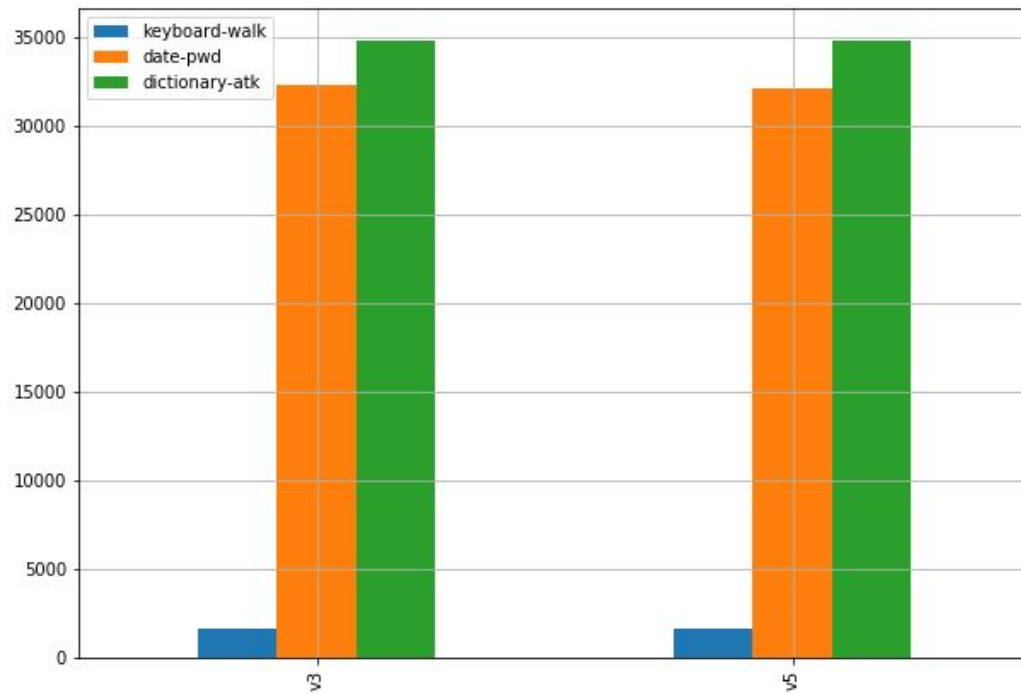
Just the uncovered passwords.

- Dominated by dictionary attacks.
- Date-format passwords are almost non-existent.
- No clear trends.
- BUT! Count is cumulative.
- Weak passwords are likely to remain on the list.
- To further look at the count delta between v3 and v5.



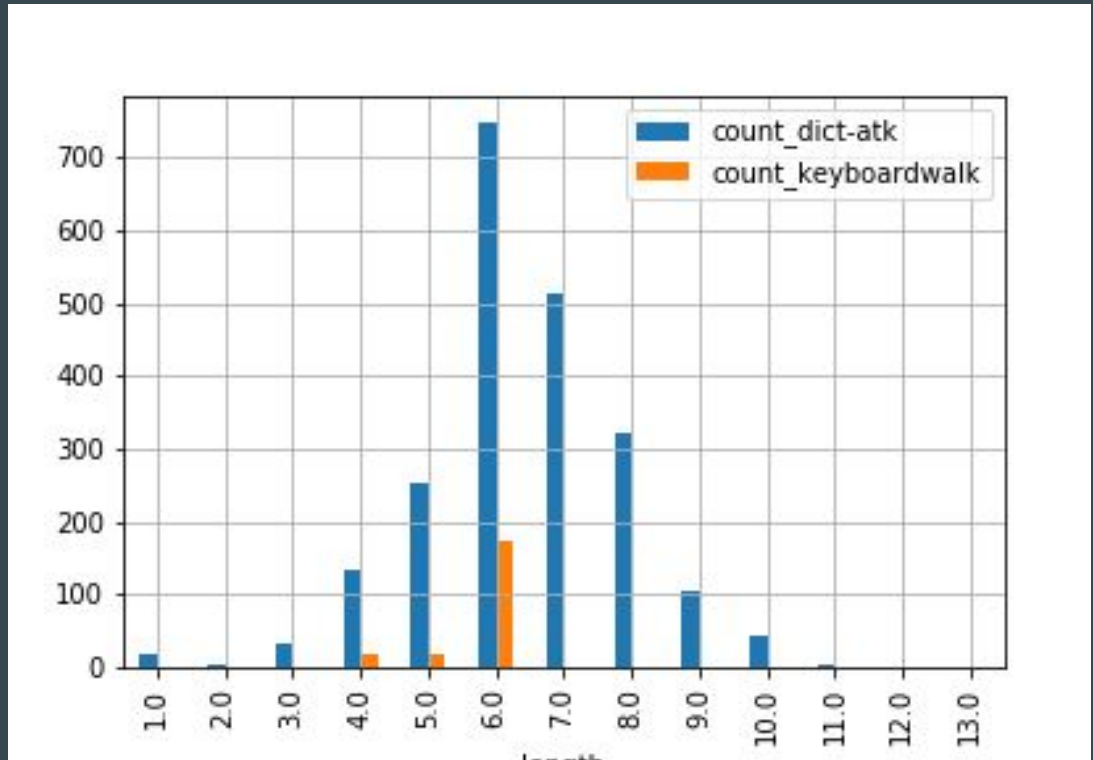
Results

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



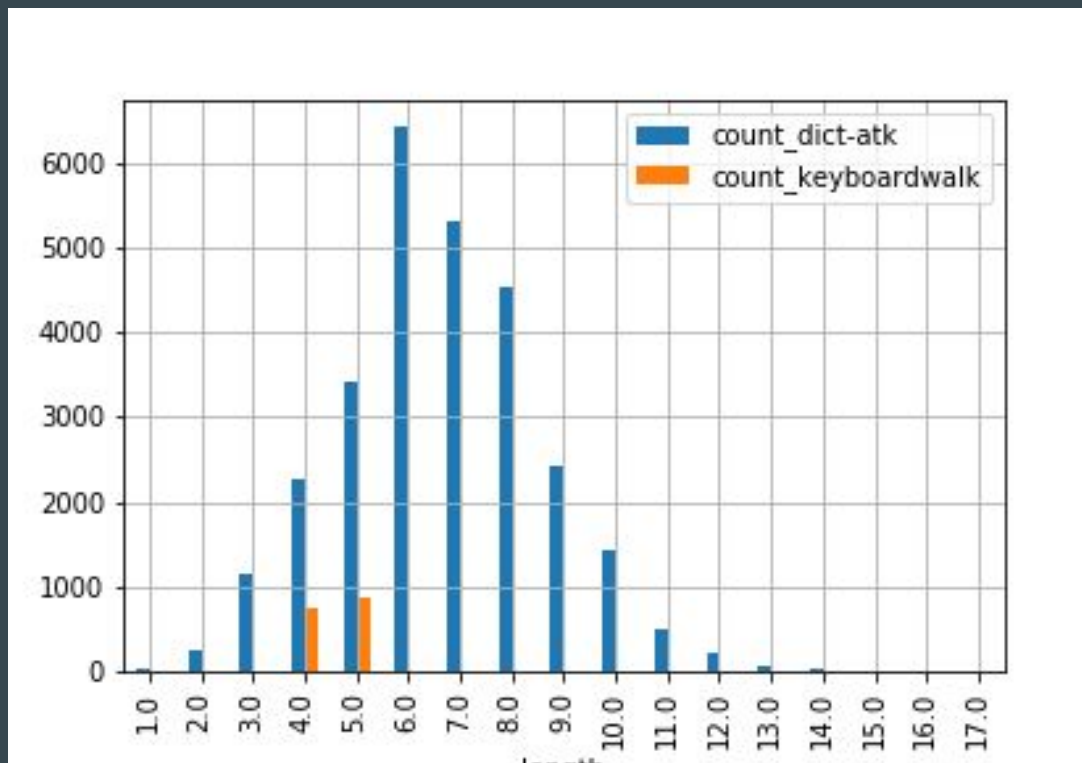
Results (bonus)

- Histogram of password length by categories.
- Random walk passwords: only length 4-6 are considered so far.
- “Date” type of passwords are omitted. They are always 8 characters long!



Results (bonus)

- Histogram of password length by categories.
- Random walk passwords: only length 4-5 are considered so far.
- “Date” type of passwords are omitted. They are always 8 characters long!



Conclusions

- The leaked passwords are unexpectedly complex.
 - No noticeable movement in each type of passwords.
 - Dictionary attack appears to be the most effective so far.
 - Most common length of compromised password is ~6.
 - Length of the password increases entropy exponentially, making brute force attacks REALLY hard.
- (First hand experience)

UNCOMMON (NON-GIBBERISH) BASE WORD

ORDER UNKNOWN

Tr0ub4dor &3

CAPS?

COMMON SUBSTITUTIONS

NUMERAL

PUNCTUATION

(YOU CAN ADD A FEW MORE BITS TO ACCOUNT FOR THE FACT THAT THIS IS ONLY ONE OF A FEW COMMON FORMATS.)

~28 BITS OF ENTROPY

$2^{28} = 3 \text{ DAYS AT } 1000 \text{ GUESSES/SEC}$

(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.)

DIFFICULTY TO GUESS: EASY

WAS IT TROMBONE? NO, TROUBADOR. AND ONE OF THE 0s WAS A ZERO?

AND THERE WAS SOME SYMBOL...

DIFFICULTY TO REMEMBER: HARD

correct horse battery staple

FOUR RANDOM COMMON WORDS

~44 BITS OF ENTROPY

$2^{44} = 550 \text{ YEARS AT } 1000 \text{ GUESSES/SEC}$

DIFFICULTY TO GUESS: HARD

THAT'S A BATTERY STAPLE.

CORRECT!

DIFFICULTY TO REMEMBER: YOU'VE ALREADY MEMORIZED IT

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.

Creative Commons Attribution-NonCommercial 2.5 License.

<https://xkcd.com/936/>