Jarod Gowgiel

Comp 116: Introduction to Computer Security

11 December 2018

PassGAN: Password cracking for the 21st Century

Abstract

The use of machine learning in a variety of technical fields has exploded in the past few years. Every few months, there seems to be another industry being disrupted by a new application of machine learning technology. Hospitals have begun using machine learning to aid technicians in identifying tumors and anomalies in x rays and other scans, with these applications often outperforming their human counterparts [1]. Social media corporations already use machine learning to provide targeted and relevant advertisements to their users, increasing the effectiveness of ads in general [2].

Notably, there have also been recent advancements in the use of machine learning programs to guess and crack encrypted passwords. Research in the past few years has shown that machine learning could be applied to improve established methods of guessing and cracking passwords. Experiments using PassGAN, a specific machine learning implementation, have shown a marked improvement over the most commonly used programs. This program can guess novel passwords that existing methods are not capable of generating without relying on costly brute force methods.

Introduction

Passwords are a critical first line of defense against attackers and towards securing data. Having strong, unique passwords is one of the easiest but often ignored steps toward protecting online information. According to most estimations [3][4][5] more than 60% of users reuse their

passwords across multiple sites. While easier to remember, this reuse of passwords means that any number of a users' accounts can be compromised by a single data leak. These leaks happen more often than one would think; just recently in late November of 2018, Marriott suffered a massive breach of up to 500 million customers' personal data [6]. Glancing through articles for this recent breach, one could be reassured by the fact that most of the stolen data was allegedly encrypted; but for any users who may have common or insecure passwords, this doesn't mean much.

Enter password cracking; the art of turning an encrypted password which resembles a random string into the original, plaintext key. As a brief example, consider the password "password". Once transformed with an encryption algorithm, MD5 in this example, it becomes the seemingly random string "5f4dcc3b5aa765d61d8327deb882cf99". With a proper hashing algorithm, there should be no way to turn this encrypted string back into the original plaintext. But attackers can try common guesses and compare them with stolen hashes, and if they guess correctly, they might as well have been given the password itself.

A number of programs have emerged over the years to automate and improve the process of trying and hashing incremental password guesses. The most famous among them are John the Ripper and Hashcat. Both employ similar methods to crack passwords: leveraging massive lists of known passwords, applying rule sets to append numbers and letters, or using statistical distributions to guess the most likely combinations of characters. Depending on the speed of an attacker's computer, millions of passwords can be hashed and compared every second. These methods are well understood and have led to a number of new suggestions for strong passwords; longer passwords are harder to guess using these methods, as are those with multiple types of characters and symbols.

To the Community

Recently, new approaches have been proposed that could be more efficient than these techniques and work in tandem with Hashcat and John the Ripper to improve the success rate of password cracking. The most recent implementation, PassGAN, uses a novel type of machine learning called "generative adversarial networks" to guess user passwords based on an input dataset of known passwords.

A generative adversarial network begins with no knowledge of the data it is trying to replicate. Over successive iterations, a generator receives feedback from a discriminator which tries to determine whether a password was produced by the generator or is from the original dataset. The generator gets better and better at imitating the target dataset as the discriminator provides feedback over hundreds of thousands of cycles.

The results speak for themselves; when Hashcat was used with the output dictionary from PassGAN, the number of cracked passwords increased by up to 50% compared to Hashcat running on the starting dataset using the original dictionary with a chosen rule set. After generating 10^8 new password guesses, PassGAN retrieved an additional 656,322 passwords from the LinkedIn data breach that Hashcat could not recover using a set of built in rules [7]. This represents an additional half a million accounts that could have been compromised in the wake of this data breach. Since the training data for a machine learning program can be used without retraining before every application, computational overhead could be invested before a leak to train the model on a large dataset, and then the resulting model could be quickly applied to guess passwords once a breach occurs.

Having a working understanding of how this emerging technology could affect and improve password cracking is important so that password rules and suggestions can be adjusted accordingly. The use of machine learning to guess passwords can be applied to evaluate the

strength of user passwords, giving a better baseline level of strength to all users whose passwords are in line with new requirements. It is also important for security-conscious people to understand how to apply these new methods to audit existing passwords, to ensure that passwords previously considered secure still hold up to emerging attacks.

Applications

One of the most promising applications of PassGAN is as a dictionary-extending tool. Given a set of known passwords, PassGAN can be used to generate additional passwords that are similar to those found in the known dictionary. These extended datasets could be used during security testing when trying to audit the strength of known user passwords. While many stolen password databases are publicly available, even more are held as closely guarded secrets by government agencies and individual attackers. Without access to these lists, using an enhanced dictionary to guess passwords could still reveal potentially weak passwords.

Another application for machine learning and password guessing is in the creation of live password-checking tools for website sign up forms. Using the model generated from a neural network, researchers have proposed a strength testing implementation that uses a proposed password as input. Rather than just requiring that a password must contain a certain number of capital letters or symbols, these implementations could be used to evaluate passwords using characteristics of known insecure passwords and assign a more deterministic “security score” [8]. While requiring certain password lengths or a certain number of symbols is a good first step, these additional requirements are often met in predictable ways. Using the output from a trained neural network to propose new rules, however, could lead to increasingly secure passwords from all users.

Conclusion

In the past few years, research has begun emerging showing how machine learning could be applied to password cracking. While machine learning has been applied with great success to other fields, including medicine and marketing, its use as a password cracking approach is still in the nascent stages. However, initial results seem promising. Using the PassGAN implementation in tandem with Hashcat resulted in a 50% increase in cracked passwords. As further research is done, and more implementations are created, these machine learning approaches have the potential to open up new realms of password-cracking possibility.

Having strong, unique, and unrepeated passwords is a critical step towards securing your online data for any user. While many of the password rules that appear commonly across the web take into account current methods of attacking user passwords, they do not take advantage of the most innovative and cutting-edge technologies available for the purpose. If attackers begin to take advantage of neural networks and machine learning to crack passwords, standards must evolve, and security professionals must be aware of the incoming changes. Using these same methods, both to audit the strength of user passwords and improve password strength verification, could be a worthwhile step for anyone concerned with online safety and security, especially in critical applications.

References

[1] Hu, Zilong, et al. “Deep Learning for Image-Based Cancer Detection and Diagnosis − A Survey.” Pattern Recognition, vol. 83, 2018, pp. 134–149.

[2] “Machine Learning for Advertising – AWS Answers.” *Amazon*, Amazon, aws.amazon.com/answers/machine-learning/machine-learning-advertising/.

[3] Truta, Filip. “59% Of People Use the Same Password Everywhere, Poll Finds.” *Security Boulevard*, 3 May 2018, securityboulevard.com/2018/05/59-of-people-use-the-same-password-everywhere-poll-finds/.

[4] Daileda, Colin. “Basically Everyone Reuses Their Passwords.” *Mashable*, Mashable, 1 Mar. 2017, mashable.com/2017/02/28/passwords-reuse-study-keeper-security/#X90u.CWSA8qS.

[5] Dellinger, AJ. “Poor Password: 92 Percent Of Millennials Reuse Login Security Identification.” *International Business Times*, 21 July 2017, www.ibtimes.com/poor-password-92-percent-millennials-reuse-login-security-identification-2568066.

[6] Ortiz, Erik. “Marriott Says Data Breach Compromised Info of up to 500 Million Guests.” *NBCNews.com*, NBCUniversal News Group, 30 Nov. 2018, www.nbcnews.com/tech/security/marriott-says-data-breach-compromised-info-500-million-guests-n942041.

[7] Hitaj, Briland, et al. “PassGAN: A Deep Learning Approach for Password Guessing.” 2017.

[8] Melicher, William, et al. “Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks.” *The Advanced Computing Systems Association*, 10 Aug. 2016.