Poster: An Analysis of Targeted Password Guessing Using Neural Networks

Huan Zhou, Qixu Liu, Fangjiao Zhang
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
liuqixu@iie.ac.cn

***Abstract*—Text-based passwords, dominant mechanism of authentication nowadays, are vulnerable to malicious attackers. Even though not recommended, users tend to use personal information (PI) when create passwords. Only a few studies have researched targeted password guessing, in which attackers guess passwords by utilizing users' PI. We propose TPGXNN, a framework that uses neural networks (NN) in targeted password guessing. The recent success applying NN to sequential data issues makes them a viable candidate on the task of password generation. Our experiments on 8 abundant real-world password sets initially demonstrate the important role of PI in password construction and the effectiveness of TPGXNN.**

Keywords—Targeted password guessing; Personal information; Neural networks.

I. INTRODUCTION

Text-based passwords still solidly remain the most prevailing method for user authentication in diverse computer systems. Even though people have raised various authentication mechanisms, no alternative can bring all the benefits of passwords without introducing any extra burden to users [1]. Owing to human-memorability demand, users are often prone to choosing popular passwords or using personal information to create passwords merely because they are convenient to memorize. Consequently, many passwords are selected within only a slight percentage of the whole password space, making users defenseless to guessing attacks.

To systematically study password security, some probabilistic guessing models, including probabilistic context free grammars (PCFGs) and Markov n-grams, have been put forward in succession. A typical characteristic of these probabilistic models is that they distinguish a trawling offline guessing hacker that chiefly performs against data breaches and intends to crack as many passwords as possible. Unfortunately, these models neglect some common behaviors when users create their passwords and are computationally intensive, requiring lots of disk space. Thus, they are generally not suitable for evaluation of password strength in practice, and sometimes have low accuracy in password guessing. Trawling online guessing primarily utilizes users' habit of choosing popular passwords. Nevertheless, targeted online guessing can take advantage of weak popular passwords as well as passwords containing personal information.

With the target of analyzing how users employ their own personal information (PI) to construct passwords and measuring the strength of text-based passwords more practically and precisely, we use more than 300 million leaked passwords which are publicly available from various popular websites. Our experiments on 8 abundant realistic password sets reveal the role of users' PI in password construction. Moreover, we propose a framework called TPGXNN which uses neural networks (NN) in targeted password guessing. Recurrent neural networks (RNN) have been proved to effectively generate novel sequences [2] and [3] applied very deep convolutional neural networks (VDCNN) in text processing successfully, so we initially choose them as our candidates on the task of targeted password guessing.

II. OVERVIEW

A. Our Datasets

Our experiments rely on eight enormous realistic password sets (see TABLE I), containing four from Chinese websites and four from English websites. They are revealed publicly online which were attacked by hackers or exposed by insiders, and a few of them have been used in large empirical analysis of passwords. Overall, these datasets comprise 355.6 million plain-text passwords and lay over a variety of popular online services.

TABLE I. INFORMATION OF OUR 8 DATASETS

Dataset	Language	Contain PI	Leak Time	Amount
LinkedIn	English	✘	May, 2016	1,000,000
Yahoo	English	✘	July, 2012	453,427
Fling	English	✓	2011	40,767,652
Neopets	English	✓	May, 2016	26,892,897
NetEase	Chinese	✘	Oct., 2015	234,842,089
Tianya	Chinese	✘	Dec., 2011	29,020,808
GFAN	Chinese	✘	Oct., 2016	22,526,334
12306	Chinese	✓	Dec., 2014	129,303

We use 1 million plain-text password cracked from 164 million LinkedIn passwords which were stored as SHA1 hashes without salt when originally hacked in 2012.

As far as we know, our collection is the biggest ever gathered for assessing the security danger of targeted password guessing.

B. *Personal Information In Passwords Construction*

Since some password lists have no PI, we correlate them with three datasets (Fling, Neopets, 12306) in TABLE I and four auxiliary PI datasets (ClixSense, Experian, Hotel, 51Job) containing PI of the same language through matching email. Consequently, eight PI-related password sets are generated.

Moreover, we adopt a brand-new metric to precisely gauge the extent of PI engagement in passwords in a precise and organized manner. Mathematically we have

$$Relevance = \log_2(1 + \sum_{i=0}^{n} \frac{{len_i}^3}{{len_{total}}^3}), \quad (1)$$

where n stands for the count of matched password sections, len_i represents the length of the relevant matched password section, and len_{total} indicates the whole password's length.

C. *TPGXNN*

We propose a targeted password guessing framework called TPGXNN. Our framework trained with users' password datasets and PI uses neural networks to calculate the probability of a password. As depicted in Fig. 1, TPGXNN contains three stages (i.e. data preparing, training and guessing). The realization of preparing and guessing stages is unambiguous, and the major effort lies in the training stage. We initially choose LSTM and VDCNN as our candidates on the task of targeted password guessing since they have been proved to be very effective at processing text.

We establish our TPGXNN on the Keras library. Our realization trains networks and guesses passwords employing Python programming language. We take the leaked data sets and users' PI as inputs and use transference learning method to train our neural networks.

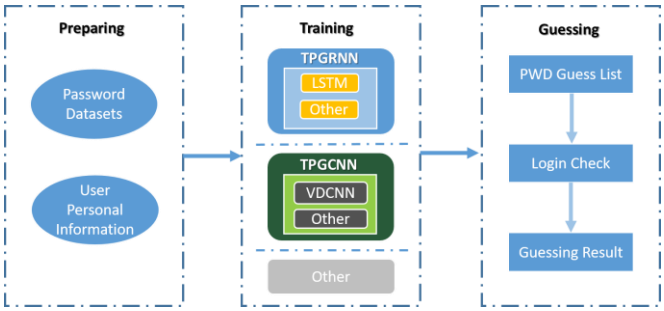


Fig. 1. An architecture overview of TPGXNN

III. PRELIMINARY RESULTS

A. *PI Role In Passwords*

We present in TABLE II how often users utilize their PI to construct passwords. As expected, PI plays an important role in the construction of passwords because of the high total proportion of different types of PI and the large amount of passwords, and users especially like to use name, birthday, username and email prefix when constructing passwords. Thus, in the training process we only use the above four kinds of PI.

TABLE II. PI ROLE IN PASSWORD CREATION

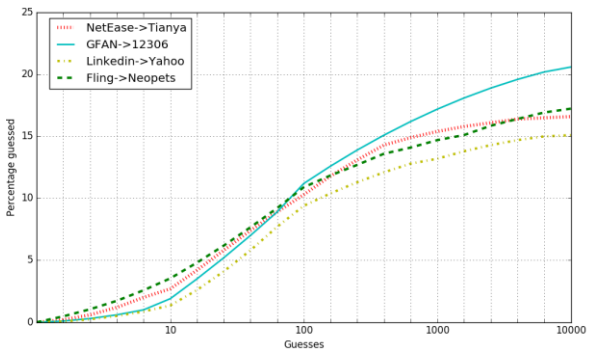
PI	Linkedin	Yahoo	Fling	Neopets	NetEase	Tianya	GFAN	12306
Name	3.41	2.07	4.35	1.47	14.32	19.21	10.34	18.15
Birthday	1.47	1.85	2.56	0.96	23.12	17.93	29.45	24.02
Username	2.79	2.83	3.96	2.64	1.43	1.07	0.76	1.96
Email prefix	1.34	2.47	0.65	3.58	6.12	4.36	5.15	3.03
Phone	0.24	0.96	1.13	0.28	0.84	0.35	0.10	0.07
NID	-	-	-	-	1.56	1.89	0.91	0.46
Address	0.00	0.04	0.13	0.09	-	-	-	-
Gender	0.62	0.58	0.04	0.96	0.96	0.07	0.14	0.12
IP Address	0.00	0.00	0.00	0.00	-	-	-	-
Relevance AVG	0.17	0.13	0.21	0.18	0.26	0.31	0.29	0.35

Decimals in the table use '%' as the unit except value of Relevance AVG.

B. *Cracking Results Using TPGXNN*

We divided the eight data sets into four groups in the course of experiment. Two groups belong to English users and the other two belong to Chinese users. The purpose of doing so is to ensure the success rate of guessing. In each group, one of the data sets is trained as a training collection, and the other one is used as a testing collection to examine the success rate of password guessing.

Fig. 2. Cracking results using TPGXNN



As shown in Fig. 2, TPGXNN is able to crack password effectively with limited guess numbers. Our preliminary results also show that it outperforms its foremost counterpart by 10.65%. In the future work, we will optimize our framework and go deep into targeted password guessing using neural networks.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their helpful comments for improving this paper. This work is supported by the Beijing Municipal Science & Technology Commission (No. Z161100002616032).

REFERENCES

[1] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in Proc. IEEE S&P, 2012.

[2] GRAVES, A. "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[3] Alexis Conneau, Holger Schwenk, Lo¨ıc Barrault, and Yann Lecun, "Very deep convolutional networks for natural language processing," arXiv preprint arXiv:1606.01781, 2016.