Available online at: http://www.ijmtst.com/vol3issue12.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 03, Issue No: 12, December 2017

Improvement on Traditional LDS Structures to Crack the Passwords Based on Individual's Information

P.Sasirekha[1] | S.Varadhaganapathy[2] | J.Premalatha[3] | A.Jeevanantham[4] | C.Visali[5]

[1,5]PG Scholar, Department of Information Technology, Kongu Engineering College, Perundurai-638060, Erode, Tamilnadu, India.

[2,3]Professor, Department of Information Technology, Kongu Engineering College, Perundurai-638060, Erode, Tamilnadu, India.

[4] Assistant Professor,, Department of Information Technology, Kongu Engineering College, Perundurai-638060, Erode, Tamilnadu, India.

To Cite this Article

P.Sasirekha, S.Varadhaganapathy, J.Premalatha, A.Jeevanantham and C.Visali, "Improvement on Traditional LDS Structures to Crack the Passwords Based on Individual's Information", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 12, December 2017, pp.-39-45.

ABSTRACT

While it is not favored, internet users wish to include personal information in their passwords for easy recognition. In this paper, we disjoint user passwords from a Chinese website to inspect the intensity to which personal information resides in a password. Then we bring out a new measurement called Depth to find the interrelation between personal information and passwords. Afterwards, based on our analysis, we propose Individual-PCFG as a supplement to the Probabilistic Context-Free Grammar method to be semantics-rich. We manifest that Individual-PCFG cracks password much faster than PCFG by generating personal information related guesses and makes online attacks much more faster to succeed.

Keywords: *Individual information, Depth, Password cracking*

I. INTRODUCTION

Authentication is a mechanism introduced to validate the identity of the users based on the comparison of data provided by the user during validation and the already existing data that was stored at the time of enrolment. The types of authentication includes password based authentication, biometric based authentication, etc. The biometric based authentication includes physical biometrics that consists of facial-scan, finer-scan, retina-scan, iris-scan. The behavioral biometrics includes voice-scan, signature-scan. Although biometric scan ensures better results, text based password authentication is dominating and it could still be irreplaceable by other authentication mechanisms.

In personal and public networks, authentication is commonly carried out in the shape of login IDs and passwords. The awareness of the login credentials is used to assure that the user is real. Each person registers first of all, using one's very own password. Password electricity meters courses the customers to set the sturdy password but still it is inconsistent and advert-hoc. On every subsequent use, the user have to keep in mind and use the formerly registered password.But password-based totally authentication isn't always encouraged to provide adequately robust safety for

a gadget that contains touchy statistics. Usernames are frequently a mixture of individual's private information which makes the attackers easy to wager. If the guidelines are not imposed through password meters, humans frequently create susceptible passwords. For this reason, net commercial enterprise and lots of different confidential transactions require a greater strict and correct authentication technique. A text based password is a combination of characters used to prove identity or access approval to gain access to to a resource. Text-based passwords still remain a dominating and irreplaceable authentication method in the foreseeable future. Although people have proposed different authentication mechanisms, no alternative can bring all the benefits of passwords without introducing any extra burden to users [1]. Problems associated with biometrics based authentication are that there is a chance of false positives and false negatives. Chances are there that a valid user is rejected and an invalid user is accepted. Often people are not comfortable with this type of authentication.

However, passwords have long been criticized as one of the weakest links in authentication. Due to human-memorability requirement, user passwords are usually far from true random strings [2]–[6]. In other words, human users are prone to choosing weak passwords simply because they are easy to remember. As a result, most passwords are chosen within only a small portion of the entire password space, being vulnerable to brute-force and dictionary attacks.

To better assess the strength of passwords, we need to have a deeper understanding on how users construct their passwords. If an attacker knows exactly how users create their passwords, guessing their passwords will become much easier. Meanwhile, if a user is aware of the potential vulnerability induced by a commonly used password creation method, the user can avoid using the same method for creating passwords.

Traditional dictionary attacks on passwords have shown that users tend to use simple dictionary words to construct their passwords [7]. Language also plays a vital role since users tend to use their first languages when constructing passwords [2]. Besides, passwords are mostly phonetically memorable [4] even though they are not simple dictionary words. It is also indicated that users may use keyboard and date strings in their passwords [5], [8], [9]. However, most studies discover only superficial password patterns, and the semantic-rich composition of passwords is still mysterious to be fully uncovered. Our work investigates how users generate their passwords by learning the semantic patterns in passwords.

In this paper, we study password semantics from the use of personal information. We utilize a leaked password dataset, which contains personal information, from a Chinese website for this study. We first measure the usage of personal information in password creation and present interesting observations. We are able to obtain the most popular password structures with personal information embedded. Next, we introduce a new metric called Depth to accurately estimate the correlation between personal information and user password. Since it considers both the length and continuation of personal information in a password, Depth is a useful metric to measure the strength of a password. Our quantification results using the Depth metric confirm our direct measurement results on the dataset, showing the efficacy of Depth. Moreover, Depth is easy to be integrated with existing tools, such as password strength meters for creating a more secure password. As long as memorability plays an important role in password creation, the correlation between personal information and user password remains, regardless of which language users speak. We believe that our work on personal information quantification, password cracking, and password protection could be applicable to any other text-based password datasets from different websites.

To demonstrate the security vulnerability induced by using personal information in passwords, we propose a semantics rich Probabilistic Context-Free Grammars (PCFG) method called Individual-PCFG, which extends PCFG [11] by considering those symbols linked to personal information in password structures. Individual-PCFG is able to crack passwords much faster than PCFG. It also makes an online attack more feasible by drastically increasing the guess success rate.

II. LABELING OF PERSONAL INFORMATION IN PASSWORDS

A) Introduction to Dataset

A leaked password dataset, which includes personal information from a chinesewebsite had been applied. It's miles referred to as 12306 dataset due to the fact that all passwords are from a website www.12306.cn, that's the respectable website online of the web railway ticket reservation

in China. There's no statistics available on the exact number of customers of the 12306 website but we infer as a minimum tens of hundreds of thousands of registered customers inside the device, since it is the legit internet site for the complete chinese railway reservation. The dataset consists of more than a millions of chinese language passwords which might be plaintext passwords, and it also includes several forms of personal statistics, along with a user's name and the authorities-issued particular identification wide variety. As the website requires a actual id range to sign up and people need to provide correct non-public facts to e book a price ticket, the statistics in this dataset is considered dependable and accurate.

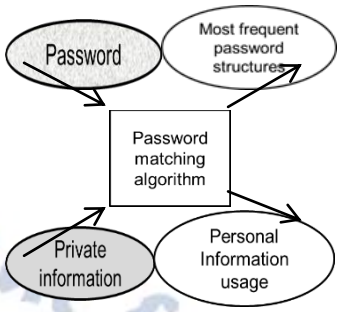
We first perform a simple analysis to expose some trendy characteristics of the 12306 dataset. For information consistency, we remove users whose id range is not 18-digit long. These users may additionally have used different IDs to check in on the device .The dataset consists of 1000 passwords is taken for analysis after being cleansed.

TABLE 1
Personal Information

Type	Description
Name	User's Name
Email address	User's Registered Email address
Mobile number	User's Registered Mobile Number
Account Name	The username used to log in to the system
ID Number	Government issued ID number

B) Personal Information

12306 dataset also includes multiple types of personal information. The personal information is tabulated in TABLE 1.The government-issued ID number is a unique 18-digit number, which intrinsically includes the owner's personal information. Specifically, digits 1-6 represent the birthplace, digits 7-14 represent the birthdate, and digit 17 represents the gender—odd being male and even being female. The 8-digit birthdate is treated separately since it is a critical piece of personal information in password creation. Thereby, six types of personal information are considered: name, birthdate, email address, cell phone number, account name, and ID number (birthdate excluded). The algorithm process is figured in fig 1.



1.New Password Representation

To illustrate the personal information correlation with user passwords, a new representation of a password is developed by adding more semantic symbols besides the conventional "D," "L," and "S" symbols, which stand for digit, letter, and special character, respectively. The password is first matched to the six types of personal information under this new representation. For example, a password "sasi1994xyz" can be represented as [Name][Birthdate]L_3, instead of $L_3D_4L_3$ as in the traditional representation. The matched personal information is denoted by corresponding tags—[Name] and [Birthdate] and for segments that are not matched, use "D," "L," and "S" to describe the symbol types. Representations like [Name][Birthdate]L_3 are more accurate than $L_4D_4L_3$ in describing the composition of a user password by including more detailed semantic information.

2. Matching Process

A matching process is proposed to locate personal information in a user password. The idea is to generate all substrings of the password and sort them in descending length order. Then match these sub strings from the longest to the shortest, to all types of personal information. If a match is found, the match function is recursively applied over the remaining password segments until no further matches are found. The segments that are not matched to any personal information will still be labeled using the traditional "LDS" tags.

III DEPTH ESTIMATION

A novel metric called Depth is introduced to evaluate the involvement of personal information in the creation of an individual password in an accurate and systematic fashion. The value of Depth ranges

from 0 to 1. A larger Depth implies a stronger correlation. Depth"0" means no personal information is included in a password, and Depth"1" means the entire password is perfectly matched with one type of personal information. Though Depth is mainly used for measuring an individual password, the average Depth also reflects the degree of correlation in a set of passwords. The Depth algorithm process is provided in fig 2.

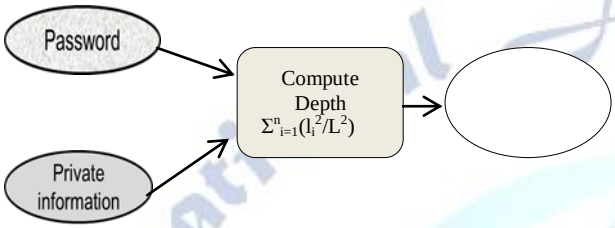


Fig 2 Depth Estimation

1.Depth Measurement

A window approach is adopted that takes word and private info as input in terms of strings. To conduct the computation, a dynamic-sized window slippy from the start to the top of the word is maintained. The initial size of the window is two. If the phase behind the window matches to a definite sort of personal info, the window size grows by one. Then we tend to strive once more to match the new

phase to the private information. If a match is found,

the window size is more enlarged till a match happens. At this time, the window resets to the initial size and continues slippy from wherever the match happens. Meanwhile, array known as tag array with a similar length because the word is employed to record the length of every matched word phase.

After eventually window through the whole word string, the tag array is employed to reason the worth of Depth that's the addition of squares of the matched word phase length divided by the square of the word length. Mathematically,

$DEPTH = \Sigma_{i=1}^{n} (l_i^2/L^2)$

where n denotes the number of matched password segments, l_i denotes the length of the corresponding matched password segment ,L is the length of the password.

2. Depth acceptance

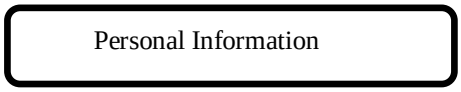
Depth might be very useful for building password strength meters that have been stated as mostly ad-hoc. Maximum meters deliver ratings primarily based on password structure and duration or blacklist normally used passwords. There also are meters that perform simple social profile analysis, together with that a password cannot incorporate the user's names or the password cannot be similar to the account call. But, those easy evaluation mechanisms can be easily manipulated through slightly mangling a password, at the same time as the password remains weak. Using the metric of intensity, password meters can be progressed to more appropriately measure the strength of a password. Furthermore, it is easy to put in force depth as part of the strength dimension. Most importantly, considering users can't easily defeat the depth dimension through easy mangling methods, they may be forced to pick greater at ease passwords. Intensity also can be included into current gear to amplify their competencies. There are numerous Markov version based gear that are expecting the next symbol while a user creates a password. Those gear rank the probability of the next image based on the Markov version discovered from dictionaries or leaked datasets, after which show the maximum probable predictions. Depth helps to decide whether or not private statistics prediction ranks high sufficient in probability to remind a user of warding off using private facts in password introduction.

IV INDIVIDUAL PROBABILISTIC CONTEXT-FREE GRAMMAR

Based on the PCFG approach, Individual-PCFG is developed as an individual-oriented password cracker that can generate personalized guesses [12] towards a targeted user by exploiting the already known personal information. Individual-PCFG leverages the basic idea of PCFG. Besides "L," "D," and "S" symbols, it features more semantic symbols, including "B" for birthdate, "N" for name, "E" for email address, "A" for account name, "C" for cell phone number, and "I" for ID number. Richer semantics make Individual-PCFG more accurate in guessing passwords. To make Individual-PCFGwork, an additional personal information matching phase and an adaptive-substitution phase are added to the original PCFG method. The workflow is given in Fig 3.

Individual-PCFG has 4 phases.

1. Personal Information Matching
2. Password Pre-processing
3. Guess formation
4. Adaptive Substitution



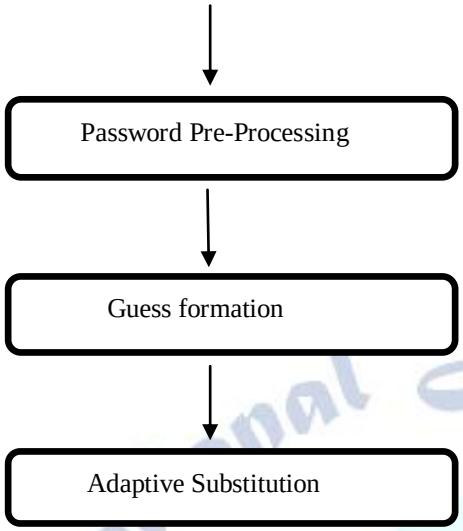


Fig 3 Workflow of Individual-PCFG

1.Personal Information Matching

Taken a password, the entire password or a substring of the password is matched to its personal information .The length of the matching segment is recorded. Then replace the matched segments in a password with corresponding symbols and mark each symbol with its length. Unmatched segments remain unchanged. For example, assume sasi was born on july 30, 1994, and her password is "hellosasi307!." The matching phase will replace "sasi" with "N_4" and "307" with "B_3". The leftover "hello" is kept unchanged. Therefore the outcome of this phase is "helloN_4B_3!".

2. Password Pre-Processing

This phase is similar to the pre-processing routine of the original PCFG. Based on the output of the personal information matching phase, the segments already matched to personal information will not be processed. For instance, the sample structure "helloN_4B_3!" will be updated to "$L_5N_4B_3S_1$" in this phase. Now the password is fully described by semantic symbols of Individual- PCFG, and the output in this phase provides base structures for Individual-PCFG.

3. Guess formation

Similar to the original PCFG, "D" and "S" symbols are replaced with actual strings learned from the training set in descending probability order. "L" symbols are replaced with words from a dictionary. The guesses keep being generated for next step. The personal information is not replaced with any symbols, so the guesses are still not actual guesses. The personal information is not handled in this step, since personal information of each user is different. Thus, the personal information symbols can only be substituted until the target user is specified. Therefore, in this phase, the base structures only generate pre-terminals, which are partial guesses that contain part of actual guesses and part of Individual-PCFG semantic symbols. For instance, the example "$L_5N_4B_3S_1$" is instantiated to "helloN_4B_3!" if "hello" is the first 5-symbol-long string in the input dictionary and "!" has the highest probability of occurrence among 1-symbol special characters in the training set. Note that for "L" segments, each word of the same length has the same probability. The probability of "hello" is simply 1/N, in which N is the total number of words of length 5 in the input dictionary.

4. Adaptive Substitution

In the original PCFG, the output of guess generation can be applied to any target user. However, in Individual-PCFG, the guess will be further instantiated with personal information, which is specific to only one target user. Each personal information symbol is replaced by corresponding personal information of the same length. If there are multiple candidates of the same length, all of them will be included for trial. In example "helloN_4B_3!," "N_4" will be directly replaced by "sasi." However, since "B_3" has many candidate segments and any length 3 substring of "19940730" may be a candidate, the guesses include all substrings, such as "hellosasi794!," "hellosasi730!," "hellosasi307!," etc. Then try these candidate guesses one by one until one candidate was found to match exactly the password of sasi. On the contrary of having multiple candidates, not all personal information segments can be replaced because same length segments may not always be available. For instance, a pre-terminal structure "helloN_5B_3!" is not suitable for sasi since hername contains only 4 characters. In this case, no guess from this structure should be generated for sasi.

V PERFORMANCE METRICS

The number of guesses are used to measure the effectiveness of Individual-PCFG compared with PCFG. The number of guesses is defined as the number of password guesses generated for cracking the passwords in the test set. In Individual-PCFG, the aggregated individual number of guesses is linearly dependent on the password dataset size.

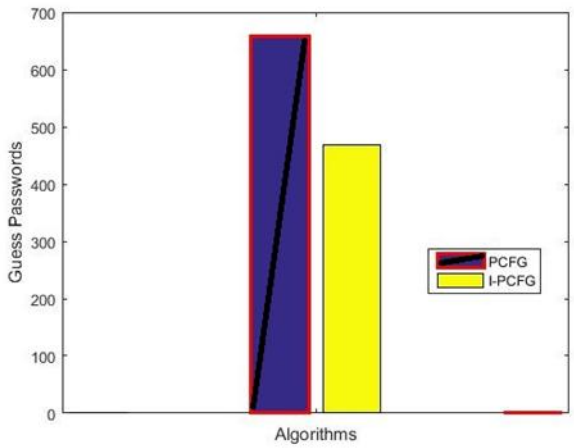


Fig 5.1 PCFG vs Individual-PCFG (Number of Guesses)

Individual-PCFG increases the guessing success rate in online and offline attacks. Online attacks are only able to try a small number of guesses in a certain time period due to the system constraint on the login attempts. The result is presented in Fig 5.1, from which it can be seen that Individual-PCFG is able to crack more passwords than the original PCFG with lower number of guesses.

Using the Chinese dataset, it had been found that PCFG generates 657 guesses and Individual-PCFG generates 459 guesses for 100 records that are taken in the test set.

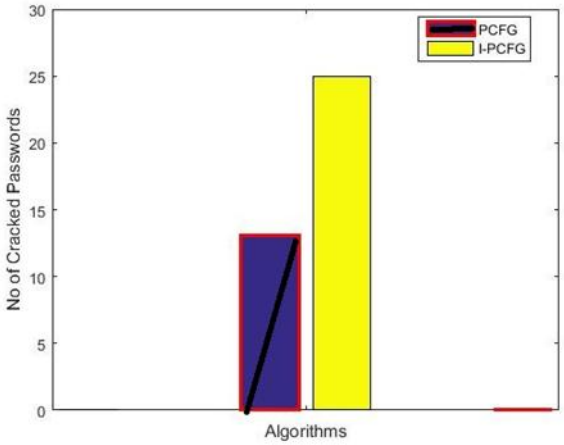


Fig 5.2 PCFG vs Individual-PCFG (Number of Cracked Passwords)

Individual-PCFG increases the number of cracked passwords in online and offline attacks. In this dataset, it had been found that PCFG cracks 13 passwords and Individual-PCFG cracks 25 passwords out of 100 in the test set. The result is presented in Fig 5.2, from which it can be seen that Individual-PCFG is able to crack more passwords than the original PCFG.

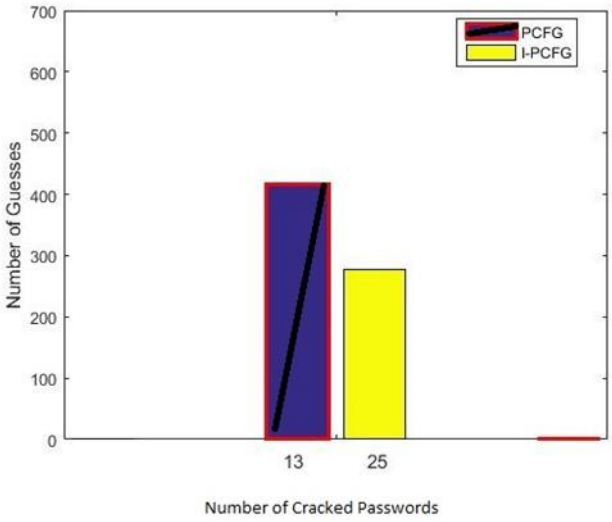


Fig 5.3 PCFG vs Individual-PCFG (Number of Guesses vs Cracked Passwords)

Given the different number of guesses, the number of cracked passwords in the entire password trial set was generated. Fig 5.3 shows the comparison result of the original PCFG and Individual-PCFG in an offline attack based on number of guesses and number of cracked passwords. Both methods have a quick start because they always try high probability guesses first. Fig 5.3 clearly indicates that Individual-PCFG can crack passwords much faster than PCFG does. As a nutshell, PCFG cracks 13 passwords out of 100 in the test set through 657 guesses and Individual-PCFG cracks 25 passwords out of 100 in the test set through 459 guesses. Individual-PCFG is able to cover a larger password space than PCFG because personal information provides rich personalized strings that may not appear in the dictionaries or training set.

Thus, Individual-PCFG is more efficient to crack the passwords within a small number of guesses. Therefore, Individual-PCFG outperforms PCFG in both online and offline attacks, due to the integration of personal information into password guessing. The extra requirement of Individual-PCFG on personal information can be satisfied by knowing the victim personally or searching on social networking sites (SNS).

VI CONCLUSION

The private facts in passwords had been analyzed and a few exciting and quantitative discoveries which include maximum of the users inside the 12306 dataset use their birthdate ,username and e-mail as a password. Then a new metric called depth is delivered to as it should examine the correlation among non-public information and a password. The

depth based quantification effects similarly verify the disclosure on the extreme involvement of personal statistics in password introduction, which makes a user password greater liable to a focused password cracking. Individual-PCFG changed into advanced primarily based on PCFG but recollect greater semantic symbols for cracking a password. Individual-PCFG generates personalized password guesses via integrating non-public records in the guesses. The experimental outcomes demonstrates that Individual-PCFG is considerably quicker than PCFG in password cracking and eases the feasibility of mounting online assaults.

VII FUTURE WORK

The distortion functions can be proposed to guard susceptible password that encompass private information for the reason that they are powerful in increasing password safety by means of greatly reducing the correlation among user passwords and private facts. Distortion functions can mitigate the problem of including personal information in user passwords without significantly sacrificing the user selected password.

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 IEEE Security & Privacy, 2012.

[2] J. Bonneau, "The science of guessing: analyzing an anonymized corpus of 70 million passwords," in IEEE Security & Privacy, 2012.

[3] D. Malone and K. Maher, "Investigating the distribution of password choices," in ACM WWW, 2012.

[4] A.Narayanan and V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," in ACM CCS, 2005.

[5] R. Veras, J. Thorpe, and C. Collins, "Visualizing semantics in passwords: The role of dates," in IEEE VizSec, 2012.

[6] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memorability and security: Empirical results," IEEE Security & Privacy Magazine, 2004.

[7] R. Morris and K. Thompson, "Password security: A case history," Communications of the ACM, 1979.

[8] Z. Li, W. Han, and W. Xu, "A large-scale empirical analysis of chinese web passwords," in Proc. USENIX Security, 2014.

[9] D. Schweitzer, J. Boleng, C. Hughes, and L. Murphy, "Visualizing keyboard pattern passwords," in IEEE VizSec, 2009.

[10] R.Veras, C. Collins, and J. Thorpe, "On the semantic patterns of passwords and their security impact," in NDSS, 2014.

[11] M. Weir, S. Aggarwal, B. De Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," in IEEE Security & Privacy, 2009.

[12] Yue Li, Haining Wang, and Kun Sun, "Personal Information in passwords and Its Security Implications," IEEE Transactions On Information Forensics And Security, Vol. 12, No. 10, October 2017.