Password Cracking by Exploiting User Group Information

Beibei Zhou[1], Daojing He[1,3], Sencun Zhu[5], Shanshan Zhu[2,4(✉)], Sammy Chan[6], and Xiao Yang[1]

[1] Software Engineering Institute, East China Normal University, Shanghai 200062, China
[2] State Key Laboratory of Public Big Data, Guizhou University, Guizhou 550025, China
zhushanshan999@126.com
[3] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China
[4] School of Economics and Management, Harbin Institute of Technology, Shenzhen 518055, China
[5] Department of Computer Science and Engineering, Pennsylvania State University, University Park, TX 19019, USA
sxz16@psu.edu
[6] Department of Electrical Engineering, City University of Hong Kong, Hong Kong 518057, China
eeschan@cityu.edu.hk

Abstract. The past research study on the characteristics of passwords has paid much attention to language, regional or cultural differences and usability. However, few studies have pointed out differences due to information such as application types, users' occupations, religious beliefs, and meanings of the digits in the culture. In this article, for the first time we put forward the concept of "group" characteristics, and found that the passwords of different groups have obviously different characteristics. For example, when dividing groups by religions of users, Christian groups like to include biblical names and words in passwords, such as "jesus", "christ", "angels" and "faith". Accordingly, we propose *gPGM*, a neural network-based password guessing method that leverages group information to increase attack success. Our experiments show that gPGM can significantly increase the password cracking rate. In addition, the cracking rates for different groups, under the same number of guesses, also vary. For example, the cracking rate of the game group is very high, but that of the hacker group is very low.

Keywords: Group password · Password analysis · Password attack

1 Introduction

It is known that users often lack security awareness when creating their passwords because they tend to select simple passwords for easy memorization [25].

H. Duan et al. (Eds.): SecureComm 2023, LNICST 567, pp. 514–532, 2025.
https://doi.org/10.1007/978-3-031-64948-6_26

In order to guarantee the security level of user-created passwords, websites use password strength meters (PSMs) to assess the security level of a given password.

We have three important observations on the creation of passwords. Firstly, different PSMs adopted by different websites may report different security levels for the same password, which is also supported by the study of Wang *et al.* [25]. In addition, the evaluation results from the PSMs are not necessarily related to whether the websites accept the password or not. For example, as a popular website in China, Sina accepts our simple test password (123456) even though its PSM evaluates it as weak. Secondly, the feedback from PSMs to users is limited, which may confuse users as they have no idea on how to improve their selected passwords. Thirdly, password creation is typically a human behavior, which may be affected by various factors [6]. For example, the religious or cultural background of the password creator may lead to the use of certain numbers or words when forming a password. Moreover, for different application types, users may tend to select different passwords. For example, users may create strong passwords for personal e-commerce accounts, but weaker passwords for entertainment accounts.

In this work, we first conduct a large-scale password analysis using real-world passwords that were leaked in previous security incidents. Analyzing such information, we observe that the passwords demonstrate different statistical characteristics among different groups (characterized by types of applications, or religious background). Then, we propose an attention-based deep neural network to realize a novel password guessing method, which can pay attention to different group information. Our evaluation experiments show that the efficiency and accuracy of our password guessing method is greatly improved. Under the same guessing method, the curves of password cracking rates for datasets in the same group have great similarity, whereas the curves of different groups are obviously different. We summarize our findings and main contributions as follows:

Password Has Group Characteristics. For example, when dividing groups by application types, the vocabulary words in the game group include "wangyut" and names of popular game characters, and adult website groups have many sexually suggestive words, such as "pussy" and "letmein". Hackers not only like to use technical words such as "admin", "administrator" and "technicians" to form passwords, "lovely" and "kawaiisensei" are also among the top five popular vocabulary words.

Religious People Password Characteristics. Among the popular passwords leaked from two Christian-oriented websites, there are vocabulary words such as "psalm" and "faith". Among the passwords leaked from an application for Tibetan people, there are vocabulary words such as "yuan", "lijie" and "xinze", which represent "practiced", "understood" and "walker", respectively in the Buddhist culture. Among the popular digits in Buddhism, "208" and "1208" are the dates when Buddha Shakyamuni became a monk and attained Buddhahood, respectively, and "219" is the birthday of Guanshiyin Bodhisattva. Our findings

can help such systems to better protect their users by purposely detecting and avoiding those religious vocabulary words/digits in user-chosen passwords.

Improvement of the Efficiency of Password Guessing. By exploiting group information, we propose a deep neural network based password guessing method called *gPGM*, which is based on a Long Short-Term Memory (LSTM) model with Attention Mechanism (AM). AM is used to make the network model pay attention to the influence of different group features on the password context-dependent features. At the same time, LSTM is used to process the sequences of text passwords. The password cracking rate of using LSTM alone is slightly higher than that of PCFG [28] and TransPCFG [10]. However, with AM, the cracking rates become 5%-7% higher than that without AM.

Differences in Group Resistance to Attacks. The cracking rates on passwords of different groups are significantly different, while the cracking rates of password datasets in the same group are very similar. In particular, we found that the password cracking rate of the hacker group is very low, whereas that of the adult website and game groups are very high.

The remainder of the paper is organized as follows: Sect. 2 reviews related work. Section 3 presents analyses of group passwords. Section 4 details our proposed method. Section 5 covers the experiments and results analysis. Finally, Section 6 concludes all the work.

2 Related Work

2.1 User Behavior in Password Creation

Because text passwords are simple and easy to implement, they are still the main authentication method on the Internet [24]. A large number of studies have pointed out the characteristics of user-created passwords [2,4,5,9,21,23,27]. Much of the understanding of passwords comes from analyzing leaked password datasets. In 2012, Bonneau [5] studied the long-term impact of character encoding on the password ecosystem based on password datasets from Chinese, English, Hebrew and Spanish users. In 2013, Mazurek *et al.* [18] analyzed passwords from an entire university. In 2014, Li *et al.* [16] mentioned users often built passwords that included personally meaningful information. In 2016, Li *et al.* [15] dissected user passwords from a leaked dataset to investigate how and to what extent user personal information resides in a password. In 2019, Alomari and Thorpe [3] explored differences in password behaviors of university students, IT professionals, and the general population (non-student and non-IT professional). In the same year, Wang *et al.* [26] performed an extensive, empirical analysis of 73.1 million real-world Chinese web passwords in comparison with 33.2 million English counterparts. In 2021, Veras [22] analyzed the semantic password model and linguistic patterns in passwords.

Few studies have explored the influence of group factors such as application types, religious beliefs, and cultural meanings on password characteristics. Compared with previous studies, our research not only considers differences in language, culture, etc., but also classifies passwords according to the application types and religious beliefs, and finds that the passwords of different groups have obvious differences, providing new insights for future password analysis.

2.2 Password Guessing/Cracking

Password guessing has received widespread attention. In 2009, Weir *et al.* [28] proposed a password guessing method based on PCFG. PCFG sorts the password structure in the password set in descending order, and generates words to guess the password according to these rules. Some work is to research and improve PCFG [10,13,17,22]. In recent years, the application of machine learning to password guessing has become a new trend. In 2016, Melicher *et al.* [19] used the Recurrent Neural Network (RNN) to guess the password to obtain the probability distribution of the password and used the Monte Carlo simulation algorithm [7] to obtain the number of password guesses. In 2018, Liu *et al.* [17] proposed a general deep learning model, named "GENPass", for password guessing, which combines PCFG Rules and adversarial generation. In 2020, Xie *et al.* [29] proposed TarGuess-I, an improved password guessing model which was capable of identifying popular passwords by generating top-300 most popular passwords from similar websites and grasping special strings by extracting continuous characters from user-generated personally identifiable information. In 2022, Yang *et al.* [30] proposed VAEPass, a lightweight password guessing model based on a Variational Auto-Encoder (VAE), which comprises an encoder and a decoder established using Gated Convolutional Neural Network (GCNN).

Previous work mainly performed password guessing from the aspects of password characteristics and network model improvement, and rarely considered the characteristics of different password groups. We applied group characteristics to password guessing, and guessed passwords based on the characteristics of different groups, and found that the password security of different groups varies greatly.

3 Large-Scale Real-World Passwords Analysis

In consideration that many human behaviors are influenced by their group characteristics [20], we speculate that password creation, which is human behavior, is also influenced by certain group characteristics. In order to verify our hypothesis, we conduct a large-scale password analysis in this section.

3.1 Definition of Group

Group members share certain characteristic. For example, they may have interactions, share the same needs, have a common goal, have the common religious

belief, or use the same application, and so on. The group we define here is "an intact social system, complete with boundaries" [6]. People with the same characteristics can distinguish themselves from nonmembers as long as the characteristics can separate the group. Users can be divided into different groups, and there is no mandatory interaction requirement among group members. For example, according to application types, users can also be divided into mail group, game group, etc. Similarly, according to the religion of the user's beliefs, users can be divided into Christian groups, Buddhist groups, etc. We use "group passwords" to represent the passwords from certain groups that might demonstrate certain statistical similarities.

3.2 Password Datasets and Ethical Considerations

The passwords can be created for different application types, which may potentially affect the users' password creation behavior. According to the application types of the accounts, we divide the passwords in our datasets into five groups, namely, mail, game, appointment, adult and hacker technology exchange (referred to as "hacker"). The details are given in Table 1. We note that hacker groups have much fewer passwords than other groups. This may be because the passwords of the hacker groups are more difficult to crack or because the hacker profession accounts for a relatively low proportion of the population. For the first time, we categorized passwords according to religious beliefs, choosing the passwords of the two major religions, Christianity and Buddhism. The Buddhism group includes Tibet, which comes from the Tibet area of the Chinese ticketing website. As Tibet is an ethnic minority autonomous region in China, there are relatively few people using ticketing websites. The number of passwords from that group is relatively small.

Ethical Considerations. Despite the fact that password datasets are ever publicly available and widely used, passwords are highly private and sensitive [8, 10, 26]. We store the datasets on computers not connected to the Internet. And we treat all user names confidentially and only report the aggregated information about passwords including some examples of weak passwords. So using them in our research does not benefit attackers in password guessing. In addition, our use is beneficial for research on password guessing and password strength evaluation. In particular, previous studies have shown that some passwords include religion-related words, but none of them studied password security for the religion group. We hope that our findings can help websites remind users to avoid including common religious words when creating passwords.

3.3 Group Passwords Statistical Characteristics

We analyze the group password characteristics from the composition characters, popular passwords, vocabulary words, digit semantics, etc.

Table 1. Password datasets.

Group		Dataset	Total #	Unique #
Application Types	Mail	163mail Hotmail	116,918,183	22,066,983
	Game	178 7k7k	113,858,458	78,401,786
	Appointment	Badoo Zoosk	28,194,094	17,401,683
	Adult	Porn Tuscl	3,273,526	1,885,372
	Hacker	Elitehacker Hack5	641,774	495,037
Religion	Buddhism	Tibet	8,209	7,017
	Christian	Faithwriter Singles	205,800	161,971

Password Composition Characters. Passwords are typically composed of letters (L), digits (D), and special characters (S). Letters can be further divided into uppercase (UL), lowercase (LL) and Mixed UL and LL (ML). The composition of passwords is an important influence factor of password strength.

Since the composition of password embodies the inherent behavior habits of users setting passwords, we postulate that passwords of different application types have greater commonality in the character distribution. From Table 2, we can see that the game group uses "LL+D" at a higher rate and the appointment group "UL+D" structure is higher than other groups. And we found that the complex structure (ML+D) of hacker groups accounted for the highest proportion. The proportion of Christian groups using complex structure (ML+D) is also much larger than other groups.

Popular Passwords. Popular passwords refer to the passwords used by many users. When hackers attack, they normally first try to use these passwords, so using popular passwords is generally considered insecure.

As shown in Table 3, in the hacker group, we can see that users' security awareness is significantly higher than other groups. For example, "trustno1" means "don't trust anyone". Although some keyboard scheme passwords appear, they are slightly different, such as "zxczxc" and "QsEfTh22". The former is a repetition of keyboard patterns, and the latter is a keyboard pattern distributed like a shape 'W' [11]. In particular, there are strange and irregular passwords among the popular words in Hack5, such as "ZjmHgC35" and "Kj7Gt65F". And the proportion of users of the hacker group using popular passwords is low, which obviously shows that hackers' security awareness is much higher than that of ordinary users. We found that most popular passwords in mail and game groups are pure digits, and the weak password "123456" accounts for a very high

Table 2. Compositions of passwords and their occurrence percentages (%).

Dataset	Group	LL	D	LL+D	UL+D	ML+D
163mail	Mail	8.93	**58.41**	29.46	0.49	0.40
Hotmail	Mail	22.14	**42.96**	27.77	0.66	0.71
178	Game	16.73	12.32	**67.62**	0.00	0.00
7k7k	Game	10.97	23.26	**60.02**	1.23	2.74
Badoo	Appointment	**82.45**	4.48	5.42	4.38	1.06
Zoosk	Appointment	**79.53**	3.25	5.15	5.91	1.43
Porn	Adult	**65.68**	12.56	17.06	0.26	1.04
Tuscl	Adult	**62.37**	8.70	19.30	1.04	3.51
Elitehacker	Hacker	**45.53**	11.62	8.16	0.11	30.30
Hack5	Hacker	18.63	5.10	**33.18**	1.02	32.03
Tibet	Buddhism	7.14	11.90	**78.57**	2.38	0.00
Faithwriter	Christian	**50.14**	6.27	33.27	0.83	3.80
Singles	Christian	**55.11**	8.37	26.48	0.82	2.34

proportion [24]. On one hand, it shows that users lack security awareness when creating mail and game accounts. On the other hand, we found that mail and game accounts have a large overlap through mail address comparison. This is because many users directly log into the game with the same credentials as their mail accounts. Among the top five popular passwords of game groups, there are passwords that contain digits related to the word "love" in Chinese, such as "5201314" and "7758521" [26]. We guess that it may be because most of the users of the game group are young people, who put their yearning for love into the password. The popular password for the appointment groups is very short, with only two simple digits. However, the popular passwords of the two datasets of the appointment group are exactly the same, and their proportions are also very similar. Among the popular passwords for the adult group, digit-only passwords are basically continuous digits without repeated digits. When dividing groups according to religious beliefs, we can see that the passwords of the two datasets in the Christian group are very similar, and many passwords contain the name "Jesus" or "Christ", while the Buddhist group and the Christian group are very different.

Vocabulary Words. Affected by different native languages, people usually use some vocabulary words as part of their passwords when creating them, such as the full spelling of the name. We use Sogou pinyin corpus and COCA corpus to match vocabulary words contained in passwords.

As shown in Table 4, the vocabulary words of the hacker group are obviously different from those of other groups. They are commonly used by computer professionals, such as "technicians", "admin", "killer", "administrator" and "random". The word "random" is used because randomness is a requirement for setting

Table 3. The most popular passwords and their occurrences per thousand (‰).

Dataset	Group	Top #1	Top #2	Top #3	Top #4	Top #5
163mail	Mail	123456(12.4)	111111(3.9)	123456789(4.1)	000000(3.4)	123123(2.8)
Hotmail	Mail	123456(18.9)	111111(4.7)	123456789(3.1)	000000(2.4)	123123(2.0)
178	Game	123456(30.9)	111111(11.3)	123123(3.3)	**7758258(2.1)**	**5201314**(7.8)
7k7k	Game	123456(35.5)	111111(9.2)	123123(5.1)	**5201314(6.6)**	**7758258(5.7)**
Badoo	Appointment	12(4.2)	11(3.6)	10(3.1)	22(3.1)	13(2.9)
Zoosk	Appointment	12(5.1)	11(4.7)	10(4.5)	22(4.3)	13(4.3)
Porn	Adult	password(2.6)	123456(2.3)	12345(2.1)	1234(2.1)	pussy (1.5)
Tuscl	Adult	password(3.1)	123456(2.7)	12345(2.6)	1234 (1.9)	tuscl(1.7)
Elitehacker	Hacker	123456(1.1)	zxczxc(1.1)	12345(1.1)	passport(1.1)	diablo(1.1)
Hack5	Hacker	123456(0.8)	QsEfTh22(0.8)	**trustno1(0.8)**	ZjmHgC35(0.8)	Kj7Gt65F(0.8)
Tibet	Buddhism	zmqamy1314(23.8)	xinze1216(23.8)	rover270134(23.8)	zhaxi2548(23.8)	fm3658512(23.8)
Faithwriter	Christian	123456(2.1)	jesus1(2.1)	christ (1.5)	jesuschrist(0.9)	blessed(0.3)
Singles	Christian	123456(1.7)	jesus1(1.5)	christ (1.2)	jesus(1.0)	princess(0.7)

a password. "Killer" is often used to describe someone with outstanding skills. "Kawaiisensei" means "lovely" in Japanese, which seems to indicate that computer technicians like lovely things. Also, the most common vocabulary word used by hackers is the keyboard mode "qwerty" in the keyboard pattern [14]. The analysis of the hacker group is very meaningful because it proves the influence of occupation on the creation of passwords by users. Most users of mail and game groups speak Chinese as their mother tongue, while users of appointment and adult groups mostly speak English as their mother tongue. We can see that users from China like to add their own last names when creating passwords, such as "wang", "zhang", "chen" and "yang", while users from Europe and America prefer to add first names, such as "alex", "david", "lucas" and "mike". Some vocabulary words in the game group are related to the game world, for example, "wangyut" means not only the character but also the game skills with death. "Jianghu" represents the world in the game, which means the user loves the game world. The most prominent feature of the adult population is that many vocabulary words are related to sex such as "strip" and "letmein". In religion groups, vocabulary words of the Christian group are all related to the *Bible*. The "yuan", "lijie" and "xinze" in the Buddhist group represent "practiced", "understood", and "walker" in the Buddhist culture. The "zhaxi" and "zxzx" in Tibet group mean happiness ("zhaxidele" in Tibetan) and are the most popular and representative blessings of the Tibetans.

Digit Semantics. Prior research mentioned that users use dates, phone numbers, love, or consecutive numbers as passwords, but digits also have cultural and customary meanings. For example, Chinese users like "6" and "8" but do not like "4" because it is a homophone for "death" in Chinese. This raises some questions: how many digits do users like to use when creating passwords? What are the most commonly used digits? Does it contain specific information?

Table 5 shows the most common digit sequence lengths and the most commonly used patterns in the passwords of each group. As an example, for the 163mail dataset, the length of top #1 (i.e., the most popular) digit sequence

Table 4. The most popular vocabulary words and their occurrences per thousand (‰).

Dataset	Group	Top #1	Top #2	Top #3	Top #4	Top #5
163mail	Mail	love(1.5)	abcd(0.8)	woaini(4.2)	wang(3.1)	zhang(2.0)
Hotmail	Mail	love(1.3)	abcd(0.5)	tianya(4.4)	null(3.6)	password(1.6)
178	Game	wang(1.3)	zhang(1.2)	chen(0.7)	yang(0.8)	jianghu(2.2)
7k7k	Game	wang(0.8)	zhang(0.7)	chen(0.6)	yang(0.5)	wangyut(1.8)
Badoo	Appointment	alex(0.1)	david(0.0)	badoo(0.2)	love(0.1)	azerty(0.0)
Zoosk	Appointment	alex(0.5)	david(0.3)	hex{[}64e(0.4)	lucas(0.4)	daniel(0.3)
Porn	Adult	love(0.5)	mike(0.6)	john(1.2)	letmein(0.6)	jeff(0.6)
Tuscl	Adult	love(0.6)	mike(0.6)	john(0.5)	tuscl(1.7)	pussy(0.7)
Elitehacker	Hacker	qwerty(3.4)	administrator(3.4)	killer(2.2)	**lovely(2.2)**	freedom(2.2)
Hack5	Hacker	qwerty(3.1)	admin(2.9)	random(1.9)	**kawaiisensei(1.8)**	technicians(1.3)
Tibet	Buddhism	**yuan(23.8)**	**zxzx(23.8)**	**lijie(23.8)**	**zhaxi(23.8)**	**xinze(23.8)**
Faithwriter	Christian	**jesus(4.1)**	**faith(2.4)**	**john(1.2)**	**psalm(2.0)**	**blessed(1.4)**
Singles	Christian	**jesus(4.0)**	**faith(1.1)**	**john(0.7)**	**christ(2.0)**	**angel(1.9)**

Table 5. Patterns of digit uses in passwords.

Dataset	Group	Top #1			Top #2			Top #3		
163mail	Mail	123456	111111	000000	**1314**	1234	2008	12345678	88888888	**31415926**
Hotmail	Mail	123456	111222	111111	1234	1004	1111	12345678	11111111	88888888
178	Game	123456	111111	123123	**5201314**	**1314520**	1234567	54545454	**25991314**	12345678
7k7k	Game	123456	111111	123123	**5201314**	**1314520**	**7758521**	12345678	1111111	88888888
Badoo	Appointment	12	11	10	2010	1234	2009	1234567	7777777	0123456
Zoosk	Appointment	12	95	11	1995	1234	1996	1234567	0123456	7777777
Porn	Adult	1234	6969	1111	1	2	4	99	69	12
Tuscl	Adult	1234	2000	6969	1	2	4	69	01	12
Elitehacker	Hacker	23456	666666	187187	1	2	6	1234	1134	1977
Hack5	Hacker	123456	147852	666666	1	3	5	1234	1337	5252
Tibet	Buddhism	123	511	136	6417	0208	1208	199100	219219	123456
Faithwriter	Christian	1	2	4	12	01	22	234	2006	2000
Singles	Christian	1	2	4	12	11	77	1234	2003	2004

included in passwords is 6, and the lengths of top #2 and top #3 digit sequences are 4 and 8, respectively, with the top 3 specific passwords listed in the table (e.g., "123456", "111111", "000000" for the length of 6). We can see that the most common lengths in different groups are quite different. Among them, the most common lengths in the game group have the longest lengths of 6, 7, and 8. According to our follow-up findings, they are exactly the passwords of users. Users in the mail group tend to use the repeated pattern of digits. In Chinese culture, the digit "8" has always been regarded as the luckiest digit. In English culture, like Chinese culture, the digit "8" also contains auspicious meanings. Therefore, the appearance of the pattern "88888888" in the table can be explained. However, the appearance of the digit "31415926" reminds us of π, which is convenient for memorization and does not have the regularity of simple digit segments. In the game group, "5201314", "1314520", "7758258" and "25991314" all have love related meanings in Chinese [26]. The pattern "54545454", "54" in the 178 dataset is homophonic in Chinese as "Samurai", which is a professional name in online games. The most popular length of digit sequence is 2 in the appointment group, and the password of length 7 contains "7777777". "7" means luck in English cul-

ture [22]. "147852" appears in the Hacker group. These are the first two columns of the number keys on the small keyboard. Many users include it in the password for easy memorization. After further searching, we found that "187" was the code-name for the murder case within the US (California) law enforcement agency, hence becoming slang for "murder". Among the popular patterns in the Buddhism group, "208" and "1208" are the dates when Buddha Shakyamuni became a monk and attained Buddhahood, respectively, and "219" is the birthday of Guanshiyin Bodhisattva.

Through the above analysis, we found that the passwords of different groups of users have significant differences in composition of characters, popular passwords, vocabulary words and digits. Such observations inspire us to propose a novel password guessing method based on group information, which has been rarely studied in the research field.

4 The Design of gPGM

In this section, we propose a password guessing method named gPGM (group based Password Guessing Method), which adds group characteristics to password guessing. It uses the AM to obtain the contextual dependency of different group characteristics in passwords. It employs the LSTM to build a neural network model to calculate the probability distribution of the passwords of different groups, based on which it guess passwords.

4.1 Process Overview

LSTM introduces the concept of gate, instead of relying only on the superposition of the original neural network. However, LSTM cannot determine the importance of different features to search tasks based on contextual information, so using LSTM alone will not be able to capture the characteristics of passwords by different groups. In view of this situation, one may build a unique LSTM model for each group based on the passwords from that specific group. Certainly, this is not a generic solution, and it will not scale well because there are numerous types of groups in real life. To solve this problem, we choose to use the AM in deep learning to help the password guessing model to pay attention to the dependence between group information and password context feature information.

gPGM performs offline guessing based on the occurrence probability of a password. The occurrence probability of the password is calculated based on our AM_LSTM model, which mainly consists of four layers as shown in Fig. 1. The *input layer* encodes the pre-processed text password data as input. The *password feature learning layer* learns user password contextual feature information. The *group weight attention layer* encodes and vectorizes the group information. In this layer, password temporal features learned by the *password feature learning layer* interact with the group information, so that the password guessing method pays attention to the group to which the password belongs. In the final *output layer*, the result of the group weight attention layer is input to the fully connected

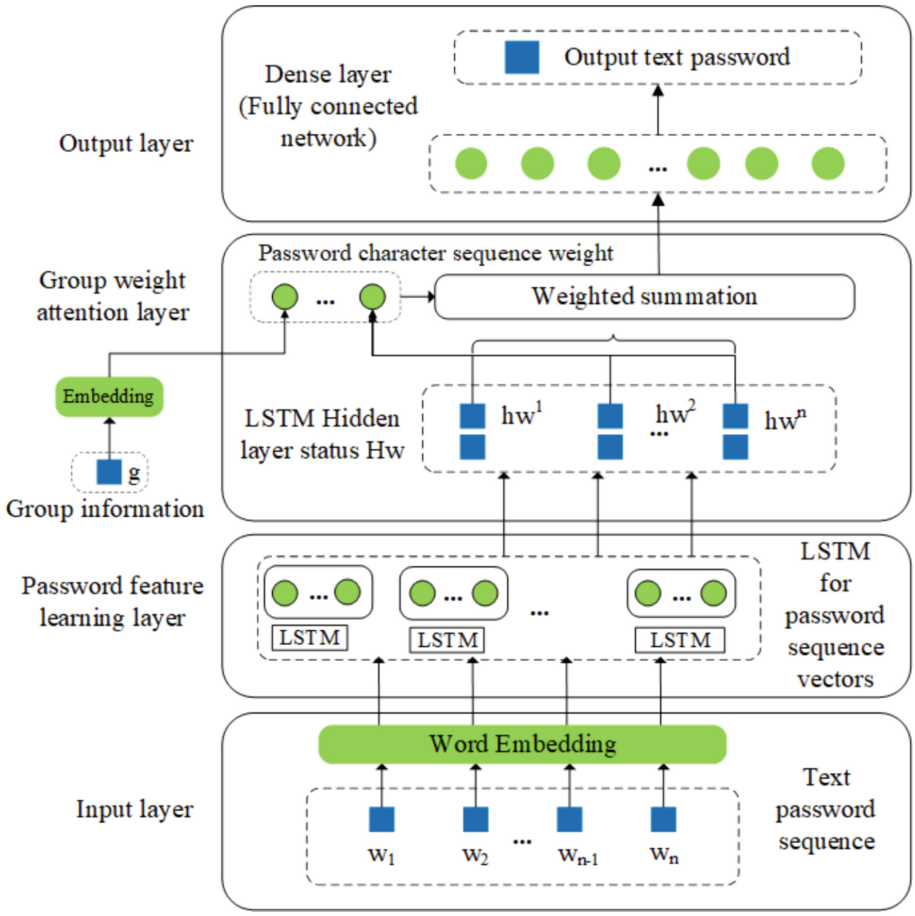


Fig. 1. AM_LSTM model.

network, and the password text probability distribution is obtained through the nonlinear activation function mapping.

Input Layer. Firstly, we carry out pre-processing to filter away invalid passwords. We only use passwords with lengths from 6 to 18, which is consistent with the major password creation policies. This pre-processing mainly focuses on printable ASCII characters and valid password lengths. The printable ASCII character set that our model takes as input in our dataset is shown in Table 6.

In order to better demonstrate the structure and the working principle of the gPGM, we formalized the variables used in the model input layer. The text password is represented as $w=\{w_1, w_2...w_n\}$, where n is the length of the password. We use numbers to represent password characters so that the password is converted into a vector that the network model can receive through Word

Table 6. Password character set.

Character Type	Character Set
UL	ABCDEFGHIJKLMNOPQRSTUVWXYZ
LL	abcdefghijklmnopqrstuvwxyz
D	1234567890
S	&*(){}';:",/^$~_#-=+!?@#%_[]

Embedding. For example, the password "abc123" is expressed as w={26, 27, 28, 52, 53, 54}.

Password Feature Learning Layer. The gates in LSTM are used to remember or forget certain pieces of information through a *sigmoid* nonlinear function and a vector dot product.

Forget Gate is to select which information can enter the memory unit. Assuming that the current time is t, the internal process of the forget gate is formulated as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

W_f represents the forget gate weight matrix, h_{t-1} represents the hidden layer state at the previous moment, x_t represents the input of the current moment, b_f is the offset, and σ is the activation function (*sigmoid* function), and its output can only be 0 or 1. 1 means the information is passed, and 0 means the information is forgotten.

Input Gate calculates how much information is learned from the context, and the internal process is formulated as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$C_t = f_t * C_{t-1} + i_t * tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

here $tanh$ is the hyperbolic tangent function, and C_t is the memory unit at time t.

Output Gate determines which information is to output. Its internal process is formulated as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = o_t * tanh(C_t) \tag{5}$$

In this work, the text password sequence w is input into LSTM to establish the connection between the password characters before and after the text password is created. Finally, the LSTM hidden layer state vector $h_w = \{h_w^1, h_w^2 \cdots h_w^n\}$ is generated, where $h_w^t \in R^{d_a}$ is the LSTM output of the password text sequence at time t, and d_a represents the dimension of the hidden layer state vector.

Group Weight Attention Layer. Although LSTM can learn the inter-character-dependent feature information among the password characters, it does not recognize the degree of influence of different group features on the weight of the text password sequence. We use the AM to help the password guess model focuses on the degree of contextual dependence between different group features and the password sequences.

Similar to the text password sequence, we first encode the group information with numbers. For example, in our study, there are seven groups, so we label them with numbers, such as mail group (0), game group (1), appointment group (2), adult group (3), hacker group (4), Buddhism group (5) and Christian group (6). Then, embedding is performed by combining the state vector dimension information of the password feature learning layer with group information. The embedded information is combined with the state vector to generate a text password character context weight sequence. The group information variable is formally defined as $g_i, g_i \in G$, where G is the group set. Then we calculate the dot product of the generated password character sequence weight and the hidden layer state vector, and output it to the next layer. The formulation of the whole process is as follows:

$$g_{embedding} = Embedding(g_i), g_i \in G \tag{6}$$

$$attention_{input} = \lambda_{reduce}(h_w * g_{embedding}) \tag{7}$$

$$attention_{tmp} = \lambda_{expand}(softmax(attention_{input})) \tag{8}$$

$$attention_{output} = \lambda_{reduce}(attention_{tmp} * h_w) \tag{9}$$

The * operator in the above equations represents the vector or matrix dot product, λ_{reduce} and λ_{expand} represents the dimensionality reduction function and the dimension addition function, the purpose of which is to reduce or increase the dimensions of the vector and matrix to ensure that the model data can obtain the feature weight. $softmax$ is a nonlinear activation function and is intended to normailze the output.

Output Layer. The output layer receives the $attention_{output}$ of the group weight attention layer, and then inputs it into the fully connected network (Dense) structure. In this layer, it generates the probability distribution $\tilde{o}$ of the password space character through the $softmax$ activation function, and selects the password character corresponding to the maximum probability value as the text output of the network model. That is,

$$\tilde{o} = softmax(W \cdot attention_{output} + b) \tag{10}$$

where W represents the nonlinear matrix weight and b is the bias.

4.2 The Calculation Method of AM-LSTM Model

We use AM-LSTM to model user password sequences and simulate the password generation process of different groups of users. The model calculates the probability of the password, and outputs the character text corresponding to the probability from high to low.

Algorithm 1. Password Probability Distribution Algorithm Based on AM-LSTM

Require: pw, AM-LSTM model M Variables:u, v, w. $G = (V, E)$
Ensure: Password probability $P(pw)$ under model M
Definition: Password initiators $s = \Delta$, The terminator is a newline
$i = 0, prob = 1, context = s$
while $i < len(pw)$ **do**
  $P = M(context)$
  $prob = prob * P(pw[i])$
  $context = context + pw[i]$
  $i = i + 1$
end while
return $prob$

We first encode the group information and the password as a vector for model training. After the model training is completed, the probability of its occurrence will be output, and the password will be guessed from the highest to the lowest through the probability. The number of password guesses refers to the number of times the attacker has guessed in descending order of probability. With a password set Γ, the probability function $P : pw \rightarrow p, pw \in \Gamma$ and $P(pw) \in [0, 1)$, then:

$$\sum_{pw \in \Gamma} P(pw) = 1 \tag{11}$$

The calculation process of the function P is shown in Algorithm 1.

If the password $pw' \in \Gamma$ and $P(pw') > P(pw)$, then the guessing number of password pw is equal to the size $|\Phi|$ of the password set Φ that may appear in the pw'. That is:

$$\Phi = \{P(pw') > P(pw), pw' \in \Gamma\} \tag{12}$$

5 Evaluation

5.1 Experiment Design and Environment

We conduct experiments to compare gPGM with a neural-network-based password guessing method that does not use the group attention feature. In this comparison, we will avoid the uncertainty caused by factors such as network model

selection and loss function. We further compare gPGM with the PCFG [12], TransPCFG [10] and Markov [1], that were commonly used by many papers for comparison. Because the Markov-based guessing rate is low, we do not list it in the comparison.

Table 7. Group model main parameters

Group	Religion	Application Types
LSTM layer #	3 layer Single	3 layer Single
LSTM neuron #	256	128
LSTM time_steps	10	10
Loss function	*categorical_crossentropy*	*categorical_crossentropy*
Optimizer	sgd	adam
Learning rate	0.0001	0.001
Batch_size	256	128

For the datasets of the application type group, we use the data shown in Table 1. There are mail, game, appointment, adult, hacker, and religion.

We choose Python 3.6 as the programming language and use the neural network framework Keras for model building. The loss function of the model is *categorical_crossentropy*. And the model is a seq2seq model. Similar to Melicher *et al.* [19], our experiment sets the *time_steps* value of LSTM to 10. We adopt the strategy of early stopping to detect the model convergence. The core idea of early stopping is to compare the accuracy with the maximum accuracy in the past 10 rounds. If the accuracy of 10 consecutive rounds is decreasing, the training is stopped. If the current accuracy is higher than the maximum accuracy, the maximum accuracy value is updated and set. The maximum number of training rounds is 100. We experimentally selected the parameters that achieve the best results. The parameters for both models are shown in Table 7.

5.2 Experimental Results

The cracking rates for the religion group and the application types group are shown in Table 8. We can see that even if there is no AM, the password guessing algorithm based on neural network has a higher cracking rate than PCFG and TransPCFG. With the addition of the AM, the cracking rates increased by 5%-7% under the same cracking times.

Furthermore, we want to understand the security strength of different groups of passwords, under the gPGM attacks. The experimental results are shown in Fig. 2.

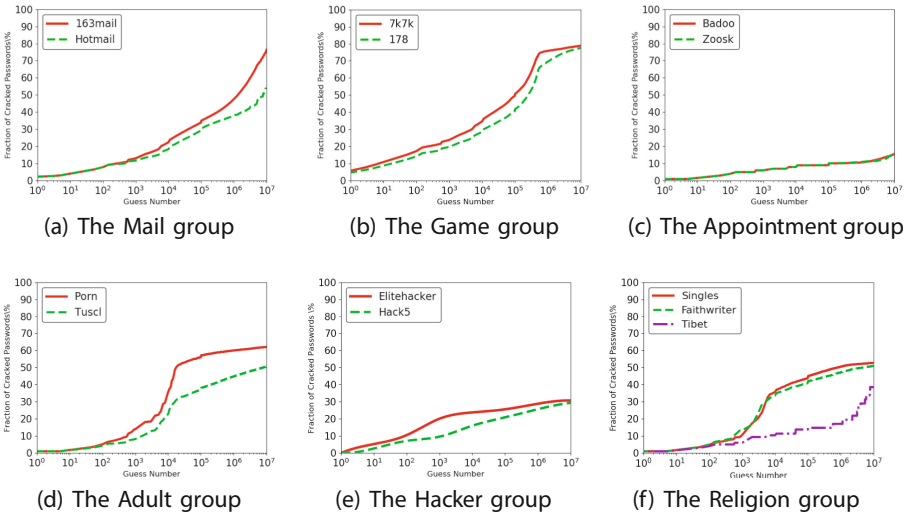
The password strength of groups to resist attacks reflects obvious group characteristics. We can observe that the password guessing curves of the websites in the same group are similar but the password guessing curves between different

groups are very different. The two datasets of the Christian group are used for writing journey of faith and blind dating, respectively. Their cracking rates are higher than that of the mail group. It shows that the password security of users' social accounts is lower than that of mail accounts. The cracking rates of the game group are higher than other groups. A possible reason is that users create accounts for adult websites and game accounts for leisure and entertainment, so they are less concerned about the security of such accounts from the perspective of cracking rates. The cracking rates of the hacker group are very low, which shows that the security awareness of this group is much higher than that of the ordinary people group. Among religion groups, the password cracking rate of Christian groups is much higher than that of Buddhist groups. A possible reason is that the application types are different.

Table 8. The percentage of cracked passwords at 10^7 guesses.

	Dataset	Group	PCFG	TransPCFG	Non Attention	Attention
Application Types	163mail	Mail	66.35	70.36	71.94	**79.43**
	Hotmail	Mail	44.02	45.16	49.60	**56.09**
	178	Game	61.81	55.47	66.80	**72.47**
	7k7k	Game	63.79	55.49	68.87	**74.81**
	Badoo	Appointment	5.43	11.06	13.51	**18.33**
	Zoosk	Appointment	6.47	12.90	11.04	**17.97**
	Porn	Adult	50.08	53.12	55.64	**62.17**
	Tuscl	Adult	37.59	44.21	42.98	**50.71**
	Elitehacker	Hacker	24.00	15.53	26.90	**30.80**
	Hack5	Hacker	18.24	10.60	23.60	**29.40**
Groups of people	Tibet	Buddhism	27.80	11.89	32.60	**39.70**
	Faithwriter	Christian	40.27	45.11	45.93	**52.40**
	Singles	Christian	42.36	44.15	47.56	**55.60**

The passwords used in the Buddhist group are from the official China Railway website, where users store their real personal identification numbers so that they can buy railway tickets on this website. Therefore, users tend to use strong passwords to protect their accounts from financial losses as well as leakage of personal information. However, the passwords in the Christian group come from sites of writing and dating, and do not involve financial information, so users may choose simple and easy-to-remember passwords. Another possible reason is that in our dataset, the number of passwords of the Christian group is large, but the Buddhist group is small.



(a) The Mail group (b) The Game group (c) The Appointment group

(d) The Adult group (e) The Hacker group (f) The Religion group

Fig. 2. The password strength of individual groups on resisting gPGM.

5.3 Recommendations for Password Generation

Based on our analysis of the characteristics of group passwords and the results of password guessing, we provide two recommendations to users who create passwords.

1. Set different passwords for the accounts of the same type of applications, such as two game accounts. According to the results of our attack experiments, the password cracking rates of the same group are very similar. If an attacker has discovered the password characteristics of one application type, it will be easier for him to launch an attack.

2. Avoid adding hobbies or religious words to passwords. We found that popular passwords of the game group contain names of game characters, and the religion group prefers using religious words and dates as part of their passwords. Attacks will be easier if attackers know users' hobbies and religions.

6 Conclusion

In this paper, we took the first exploration towards the "group" concept in password evaluation. We conducted large-scale data analysis based on the real-world password dataset and have confirmed the existence of group characteristics in password groups. Based on the group concept, we proposed gPGM, a neural network based password guessing method that takes group information into consideration. Through experiments, we have shown that the cracking rate of using neural network to guess the password is slightly higher than that of PCFG and TransPCFG, and the password cracking rates after adding the group feature AM is 5%-7% higher than that without the AM. Through the cracking rate curves, we

have demonstrated that users' security awareness of entertainment accounts such as games and adult websites is significantly lower than their business accounts, but security awareness of special groups such as hackers is significantly higher than that of ordinary users.

Acknowledgements. This research is supported by the National Key R&D Program of China (2021YFB2700900), the National Natural Science Foundation of China (Grant: 62376074), the Shenzhen Science and Technology Program under (Grant: KCXST20221021111404010, JSGG20220831103400002, JSGGKQTD20221101115655027), the Fok Ying Tung Education Foundation of China (Grant 171058), and Foundation of State Key Laboratory of Public Big Data (No. PBD2023-06). Shanshan Zhu is the corresponding author of this article.

References

1. John's markov generator. https://openwall.info/wiki/john/markov
2. Alebouyeh, Z., Jalaly Bidgoly, A.: Zipf's law analysis on the leaked Iranian users' passwords. J. Comput. Virol. Hacking Tech **18**(2), 101–116 (2022). https://doi.org/10.1007/s11416-021-00397-9
3. Alomari, R., Thorpe, J.: On password behaviors and attitudes in different populations. Inf. Secur. Appl. **45**(19), 79–89 (2019). https://doi.org/10.1016/j.jisa.2018.12.008
4. Bonneau, J.: The science of guessing: analyzing an anonymized corpus of 70 million passwords. In: Proceedings of S&P 2012, pp. 538–552 (2012)
5. Bonneau, J., Xu, R.: Of contrasenas, sysmawt, and mıma: character encoding issues for web passwords (2012). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.294.8406&rep=rep1&type=pdf
6. Canetto, S.S., Lester, D.: Gender, culture, and suicidal behavior. Trans. Psychi. **35**(2), 163–190 (1998)
7. Dell'Amico, M., Filippone, M.: Monte carlo strength evaluation: fast and reliable password checking. In: Proceedings of ACM SIGSA 2015, pp. 158–169 (2015)
8. Dong, Q., Wang, D., Shen, Y., Jia, C.: PII-PSM: a new targeted password strength meter using personally identifiable information. In: Proceedings of EAI SecureComm 2022, pp. 648–669 (2023)
9. Guo, Y., Zhang, Z., Guo, Y., Guo, X.: Nudging personalized password policies by understanding users' personality. Comput. Secur. **94**(1), 101801 (2020). https://doi.org/10.1016/j.cose.2020.101801
10. Han, W., Xu, M., Zhang, J., Wang, C., Zhang, K., Wang, X.S.: TransPCFG: transferring the grammars from short passwords to guess long passwords effectively. IEEE Trans. Inf. Forensics Secur. **16**(10), 451–465 (2021)
11. He, D., et al.: Group-based password characteristics analysis. IEEE Netw. **35**(1), 311–317 (2021)
12. Houshmand, S., Aggarwal, S.: Building better passwords using probabilistic techniques. In: Proceedings of ACSAC 2012, pp. 109–118 (2012)
13. Hranick, R., Zobal, L., Rysav, O., Kolar, D., Mikus, D.: Distributed PCFG password cracking. In: Proceedings of ESORICS 2020, pp. 701–719 (2020)
14. Li, W., Zeng, J.: Leet usage and its effect on password security. IEEE Trans. Inf. Forensics Secur. **16**(1), 2130–2143 (2021)

15. Li, Y., Wang, H., Sun, K.: A study of personal information in human-chosen passwords and its security implications. In: Proceedings of INFOCOM 2016, pp. 1–9 (2016)
16. Li, Z., Han, W., Xu, W.: A large-scale empirical analysis of Chinese web passwords. In: Proceedings of USENIX SEC 2014, pp. 559–574 (2014)
17. Liu, Y., et al.: GENPass: a general deep learning model for password guessing with PCFG rules and adversarial generation. In: Proceedings of ICC 2018, pp. 1–6 (2018)
18. Mazurek, M.L., et al.: Measuring password guessability for an entire university. In: Proceedings of ACM SIGSAC 2013, pp. 173–186 (2013)
19. Melicher, W., Ur, B., Komanduri, S., Bauer, L., Christin, N., Cranor, L.F.: Fast, lean, and accurate:modeling password guessability using neural networks. In: Proceedings of USENIX SEC 2016, pp. 175–191 (2016)
20. Sully De Luque, M.F., Sommer, S.M.: The impact of culture on feedback-seeking behavior: an integrated model and propositions. Pers. Soc. **25**(4), 829–849 (2000)
21. Veras, R., Collins, C., Thorpe, J.: On semantic patterns of passwords and their security impact. In: Proceedings of NDSS 2014, pp. 1–16 (2014)
22. Veras, R., Collins, C., Thorpe, J.: A large-scale analysis of the semantic password model and linguistic patterns in passwords. Trans. Priv. Secur. ACM **24**(3), 1–21 (2021)
23. Wang, D., Cheng, H., Wang, P., Huang, X., Jian, G.: Zipf's law in passwords. IEEE Trans. Inf. Forensics Secur. **12**(11), 2776–2791 (2017)
24. Wang, D., He, D., Cheng, H., Wang, P.: FuzzyPSM: a new password strength meter using fuzzy probabilistic context-free grammars. In: Proceedings of IEEE/IFIP DSN 2016, pp. 595–606 (2016)
25. Wang, D., Wang, P.: The emperor's new password creation policies. In: Proceedings of ESORICS 2015, pp. 456–477 (2015)
26. Wang, D., Wang, P., He, D., Tian, Y.: Birthday, name and bifacial-security: understanding passwords of Chinese web users. In: Proceedings of USENIX SEC 2019, pp. 1537–1555 (2019)
27. Wang, D., Zhang, Z., Wang, P., Yan, J., Huang, X.: Targeted online password guessing: an underestimated threat. In: Proceedings of ACM CCS 2016, pp. 1242–1254 (2016)
28. Weir, M., Aggarwal, S., Medeiros, B.d., Glodek, B.: Password cracking using probabilistic context-free grammars. In: Proceedings of IEEE S&P 2009, pp. 391–405 (2009)
29. Xie, Z., Zhang, M., Yin, A., Li, Z.: A new targeted password guessing model. In: Liu, J.K., Cui, H. (eds.) ACISP 2020. LNCS, vol. 12248, pp. 350–368. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-55304-3_18
30. Yang, K., Hu, X., Zhang, Q., Jianghong, W., Wenfen, L.: VAEPass: a lightweight passwords guessing model based on variational auto-encoder. Comput. Secur. **114**(1), 102587 (2022). https://doi.org/10.1016/j.cose.2021.102587