How does personal information reside in human-chosen passwords? –A quantitative study

Yue Li
College of William & Mary
yli@cs.wm.edu

Haining Wang
University of Delaware
hnw@udel.edu

Kun Sun
College of William & Mary
ksun@cs.wm.com

ABSTRACT

Left blank.

Categories and Subject Descriptors

[**Security and privacy**]: *Human and societal aspects of security and privacy*; [**General and reference**]: *Metrics*

General Terms

Security

Keywords

passwords, password cracking, data processing, password protection

1. INTRODUCTION

Left blank

2. PERSONAL INFO IN PASSWORDS

Human-generated passwords are long criticized to be weak. Numerous works have shown that due to memorability requirement, users are more likely to use meaningful strings as their passwords. Therefore, passwords are usually very different from real random strings. For example, "password" is more likely a password than "ziorqpe". As a result, most passwords are within only a small portion of the large password space, making password guessing a lot easier. A natural question is: how do users choose their passwords so that they are different from random strings? The answer to the question has great significance for it has strong security implication to both users as well as systems. If an attacker knows exactly how users construct their passwords, cracking their passwords will become an easy task. On the other hand, if a user knows how other users construct their passwords, the user can easily improve his/her password strength by avoid using these password construction methods.

To this end, researchers have done much to unveil the composition of passwords. Traditional dictionary attacks on passwords have shown that users tend to use dictionary words to construct their passwords. [][] claims that the distribution of characters in passwords is very similar to that in their native languages and people are prone to use words in their languages. [] Shows password words distribution is not the same as word distribution in the language. [Markov] shows that passwords are phonetically memorable. [PCFG] shows that using dictionary words to guess passwords is effective. [][][] indicate that users use keyboard strings such as "qwerty" and "qweasdzxc", trivial strings such as "password", "123456", and date strings such as "19951225" in their passwords.

As far as we see, most studies are done at a macro level. We now study user passwords in an individual base. We would like to show that user personal experience plays an important role when users create their passwords. Intuitively, people tend to choose their passwords based on their personal information because human beings are limited by their memory – totally unrelated passwords are much less memorable.

2.1 12306 Dataset

In recent years, many password datasets are exposed to the public. Recent works on password measurement or password cracking are usually based on these datasets. Some of these datasets, such as Rockyou, are very large such that they even constitute millions of passwords. Now we are going to use a dataset which we call 12306 dataset to illustrate how personal information is used in user passwords.

2.1.1 *Introduction to dataset*

At the end of year 2014, a Chinese dataset is exposed to the public by anonymous attackers. It is said that the dataset is obtained using social engineering[], in which attackers use datasets at hand to try other websites. We call this dataset 12306 dataset because all passwords are from a website www.12306.com. The website is the official website for online railway ticket booking for Chinese users.

12306 dataset contains over 130,000 Chinese passwords. Having witnessed so many large datasets been leaked out, the size of 12306 dataset is just medium. What makes it special is that together with plain text passwords, the dataset also carries several types of user personal information. For example, user's name, ID number, etc. As the website needs real ID number to register and people need to provide real information to book a ticket, information in the dataset is considered reliable.

Table 1: Most Frequent Passwords

Rank	Password	Amount	Percentage
1	123456	389	0.296%
2	a123456	280	0.213%
3	123456a	165	0.125%
4	5201314	160	0.121%
5	111111	156	0.118%
6	woaini1314	134	0.101%
7	qq123456	98	0.074%
8	123123	97	0.073%
9	000000	96	0.073%
10	1qaz2wsx	92	0.070%

Table 2: Most Frequent Passwords

Rank	Structure	Amount	Percentage
1	D_7	10893	8.290%
2	D_8	9442	7.186%
3	D_6	9084	6.913%
4	L_2D_7	5065	3.854%
5	L_3D_6	4820	3.668%
6	L_1D_7	4770	3.630%
7	L_2D_6	4261	3.243%
8	L_3D_7	3883	2.955%
9	D_9	3590	2.732%
10	L_2D_8	3362	2.558%

”D” represents digits and ”L” represents English letters. The number indicates the segment length. For example, D_7 means the password contains 7 digits in a row.

2.1.2 *Basic Measurement*

We do fundamental measurement to reveal some characteristics of 12306 dataset. After appropriate cleansing, we remove a minor part of passwords (0.2%), with 131,389 good passowrds left for analysis. Just like previous works, we first show the most common passwords in 12306 dataset. They are listed in Table 1

From Table 1 we can see that the dominating passwords are trivial passwords (123456 and mangled 123456 like a123456), keyboard passwords (1qaz2wsx and 1q2w3e4r), and ”I love you” passwords. Both ”5101314” and ”woaini1314” means ”I love you forever” in Chinese. The most commonly used Chinese passwords are similar to previous studies [fudan]. However, 12306 dataset is much less congregated. The most popular password ”123456” accounts less than 0.3% of all passwords while the number is 2.17% in [Fudan]. We believe that the sparsity is due to the importance of the website so that users are less prone to use trivial passwords like ”123456”, etc.

Then, we show the basic structure of passwords. The most popular password structures are shown in Table 2. Our result again shows that Chinese users prefer to use digits in their passwords instead of letters as in English-speaking users. The 5 top structures all have significant portion of digits, in which at most 2 or 3 letters are appended in front.

We reckon that the reason behind may be Chinese users lack vocabulary because Chinese use non-ASCII character set. Digits seem to be the best choice when creating a password.

In conclusion, 12306 dataset is a Chinese password dataset that has general Chinese password characteristics. However, its passwords are more sparse than previously studied datasets.

2.2 Personal Information

As we have mentioned, 12306 dataset not only contains user passwords, it also carries multiple types of personal information. They are:

1. Name: User’s Chinese name
2. Email address: User’s registered email address
3. Cellphone number: User’s registered cellphone number
4. Account name: the account used to log on the system.
5. ID number: Government issued ID number.

Note that the government issued ID number is an 18-digit powerful number. These digits actually show personal information as well. Digit 1-6 represents the birth place of the owner, Digit 7-14 represents the birthday of the owner, and digit 17 represents the gender of the owner – odd number means male and even number means female. We take out the 8-digit birthday information and treat it separately because birthday information is very important in a password. Therefore, we finally have 6 types personal information - 1)Name, 2)Birthday, 3)Email, 4)Cellphone 5)Account name, and 6) ID number (birthday not included).

2.2.1 *New Password Representation*

To better illustrate how personal information correlates to user passwords, we develop a new representation of password which add more semantic symbols beside the conventional ”D”, ”L” and ”S” symbols, which means digit, letter, and special symbol accordingly. We try to match password to the 6 types of user personal information, and express the passwords with these personal information. For example, a password ”alice1987abc” may be represented as $[Name][Birthday]L_3$ instead of $L_3D_4L_3$ in a traditional measurement. We substitute personal information with corresponding tag ([Name] and [Birthday] in this case). For the segments that are not matched, we still use ”D”,”L”, and ”S” to describe the types of characters.

We believe representation like $[Name][Birthday]L_3$ is better than $L_5D_4L_3$ since it more accurately describe the composition of user passwords. We apply the matching to the whole 12306 dataset to see how these personal information tag appear in such password representations.

2.2.2 *Matching Method*

In order to make personal information password representations, an essential question will be: How do we match the personal information to user passwords? To answer this question, we show the algorithm we used in Algorithm 1. The high level idea is that we find all substrings of the password and sort them in descending length order. Then we try to match the substrings from longest to shortest to all types of personal information. If one match is found, the leftover password segments are recursively applied the match function until no further match is found. Segments that are not matched to personal information will be processed use the traditional ”LDS” method.

Note that we did not show specific matching algorithm to each type of the personal information (line 10 and line 16). To keep Algorithm 1 clean and simple, we describe the matching methods as follows.

Algorithm 1 Match personal information with password

```
procedure MATCH($pwd$,$infolist$)
    $newform$ = empty_string
    if l thenen($pwd$) == 0
        return empty_string
    end if
    $substring \leftarrow$ get_all_substring($pwd$)
    reverse_length_sort($substring$)
    for $eachstring \in substring$ do
        if len($eachstring$) $\geq 2$ then
            if matchbd($eachstring$,$infolist$) then
                $tag \leftarrow$ "[BD]"
                $leftover \leftarrow pwd$.split($eachstring$)
                break
            end if
            ...
            if matchID($eachstring$,$infolist$) then
                $tag \leftarrow$ "[ID]"
                $leftover \leftarrow pwd$.split($eachstring$)
                break
            end if
        else
            break
        end if
    end for
    if $leftover$.size() $\geq 2$ then
        for i $\leftarrow$ 0 to $leftover$.size()-2 do
            $newform \leftarrow$ MATCH($leftover[i]$,$infolist$) + $tag$
        end for
        $newform \leftarrow$ MATCH($leftover[leftover.size()-1]$)+$newform$
    else
        $newform \leftarrow$ seg($pwd$)
    end if
    return $newform$
end procedure
```

Table 3: Most Frequent Passwords

Rank	Structure	Amount	Percentage
1	D7	7105	5.407%
2	[ACCT]	6103	4.644%
3	[NAME][BD]	5410	4.117%
4	D6	4873	3.708%
5	[BD]	4470	3.402%
6	D8	4233	3.221%
7	L1D7	3286	2.500%
8	[NAME]D7	2941	2.238%
9	[NAME]D3	2363	1.798%
10	[NAME]D6	2241	1.705%

Table 4: Most Frequent Passwords

Rank	Information Type	Amount	Percentage
1	[BD]	30674	23.34%
2	[NAME]	29653	22.56%
3	[ACCT]	17065	12.98%
4	[EMAIL]	4229	3.218%
5	[ID]	2918	2.220%
6	[CELL]	529	0.402%

First we make sure the password segments are at least length of 2. For segment of length 1, we directly map it to digit, letter, or special character.

For name information, we first convert Chinese names into Pinyin form, which is alphabetic representation of Chinese. Then we compare password segments to 10 possible permutations of the names, which include $lastname + firstname$, $lastinitial + firstname$, etc. If the segment is exactly same as any of the permutations, we consider a match is found. We list all the 10 permutations in the Appendices.

For birthday information, we list 17 possible permutations and compare password segments to each of the permutation, if the segment is same as any permutations, we consider a match is found. We list all the birthday permutations in the Appendices.

For account name, cellphone number, and ID number, we further restrain the length of segment to be at least 4 to avoid coincidence. We believe a match of length 4 is very likely to be an actual match. If the segment is a substring of any of the 3 personal information, we regard it a match to the corresponding personal information.

2.2.3 Matching Result

After applying Algorithm 1 to 12306 dataset. We found that 71,037 out of 131,389 (54.1%) of the passwords contain at least one of the 6 types of personal information. Apparently, personal information is an essential part of user passwords and most users put certain personal information in their passwords. We believe the rate could be higher if we have more personal information at hand. However, this percentage has served its purpose properly. Then We present the top 10 password structures in Table 3 and most commonly used personal information in Table 4. Based on Table 3 and Table 4, we have the following observations

1. The second and third structures are perfectly matched to personal information. 3 out of the top 10 structures are composed by pure personal information and 6 out of the top 10 structures have personal information seg-

ment. The dominating structures D_7, D_6, and D_8 in Table 2 still rank fairly high.

2. Birthday, name, and account name are most popular personal information in user passwords.
3. Set aside personal information, digits are still dominating user passwords. The result confirms again that Chinese users prefer to use digits in their passwords.

2.3 Service Information

2.4 Ethical Consideration

We do realize that studying leaked datasets involves much ethical concern. Like many other works, we only use this dataset for researching purpose. We will not expose any user personal information of in this dataset.

3. CORRELATION QUANTIFICATION

3.1 Coverage

4. PERSONAL-PCFG, AN INDIVIDUAL-ORIENTED PASSWORD CRACKER

4.1 Attack Scenarios

5. PASSWORD PROTECTION

6. RELATED WORKD

Left blank

7. REF

7.0.1 *Inline (In-text) Equations*

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin. . .\end` construction or with the short form `$. . .$`. You can use any of the symbols and structures, from α to ω, available in LaTeX[5]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\rightarrow\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

7.0.2 *Display Equations*

A numbered display equation – one set off by vertical space from the text and centered horizontally – is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\rightarrow\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_{0}^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

Table 5: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

7.1 Citations

Citations to articles [1, 3, 2, 4], conference proceedings [3] or books [6, 5] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [5]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide*[5].

This article shows only the plainest form of the citation command, using `\cite`. This is what is stipulated in the SIGS style specifications. No other citation format is endorsed or supported.

7.2 Tables

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material is found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed dvi output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed dvi output of this document.

7.3 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of

Table 6: Some Typical Commands

Command	A Number	Comments
`\alignauthor`	100	Author alignment
`\numberofauthors`	200	Author enumeration
`\table`	300	For tables
`\table*`	400	For wider tables

Figure 1: A sample black and white graphic (.eps format).

Figure 2: A sample black and white graphic (.eps format) that has been resized with the `epsfig` command.

the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of **.eps** and **.ps** files to be displayable with LaTeX. More details on each of these is found in the *Author's Guide*.

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption. and don't forget to end the environment with figure*, not figure!

Note that either **.ps** or **.eps** formats are used; use the `\epsfig` or `\psfig` commands as appropriate for the different file types.

7.4 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. There are two forms, one produced by the command `\newtheorem` and the other by the command `\newdef`; perhaps the clearest and easiest way to distinguish them is to compare the two in the output of this sample document:

This uses the **theorem** environment, created by the `\newtheorem` command:

Theorem 1. *Let f be continuous on $[a,b]$. If G is an antiderivative for f on $[a,b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

The other uses the **definition** environment, created by the `\newdef` command:

Definition 1. If z is irrational, then by e^z we mean the unique number which has logarithm z:

$$\log e^z = z$$

Two lists of constructs that use one of these forms is given in the *Author's Guidelines.*

There is one other similar construct environment, which is already set up for you; i.e. you must *not* use a `\newdef` command to create it: the **proof** environment. Here is a example of its use:

Proof. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)}\right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

Complete rules about using these environments and using the two different creation commands are in the *Author's Guide*; please consult it for more detailed instructions. If you need to use another construct, not listed therein, which you want to have the same formatting as the Theorem or the Definition[6] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

A *Caveat* for the TeX Expert

Because you have just been given permission to use the `\newdef` command to create a new form, you might think you can use TeX's `\def` to create a new command: *Please refrain from doing this!* Remember that your LaTeX source code is primarily intended to create camera-ready copy, but may be converted to other forms – e.g. HTML. If you inadvertently omit some or all of the `\def`s recompilation will be, to say the least, problematic.

8. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

9. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

10. REFERENCES

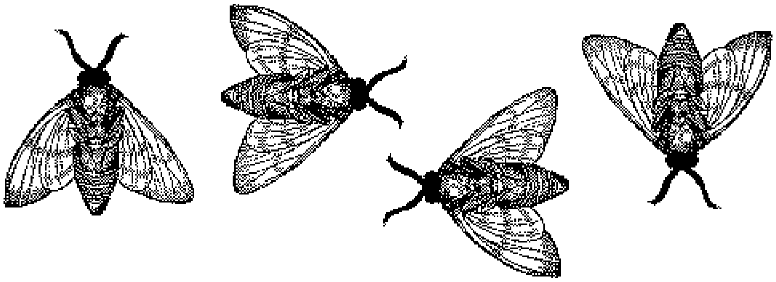

Figure 3: A sample black and white graphic (.eps format) that needs to span two columns of text.

[1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[2] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[3] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[4] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[5] L. Lamport. *LaTeX User's Guide and Document Reference Manual*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

[6] S. Salas and E. Hille. *Calculus: One and Several Variable*. John Wiley and Sons, New York, 1978.

APPENDIX

A. HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e. the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

A.1 Introduction

A.2 The Body of the Paper

A.2.1 Type Changes and Special Characters

A.2.2 Math Equations

Inline (In-text) Equations.

Display Equations.

A.2.3 Citations

A.2.4 Tables

A.2.5 Figures

A.2.6 Theorem-like Constructs

A Caveat for the TEX Expert

A.3 Conclusions

A.4 Acknowledgments

A.5 Additional Authors

This section is inserted by LATEX; you do not insert it. You just add the names and information in the `\additionalauthors` command at the start of the document.

A.6 References

Generated by bibtex from your .bib file. Run latex, then bibtex, then latex twice (to resolve references) to create the .bbl file. Insert that .bbl file into the .tex source file and comment out the command `\thebibliography`.

B. MORE HELP FOR THE HARDY

The sig-alternate.cls file itself is chock-full of succinct and helpful comments. If you consider yourself a moderately experienced to expert user of LATEX, you may find reading it useful but please remember not to change it.