Uncovering strategies for password creation and updation for Indian users

Yash Saraswat
Department of Chemical Engineering
Indian Institute of Technology, Kharagpur, India
yashsaraswat@gmail.com

Mukul Mehta
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India
mukul.csiitkgp@gmail.com

Rashil Gandhi
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India
rashil2000@gmail.com

MOTIVATION

The Indian Computer Emergency Response Team (CERT-In) observed over 0.6 million cyber security incidents in the first six months of 2021, of which about 12,000 incidents were related to government organisations. According to data from CERT-In, which is mandated to track and monitor cyber security incidents in the country, a total of 607,220 cyber security incidents were observed during 2021 June. This number stood at 208,456 in the year 2018; 394,499 in 2019; and 1,158,208 in 2020.

Password-based authentication is the most extensively used method of safeguarding users' accounts and personal data on web-based services throughout the world. Passwords are used by millions of individuals every day to authenticate and authorise them to access their email, conduct financial transactions, interact with government agencies, convey confidential data, conduct sensitive activities, and more. Making password creation and updation one of the most essential components in security analysis.

As with most of the areas under the Human-Computer Interaction umbrella, password creation and updation is highly influenced by the user's demographic variables like the country of origin, cultural background, primary language spoken, etc. The rise in the number of internet users in India and the subsequent rise in the number of registered Indian users on numerous websites demands a study that looks into how the country's users deal with their passwords.

This study aims to understand how factors like cultural references, local languages, current affairs, etc. affect the choice of passwords for Indian users. We will also look into how this affects the perceived strength of the passwords when compared to existing databases and studies which aren't rich with Indian accounts.

1 Research Questions

1. **What patterns/entities from real life are Indian people most likely to base their passwords on?**
   People are most likely to enter the text they find the easiest to remember in the long run. As such, passwords can be broadly classified into patterns of name-date combinations, phone numbers, popular sayings, names of famous people etc. A linguistic analysis of the password strings can help us quantify what patterns are the most common.

2. **What are the similarities or differences between the linguistic patterns of passwords created by people of Indian origin and those of foreign origin?**
   Does the extreme diversity of culture and languages in India contribute to the strength of passwords? And what patterns/combinations of password strings are more common in the Indian context, compared to others?

2 Dataset Information

The study is based on a dataset that was obtained from a Raid Forum which claimed to have Indian Email-Password pairs in cleartext. The text file contains around 400,000 such pairs in its raw form.

2.1 Ethical considerations

Given the fact that the study is performed on a cleartext database that contains personal information like email addresses, we haven't made the database available publicly (even though the Raid Forum databases can be openly accessed).

The steps that were taken to ensure an ethical study:

- Following industry best practices when it comes to storing data in an encrypted way.
- Not spreading the disclosed information in any way, including sharing or submitting passwords or hashes to third-party services.
- Not releasing any information that might lead to the identification of any of the people involved in the leak.

2.2 Sampling the data

Given the source of the dataset, it is prudent to perform some sort of manual verification that the data is indeed legitimate.

Since there is no conclusive way of verifying that the data is from users of Indian origin, prior to the study of the dataset, we sampled the given data into several randomly generated samples [code] of sample size 100. These samples were then be checked by people of Indian origin to receive a satisfactory nod on the data points having common phrases or words from the local languages, which points towards the fact that the data points are most probably from users of Indian origin.

The results from the study performed, further confirm the fact that the passwords are in fact from users of Indian origin.

2.3 Sanitising the data

The data was first cleaned [code] by eliminating data points with the following properties:

- Pure duplicate entries, i.e. entries having the same Email ID and password.
- Data points with having the same Email IDs but different passwords were not removed to study the user's behaviour when creating accounts at multiple sites or just updating their passwords.
- Email IDs are not present in the usual format.
- With either a NULL Email ID or password.

After this preliminary round of cleaning, we found that there were many unusual domains that were present in the dataset. These domains often had only one occurrence in the file but passwords like "x4ivyga51f" were common for all such entries. These domains were most probably temporary instance accounts created on various forums, explaining the common password and unusual domain names.

To clear such domains, we created a whitelist [file] of domains that were significant for the purpose of the study. The whitelisted domains included:

- Top 20 most frequent domains in the database. This covered around 90.3% of the preliminary cleaned data.
- All domains were registered under the .in name.

The above sanitation gave us a total of 329,142 data points that were used for the study.

2.4 Zipf-y data

Zipf's law is an empirical statement stated using mathematical statistics that states that the rank-frequency distribution is inverse for many types of data analysed in the physical and social sciences. The Zipfian distribution is a discrete power law probability distribution that belongs to a family of related discrete power law probability distributions.

A distribution is said to follow the Zipf's distribution if the frequency (f_r) and the rank (r) follow the following relation:

$$f_r = \frac{C}{r^s}$$

$$\ln f_r = \ln c - s \times \ln r$$

As was shown in a previous study [7], Zipf's law is known to perfectly exist in user-generated passwords. Therefore, we investigated the distribution of the password frequency with respect to the rank of the password.

Figure 1 [code] shows that even though the data seems to follow a linear pattern in the mid sections of the distribution, it does not follow a linear fit with a reasonable accuracy ($R^2 = 0.734$). The best fit is given by:

$$\ln f_r = \ln 26.37 - 0.27461 \times \ln r$$

We assumed this might be the case because of the extensive use of pure numerical data in the dataset (discussed in Section 3.5), and hence plotted the linear fit for passwords excluding the class of passwords that were 10 digit long numbers:

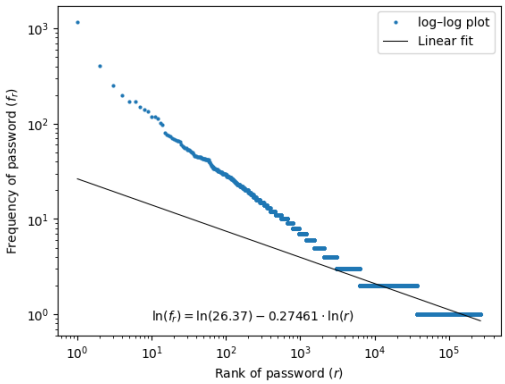

Figure 1: **Zipf linear modelling for the unaltered dataset**

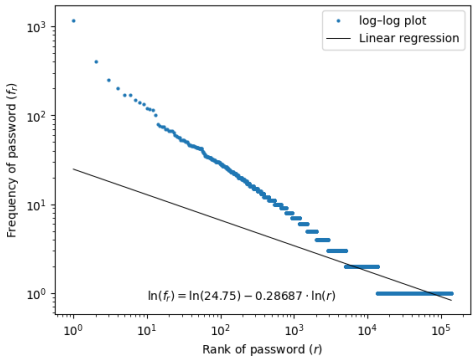

Figure 2: **Zipf linear modelling for the altered dataset**

To our surprise, the fit worsened ($R^2 = 0.664$) this time with the it being modelled on the following relation:

$$\ln f_r = \ln 24.75 - 0.28687 \times \ln r$$

This provides us with an interesting insight into the usage of phone numbers, which is that most of them are used multiple times. Pointing to the fact that users use their phone numbers as passwords at not only one, but multiple accounts. We will discuss the usage of phone numbers at length in Section 3.5.

We then tried various polynomial models to fit the data. Starting degree of polynomial being two, upto four. Polynomial fits of degree greater than four, overfitted the model.

Figure 3 shows that a model with polynomial degree [code] of 4, fits the data with reasonable accuracy ($R^2 = 0.880$). Even though this is our best polynomial fit, it is nowhere close to the accuracies reported in the previously mentioned study [7]. We would need a more sophisticated model to try and fit the data, which is beyond the scope of the study.

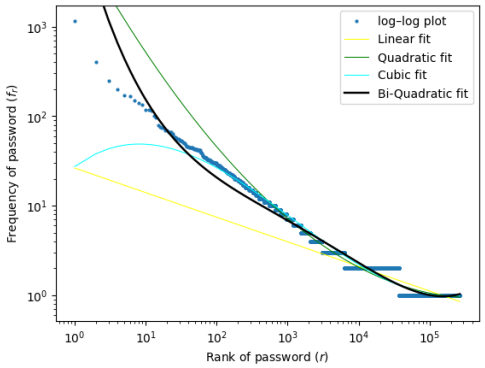

Figure 3: **Zipf modelling using polynomial fits**

The Bi-Quadratic model is of the following form:

$$\ln f_r = \ln C_0 + C_1 \ln r + C_2 \ln^2 r + C_3 \ln^3 r + C_4 \ln^4 r$$

The model that best fits the data follows the relation

$$\ln f_r = \ln 0.0012 + 0.95 \ln r + 0.56 \ln^2 r - 3.18 \ln^3 r + 9.94 \ln^4 r$$

This conclusively proves that the dataset we are operating with in the study does not follow a Zipf's distribution. This is attributed to the huge chunk of password strings being only used once, which are not necessarily the 10 digit numerical passwords.

3 Demographic Based Analysis

We give a range of typical password analysis in this part, but we break down the results using the demographic significance of the passwords we established earlier.

3.1 Most frequent base words

Along with the most used passwords [code], we extract the most used base words and display them in Table 1 to compare these groups in terms of the most commonly used tokens or patterns. We used pipal [1], a well-known tool for password analysis, to carry out the aforesaid process.

A base word is a piece of string that is common between multiple passwords, which does not have to be consecutive in the password. For example, the password "pass123", "p1a2s3s" and "pass@345" have a common base word "pass". Pipal is a powerful, yet lightweight tool that helped us in enumerating these base words from the database.

Base Words	Frequency
password	650 (0.2%)
abhi	433 (0.13%)
h54rsjrf5j	421 (0.13%)
welcome	408 (0.12%)
aditya	389 (0.12%)
india	379 (0.12%)
qwerty	350 (0.11%)
amit	299 (0.09%)
abcd	273 (0.08%)
shubham	270 (0.08%)
krishna	257 (0.08%)
pass	255 (0.08%)
sachin	234 (0.07%)
ashish	230 (0.07%)
shreya	224 (0.07%)
ayush	215 (0.07%)
yash	214 (0.07%)
ankit	212 (0.06%)

Table 1: **Top 20 Base Words**

From Table 1, we can see that along with base words such as “password”, “welcome” and “qwerty” which are commonly seen throughout password studies, we observe words such as “abhi”, “aditya”, “amit”, etc. which are names commonly used in India.

As we show in Section 4.1 state-of-the-art password metres do not perceive such base words as common, resulting in exaggerated entropy readings for passwords including certain contextually frequent base words.

Another interesting observation is that words such as “delhi” and “mumbai”, which are the names of prominent cities in India, are used less frequently than names for setting a password. We investigate this observation at length in Section 3.4.

This also cements our belief in the fact that the database has been taken from users of Indian origin.

3.2 Password length

Password length is one of the most essential components of the composition of the password. It gives us an insight into how much of an “effort” the user makes to remember the password and what is ideally length of a password that is easy to remember.

Database studied	RockYou
10 = 183635 (55.79%)	6 = 8497562 (26.06%)
8 = 50288 (15.28%)	8 = 6504916 (19.95%)
9 = 33985 (10.33%)	7 = 6284712 (19.28%)
11 = 16189 (4.92%)	9 = 3938519 (12.08%)
6 = 15140 (4.6%)	10 = 2943315 (9.03%)
7 = 10866 (3.3%)	5 = 1343832 (4.12%)

Table 2: **Password lengths**

Table 2 compares the length of passwords present in the database to that present in the RockYou [2] database. We clearly observe a huge percentage of passwords being of length 10 characters (55.79%), followed by 8 characters (15.28%). In the case of the RockYou database, the password lengths were prominently in the range of 6-8 characters (71.88%).

This anomaly is observed due to a huge part (46.75%) of the password dataset being 10-digit numerical values. It is shown in Section 3.5 that these are most likely phone numbers of the users set as passwords.

3.3 Password composition

We analysed how the passwords are structured and what combinations of string, numeric and special characters are used for their creation. In Table 3, string values (a-z, A-Z) are referred to as S, numerical values (0-9) are denoted by N and special characters by SC. Therefore S_SC_N refers to a password that starts with a string followed by a special character and ends with a numerical value. Table 3 also contains these compositions for the passwords retrieved from the RockYou dataset for comparison.

Database studied	RockYou
N: 168862 (51.3%)	S: 14446520 (44%)
S_N: 81232 (24.68%)	S_N: 9833381 (30%)
S: 37214 (11.31%)	N: 15194466 (15%)
S_SC_N: 19202 (5.83%)	N_S: 896003 (2%)
Other: 12896 (3.92%)	Other: 631963 (1%)
S_N_S: 4268 (1.3%)	S_N_S: 597558 (1%)

Table 3: **Password composition**

As discussed earlier in Section 3.2, purely numerical values dominate the database. An interesting observation is the presence of S_N and S_N_SC type passwords. Upon further investigation, these passwords mostly comprise of a string followed by a four or digit number (in the case of S_N), or separated by a special character like “@”. The numbers are generally “19”, “05” or “2019”, “2005” for two and four-digit types respectively. This was an obvious indication towards users using years followed by a string in their passwords.

Year	Frequency
2019	3271 (0.99%)
2005	1361 (0.41%)
2004	1293 (0.39%)
2006	1114 (0.34%)
2007	1103 (0.34%)
2008	1003 (0.3%)

Table 4: **Year wise distribution**

Table 4 shows our analysis of the entire dataset for the presence of “year-like” numbers at the end of passwords. The years could possibly be the user’s year of birth or the year of creation of the account.

There can be two possible reasoning for the distribution observed in Table 4:

- Assuming that the database contains data from one leak, one can deduce that 2019 is probably the year used in the context of year of account creation, whereas the other years are most probably the year of birth of the users.
- The dataset can also be a dump accumulating password leaks over the years, explaining the spread out distribution in the years.

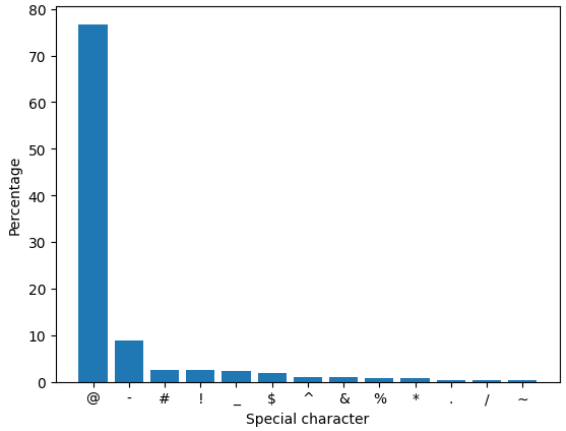

Figure 4: **Most used Special Characters**

Due to the important role the special characters play in this context, Figure 4 shows the distribution of various special characters used in the passwords. As expected from the above observations, "@" is the most used special character.

3.4 Use of names

As discussed earlier in Section 3.1, the most used base words are commonly used Indian names. To investigate this further, we acquired a dataset [3] with the most commonly used Indian names which contained nearly 6,500 names after processing. We eliminated numerals and special characters from each password in the datasets and searched the name list for full string matches (ignoring case because it is unnecessary) whose lengths are greater than or equal to σ. Figure 5 shows the percentage (with respect to passwords that have at least one alphabet, i.e. eliminated pure numerical values) of passwords having a hit on a name with σ values between 4-6. [code]

σ values two and three were dropped to reduce the number of false positives, as there is a greater chance that names of this length have a match with a password even though they are not used in that context. For example, the name "ali" will have a hit with the password "alisha".

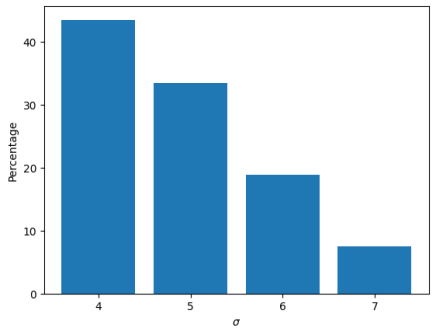

Figure 5: **Most used Special Characters**

In contrast to a prior study [4] that found just 14% of passwords contained names, our findings reveal that names appear in a significant number of passwords. Another research that looked at a stolen dataset of Chinese users [5] discovered that 22% of people use their own names as passwords.

3.5 Use of phone numbers

As discussed earlier in Section 3.2, a major chunk (55.79%) of the dataset is composed of purely numerical values that are 10 digits long. Almost all (99.52%) of such passwords start with the numbers 6-9 which are usually the digits with which phone numbers start in India. This confirms the fact that these 10 digit long numbers that make up more than half of the dataset are phone numbers. This is another important insight into password creation patterns for Indian users. [code]

As observed in Section 2.4, most of the phone numbers are used multiple time by users to create passwords for their multiple accounts. One can pose a threat to the safety of more than one of their accounts, by just knowing their phone numbers. This is a serious security threat to numerous users across the country.

4 Strength Based Analysis

In this section, we consider the following adversarial model.

A: The attacker attempts to crack a password using a dictionary or brute-force assaults. To recover a set of unknown cryptographically hashed passwords, the attacker uses external datasets (training sets). We utilise zxcvbn [6], a state-of-the-art password strength estimator, to determine the entropy in bits to assess the guessability of our datasets against this attacker model.

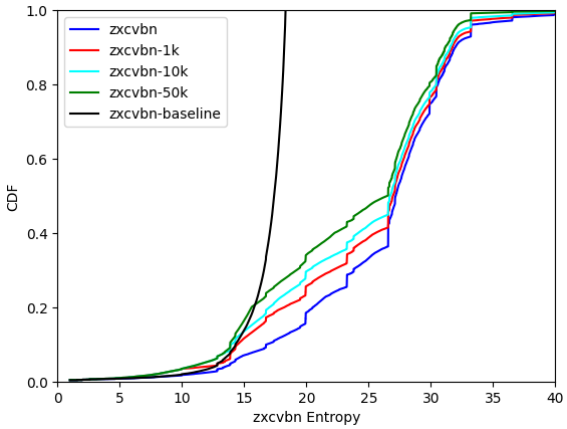

Figure 6: **zxcvbn Entropy plots for varied training**

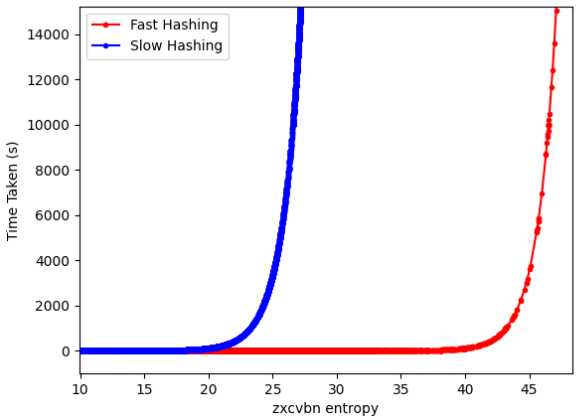

Figure 7: **Estimated time taken for unlashing passwords with zxcvbn predicted Entropy**

4.1 The adversary model

We use zxcvbn for estimating the entropy in bits of the passwords present in the dataset.

In three steps, zxcvbn calculates the strength of a password. In order to locate a collection of S overlapping substrings in the password, a matching step is done first. The dictionaries (of popular names, passwords, keyboard walks, patterns, and so on) that zxcvbn develops based on frequency are used to find such matches. Following that, a scoring phase awards a guess attempt estimation to each match. Finally, look for a sequence S′ of non-overlapping contiguous matches taken from S that completely covers the password while lowering the total guess attempt score.

To improve the zxcvbn findings, we used pipal extracted base words used in Section 3.1 to train the tool. When zxcvbn is fed with increasing sizes of base word sets, the entropy distribution findings for each of the four groups are shown in Figure 6. To put the results in context, we used zxcvbn to generate baseline entropy distributions for each group by feeding it the whole passwords dataset. [code]

As expected, with increased training zxcvbn produces better results over the passwords. For example, the password “jaymataji” was given an entropy value of 27.50 before training, but the value dropped to 17.79 after training. *jaymataji* is a commonly used term in India, being culturally significant as it is associated with religious practices. Therefore, the password’s perceived strength was much higher despite its common usage due to the lack of Indian words present in the default dictionary of zxcvbn.

Figure 7 gives an estimate of the time required to retrieve the passwords from their hashes given their entropy values. The methods used for recovery are Fast Hashing (offline fast hashing: 1e10 per second) and Slow Hashing (offline slow hashing: 1e4 per second). From Figure 5 and Figure 7 we can see that 50% (entropy ≈ 25) of the passwords can be recovered within 3,000 seconds even when using slow hashing offline.

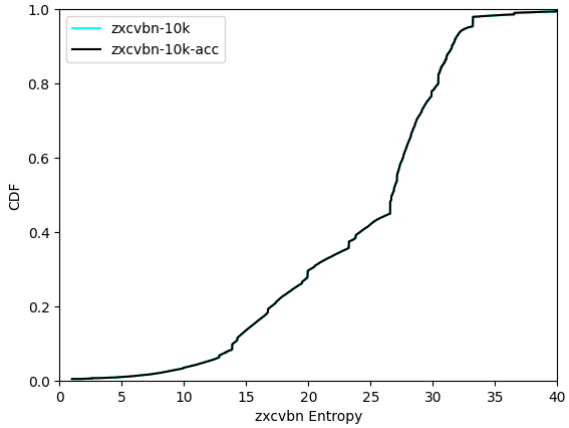

Figure 8: **zxcvbn Entropy plots for when email ID is provided to the tool**

We also analyse the entropy values given by zxcvbn when it is also given with the email ID for the individual passwords. As seen in Figure 8, there is practically no difference in the entropy values for the dataset. In fact, only 4,239 passwords have different entropy values. This demonstrates the lack of a correlation between email ID naming and password creation. [code]

5 Conclusion and Future Work

This study extensively shows how for a region like India, the cultural influence on password creation is massive. The results we have obtained highlight some of the most key rules Indian users follow while creating passwords.

We have also shown the lack of Indian words in the dictionaries used to train password strength meters like zxcvbn. We have created a prototype for a zxcvbn on-the-go trainer that trains on custom dictionaries and operates with a pre loaded base word dictionary from our study. [code].

6 Work distribution

Yash performed the cleaning, and analysis over zxcvbn models.

Rashil performed the whitelisting, Zipf analysis and frequency analysis

Mukul handled the pipal analysis and all of the linguistic analysis and classification.

References

[1] Pipal, GitHub, https://github.com/digininja/pipal

[2] RockYou dataset, kaggle, https://www.kaggle.com/wjburns/common-password-list-rockyoutxt

[3] Indian name dataset, kaggle, https://www.kaggle.com/ananysharma/indian-names-dataset

[4] T. Hunt, The Science of Password Selection, https://www.troyhunt.com/science-of-password-selection/

[5] Y. Li, H. Wang, K. Sun, A study of personal information in human-chosen passwords and its security implications, in 35th Annual IEEE International Conference on Computer Communications, INFOCOM 2016, San Francisco, CA, USA, April 10-14, 2016, 2016, pp. 1–9.

[6] D. L. Wheeler, zxcvbn: Low-budget password strength estimation, in 25th USENIX Security Symposium (USENIX Security 16), USENIX Association, Austin, TX, 2016, pp. 157–173. Zxcvbn, GitHub, https://github.com/dropbox/zxcvbn

[7] D. Wang, H. Cheng, P. Wang, X. Huang and G. Jian, "Zipf's Law in Passwords," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2776-2791, Nov. 2017, doi: 10.1109/TIFS.2017.2721359.