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 up from June. This number stood at 208,456 in the year 2018; 394,499 in 2019; and 1,158,208 in 2020.

Passwords are one of the most used ways of protection employed at various websites. The rise in the number of internet users in India and the subsequent rise in the number of registered Indian users on numerous websites asks for a study that looks into how the country's users deal with their passwords. While there are several studies that deal with analysing password dumps, there are very few that are conclusive enough when it comes to how the user's region affects their choices of 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 compares to existing studies which aren't region-based.

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 Study Design

The study is performed on a dataset of around 400,000 raw account-password clear text pairs. The dataset was acquired from a Raid Forum. We start the study by sanitising the data by removing the following classes of data points:

- Pure duplicate entries, i.e. entries having the same Email ID and password.
  Data points with having 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.
- With Email IDs having no @ [at] symbol present.
- With either a NULL Email ID or password.

After the data is sanitised, we will be analysing the clean data set to study the patterns and common occurrences in order to answer the research questions.

This will be the general workflow for the analysis

2.1 Sampling the data

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 sample the given data into several randomly generated samples of sample size 100. These samples will then be checked by people of Indian origin to get a satisfactory nod on the data points having common phrases or words from the local languages, which will point towards the fact that the data points are most probably from users of Indian origin.

2.2 Frequency of strings

We start by performing several preliminary runs, like finding out the most common password strings, domains and Email ID occurrences.

This will help us draw conclusions about the password frequency list on the dataset and how similar it is to the public lists of most used passwords.

2.3 ID-Password relationship

We then identify the entries that have a clear relationship between the Email ID and the password string: common substrings.

This helps us identify user behaviour patterns when setting up their passwords. Despite the string strength, it is easier to break passwords that are related to the IDs when it comes to targeted attacks.

2.4 Linguistic analysis

We can now start categorising and identifying password strings that are based on cultural references, local languages, current affairs, etc. This will provide us with an insight as to how regionality affects the user's choice of passwords. Linguistic analysis to find the patterns and categorising the strings to find out the most used categories of passwords is perhaps the core of the study while finding out the demographic influence over the password strings.

To categorise the data by commonly used local strings, we first do a base word frequency check on it. After having a comprehensive base word data we classify the most used "kinds" of strings, based on local language occurrences and cultural relevance.

2.5 Password strength and comparison

We will finally perform a test as to how various password cracking tools, perform against the passwords in the dataset when compared to a dataset that is void of Indian phrases.

This will help us distinguish between the strength of the strings of the passwords and how commonly used it actually is in the user's region.

3 Study Instruments

The study will be carried out using a series of scripts written in the Python programming language. The following methods will be used to perform tasks mentioned in the Study Design.

3.1 Linguistic analysis

We do the categorisation of substrings using commonly used algorithms like the Maximum-Entropy model for part-of-speech tagging.

We can also use common categories like the following for tagging the data:

- **Common passwords**: common words, number sequences, popular names, places etc.
- **Password Character Composition**: most commonly - words followed by digits.
- **Use of names**: analysing substrings between the email ID and the passwords.
- **Keyboard walks**: patterns found on a computer keyboard.
- **Phone numbers**: along with other all-numeric combinations (like date of birth)

3.2 General string strength

We analyse the strength of individual password strengths using tools that measure the entropy of the password string. This can be done using the state-of-the-art password strength estimator, zxcvbn. This helps us evaluate if these tools overestimate the strength of the passwords created by non-English speaking backgrounds.

4 Justification

The categorisation of passwords into pattern groups will help us gauge how much culture and regionality affects password choices. Based on experience, we expect the passwords to be broadly classified into patterns of name-date combinations, phone numbers, popular sayings, names of famous people etc. This will help us answer the first research question.

To answer the second research question, we compare the strength (as gauged by password cracking tools) of the commonly used passwords from the dataset to that from a publicly available password dump. We aim to differentiate between the common use of the password string in the region and its apparent strength.

How people mix personal information with literary devices to help them remember the passwords is something we expect to learn and contrast these outcomes with those of foreign cultures. Our overall analysis will also help us substantiate our hypothesis that a culturally diverse demographic is more likely to create a (lexically) stronger password

5 Work distribution

Currently, the work of obtaining and sanitising the dataset was performed by Yash. Preliminary analysis and obtaining frequencies, sorting, and another basic text analysis was done collectively.