Name: Gan Chong Yee
Class: CS 5435 – Security and Privacy Concepts in the Wild
Title: Homework 3

The honeyword generation strategies were described in Section 1, Section 2 and Section 3 respectively. Each algorithm were instances of a single master honeyword algorithm, with slight alterations in the approach based on the training set, T, used in each case.

Section 1 – T is the empty set

Given an input file with a list of true passwords, the algorithm performed 2 operations on each true password to generate n number of honeywords.

The first alteration to the true passwords involved generating 2 digits to be prepended at the head or appended at the tail of the true passwords. These 2 digits were meant to simulate the birth year of the users. Hence, we had chosen to generate the digits with a range from 60 – 99 to reflect the distribution of internet users by age groups[1]. The decision to prepending/ appending/not doing anything to the true passwords were chosen at random, with N(a) of the honeywords being appended, N(p) being prepended and N(n) having no action done on them. This reduced the guessing probability to 2.44% (no changes and 40 possible digits) in the case of word passwords, whilst increasing the entropy of the sweetwords by creating honeywords of different lengths in the case of full digit or hybrid (half digits half words) true passwords.

Next, each letter within each true password was capitalizedwith a random probability once again. For instance, should we choose to generate three honeywords, the true password, "Admin" will output aDMin, admiN and AdMin. This alteration would be skipped should the true password contained full digits, such as "123456". Whilst this method did not increase the entropy significantly, it managed to reduce the guessing probability in cases of compound words (ChongYee, RockYou), or just to distract adversaries from true passwords who had the same aforementioned format (random capitalizations)

Section 2 – T is the top 100 RockYou Passwords

In this case, the honeywords were generated via a base chosen from either the user's inputted true password, or from the top 100 most common RockYou passwords. The operation mentioned in Section 1 was then performed on each of these bases.

This meant that given n number of honeywords, each honeyword had a probability of N(u) of being generated from the user's input, and N(ry) of being generated from the top 100 RockYou passwords. This approach significantly reduced the guessing probability, the top 100 RockYou password had a high cumulative probability of being a true password used by a user.

[1] http://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/

Section 3 – T is the fullRockYouDataset

The approach used in Section 3 was highly similar to Section 2, just that there was now N(ry) probability that the honeyword base was chosen randomly from the full RockYou dataset rather than just the top 100 most common RockYou passwords. This approach had a higher probability of guessing the true password, as there was a chance that the RockYou password was chosen at the bottom of the list, thus making it more obvious to the adversary that said password was a honeyword (fake).