CS 6431: Security and Privacy Technologies
Homework #1

Instructors: Tom Ristenpart and Vitaly Shmatikov

Due: February 23th, 2018

To start, obtain the Rockyou password leak, it is publicly available at various websites, including:

`https://wiki.skullsecurity.org/Passwords`

Particularly check out rockyou-withcount.txt. Get at least one additional large password dump online that will be used as a test set. Your responses to the below should be a short PDF written in the style of a research paper. Answer as many parts as you can; you'll also get partial credit for reasonable attempts at answers — give a discussion of where the approach seems to fail. Please use Latex, one column format with standard margins. No need for an abstract. Each student should provide their own writeup. Feel free to leverage any known results from the literature. As with any research paper, you must cite appropriately any approaches used from previous sources, and other students who gave you ideas.

1. Implement an n-gram Markov model, trained using the Rockyou leak. The n value is your choice, though you should justify your choice. If you want to use a different model type, feel free to do so.

2. Explain an algorithm for estimating the guess rank of a given password according to your model. If you use the Monte-Carlo methods mentioned in class, then justify your choice of parameters for it (the number of pre-computed passwords).

3. Provide a graph in the style of the Melicher et al. paper showing the brute-force guessing efficacy of your model against the test set. Also include

in the graph the trivial attacker that simply uses the RockYou empirical distribution to make guesses. Explain how you computed the data for the graph. Use the pgfplots package for Latex (or something that looks equally professional — no grainy images, please!).

4. Describe an efficient algorithm for computing a list, in order, of the k most-probable passwords according to your probability model. Try to give both a model-specific algorithm that leverages the structure of the model (i.e., that it is a Markov model) and a generic algorithm that just uses the ability to sample a password according to the model's learned distribution and the ability to compute a guess rank for any given password. Give a run time analysis of your algorithm.