Cryptographic Hash Algorithm Identifier

John Blodgett, Harshit Mittal, Shubham Sengar

Summary Statement

The goal of this project is to use machine learning to be able to differentiate between hashed files and sequences of random bits. We would also like to determine which hashing algorithms are most indistinguishable from random and if certain file types abet such distinguishing.

Specific Aims

- Are hashing algorithms as random as they were intended to be?
- Are there certain types of files that are hashed in a way that is less random than others?
- What kind of algorithm is best served to differentiate hashes from randomness?

Timeline

Note that we already code that runs a very rudimentary, basic simulation

- Apr 30
    - Conduct research on different hashing algorithms to use and which would be appropriate (are there some that have security vulnerabilities) (All)
    - Learn how to use the hashing algorithms to create data (Harshit)
    - Populate data with hash outputs of files using an assortment of hashing algorithms (John, Harshit)
    - Transform, clean, and manipulate data (Shubham, Harshit)
- May 7
    - Separate data into training and testing sets (Shubham)
    - Perform preliminary data analysis using standard processing tools and libraries (pandas, numpy, matplotlib) (All)
        - Determine the distribution of bits: does it differ based on groupings of algorithm, file type
        - Metrics such as average distance of a bit sequence from the others, etc. for each algorithm, file type grouping
- May 14
    - Research and create basic models and validate them using results from preliminary analysis; use standard libraries (scikit learn, etc.)
        - Using neural networks (Shubham)
        - Using model algorithm based on research (John)
        - Using KNearestNeighbors (Harshit)
- May 21
    - Pick an algorithm for the model
        - Perform cross validation, hyperparameter tuning, etc (John, Shubham)

    - Create ensemble models and determine whether they are better (Harshit, Shubham)
    - Bare bones website which asks for hash and calls model (John)
- May 28
    - Testing and finishing touches (All)
    - Writing report (All)

Deliverables

Website which provides an interface to use the model and determine which hashing algorithm a specific hash came from (if model has reasonably low error). Report with findings and answers to the questions laid out in the aims.