Malware Analysis at Scale with ElasticSearch

a.k.a playing around with a whole lot of malware

Summer 2021

Team Members

Taylor McCampbell
- Junior, COSC Major
- INL Intern

Rafer Cooley
- COSC Ph.D. Student
- INL Intern

Background

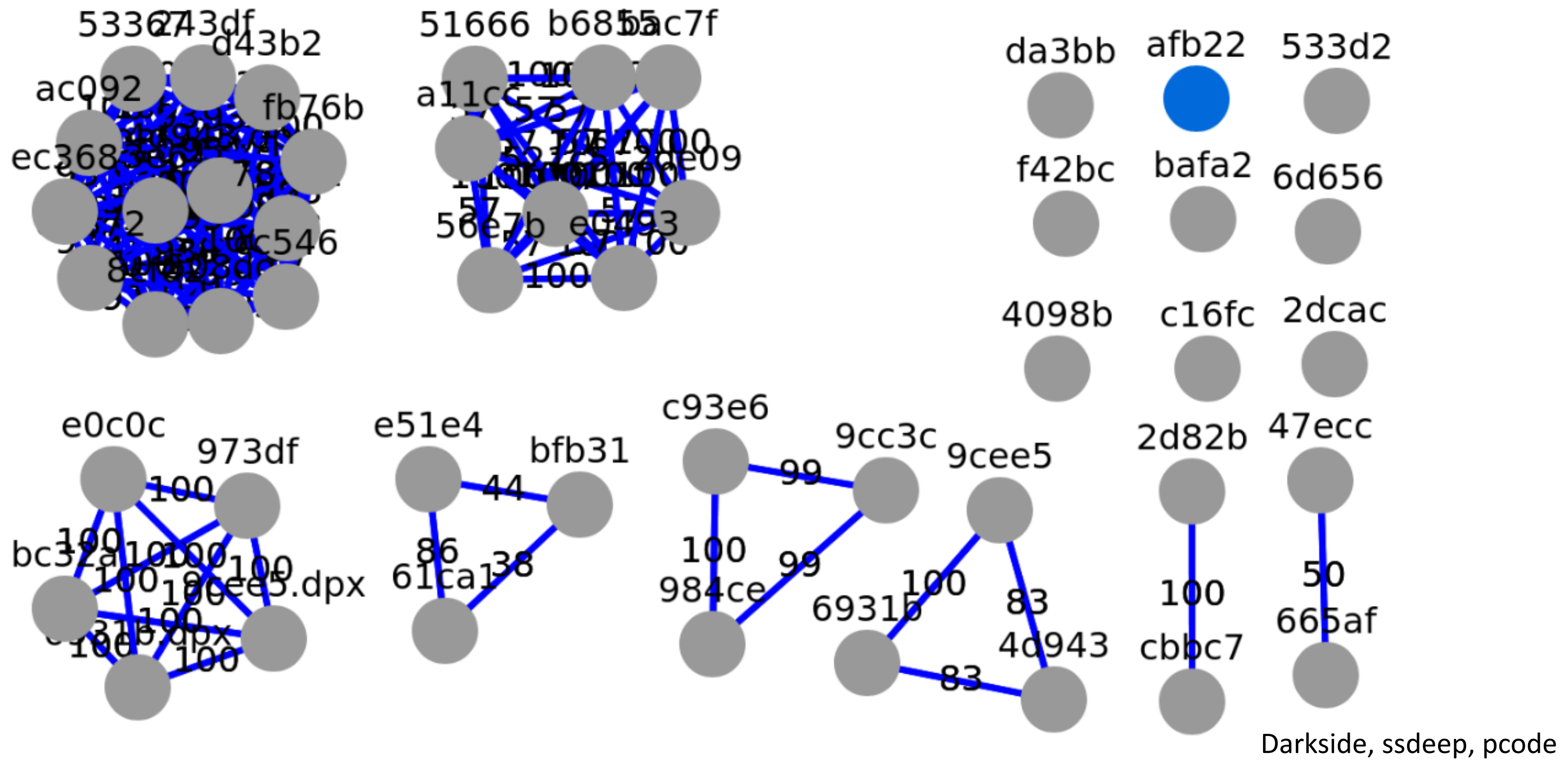
- Amount of malware increasing at exponential pace
- Modern indicators are easily bypassed by changing minor behaviors in malware programs
- Big data problems have generated Big Data solutions
- Search engines have been developed to efficiently search through large streams of information

Problem Statement

With the exponential increase of malware being released everyday, an efficient and resilient scheme for malware indicators is needed. In order to store, analyze, and develop future indicators an efficient data management system is needed that can handle the amount of information necessary to accurately test said indicator schemes.

Methods

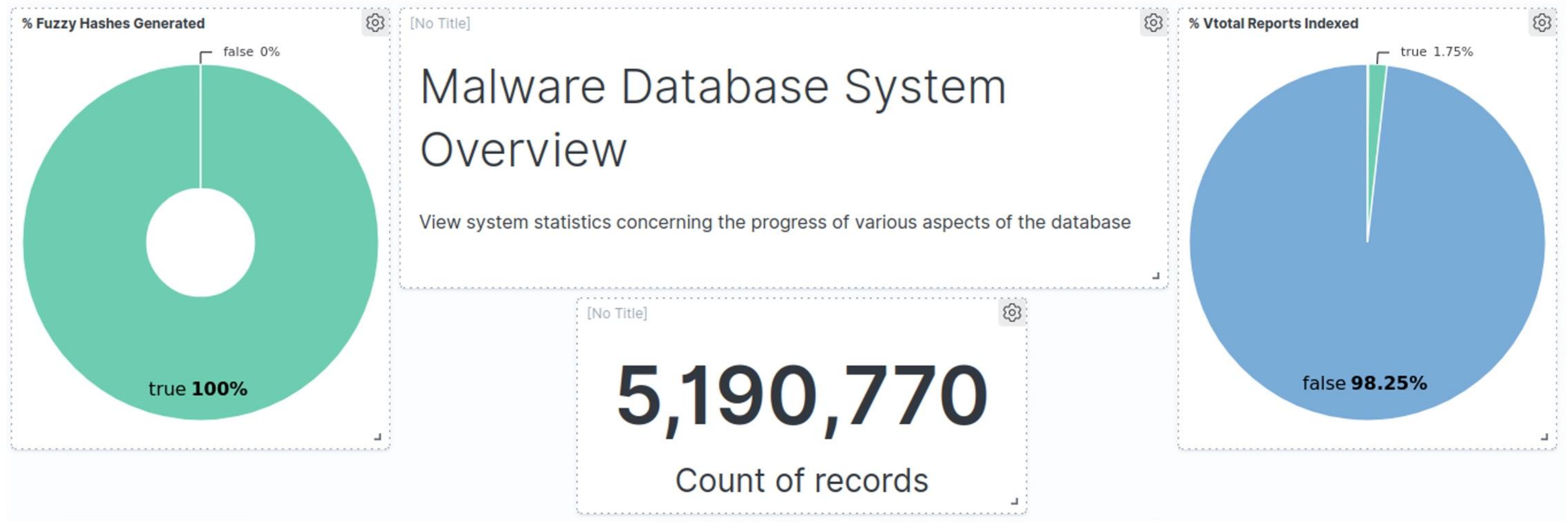
- 27 TB of malware gathered from public sources
- VirusTotal reports gathered for each corresponding sample and uploaded into ElasticSearch
- SSDEEP and TLSH hashes and comparisons generated from sample set and stored in Neo4J
- Python based CLI program developed to interact with every feature



Darkside, ssdeep, pcode

Results

- 5,190,770 Individual samples of Malware
- LSH used to detect similar malware with Neo4j
- Efficient and effective interaction with large scale data with ElasticSearch
- Automatic data collection and logging from open-source repositories
- Campaign visualization created to manage individual projects



% Fuzzy Hashes Generated

false 0%

true 100%

[No Title]

Malware Database System Overview

View system statistics concerning the progress of various aspects of the database

[No Title]

5,190,770

Count of records

% Vtotal Reports Indexed

true 1.75%

false 98.25%

Challenges & Future Work

- Develop Neo4j to develop visualizations of custom indicator comparisons (custom LSH)
- Mathematically verify the integrity of LSH's
- Automatic VirusTotal report transforms to other threat indicator types (i.e. STIX)
- Develop search by library, function, etc capability
- Create easily deployable program (i.e. Dockerize)

Advisor: Dr. Mike Borowczak(Uwyo), Rita Foster(INL), Jed Haile(INL)
Group Members:
- Taylor McCampbell (tmccampb@uwyo.edu)
- Rafer Cooley (rcooley2@uwyo.edu)
- Mike Borowczak (mike.borowczak@uwyo.edu)

Thanks to Department of Energy, EERE: Building Technology Office, Wind Energy Technology Office and Solar Energy Technology Office, nuclear Office – Cybersecurity Energy Systems & Emergency Response (CESER) funded the Grid Modernization Laboratory Consortium (GMLC) which funded the Deep Learning Malware (DLM) Project responsible for ongoing development of the Malware Database.