UNITED STATES MILITARY ACADEMY

STATEMENT OF THE PROBLEM

XE 401

SECTION H2I2

MR.KAPRALOS

By

CADET JAROD MOCKUS ’21, CO D3

CADET ALEX HELDSTAB ’21 CO H1

CADET HARRY HERNBERG ’21 CO B3

CADET AARON MULLALLY ’21 CO E1

CADET ORION HOLLIN ’21 CO H4

CADET KENDYL MACFARLAND ’21 CO G1

WEST POINT, NEW YORK

24 SEPTEMBER 2020

_____ OUR DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT

_X_ WE DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT

SIGNATURE: HOLLIN.ORION.CLAY.1537926585 Digitally signed by HOLLIN.ORION.CLAY.1537926585 Date: 2020.09.28 09:32:33 -04'00' MULLALLY.AARON.STEPHEN.1538068213 Digitally signed by MULLALLY.AARON.STEPHEN.1538068213 Date: 2020.09.28 09:34:45 -04'00' MCFARLAND.KENDYL.TAYLOR.1537937692 Digitally signed by MCFARLAND.KENDYL.TAYLOR.1537937692 Date: 2020.09.28 09:37:17 -04'00' MOCKUS.JAROD.ETIENNE.1272907834 Digitally signed by MOCKUS.JAROD.ETIENNE.1272907834 Date: 2020.09.28 09:39:54 -04'00' HERNBERG.HARRY.FRANCIS.1537954368 Digitally signed by HERNBERG.HARRY.FRANCIS.1537954368 Date: 2020.09.28 09:50:12 -04'00'

## *Statement of the Problem*

---

**Background:**

The motivation for this project's creation was Detective Napoli with the New York Police Department (NYPD) working on cryptocurrency and the dark web, came to USMA to give a talk on their process. During this presentation, EECS faculty and students realized that there could be a better way of accomplishing their research tasks. To help facilitate their ability to find key terms on the dark web we began building an intuitive search engine which can parse and categorize plaintext on the dark web sites which have been found through the clear web as the dark web is not able to be crawled in traditional methods. By working with the NYPD and seeing their methods, our team is able to work towards the functionality that they need and will speed up their investigative process such as site grouping by keyword and seeing the up status of different sites so they can tell when certain marketplaces go down or where they move to. Ultimately, our group wants to help the NYPD track dark web activity and identify avenues of approach to solve dark web crimes -- the NYPD's main problem regarding this sector.

The DREAD project will deliver an intuitive search engine, hosted on a web server that accesses a database tracking illicit dark web activity. Our main deliverable is the search engine, but many supplementary deliverables will support that main deliverable. We have assigned different team members to work on such supplementary deliverables. CDT Heldstab is working on creating a parser to create the deliverable handling common crawl. CDTs Hernberg and Mockus are focused on the database deliverable. CDTs Hollin and Mullally are working on the Virtual Machine to host the dark web searching. Lastly, CDT MacFarland is working on the web server and ensuring that main deliverable is a presentable, appealing product.

**The Design:**

- The DREAD project design spans 5 distinct systems over 3 network spaces, divided into 4 areas of responsibility (AOR).
- The accumulation, transfer, validation, manipulation, storage, and display of data required a meticulous design plan. This is particularly true when bringing large quantities of data across multiple networks and systems.
- AOR 1 operates in the Clearnet and EECSNet spaces. A clearnet crawler accumulates as much data as possible from the web, which is received in .wet form by the Aquamentus machine hosted in EECSNet.
- Aquamentus manages the considerable amount of raw data by moving it through a parser which filters and collects tor addresses appended with .onion, which are then compiled in .csv format.
- This parser needs to be able to finish a single run of compressed .wet files in less than 2 days.
- AOR 2 operates solely in the Darknet space. By way of VNC, the .csv containing addresses is received from Aquamentus and the addresses therein are checked for status. Aggregation of site status (up/down) and metadata is appended to the .csv before a handoff via VNC to the next system.
- AOR 3 operates once again in EECSNet and the Clearnet spaces; .csv data is transferred to and managed in a database within EECSNet, then integrated into the Web Application hosted on the Clearnet Web Server.
- AOR 4 overlaps AOR 3, focused on Web Application Design. SQLite 3 Queries from the Web Server retrieve data from the EECSNet hosted database as needed.
- The final component of system / network design is the Clearnet Client who would access the Website and make the queries. Web App design elements manage user permissions and facilitate accessibility functions for the Client.

· A visual representation of this network design is located in Appendix A after the description of the deliverables of the design project.

**Deliverables of the Design Project:**

The main deliverable our project will produce is a website, which is currently hosted at dargle.net, that will serve as a public interface to our database. As of right now, dargle.net is serving as the prototype web server. We are looking to transition this prototype into our own website “dread.wtf.” Behind this website we will have a robust database that will be able to handle search requests and categorize websites by keywords. We will also have a GitHub repository that contains our scripts that will be used on our Dark Web VM. These scripts will be able to take our database of .onion addresses and check their online status and remove them from the database if they are down. We will also have a script that compares our keyword lists to the text of the online Dark Web pages, and categorize our database based on these results. Our repository should also contain documentation on how to use our programs for continued use. This documentation will include our research findings with what sites hold the most dark web information, other helpful statistics from our data, and an explanation of how we solved some of our hard problems in our project. Lastly, we will provide code that can automate this entire process except the pulling of the commoncrawl, and maybe a means to automate that as well.

*APPENDIX A*

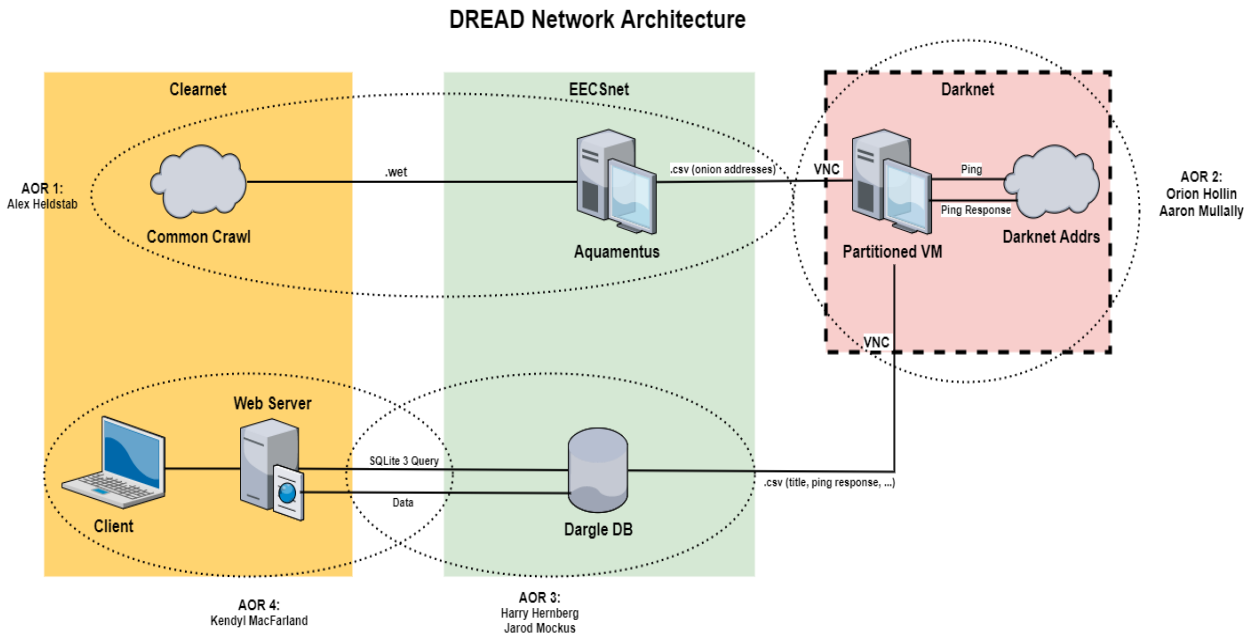


Fig. 1 Areas of Responsibility for the Design of the Project