*{date}*

CIS 4911 Senior Project

Intelligence Inference Engine

Group Members

Jose Acosta

Lazaro Herrera

Mentor

Eric Kobrin

Instructor

Masoud Sadjadi

*{Copyright}*

Abstract

Table of Contents

1. Introduction
   1. Problem definition

As the “Internet of Things” becomes larger and more connected, more and more of our devices become connected. Toasters, coffee makers, smartphones, items that were never connected are now as connected as your personal computer. As always, with greater connectivity comes greater risk at attack. Cyber-attacks on critical infrastructure can occur at any time and the new generation of tools to protect our domestic infrastructure cannot keep up with demands. What if a cyber-attack takes out power to the continental US, shuts down the traffic system of a city, disables the water processing plants of a state? Security researchers around the world are constantly monitoring the Internet for today and tomorrow’s threats but despite all of their knowledge, the information is not shared as easily as possible. What if you know that someone will be targeting critical infrastructure while a researcher in Cambridge finds the code that will make it happen? How do you make a cross-company, cross-country connection that could prevent a disaster?

* 1. Background

As more devices come online and more of our infrastructure joins the Internet, some of the infrastructure that powers our lives becomes more and more reliant on the Internet and any vulnerabilities in them can cause issues. Power, water, traffic, financial, educational and many others are just some of the infrastructure that is constantly under attack by our enemies both foreign and domestic. A well-placed attack could cripple one or many of these necessary of society. Although the online communities of security researchers have many tools at their disposal, there is not one tool that allows connections to be made when you don’t know exactly what is being looked for. As an example, what if a researcher sees an attack that occurs every third Wednesday of every month that attacks only financial industries? They may not even be aware that a different researcher has seen the same attack originating from a single computer that directs others to attack, or that the attacks come from a specific sub-net. All of this data when properly connected in a single location can allow more flexibility when attempting a cyber defense. Knowing where the attack is coming from, and at what time can allow our more critical infrastructure to be more resilient to the future of cyber-attacks.

* 1. Definitions, Acronyms, and Abbreviations

OSINT - Open-source intelligence

Cyber-attack - Any type of offensive maneuver employed by individuals or whole organizations that targets computer information systems, infrastructures, computer networks, and/or personal computer devices by various means of malicious acts usually originating from an anonymous source that either steals, alters, or destroys a specified target by hacking into a susceptible system {ref}

Triple store - {def} {ref}

* 1. Overview of document

This document introduces the problem that the system will be tyring to solve and gives background information on why this problem is something that should be tackled. This document also covers the basic definitions and abbreviations in field relating to the problem. Within this document, The current way that the problem is tackled is described and also there will also be a description of the new system that will be implemented in order to combat the problem. The high level user requirements that specify what the system must be able to do are stated. There is also some investigation in alternative solutions and implementations to this problem and those alternatives are are compared to the suggested solution and analyzed. Finally, this document provides what the necessary hardware and software is and describes the team structure. A project plan is also included which describes that has been accomplished and what work still needs to be done.

1. Feasibility Study
   1. Description of Current System (Limitations and Constraints)

Currently, there exist many websites that catalog cyber-attacks, data breaches and vulnerabilities. Although all of these websites are critical to today’s security researchers, missing any of them can leave important data behind as well as not knowing the correct term to search for can cause some of the important information to be left behind. Consider the possibility that an attack may be filed as targeting “banks” yet the security researchers might be searching for an attack that is targeting “financial institutions”. Similar terms being used but a standard search engine does not take these into account.

* 1. Purpose of New System

The purpose behind the new system that will be created is to allow users to be able to easily share information with the public while also being able to access said information about OSINT data as well as internal cyber-attack data using semantic web techniques. The front end of the system will be a web interface to facilitate the access and use of the system. From the front end, the user will be able to easily submit data about recent cyber-attacks or possible cyber-attacks into the system through an easy to us web form. The data that users provide will be stored in a triple store. Once the data is being stored, users will be able to query the data that has been collected to see any relevant data to that user. Also, the system will be able to make inferences about future data from current data that is being stored. The system should provide storing for intelligence-critical metadata such as assertion provenance and confidence of assertions. The system should also be able to collect information that is not directly provided by users through the use of web crawlers.

* 1. High-level Definition of User Requirements (must include security/privacy requirements)

- A user must be able to submit data to be stored through some type of web form

- The data must be stored in one of the existing semantic web triple- or quad-stores such as Mulgara or Jena

- The data should be query-able directly using Sparql or Datalog

- A user must be able to set up predefined queries which are accessible by other users

- The system must be able to gather data by itself through the use of web-crawlers

- The system must be able to set a confidence interval for the data that is being collected

* 1. Alternative Solutions
     1. Description of Alternatives

An alternative to the proposed solution would be to change the entry and query of data from a web form to an independent client that a user would install. The system could also use a relational database rather than a triple store to store all the data the system will receive.

* + 1. Selection Criteria (Briefly describe the feasibility criteria used in the analysis component)

1) Ease of use

2) Ease of development

3) Extensibility (support newer technology)

4) Reliability

5) Continued Support

* + 1. Analysis of Alternatives (refer to Appendix C – Feasibility Matrix) – you should provide a score so that the alternatives can be compared.

For our project we will be using a boostrap-based web form built on Mulgara. A web form is being used for easy user access from both desktops and mobile devices instead of providing a native client. Mulgara’s triple-store along it’s REST interface allows us to perform queries directly from the web form in multiple languages (Sparql/Tql).

Mulgara was chosen over sqlLite for the inference engine mainly because of its ease of development. Mulgara is significantly more efficient for what this system needs to do as opposed to using sqlLite. A bootstrap web form was chosen because it is much easier to develope a single web client that anyone with a web browser can access as opposed to native clients for each operating system. Also with a web client there is no installation so its a little easier to use than a web client.

* 1. Recommendations

1. Project Plan
   1. Project Organization
      1. Project Personnel Organization

Lazaro Herrera - Developer / Test Engineer

Jose Acosta - Developer / Web Design

Eric Kobrin - Project Manager / Client

* + 1. Hardware and Software Resources

Hardware - Dual core CPU / 4 GB Ram / 50 GB hard drive

Software - Mulgara / Easy Dev PHP / Bootstrap / SQL lite / Selenium / Web Crawler Framework

* 1. Identification of Tasks, Milestones and Deliverables (work breakdown)

For our Trello board, we have decided to use color tagging to both identify the assignments of tasks and the completion status of tasks

Colors and Definitions for Priority

Blue - Card created by students

Purple - Card created by client

Colors and Definitions for Completions (only one active at a time)

Green - This task has a completion of 25% or more

Yellow - This task has a completion of 50% or more

Orange - This task has a completion of 75% or more

Red - This task has a completion of 100%

The following epics currently exist on our board.

Some of these are receiving checklists to be converted into stories.

- Setup Development Environment (Mulgara)

- Setup Development Environment (Web Server + Bootstrap)

- Develop Primary Feature (Data Entry Web Form)

- Test Primary Feature (Data Entry Web Form)

- Develop Primary Feature (Data Retrieval Web Form)

- Test Primary Feature (Data Retrieval Web Form)

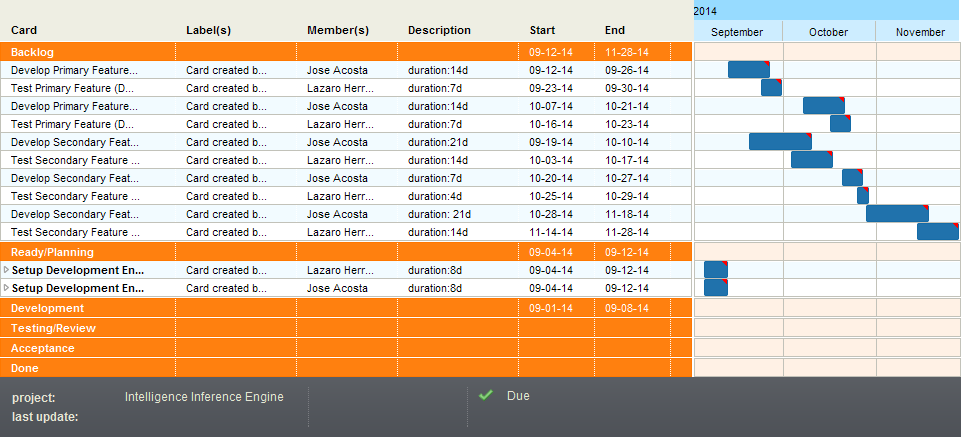
- Develop Secondary Feature (RDF Web Crawler)

- Test Secondary Feature (RDF Web Crawler)

- Develop Secondary Feature (Custom Reusable Queries)

- Test Secondary Feature (Custom Reusable Queries)

- Develop Secondary Feature (Confidence Ranking)

1. Appendix
   1. Appendix A - Project schedule 
   2. Appendix B – Feasibility Matrix

|  |  |  |
| --- | --- | --- |
|  | Mulgara | SqlLite |
| Ease of Use | n/a | n/a |
| Ease of development | 10 | 1 |
| Extensibility | 6 | 8 |
| Reliability | 9 | 9 |
| Continued Support | 3 | 9 |
|  | Bootstrap web form | Native Client |
| Ease of Use | 9 | 8 |
| Ease of development | 10 | 5 |
| Extensibility | 5 | 5 |
| Reliability | 7 | 7 |
| Continued Support | 9 | 9 |
|  | Easy Dev PHP | Custom Webserver |
| Ease of Use | n/a | n/a |
| Ease of development | 10 | 1 |
| Extensibility | 4 | 9 |
| Reliability | 9 | 9 |
| Continued Support | 9 | 10 |
|  | Selenium | Manual Testing |
| Ease of Use | n/a | n/a |
| Ease of development | 8 | 1 |
| Extensibility | 5 | 9 |
| Reliability | 8 | 4 |
| Continued Support | 8 | 9 |

* 1. Appendix C – Cost Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cost Matrix | Weeks Required | Jose - $30 hour / 12hr week | Lazaro - $30 hour / 12hr week | Total Cost |
| Data Entry Form Implementation | 2 | $720.00 | $720.00 | $12,672.00 |
| Data Entry Form Testing | 1 | $360.00 | $360.00 |  |
| Data Retrieval Form Implementation | 2 | $720.00 | $720.00 |  |
| Data Retrieval Form Testing | 1 | $360.00 | $360.00 |  |
| RDF Crawler Implementation | 3 | $1,080.00 | $1,080.00 |  |
| RDF Crawler Testing | 2 | $720.00 | $720.00 |  |
| Custom Queries Implementation | 1 | $360.00 | $360.00 |  |
| Custom Queries Testing | 0.6 | $216.00 | $216.00 |  |
| Confidence Ranking Implementation | 3 | $1,080.00 | $1,080.00 |  |
| Confidence Ranking Testing | 2 | $720.00 | $720.00 |  |

* 1. Appendix D - Diary of Meetings

{Copy from moodle}

{NOTE: We have a forum post that is used to keep track of meetings, when were are nearing the final version of this document we will fill this section from moodle}

1. References (you should reference any work that is not your own)