Date: 13 September 2018
To: Romcholo Macatula
From: Cyber Warriors
Subject: Technical Memo 2: Project Components and Criteria

1 Summary

This memo shall enumerate and describe the components of the project: Mathematical Modeling, Data Collection, and Data Analytics. A diagram with each of the components shall describe their dependencies on one another. A description of the criteria for each of the components shall follow the dependencies diagram. In addition, this memo addresses the prioritization of the component criteria through decision matrices and explanations of those tables.

2 Project Components and Ideal Criteria

This section shall address the components of the project. In addition, this section describes the criteria each component shall meet and which criteria take priority.

2.1 Project Components

The project shall consist of the three components below:

(1) Mathematical Modeling;

(2) Data Collection;

(3) Data Analytics.

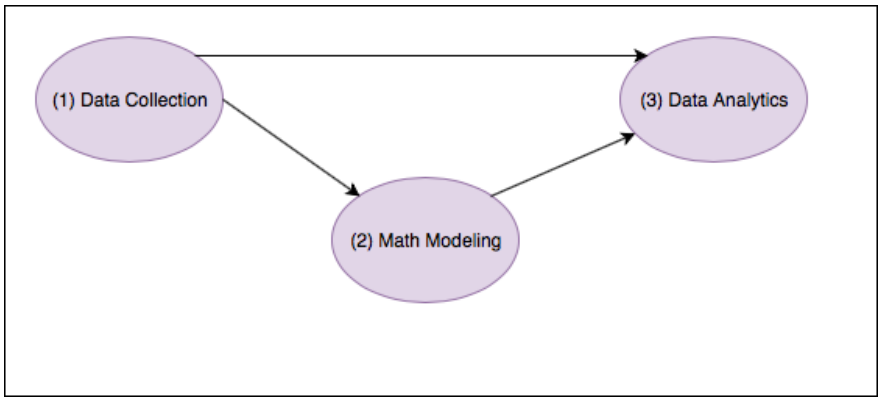

Figure 1: Component Diagram

The diagram above illustrates the three main components of the project and their dependencies on one another. The team shall collect data first. The factors of the mathematical model shall depend on the usable data that the team finds. The data analytics portion of the project shall depend on both the data collected and the and the mathematical model.

2.2 Component Criteria

For each component of the project, we establish a list of criteria that characterize an ideal solution. See below the list of criteria for each component:

Criteria for (1) Data collection.

- *Number of data sources.* Collection of 5-10 data sources, the range specified by the client.
- *Data format.* Data collected through webscraping and other means that exists in a usable file format (.csv, .txt, etc.).
- *Access.* Data collected through public sources that require zero additional cost.
- *Accuracy.* Data collected from reputable sources with 0 data falsification accusations.
- *Integration.* Data collected that varies no more than 50% from all other sources.
- *Ease.* Data collection that involves data that requires no more than 4 weeks to find.

Criteria for (2) Mathematical Modeling

- *Language.* Python or R, as requested by the client.
- *Speed.* The model shall take no longer than 10 minutes to run on a standard personal computer.
- *Accuracy.* The model should produce no more than 10 percent false positives.
- *Factors.* The team will use 5-10 factors specified by the client.
- *Ease.* Simple usability and implementation given the current knowledge of the team and the single semester time constraint.

Criteria for (3) Data Analytics.

- *Language.* Language specified by the client. Develop in R and translate to Python if requested.
- *Visualization.* Visualization of data through spider charts that illustrate the importance of each factor.
- *Scope.* Analysis of data related to privileged account escalation.
- *Accuracy.* Over estimation of risk (false positives no more than 10% of the time) through statistical inferences.
- *Interpretation.* Analysis of data through regression models to understand the relationships between coefficients and their importance.

2.3 Criteria Prioritization

To prioritize our efforts in the rest of the project, we compare the relative merits of each criteria. Engineers do this prioritization by a pairwise comparison chart. We created a table with the major criteria listed down each column and across the top. To prioritize, we add a "score" column, and add across the rows. These scores then give an ordered list of our priorities for this component and list the criteria from the highest score to the lowest [1].

(1) Data collection

	# of Data Sources	Data Format	Access	Accuracy	Integration	Ease	*score*
# of Data Sources	–	1	1	1	1	1	5
Data Format	0	–	0	0	0	0	0
Access	0	1	–	1	0	1	3
Accuracy	0	1	0	–	0	1	2
Integration	0	1	1	1	–	1	4
Ease	0	1	0	0	0	–	1

1. Number of Data Sources
2. Integration
3. Access
4. Accuracy
5. Ease
6. Data Format

Number of data sources claims the top priority for the Data Collection component of the project because the client requested 5-10 data sources in order that the team assesses the proper amount of data required to perform analytics. Next, the team focuses on integration of data sources, data sources must parallel one another so that logical integration occurs. No budget exists for this project so public source data exists as the option for data collection. The team shall avoid any falsified data, however, the team acknowledges that inaccurate data exists and shall address any occurrence of inaccurate data as model creation occurs. A semester time constraint prompts the team to attempt to find data sources in a short time span. While preferable, a usable data format only provides convenience. The team shall translate any data in an unusable file format into a usable file format.

(2) Math Modeling

	Language	Speed	Accuracy	Factors	Ease of Use	*score*
Language	–	1	0	0	0	1
Speed	0	–	0	0	0	0
Accuracy	1	1	–	1	1	4
Factors	1	1	0	–	0	2
Ease of Use	1	1	0	1	–	3

1. Accuracy
2. Factors
3. Language
4. Speed

The creation of an accurate mathematical model to locate weaknesses in the clients network is the teams first priority. The model will be based on relevant factors given to the team by the client or found through independent research. The team shall construct the model in either R or Python so that it takes no longer than 10 minutes to run on a standard personal laptop.

(3) Data Analytics

	Language	Visualization	Scope	Accuracy	Interpretation	*score*
Language	–	0	0	0	0	0
Visualization	1	–	0	0	0	1
Scope	1	1	–	0	0	2
Accuracy	1	1	1	–	1	4
Interpretation	1	1	1	0	–	3

1. Accuracy
2. Interpretation
3. Visualization
4. Scope
5. Language

The team plans to prioritize accuracy to ensure results are represented in an appropriate manner to the user and client. The team shall focus on ways to interpret the data through relationships between factors instead of predictive measures. The team shall create a spider web visualization to aid in the ability of the user to assess the risk of their network. The team shall ensure the scope of attributes and features used represents the compromised network in an accurate way. The team shall proceed with development in R with a future goal to convert the code to another language such as python.

References

[1] M. Embree, *Assignment for technical memo 2: Project components and criteria*, tech. rep., Virginia Tech, 2018.