Cyberlytic interview:
Data science technical questions

Introduction

The purpose of this activity is to test your initiative and technical skills with the analysis of web-based data. This data is similar to what you will be confronted with as a *Data Scientist* at *Cyberlytic*.

Specifically, you have been supplied with a *.tsv* (tab-separated value) file containing a representative HTTP traffic log from a mock website that has been run on a local server. The data was collected over a 3 hour period using ModSecurity, and was initially stored in a Mongo database. The data includes information regarding the request made by each visitor to the website and the servers response to each visitors request. Indeed, each observation contains information on both a request from an individual IP address and a response.

The data supplied must be used to answer the following questions. You can use whatever tool or programming language that you feel most comfortable with. You are required to provide your code and workings out alongside your answer to the questions below. Please generalise your code and process as much as possible.

Technical questions

1. Present summary statistics regarding the traffic to the web service.

   (**Hint:** For example, you may want to note all visitors to the web service and discuss their characteristics, such as: the frequency of interaction, the fields that they are accessing, the browser that they are using, the distribution of traffic from each user, etc.)

2. Are there any malicious activities present in the dataset? If so, elaborate on the malicious activities.

   (**Hint:** You may want to elaborate on the type of malicious activity you think is contained in the dataset, when it began to emerge, the type of malicious activity that you detected, the user(s) that engaged in the malicious activity, etc.)

3. An important aspect of anomaly detection is representing traffic data in the form of a weighted directed network or graph. This shows the typical flow of traffic. A graph can be created for each visitor to the web service and combined to analyse typical traffic patterns. Indeed, the weights on the graph are inferred from the aggregation of the visitors traffic flows. This is a technique known as '*graph mining*.'

Represent the traffic data in terms of a weighted directed network and discuss the properties of typical traffic flow. Can you detect any anomalous traffic from your analysis? You may use plots to visualise the graph(s) that you have developed.

(**Hint:** You can represent the traffic as a graph in whatever form you find the most effective. A common method is to represent individual webpages as nodes and the flow of a visitor from one webpage to another as a directed edge from one node to another. This can be represented as an adjacency matrix or sparse matrix with weights attached.)

4. Using your own independent research, provide a discussion of other techniques for anomaly detection within the cyber security context and apply these techniques to the synthetic data provided.

You may present your answers in any way you feel is most effective, For example, you can choose to present your results in the form of a Word or PDF document or Powerpoint presentation. Please send your answers and code by email to owen.sims@cyberlytic.com.

Owen Sims
Data Scientist @ Cyberlytic