The Advantages to Using the NSFCloud for Data Mining in the Cybersecurity Domain

Dallas H. Snider, Ph.D.
University of West Florida

1. Research Statement

We would like to answer the following research question: What are the minimum features needed to construct a Bayesian Belief Network that will help an organization to comprehend threats to their computer systems? One of my research projects at the University of West Florida (UWF) involves exploring ways to assist organizations to better understand their systems and how to communicate threats to these systems across all levels of an organization. My collaborators and I plan to build Bayesian Belief Networks from large quantities of computer network traffic data. Once these Bayesian networks are built, we plan to construct concept maps of an organization's system and augment these maps with values from the Bayesian Network's conditional probability tables to help provide clear, actionable statements to improve the communication of cybersecurity threats across all levels of the organization. Preliminary results from our research were presented at the recent 2014 CEWIT Conference [1].

2. Experimental Procedures

The experiments we plan to run involve data from large amounts of anonymized network logs and other data provided by organizations that cannot be disclosed at this time due to current negotiations with memoranda of understanding and nondisclosure agreements. The experimental process will follow the knowledge discovery in databases (KDD) process described by [2] and [3]. There will be processes created to clean and integrate the disparate data sources, to perform any aggregations or transformations, and to select the proper features as input for the Bayesian network algorithm. Our plan also calls for using 10-fold cross validation to ensure there is no overfitting of the network. To validate the accuracy of the network, we will use naïve Bayes and decision tree algorithms to ensure that our Bayesian network is sound. As part of the KDD process, we will run the Bayesian network algorithm repeatedly in a hill-climbing manner with varying features in the input vector to find the most accurate network with the least number of features. Future experiments could also include clustering the data to examine outliers which could indicate a cyber-attack or suspicious behaviors.

3. Infrastructure Needs

Data analytics in the cybersecurity arena requires a substantial amount of data processing and storage, which a cloud environment can provide. In the NSFCloud, we will need the ability to install Hadoop and supporting software to manage both administrative and data nodes. We will also need the capability to install and administer NoSQL databases such as Cassandra or MongoDB as needed. Finally, there is a need to install machine learning software such as Apache Mahout or Cloudera's Oryx. With all of these installation and management capabilities, there is also a requirement to apply patches and updates.

4. Position

There are cloud service providers that are quite capable of handling the storage and processing requirements for our experiments. However, cloud service subscribers can be limited by the provided software and administrative tasks. UWF could also acquire the necessary hardware, but the costs of acquisition, maintenance and infrastructure would be overwhelming to our research budget since UWF is a primarily undergraduate institution and we do not have as great an influx of research monies as Ph.D. granting institutions. It is the author's position that the NSFCloud will provide us the required data processing and storage capacities, lower costs and administrative privileges needed for success in our research.

5. References

[1] S. Pramanik and D. Snider, "Defining Threats across Organizational Boundaries," in *The 11th International Conference and Expo on Emerging Technologies for a Smarter World (CEWIT 2014)*, Melville, NY, 2014.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, Cambridge: MIT Press, 1996.

[3] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques, 3rd ed.*, Waltham, MA: Morgan Kaufmann, 2011.