University of Nevada, Reno

Network Security Monitoring and Analysis based on Big Data Technologies

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in
Computer Science and Engineering

by

Bingdong Li

Dr. Mehmet Hadi Gunes / Dissertation Co-Advisor
Dr. George Bebis / Dissertation Co-Advisor

December, 2013

Abstract

Network flow data provide valuable information to understand the network state and to be aware of the network security threats. However, processing the large amount of data collected from the network and providing real time information remain as big challenges. *Big data* technologies provide new approaches to collect, store, measure, and analyze the large amount of data. This dissertation aims to provide a system of network security monitoring and analysis based on the *big data* technologies.

First, I present an extensive survey of the network flow applications that covers past research perspectives, methodologies, and a discussion of challenges and future works. Then, I present system design of the network security monitoring and analysis platform based on the *Big Data* technologies. Components of this system include *Flume* and *Kafka* for real time distributed data collection; *Storm* for real time streaming distributed data processing; *Cassandra* for NoSQL data storage, data processing, and user interfaces. The system supports real time continuous network monitoring, interactive visualization, network measurement, and modeling to classify host roles based on host behaviors and to identify a particular user among the other users.

It is critical to continuously monitor the network status and network security threats in real time, but it is a challenge to process the large amount of data in real time. I demonstrate how the *big data* security system designed in this dissertation supports such features. Another usage of the network flow data is to measure the contents of the network. I demonstrate how this *big data* system provides understanding of the usage of anonymity technologies on the campus Internet. Then, I present methods and the results of classification and identification of network objects based on the *big data* system designed in this dissertation. Finally, I use *Decision Tree* and *Support Vector Machine* to model host role behaviors and user behaviors. Sample results indicate very high accuracy of host role classification and user identification.