

# Machine Learning Project

April, 2006

## References

- [1] James P. Early, Carla E. Brodley, and Catherine Rosenberg. Behavioral authentication of server flows. In *Proceedings of 19th Annual Computer Security Applications Conference*, 2003.

Understanding the nature of the information flowing into and out of a system or network is fundamental to determining if there is adherence to a usage policy. Traditional methods of determining traffic type rely on the port label carried in the packet header. This method can fail, however, in the presence of proxy servers that re-map port numbers or host services that have been compromised to act as backdoors or covert channels. We present an approach to classifying server traffic based on decision trees learned during a training phase. The trees are constructed from traffic described using a set of features we designed to capture stream behavior. Because our classification of the traffic type is independent of port label, it provides a more accurate classification in the presence of malicious activity. An empirical evaluation illustrates that models of both aggregate protocol behavior and host-specific protocol behavior obtain classification accuracies ranging from 82-100%.

- [2] Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining anomalies using traffic feature distributions. In *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, 2005.

The increasing practicality of large-scale flow capture makes it possible to conceive of traffic analysis methods that detect and identify a large and diverse set of anomalies. However the challenge of effectively analyzing this massive data source for anomaly diagnosis is as yet unmet. We argue that the distributions of packet features (IP addresses and ports) observed in flow traces reveals both the presence and the structure of

a wide range of anomalies. Using entropy as a summarization tool, we show that the analysis of feature distributions leads to significant advances on two fronts: (1) it enables highly sensitive detection of a wide range of anomalies, augmenting detections by volume-based methods, and (2) it enables automatic classification of anomalies via unsupervised learning. We show that using feature distributions, anomalies naturally fall into distinct and meaningful clusters. These clusters can be used to automatically classify anomalies and to uncover new anomaly types. We validate our claims on data from two backbone networks (Abilene and Geant) and conclude that feature distributions show promise as a key element of a fairly general network anomaly diagnosis framework.

- [3] Terran D. Lane. *Machine Learning Techniques for the computer security domain of anomaly detection*. PhD thesis, Department of Electrical and Computer Engineering, Purdue University, 2000.

Abstract: In this dissertation, we examine the machine learning issues raised by the domain of anomaly detection for computer security. The anomaly detection task is to recognize the presence of an unusual and potentially hazardous state within the activities of a computer user, system, or network. “Unusual” is defined with respect to some model of “normal” behavior which may be either hard-coded or learned from observation. We focus here on learning models of normalcy at the user behavioral level.

- [4] Marcus A. Maloof. *Machine Learning and Data Mining Computer Security*. Springer, 2005.

Machine Learning and Data Mining for Computer Security provides an overview of the current state of research in machine learning and data mining as it applies to problems in computer security. This book has a strong focus on information processing and combines and extends results from computer security. The first part of the book surveys the data sources, the learning and mining methods, evaluation methodologies, and past work relevant for computer security. The second part

of the book consists of articles written by the top researchers working in this area. These articles deal with topics of host-based intrusion detection through the analysis of audit trails, of command sequences and of system calls as well as network intrusion detection through the analysis of TCP packets and the detection of malicious executables. This book fills the great need for a book that collects and frames work on developing and applying methods from machine learning and data mining to problems in computer security.

- [5] David J. Marchette. *Computer Intrusion Detection and Network Monitoring*. Springer, 2001.

This book covers the basic statistical and analytical techniques of computer intrusion detection. It is aimed at both statisticians looking to become involved in the data analysis aspects of computer security and computer scientists looking to expand their toolbox of techniques for detecting intruders. The book is self-contained, assuming no expertise in either computer security or statistics. It begins with a description of the basics of TCP/IP, followed by chapters dealing with network traffic analysis, network monitoring for intrusion detection, host based intrusion detection, and computer viruses and other malicious code. Each section develops the necessary tools as needed. There is an extensive discussion of visualization as it relates to network data and intrusion detection. The book also contains a large bibliography covering the statistical, machine learning, and pattern recognition literature related to network monitoring and intrusion detection.