SANDIA REPORT

SAND2011- 78
Unlimited Release
Printed May 2011

Analytics for Cyber Network Defense

Todd D. Plantenga, Tamara G. Kolda

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.aspx

SAND2011-3786
Unlimited Release
Printed May 2011

Analytics for Cyber Network Defense

Todd D. Plantenga
Systems Analytics Department (8958)
Sandia National Laboratories
Livermore, CA 94551
tplante@sandia.gov

Tamara G. Kolda
Informatics and Systems Assessments Department (8966)
Sandia National Laboratories
Livermore, CA 94551

Abstract

This report provides a brief survey of analytics tools considered relevant to cyber network defense (CND). Ideas and tools come from fields such as statistics, data mining, and knowledge discovery. Some analytics are considered standard mathematical or statistical techniques, while others reflect current research directions. In all cases the report attempts to explain the relevance to CND with brief examples.

This page intentionally left blank.

Contents

This page intentionally left blank.

1 Overview

This report provides a brief survey of analytics tools considered relevant to cyber network defense (CND). Ideas and tools come from fields such as statistics, data mining, and knowledge discovery. Some analytics are considered standard mathematical or statistical techniques, while others reflect current research directions. In all cases the report attempts to explain the relevance to CND with brief examples. This document is only an overview; Sandia can provide greater depth on many topics.

CND data for this application is assumed to be passively collected network data distilled into a summary format. For example, netflow records might consist of tuples containing <timestamp, source, destination, protocol ID, packet size>. More sophisticated processing for level 4 protocol-based sessions might yield records that summarize an HTTP request and reply, or metadata for an SMTP send. This report does not consider focused analysis of full packet capture data; for instance, tools that identify suspicious Javascript. However, the output of focused analytics can certainly be used as part of the summary data.

In general, analytics for CND data strives to characterize normal and abnormal behavior of users and networks. The mathematical algorithms behind an analytic can be simple (counting the top 10 visited sites) or complex (tensor decomposition to detect user behavior patterns).

2 Scalability Challenge

A special challenge is to handle large amounts of data (terabytes to petabytes) on a routine basis. This is a primary reason the CND data is assumed to be in summary form. Some analytics research is aimed at identifying what data to keep; for instance, anomaly detectors might classify a large percentage of incoming data as uninteresting so it can be discarded.

Most analytics discussed in this report were designed for small problem sizes. Additional thought and effort is required to scale up to CND data sizes. In general, there are two approaches for achieving scalability:

- Streaming analytics examines data on the fly without the benefit of large-scale storage. Streaming algorithms are sometimes defined as being able to see the data only once, though in many cases this condition can be relaxed. For example, counting the frequency of all distinct IP addresses becomes nontrivial when there are millions of addresses. At Sandia, one approach is to develop streaming algorithms as QK modules; for instance, using hash tables or Bloom filters for the counting problem.
- Archived analytics operate on large-scale stores of data, typically a time history of CND records. Commercial companies (Google, Yahoo, Facebook, Twitter, Netflix, etc.) have pushed the development of "Big Data" frameworks, utilizing commodity cloud clusters and open source software. At Sandia, one approach is to develop algorithms in MapReduce to analyze CND data stored in distributed databases such as the Hadoop Distributed File System (HDFS).

3 Types of Analytics

The list below attempts to describe categories of algorithms appropriate for CND analytics. Most can be implemented as streaming or archived analytics, though design principles may be quite different for the two cases. Some specific CND applications are provided as examples.

3.1 Counting and histograms

These methods are the simplest mathematically, but sophistication is required to operate efficiently with huge amounts of data. CND examples: find the Top K web sites / source IPs / etc. over a period of time; find the distribution of activity against a client by port; find the most frequent source-destination pairs; alarm any of these against threshold values.

3.2 Regression models

Regression methods are used for extrapolation and prediction. CND example: predict traffic based on season, week day, and hour.

3.3 Clustering methods

Typical algorithms are k-means clustering or hierarchical methods. Clustering is often based on attributes which are generated by other analytics. CND example: cluster user types based on HTML habits.

3.4 Supervised learning for classification

Analytic methods include Bayesian classifiers, decision trees, support vector machines, association rules, Markov models, etc. Supervised methods require a labeled set of training data which is often difficult to obtain; for example, manually classifying a list of web sites as compromised or safe. CND examples: classify web sites as search/blog/games/adult/etc based on URL information; detect a botnet from traffic patterns.

3.5 Text mining

CND example: spam email detection.

3.6 Unsupervised learning and pattern detection

These methods include principle component analysis, tensor decomposition, and latent Dirichlet allocation. Methods look for patterns that explain typical behavior, and then identify outliers.

CND examples: identify unusual user visit patterns to specific web sites, find anomalous packet size changes between specific IPs over time.

3.7 Graph algorithms

Data can often be represented as a graph of vertices connected by edges; for instance, HTML links between web pages, or individuals linked by email. Graph algorithms can find social communities (common components, eigenvector analysis), vertices of influence (page rank), hubs and chokepoints (commute distance, centrality measures), and patterns of interest (subgraph isomorphism). CND examples: detect botnet communities, measure the influence of a known malware site.

4 References

This informal report does not include references. Certainly a more complete report could include references from textbooks and the technical literature.


Sandia National Laboratories