Cyber Security

Big Data Big Think II Working Group Meeting

April 21, 2015

Thomas Hinke and Derek Shaw, CISSP

Thomas H. Hinke, CISSP, Ph.D.
NASA Ames Research Center, Moffett Field, CA
Thomas.H.Hinke@nasa.gov
(650) 604-3662

Agenda

- The Big Data Problem in cyber security
- The cyber security environment for a NASA supercomputer center and its associated big data (referred to as a Data Computation Center in this presentation)
- General cyber security objectives
- Focus on identification of actionable security events from the big mountain of data that flows into and out of a Data Computation Center

The Big Data Problem in Cyber Security

- Internet-accessible Data Computation Centers are under constant attack, which come from the Internet
- Attacks are intermixed with a large volume of legitimate Internet traffic flowing into and out of the Data Computation Center, which leads to a big mountain of data
- Cyber security tools are used to monitor for attacks, but these generate a big mountain of data that is added to the mountain of legitimate data traffic
- To keep from being overwhelmed, security analysts are interested in knowing about only those attacks for which manual or automated action is required to counter them
- The problem addressed by this presentation is how to extract actionable security data out of this big data mountain

The NASA Big Data Supercomputer Cyber Security Environment

- The Data Computation Center is accessible over the Internet
- Legitimate users must be identified and authenticated
- Users can download their own programs and any additional data that is needed
- Security should not be so onerous that it impedes the use of the system
- Actionable security events that could lead to a misuse of the system should be identified and acted upon
- Need to discover actionable events from the big mountain of data in which they are embedded

General Cyber Security Objectives

- Identify and authenticate legitimate users prior to granting them access to the Data Computation Center
- Ensure that users access only data to which they are authorized
- Identify attacks and block them while minimizing
  - False positives – so that legitimate users are not blocked
  - False negatives – so that dangerous attacks are not missed
- Ensure that malicious code and backdoors
  - Are not planted in system or
  - Are detected if they are

Two Broad Strategies for Detecting Attacks

- Direct detection approaches applied directly to the Data Computation Center data flowing to and from the network
- Detection approaches applied only after reducing the big mountain of Data Computation Center data flowing to and from the network to a more manageable level

Direct Detection Can Be Used To Detect Malicious Delivery Sites

- Phishing and malware delivery sites can hide behind constantly changing addresses
- Detection approaches are compute intensive for the big mountain of DNS data flowing to and from the network
- Changing IP addresses over a short time period is called Fast-Flux
  - Detection based on counting number of IP addresses associated with domain and time-to-live assessment
  - Important to add other parameters when making a decision as to whether this is malware, since content delivery networks use this for load balancing
- Randomly generated domain names at regular intervals in large numbers is called Domain Flux
  - Detection based on number of domain name failures returned by the host making DNS request

Direct Detection Can Be Used To Detect Malicious Delivery Sites

- Domain Generation Algorithms (DGAs) can automatically generate domain names
  - Generated names tend to be gibberish
  - Generation of names is deterministic that so that Command and Control site and embedded malware can communicate after change
- Use Machine Learning to Detect Domain Generation Algorithm Attacks

Use Machine Learning to Detect Domain Generation Algorithm Attacks

- Use a combined 2-gram to 3-gram Markov chain to detect deviation from “normal” domain names
- While this may not detect newer DGA-generated domains name such as those that pair real words in a noun-verb-noun-verb combination, it is still very useful
- Combining with other techniques will reduce false-positives and false-negatives
- Combining a 2-gram Markov chain (which is less strict on detecting gibberish) with an organization check and fast-flux domain flux detection may help detect the noun-verb-noun-verb DGA domains.

Detection After Reducing the Big Data Mountain to a More Manageable Volume

- Network flow data is the largest component of the big data mountain
- Want to categorize network flows into three risk categories:
  - Acceptable Risk: Flows that are unlikely to contain any security attacks – and thus can be discarded from further analysis
  - High Risk: Flows that are likely to contain security attacks – and thus must be extensively analyzed
  - Unknown Risk: Flows for which it is unknown whether or not they contain a security attack – which also must be analyzed

Categorization of Flows Based on the Source of the Flow

- Acceptable Risk flows:
  - Flows that come from IP addresses that fall under an organization from which an authorized user has logged on to a Data Computation Center system
- High Risk risk flows:
  - Flows that come from IP addresses and associated organizations that are on watch lists, have mounted an attack, or have scanned dark space
- Unknown Risk flows:
  - Flows that cannot initially be placed into the acceptable or high risk categories

Detection of Malicious Command & Control Channels Using High Risk & Unknown Risk Flows

- Detect Command and Control activity for any malicious code that may have been installed in a system
  - Network beaconing activity could indicate presence of
    - Botnet slave checking in for assignments
    - Advanced Persistent Threat (APT) checking in for assignments or delivering intellectual property
  - Detection by looking for ~~a~~ recurring flows that provides a pattern of communications
  - Need big data trending with flows remaining from data mountain reduction

Detect Attempted or Actual Intrusions by Looking at High Risk and Unknown Risk Reduced Flows

- Normally a large number of Intrusion Detection System (IDS) events for a large data center with heterogeneous systems
  - Big data mountain reduction eliminates all but High and Unknown risk flows, helping to focus on events of most concern
- Still need significant computation to de-clutter the IDS results to identify only actionable data

Correlation Provides a Useful Method to Eliminate False Positives

- Correlate IDS events with the flow that caused the IDS to trigger
  - The flows that correlate with an IDS event are the flows on which to focus attention
- Correlate IDS-flows with vulnerabilities from vulnerability scanner
  - The flows where the IDS correlates with a vulnerability are the events on which to focus
- Correlation reduces false positives, but is computationally intensive
- Use pattern discovery to detect distributed, targeted attacks that may involve recurring flows from multiple sources

Identify Indicators of Possible Data Exfiltration

- Aberrant user behavior may indicate insider threat or user-level compromise.
  - Unusual login time for a particular user
  - New source host or organization
  - Multiple logins for same user from different locations within a small time window

Identify Indicators of Possible Data Exfiltration

- Unusual outbound activity may indicate unauthorized data exfiltration
  - Unauthorized access or attempting to access unauthorized hosts
  - Unauthorized privilege escalation or attempting privilege escalation
  - Unusual file transfer protocol
  - File transfer to host or organization not seen before or to an IP address on a watch list
  - File transfer to unauthorized foreign county
  - Unusual emails such as emails with extremely large attachments and email bursts to free email providers

Visualization Is Important to Cyber Security Since a Picture is Worth 1000 Words

- Visualization allows the security analyst to see events and relationships that may be lost in rows of data as well as summarize security event activity
- Provide visualization of the various flows that are of most concern to help identify patterns not before seen

Questions?