A system for end to end brokering of diverse data types

Blake Matheny

March 2003

A LaTeX Presentation

Introduction

- What’s the problem space?
- Is there a solution?
- Addressing the issues
- Next generation Intrusion Detection
- Other Applications

Motivations

- What is the purpose of a logging mechanism?
  - To relay relevant information to an appropriate agent in the amount of time most suitable for that event.
- What's wrong with current mechanisms?
  - They don't do this, they relay data to any agent that understands a certain data format and speaks a certain protocol regardless of content

Why is that a problem?

- Heterogeneous application layer data formats lead to information loss or at best information overload
- Protocol diversity leads to unusable data
- Not all data is needed by every agent
- Makes forensics and correlation difficult or impossible

What does solving these problems give us?

- Scope resolution for events
- Better correlation through reduction
- Better detection by directing higher level events to appropriate systems
- Detection of unknown attacks through system modeling

What’s the solution?

What's the solution?

A system for end to end brokering of diverse data types

Some good ideas

- Firewall paradigm
  - Network filters are good, so are application filters (content inspection)
  - **The combination is better**
- Routing paradigm
  - Source routing can be useful
  - How about information routing? (this is where Babe comes in)

Babe architecture

- Modules
  - Input - Collect data from sources
  - Data Format - Convert ASCII data to an intermediary language or from intermediary language to ASCII
  - Output - Send data to requested destinations
- Core
  - When new input data is received from an input module, pass it to a format module or process ourselves with FSM
  - Use filters on intermediary language
  - Filters evaluate as boolean expressions
  - Send to one or more output modules, or discard based on rule-set

Addressing the issues

- Need to model events
- Language to model high level scenarios
- Data format transformation is hard
- Complex systems are vulnerable to SADS (SysAdmin Deficiency Syndrome)

Data Types

- Data abstraction
- Three types of data formats
  - Unstructured (NL, free text, etc)
  - Semi-structured (html, syslog)
  - Structured (xml, uml, most programming languages)
- Most security data formats are either semi-structured or structured
- We believe that we can model the second two types of data formats, while staying within our requirements, with one common language

Requirements for Transformations

- Need bijection between original data and intermediary language
  - Must maintain structure
  - Transformation must be lossless
- Mappings must be user definable
- Mapping definitions are two-way functions (i.e. $f(x) = y$ and $f(y) = x$)

Event Modeling

- Allen's Temporal Algebra
  - Temporal relations between events
- McCarthy's Situational Calculus
  - Causal relations between events

What's right with current Intrusion Detection Systems?

- They gather lots of information
- Sometimes they actually detect intrusion attempts
- Some of them sort of do anomaly detection
- A couple of them are open source
- more...

What’s wrong with current Intrusion Detection Systems?

- They gather lots of information
- Primitive
- Deal with mostly homogenous sources of information
- Anomaly detection isn’t
- False positives lead to habituation in system admins
- The knowledge that a system administrator has often isn’t taken into account
- much more...

How this system fits in

- Reduce and correlate events
- Send primitive and aggregate events to systems that can use them
- Detection of unknown attacks through system modeling
- Take event information from multiple events for better detection
- more..

Where do things stand?

- What's done?
  - We have a prototype system that does:
    - We do data proxying (smtp->smtp)
    - We can do network filtering (if source == x send to y)
  - We've mostly worked out the details for our event logic
- What's not done?
  - Transformations
  - Lots more

What’s up with the transformations?

- We found a solution!
- We threw that solution in the trash
- Currently investigating FST
  - AT&T has done some interesting work, as has Xerox
  - Looks good, operates as a two way ‘function’

Research ’opportunities’ for budding scientists

- Compiler design
- Mathematical modeling of attacks
- Ontological systems and their role in NLP and NLG
- Data mining
- Digital Libraries

Questions?

http://www.nongnu.org/babe/