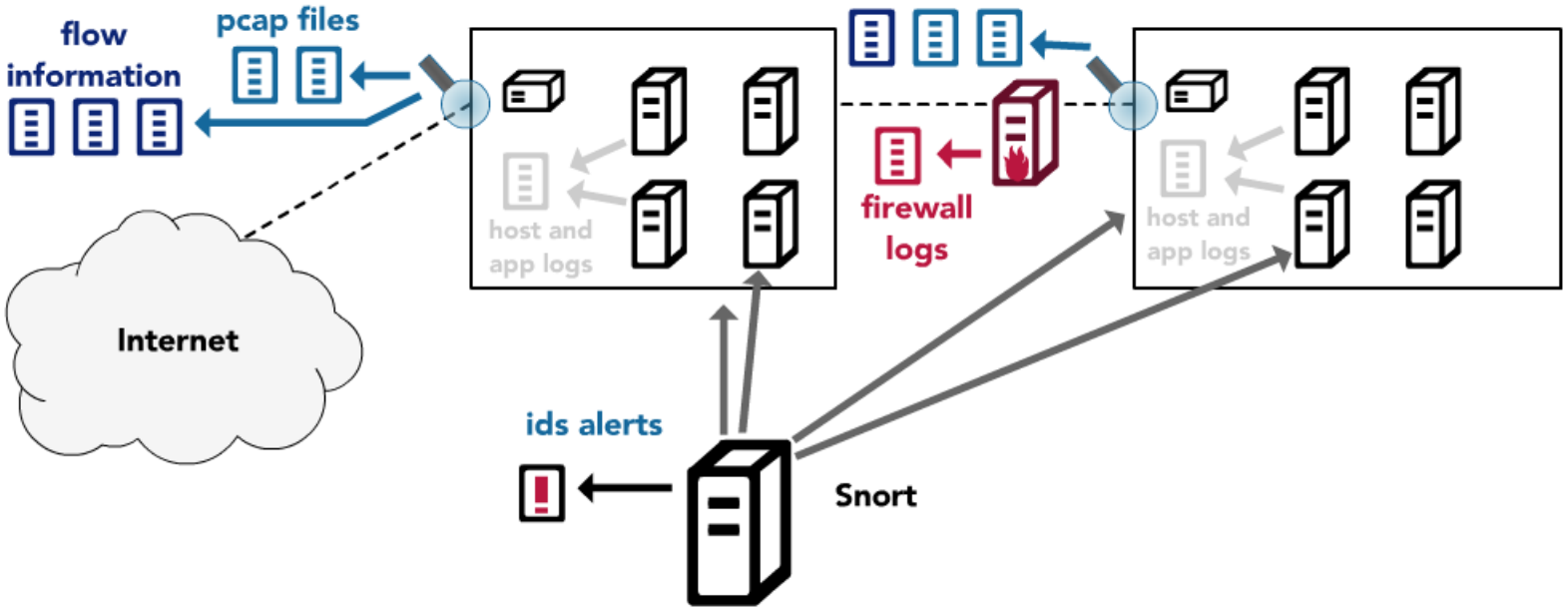
Representing Reality in a Research Environment

Sam Gorton <sam@skaion.com>
Skaion Corporation
North Chelmsford, MA
http://www.skaion.com

ACSAC 2005
December 13, 2005

Introduction

- Skaion provides realistic sets of sensor streams
  - Supporting Cyber Situational Awareness systems for ARDA's P2INGS program, data for official use only
  - Labeled training sets and unlabeled testing sets



Why do we need realistic data?

- Problem sets embody domain-specific challenges
  - “Show, don't tell”
  - Researchers may not be domain experts
- Test and validate approaches
  - Researchers have multiple ideas
  - Some researchers use data mining to discover new techniques
  - Fix potential problems during development
- Goal: simplify transition from research lab to the real world

Skaion Corporation

ARDA

Data Sets Represent the Problem

- “Foreground” scenarios of interest – the needle
  - Both malicious and non-malicious traffic
  - Include the full range of the problem set: don't overspecialize
- Scenarios should represent customer's requirements
  - Customers have to codify their desires
  - Give researchers guidance early, possibly in a paper
- Model specific networks – don't try for an “average”
  - Skaion chose the Open Source Information Network (OSIS), an unclassified, Internet-connected intelligence community network
  - Modeled OSIS only using open source information

Skaion Corporation

ARDA

Data Set Generation Tools

- Skaion Traffic Generation System
  - Real traffic generated using synthetic users
- Metasploit Framework
  - Flexible, runs on multiple platforms
  - Only a few exploits valid on our testbed
- CORE Impact
  - Windows-only, not as flexible or scriptable, unsubtle
  - Very useful for post-exploit behavior
  - Client-side exploits

Skaion Corporation

ARDA

What About Real Data?

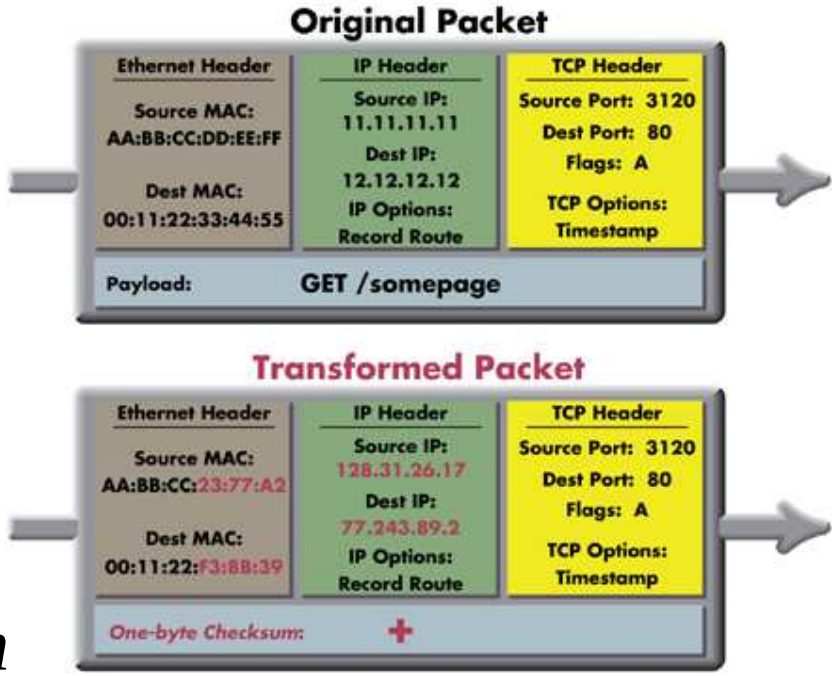
- Real data is great – if you can get it
  - If it represents the target environment
  - And if you can distribute it
  - And if you know the ground truth
- Real incidents are particularly sensitive
  - Honeypots record attacks, but with no “normal” traffic
- **Real b**ackground traffic feeds synthetic traffic generators
  - Defensible model for protocols, statistics, proportion

Skaion Corporation

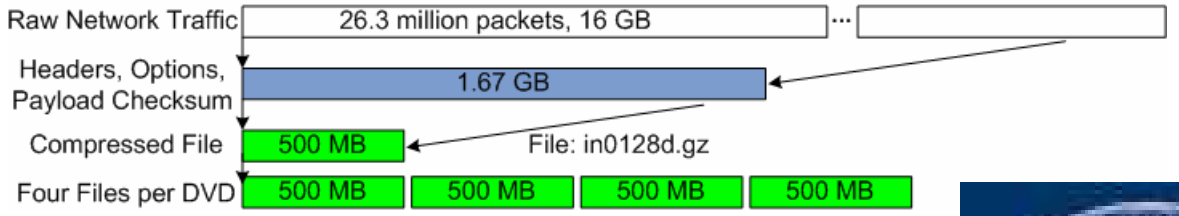
ARDA

Traffic Characterization

- Collected (anonymized) real packet headers
  - Consistent mapping
  - Headers compress well
  - No private information
  - ... but no *attack information*

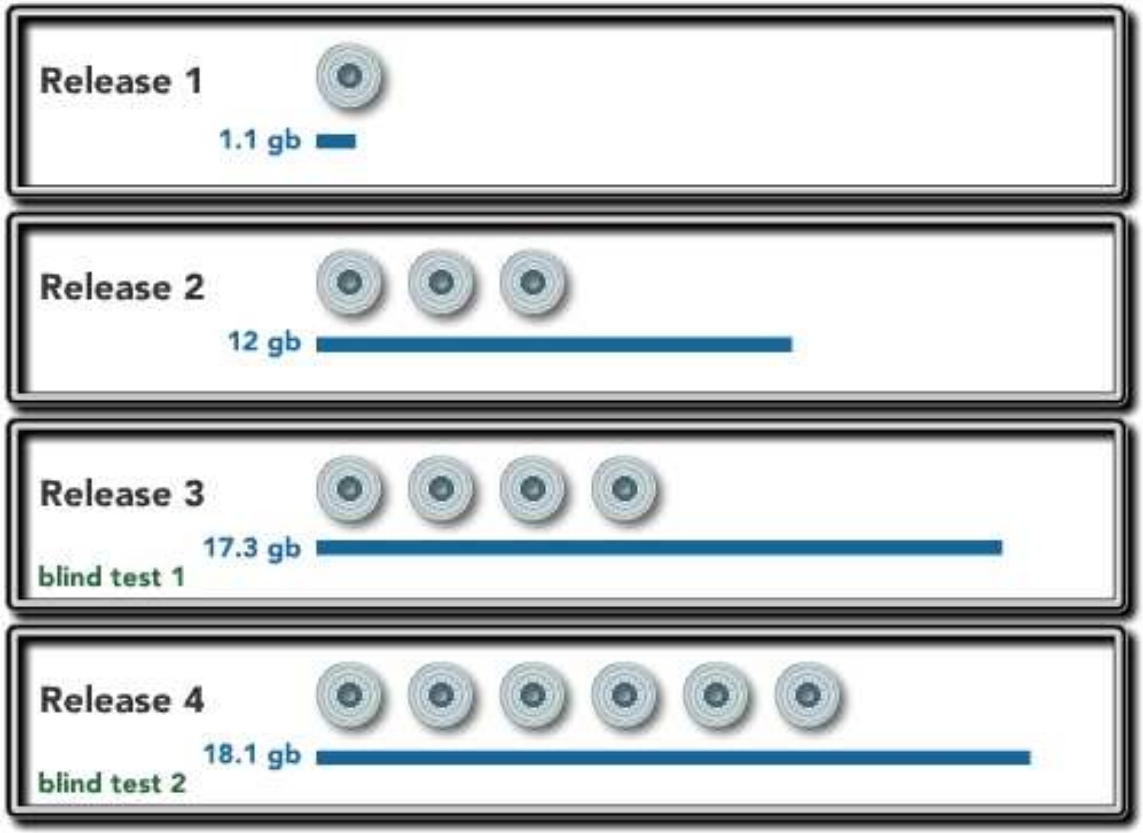


- Characterized with CERT SiLKtools, perl scripts



History

- Four data sets over two years, two blind tests
  - Blind test 2 is currently ongoing, shows major improvement
- Iterative releases
- Collaborative approach
- Increasing in scale and complexity



What did we gain?

- Reality check for research systems
  - A.K.A. “A kick in the pants”
- Measures of improvement
- Understanding of the research systems
  - Review process can be more useful than the results
- Lessons learned, improving data generation and research systems

Results from Labeled Data Releases

- First data set was not complex enough
  - Opening a novel backdoor port on a target system is too easy to detect
  - Less is more: too many background attacks are not anomalous
- Artificially clean “role separation” between client and servers made attacks easy to detect
- Collaborative, iterative approach improved data set generation
  - Researchers could spot flaws early
- Discovered data sets and ground truth weren't fully utilized until government-run blind testing

Results from Blind Test 1

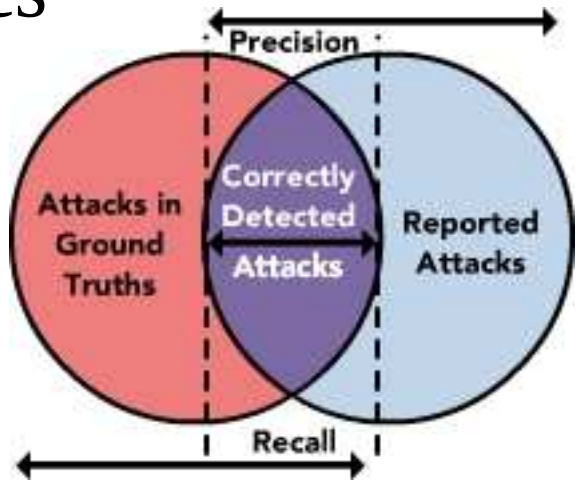
- Defined desired output as a ranked list of suspicious connections
- Discovered conceptual flaws in systems that:
  - Caught attack, but ranked it as part of a 672-way tie
  - Thought all alerts were attacks
  - Ignored source IPs during correlation
- Many systems confused failed and successful attacks
- Realized Skaion staff needed to attend each test
- Discovered at least one bug in each system
  - Including ours: needed to “close the loop” for ground truth

Ground Truth

- What happened during the scenario?
- What parts of the traffic are valid, what are not?
- Ground truth is not just the foreground scenario
  - Research systems need to understand false positives as well as false negatives
  - Anomaly detection systems need to understand the true context of detected anomalies
- Ground truth broken down by IP address and sensor, all alerts stored in a relational database
  - Different versions of an IDS are essentially different systems

Metrics

- Characterizing the difficulty of datasets
  - Complexity and scale are related, but not identical
- Situational Awareness metrics
  - Recall and Precision
  - Fragmentation or Overcorrelation
  - Data/Information Ratio



- SA metrics published as "Achieving Situation Awareness in a Cyber Environment" in Milcom 2005 by John Salerno *et al, contact <john.salerno@rl.af.mil>*

Known Concerns

- Some research systems can't take prepackaged input
- Finding “level 1” fusion info for host, web logs
  - Correlators consume interesting events, not raw logs
- Differentiating between successful and failed attacks
  - We argue it can't be done (today) from IDS logs alone
- Real world has data sets with millions of alerts
  - Can sacrifice some detail and realism to increase scale
- Classified environments are harder to characterize
  - Ad-hoc coalition networks are complex and difficult to defend

Recommendations for Future Research Programs

- Start work on providing test data early
- Record real traffic in target environments
  - Real traffic can seed generated traffic
  - Abstracted descriptions can be distributed
- Blind tests can detect problems long before deployment
- Researchers should also have internal blind tests