What You Don't Know Can Hurt You!

An Overview of Scalable Security Data Management for *Internal/External* Data Sharing

William Yurcik* Adam Slagell Jun Wang

NCSA Security Research
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign

ISMA Data Catalog Workshop
3 June 2004 SDSC

Outline

- **Log Problem Overview**
- **Incentives**
- **Log Management @ NCSA**
- **Log Visualization @ NCSA**
- **Discussion**

Log Problem Overview

My Personal Motivation

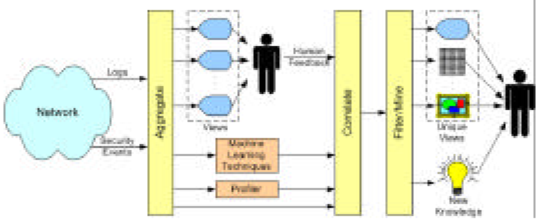
N-Dimensional Security Solution Space:

- large networks
  - Class B IP address space, 65,000 devices
- complex networks:
  - 130K ports per computer (tcp/udp)
  - heterogeneous hw platforms (intel,mac,sgi,sun)
  - heterogeneous sw (OSs, applications)
  - many services & protocols (web, mail, ftp, streaming,..)
- many types & dynamic nature of both
  - vulnerabilities (hw, sw (OS/application), network…)
  - attacks (worms, viruses, DoS, intrusions, …)

BOSS: enable situational awareness of a large & complex environment by leveraging human visual processing capabilities (interactivity & measurement)

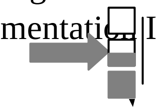
The NCSA SIFT Project Approach

Improved intrusion detection process and visualization

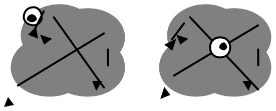


Streaming Data Instrumentation Issues

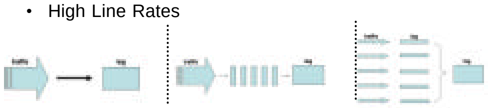
- Logs in time



- Vantage Point



- High Line Rates



Commonly Available Logs

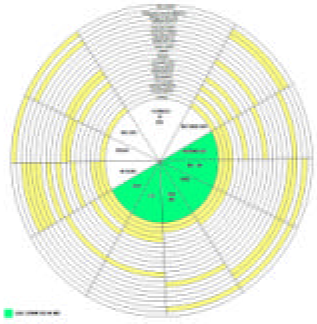
1) NetFlows Logs
2) Packet Traces - tcpdump
3) Network IDS- BRO,Snort
4) Host IDS - Tripwire
5) Syslogs (general)
6) Kerberos Logs
7) DHCP Server Logs
8) Firewall logs
9) Mail Server Logs
10) Vulnerability Scan Logs
11) Nameserver DNS Cache
12) SNMP Logs
13) BGP tables
14) Dial-Up Server
15) ARP Cache
16) Workstation Logs
17) Process Accounting Logs
18) Trace Route Logs

Principles of Log Selection

- Logs must be commonly available
- Accuracy to detect specific known attacks
- Coverage over many different attacks
- Extensible to detect new attacks
- Orthogonal (independent) attribute information
  - **Our Selection:**
    - **system logs (specifically syslog but others available)**

    versus
    - **network logs (specifically NetFlows but others available)**
  - **other possibilities**
    - **storage logs**
    - **application logs**
    - **human user logs (video cameras, biometrics)**
    - **hardware logs**

Attributes Across Logs



Challenges

Incentives

- Time/Effort
- Economic (probably not)
- Law (regulation possible)
- Altruism & PhD Research (fringe)
- Security may be the key

Data Management

- Huge data volume!
- Data distributed all over
- Data sources change over time

Security

CIA
- Confidentiality (anonymization vs key management)
- Integrity (checksums)
- Availability (access control)

Only cooperation will make us less vulnerable

Incentives

Question:

What is the profile of
who would not share data?

Where Does Data Sharing Take Place Now?

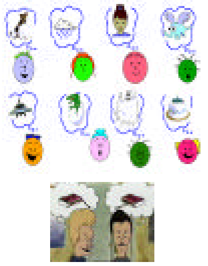
<http://www.first.org/>

Forum of Incident Response and Security Teams

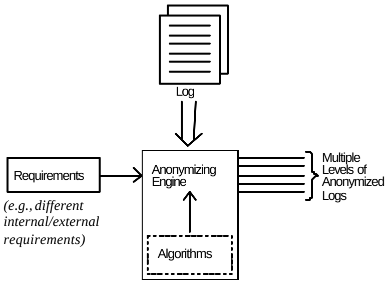
CIC-SWG

Committee on Institutional Cooperation
- IT Security Working Group
(Big Ten Universities plus the University of Chicago)
<http://www.cic.uiuc.edu/groups/ITSecurityWorkingGroup/>

VoIP



Log Anonymization



Log

Requirements

(e.g., different internal/external requirements)

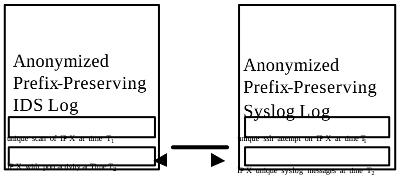
Anonymizing Engine

Algorithms

Multiple Levels of Anonymized Logs

Known Plain-Text Attacks

Statistical Inference



Anonymized Prefix-Preserving IDS Log

unique scan of IP X at time T_1

IP X with p2p activity at Time T_2

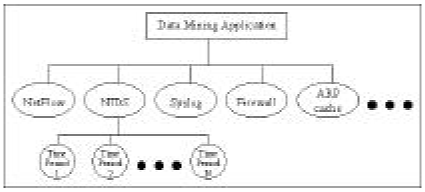
Anonymized Prefix-Preserving Syslog Log

unique ssh attempt on IP X at time T_1

IP X unique syslog messages at time T_2

Log Management @ NCSA

The Data Management Problem



Data Mining Application

NetFlow

NIDS

Syslog

Firewall

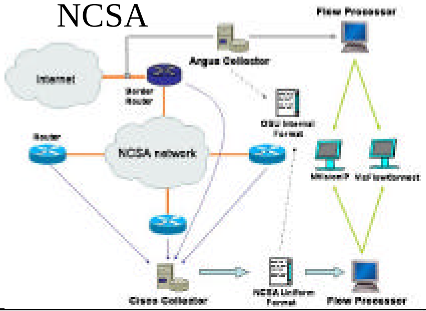
ARP cache

Time Period 1

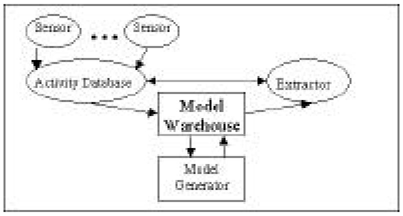
Time Period 2

Time Period N

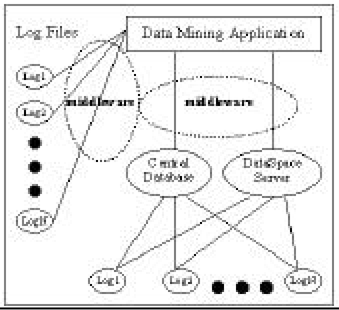
Four (4) Parallel Data Management Efforts @ NCSA



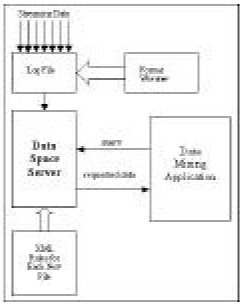
(1) Central Database Architecture



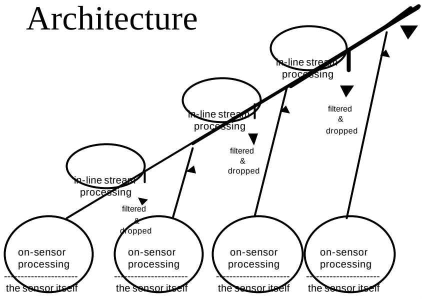
(2) Middleware Architecture



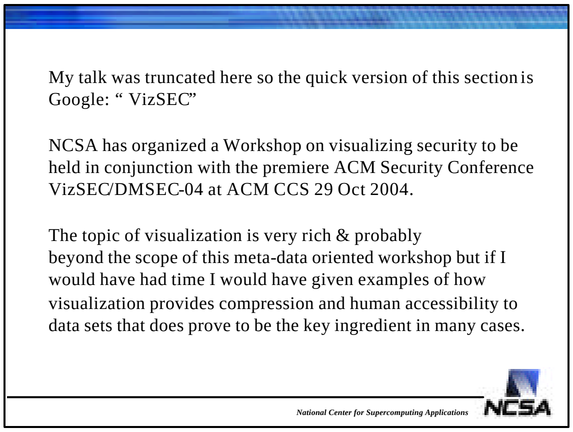
(3) DataSpace Architecture



(4) Datamines Distributed Agent Architecture



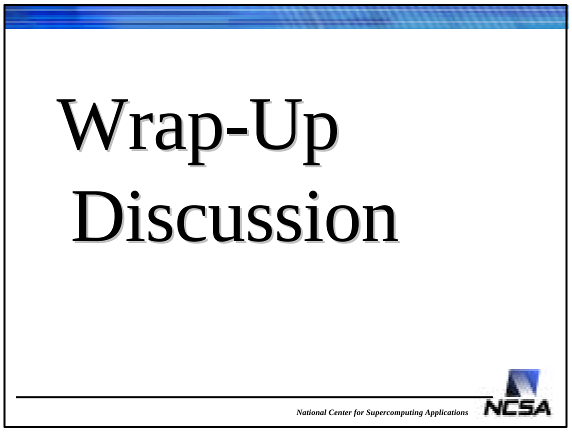
Log Visualization @ NCSA



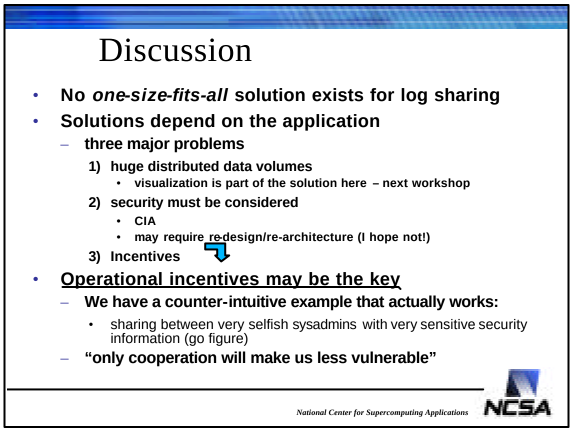
My talk was truncated here so the quick version of this section is Google: " VizSEC"

NCSA has organized a Workshop on visualizing security to be held in conjunction with the premiere ACM Security Conference VizSEC/DMSEC-04 at ACM CCS 29 Oct 2004.

The topic of visualization is very rich & probably beyond the scope of this meta-data oriented workshop but if I would have had time I would have given examples of how visualization provides compression and human accessibility to data sets that does prove to be the key ingredient in many cases.



Wrap-Up Discussion



Discussion

- **No *one-size-fits-all* solution exists for log sharing**
- **Solutions depend on the application**
  - **three major problems**
    1) **huge distributed data volumes**
       - **visualization is part of the solution here – next workshop**
    2) **security must be considered**
       - **CIA**
       - **may require re-design/re-architecture (I hope not!)**
    3) **Incentives**
- Operational incentives may be the key
  - **We have a counter-intuitive example that actually works:**
    - sharing between very selfish sysadmins with very sensitive security information (go figure)
  - **"only cooperation will make us less vulnerable"**