The Changing Internet Ecology: Confronting Security and Operational Challenges by Mining Network Data

Farnam Jahanian
University of Michigan and Arbor Networks

Workshop on Mining Network Data (MineNet-05)
August 26, 2005
SIGCOMM 2005

Security and operational challenges and ... a few trends
Emerging Trends in Security Threats

- **Globally scoped**, respecting no geographic or topological boundaries.
  - At peak, 5 Billion infection attempts per day during Nimda including significant numbers of sources from Korea, China, Germany, Taiwan, and the US. [Arbor Networks, Sep. 2001]

- Exceptionally **virulent**, propagating to the entire vulnerable population in the Internet in a matter of minutes.
  - During Slammer, 75K hosts infected in 30 min. [Moore et al, NANOG February, 2003]

- **Zero-day** threats, exploiting vulnerabilities for which no signature or patch has been developed.
  - In Witty, “victims were compromised via their firewall software the day after a vulnerability in that software was publicized” [Symantec Security Response Mar 2004]

- Profound **transformation** underway: from attacks designed to disrupt to attack that take control.
  - Over 900,000 infected bots as phishing attacks are growing at 28% per month [Anti-Phishing Working Group 2005]

The Crumbling Perimeter

Much of perimeter security problem addressed by making perimeter vulnerability-aware (IDS, smart firewall, VA)

With crumbling perimeter (wireless, tunnels, etc) and near-zero visibility, internal network security has emerged as the most pressing IT security issue
Yesterday ... Availability Attacks

Worms

These attacks disrupt infrastructure

DoS

Viruses

A Dramatic Transformation and Escalation

ID Theft

Phishing

These attacks directly target people

SPAM

Spyware

Yahoo Attributes a Lengthy Service Failure to an Attack

San Francisco, Feb. 7 -- Yahoo Inc. blamed a "planned attack" by computer hackers for a service failure that lasted nearly three hours today, in a rare interruption of one of the most popular and best performing sites on the World Wide Web.

FTC, Partners Launch Campaign Against Spam "Zombies"

The Associated Press

"Phishing" e-mails widespread, survey finds
Rise of the Botnets (Zombie Armies)

- 1000's of new bots each day [Symantec 2005]
- Over 900,000 infected bots as phishing attacks are growing at 28% per month [Anti-Phishing Working Group 2005]
- A single botnet comprised of more than 140,000 hosts was observed “in the wild” [CERT Advisory CA-2003-08, March 2003]
- A study conducted by the UofM showed that an out of the box Windows 2000 PC was recruited into 3 discrete botnets within 48 hours

**Attackers have learned a compromised system is more useful alive than dead!**

- Significant more firepower: Broadband (1Mbps Up) x 100s == OC3!!!
- An entire economy is evolving around bot ownership
  - Sell and trade of bots ($0.10 for “generic bot”, $40 or more for an “interesting bot; e.g., a .mil bot)
  - Bots are a commodity - no significant resource constraints

What Threats are Providers Concerned About?

- Recent Arbor/UM survey of 40+ tier1/tier2 providers

Top Two Threats

- DDOS
- Worms
- Compromise
- DNS Poisoning
- BGP
Network Managements & Traffic Engineering

- Transit/Peering Management
- Backbone Engineering
- Capacity Planning / Provisioning
- Root-cause Analysis / failure diagnosis
- Routing Anomalies
- Abuse and Misuse
- Distributed Denial of Service

BGP Address Hijacking

- Though providers filter customer BGP announcements, few filter peers
  - Memory, line-card limitations
  - Maintenance problem
- More specific announcements wins
- Injection attack requires compromised commercial or PC-based router
  - man-in-middle session attacks rare
ISP Network Architecture

Operational and security issues are increasingly global crossing provider and customer boundaries

A Crash Course in Data Mining Terminology

- What is data mining?
  “Data mining is the process of automatically discovering useful information in large data sets.” [Tan, Steinbach and Kumar 2006]

  “Concerned with uncovering patterns, associations, changes, anomalies, and statistically significant structures and events in data.” [RL Grossman 1997]

- Descriptive Analysis: Derive patterns (correlations, trends, clusters, trajectories) that capture the underlying relationships in data.

- Predictive Analysis: Predict the value of a target variable based on the values of explanatory variables.

Data Mining Concepts

• Data Exploration

• Association Analysis

• Cluster Analysis

• Predictive Modeling
  • Classification
  • Regression

• Anomaly Detection

Data Exploration

• Preliminary investigation of data to better understand its characteristics

• Informs the selection of data analysis techniques

• Summary statistics
• On-line analytical processing
• Visualization
**Association Analysis**

- Association analysis is used to discover patterns and relationships hidden in large data sets
  - Association rules or sets of frequent items (binary attributes)
  - Association analysis for categorical and continuous attributes, and more complex entities (hierarchies, sequences, subgraphs)

**Cluster Analysis**

- Cluster analysis divides data (or objects) into groups (classes) that share certain characteristics or closely related attributes.
  - K-means (prototype-based clustering)
  - Hierarchical agglomeration (graph-based clustering)
  - DBSCAN (density-based)
Predictive modeling refers to the task of building a model for the target variable as a function of explanatory variables.

• Classification: for discrete targets --- task of assigning objects to one of several predefined categories called class labels.

• Regression: for continuous targets --- task of learning a function that maps attributes into a continuous-valued target variable.

• Decision trees, rule-based, nearest-neighbor, Bayesian classifiers, neural networks.
Anomaly Detection

• Anomaly detection is the task of identifying observations whose characteristics are measurably and significantly different from the rest of the data.

• High detection rate and low false positive rate

• Major categories of anomaly detection approaches: statistical, proximity-based, density-based, and cluster-based.

Challenges of Data Mining

• Instrumentation and Measurement
• Scalability
• Dimensionality
• Complex and Heterogeneous Data
• Data Ownership and Distribution
• Privacy Preservation
• …
Raw Traffic

- Getting the traffic
  - Span Port
  - Static Routing
  - NBAR (Cisco)
  - AS-PIC (Juniper)
  - Fiber Tap

- Reading the traffic
  - Roll your own (hardware) with a network processors like IXP
  - Buy a DAG (e.g. Endace)
  - Roll your own (software) with a PC and NICs

Instrument or Monitor Devices

- Core infrastructure devices
  - Routers
  - SNMP
  - DNS

- Application Servers
  - Web
  - Mail

- Security devices
  - Firewalls
  - IDS
  - AV
Blackhole Monitoring Sensors

- CAIDA - Network Telescope
- Internet Motion Sensor (IMS)
- Team Cymru - DarkNets
- IUCC/IDC Internet Telescope
- iSink
- BGP off-ramping techniques (CenterTrack, SinkHoles)

⇒ Investigating DDoS
⇒ Tracking worms
⇒ Characterizing emerging Internet threats

Distribute Sensors
(Not All Blackholes are Created Equal)

- Clearly more addresses are better, but…
- Normalized by /24
- Includes all protocols
- Month long observation period

Each sensor block sees a very different traffic rate

Different Perspectives
(In Search of Network-wide Visibility)

- Worms can have a local preference
- Local service scanning
- Local mis-configuration

Each sensor block sees very different local preference

Analyzing Global Events

- Different sensors see different things
  - Just because an event is globally scoped, doesn’t mean that all parts of the network have the same view of an event.
- Many sensors are dominated by targeted attacks and local activities
  - Just because an event is very prevalent at 1 or a small number of locations does not mean the event is global
- The challenge with network-wide view
Sources are not Prevalent Across Locations

- For random scanning events, monitored block size and scanning rate define the time to detection.
- Even larger blocks don’t see all IP addresses, some attacks are targeted.
- As an example, compare the average daily source IP address overlap between sensors

<table>
<thead>
<tr>
<th></th>
<th>A/18</th>
<th>B/16</th>
<th>C/16</th>
<th>D/23</th>
<th>D/8</th>
<th>E/22</th>
<th>E/23</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/18</td>
<td>100(0)</td>
<td>25(2)</td>
<td>58(5)</td>
<td>4(0)</td>
<td>78(5)</td>
<td>2(0)</td>
<td>4(0)</td>
</tr>
<tr>
<td>B/16</td>
<td>23(3)</td>
<td>100(0)</td>
<td>38(5)</td>
<td>3(0)</td>
<td>54(8)</td>
<td>1(0)</td>
<td>3(0)</td>
</tr>
<tr>
<td>C/16</td>
<td>23(2)</td>
<td>17(1)</td>
<td>100(0)</td>
<td>3(0)</td>
<td>78(6)</td>
<td>0(0)</td>
<td>2(0)</td>
</tr>
<tr>
<td>D/23</td>
<td>10(0)</td>
<td>10(1)</td>
<td>20(1)</td>
<td>100(0)</td>
<td>30(1)</td>
<td>0(0)</td>
<td>1(0)</td>
</tr>
<tr>
<td>D/8</td>
<td>2(0)</td>
<td>1(0)</td>
<td>5(0)</td>
<td>0(0)</td>
<td>100(0)</td>
<td>0(0)</td>
<td>0(0)</td>
</tr>
<tr>
<td>E/22</td>
<td>10(2)</td>
<td>8(1)</td>
<td>13(1)</td>
<td>1(0)</td>
<td>12(1)</td>
<td>100(0)</td>
<td>3(0)</td>
</tr>
<tr>
<td>E/23</td>
<td>25(4)</td>
<td>20(3)</td>
<td>33(5)</td>
<td>3(0)</td>
<td>34(5)</td>
<td>5(1)</td>
<td>100(0)</td>
</tr>
</tbody>
</table>

Destination ports are not equivalent across Locations

- Examine the top ten ports over a day, week and month time frame.
- Determine how many of those ports appear at each of the 31 blackhole sensors.
- Only a few ports are visible at all sensors. Many are only visible at one.

Scalability

- Tier-1 Service Providers generally see O(GB) of traffic at a single interface. Hundreds or thousands of such interfaces exist in their backbone.
- Network-wide data collection and analysis

Case Study: Flow Based Abstractions

- A flow is a traffic stream with a unique [source-IP-address, source-port, destination-IP-address, destination-port, IP-protocol] tuple.
- Contains a wealth of information about the conversation:
Flow Export

Export Packets
- Approximately 1500 bytes
- Typically contain 20-50 flow records
- Sent more frequently if traffic increases on NetFlow-enabled interfaces

Collector
NFC, cflowd, flow-tools, Arbor

Application GUI
Arbor, FlowScan

Flow Deployment Issues

- Impact of sampling on flow accuracy
  - Most deployments use 1/100 -1/1000 sampling

- For what purpose netflow is being collected?
  - e.g., Anomaly detection vs. traffic profiling
  - Achieving network-wide visibility

- Vendor sampling algorithms / knobs vary

- Backhaul versus local collection
  - Bandwidth usually 1% of offered traffic
  - Significant deployment issue

- Wide range of capabilities across router, engine cards and IOS/JUNOS/*OS version
  - Generally available in ASIC on most modern Cisco cards
  - Some support on Juniper RE w/additional capabilities from AS PIC, support by other vendors vary as well
Flow Accuracy versus SNMP Octets

- Flow does not include link layer header (only IP)
- Ground truth sometimes difficult -- significant implementation issues and inaccuracies with SNMP

Flow Accuracy at Low Traffic Rates
Measuring CBR 15Kbps on OC12 Link via tcpdump and 1/100 sampled NetFlow

- 1/100 Sampled flow v.tcpdump octet count associated with 15Kbps CBR microflow
- SNMP ifOctet count for interface during measurement period

Source: 2004 Arbor Networks Technical Report
High Dimensionality

- Data sets with large number of dimensions (features)
  - e.g. frequency with which each word occurs in a document
  - e.g. origin-destination pairs in flow analysis
- Dimensionality reduction can lead to more efficient data mining algorithm, can potentially eliminate irrelevant features and noise, may allow better visualization.

Flamingo: Visualizing Internet Traffic
Manish Karir, Merit Network
http://flamingo.merit.edu/

Techniques for Dimensionality Reduction

- Principal component analysis (PCA)
- Singular Value Decomposition (SVD)
- Multi-dimensional Scaling (MDS)
- Feature Subset Selection
- Factor Analysis
- Locally Linear Embedding
- 

[Patwari, Hero, and Pacholski, MineNet 2005.]
[Lakhina, Crovella, and Diot, SIGCOMM’05]
[Xu, Chandrashekar, and Zhang, MineNet 2005.]
Heterogeneity and Complex Data

• HTTP is a simple protocol
  • HTTP-message = Request | Response
• With lots of fields.
  • 8 different methods
    • OPTIONS
    • GET
    • PUT
    • …
  • 47 different header fields
    • ACCEPT
    • AGE
    • ALLOW
    • …
  • Complex message bodies

Network Data: Deep Packet Inspection

• Hundreds if not thousand of applications running.
• Top ten known applications only make up 50% of the total traffic.

1. HTTP
2. Bit-torrent
3. Edonkey
4. Gnutella
5. HTTPS
6. NNTP
7. SMTP
8. RTSP
9. SSH
Encryption and Tunneling Issue

• Lack of visibility for ISP traffic modeling and engineering

• Impact on behavioral modeling based on deep packet inspection

• Application classification

Data Ownership, Distribution, and Privacy

• Over the last four weeks, Merit (regional ISP) saw traffic from ~1,600 of the roughly ~17,000 origin AS

• Top ten percent of ORGIN AS’ only result in 20% of the traffic
Privacy

• Most Acceptable use policies of network prohibit the network operators for from collecting data other than for debugging the network.

• Any data collected must be sure to not identify individuals.
  • This most often means masking source IP address
  • May limit access to payloads.

• Protecting the sensor network is also problematic as fingerprinting may lead to reduced utility
  • Active fingerprinting techniques such as scanning
  • Passive fingerprinting of sensors through data publication

Predictive Modeling of Worms

• Many worms follow a 4-phases lifecycle of latency, growth, decay and persistence.

“The Blaster Worm: Then and Now”
Worm Growth

- The authors describe worm growth using a classic SI epidemic model.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>size of the total vulnerable population</td>
</tr>
<tr>
<td>$S(t)$</td>
<td>susceptibles at time $t$</td>
</tr>
<tr>
<td>$I(t)$</td>
<td>infectives at time $t$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>contact rate</td>
</tr>
<tr>
<td>$s(t)$</td>
<td>susceptibles ($S(t)$) / population ($N$) at time $t$</td>
</tr>
<tr>
<td>$i(t)$</td>
<td>infectives ($I(t)$) / population($N$) at time $t$</td>
</tr>
</tbody>
</table>

\[
\frac{dI}{dt} = \beta \frac{IS}{N}
\]

\[
\frac{dS}{dt} = -\beta \frac{IS}{N}
\]

Worm Decay

- Worm decay is often described in terms of half life and modeled with exponential decay (e.g. Witty)
Worm Persistence

• Decay is a slight misnomer as, unlike witty, most worms persist even years after their release.
• Most persistent populations follow a circadian pattern.

Slammer Data Plane Impact – A European SPs View

• Some DDOS/worms easier to detect than others…
Slammer Control Plane Impact – Temporal Correlation

Root-Cause Analysis?

Event Correlation: Is this a DoS attack?

BGP Flaps

Turns out that configuration/policy change lead to high CPU, which lead to BGP dropping, which lead to path change and new traffic on router.

CPU

Root-Cause Analysis?
Easy Anomaly Detection: MySQL Worm

Unique source IPs contacting TCP port 3306 over 4 weeks

23 unique IMS address blocks at 14 sites normalized by /24

Noise on TCP/445

Signal: A Sasser signature

Outbreak of the Sasser worm observed at a /24
Feature Entropy Timeseries

Anomalies can be detected & classified by inspecting *traffic features*

- # Bytes
- # Packets
- H(Dst IP)
- H(DstPort)

Port scan dwarfed in volume metrics…

But stands out in feature entropy, which also reveals its structure


Lessons of IDS / IPS / SEM

- Flood of low quality information
  - Too many alerts
  - Too many false positives
  - No “context”
  - Not adaptive to new threats
  - Just reporting

- As a result
  - IDS mainly serve forensic purpose
  - Emergence of Security Event Management market (SEM)
    - Event Correlation Engines: netForensic, ArcSight, GuardedNet
  - Emergence IPS (think IDS with a gun)
Network-Wide: Measurement, Detection, Backtracing and Mitigation

Wrap Up

- Network-wide view: Challenges of measuring, aggregating, storing and analyzing network-wide data
- Scalability and high dimensionality
- Off-line Analysis != On-line Analysis
  - Extending data mining techniques to real-time decision support
- Congratulations! You’ve detected an anomaly!
  - Anomaly Detection != Actionable Event
  - Anomaly Detection != Mitigation
- Correlation != Root-Cause Relationship
- Internal security: Deep packet inspection
- More predictive, less descriptive data mining techniques
- Hybrid model: mining network and host data
- Data mining as an interactive process
  - Contextualization of alarms and events
  - Informing security and network operations