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

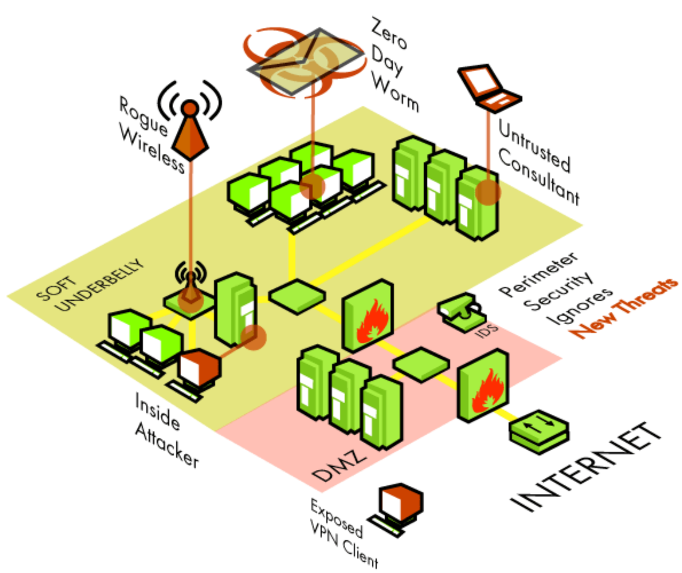
Security and operational challenges and … a few trends

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

A Dramatic Transformation and Escalation

Rise of the Botnets (Zombie Armies)

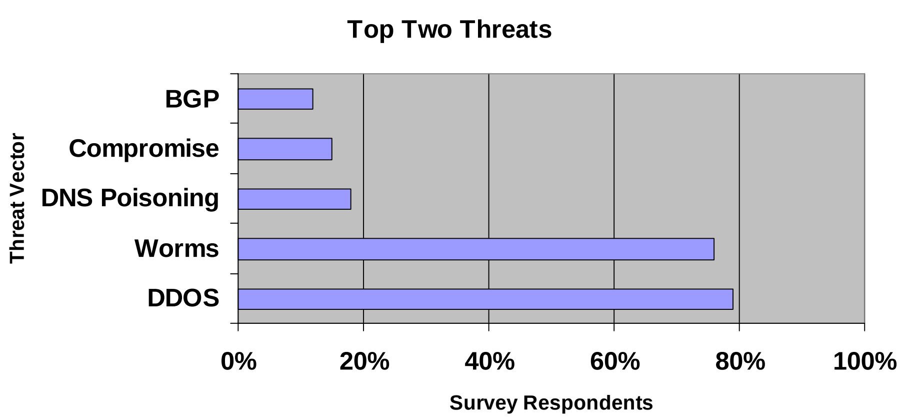
- 1000’s of new bots each day [Symantec 2005]
- Over 900,000 infected bots as phishing attacks are growing at 28% per month [Anti-Phishing Working Group 2005]
- A single botnet comprised of more than 140,000 hosts was observed “in the wild” [CERT Advisory CA-2003-08, March 2003]

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

  - (evasion/economics?)
  - 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?

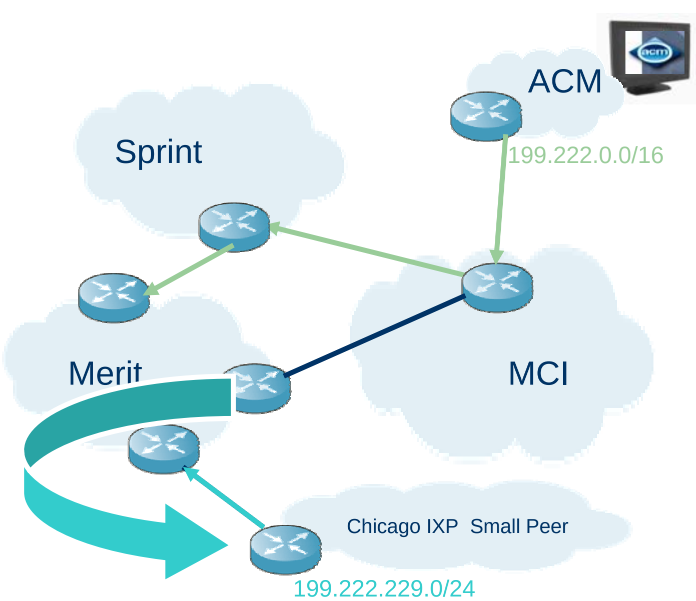
- Recent Arbor/UM survey of 40+ tier1/tier2 providers



Network Managements & Traffic Engineering

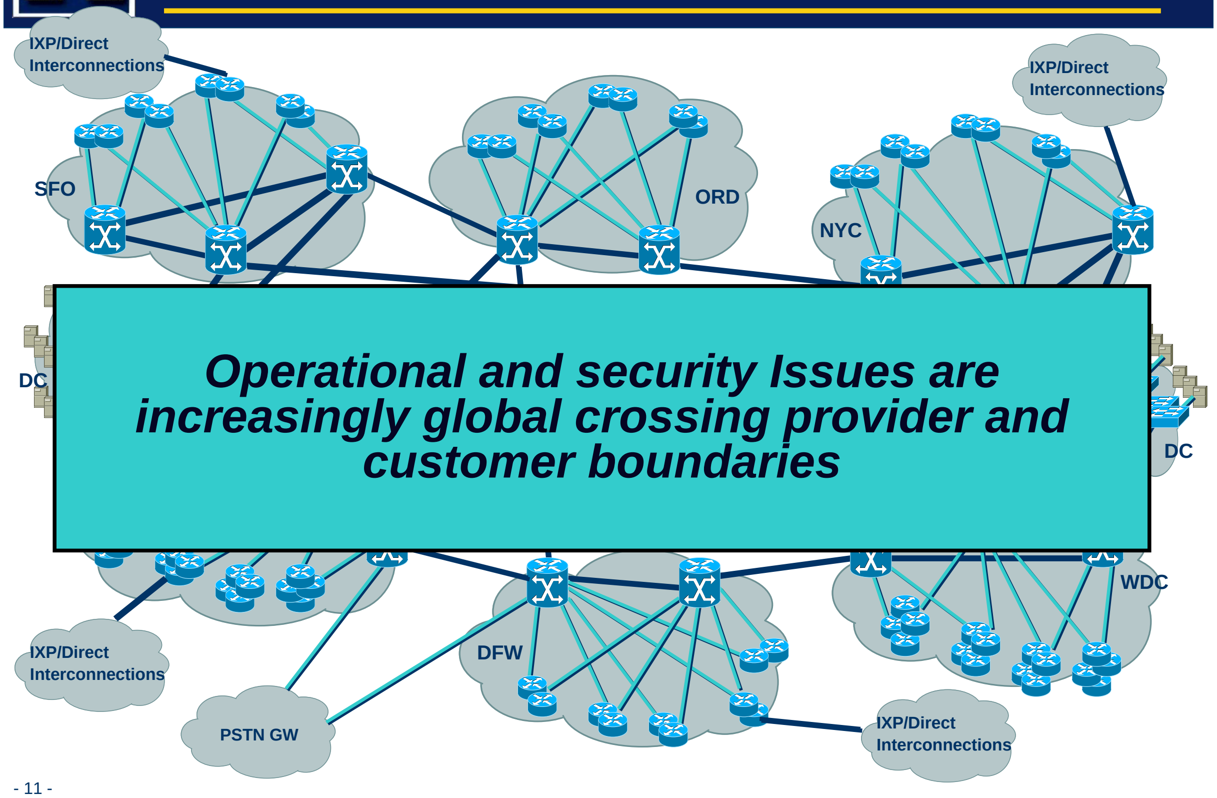
- Transit/Peering Management
- Backbone Engineering
- Capacity Planning / Provisioning
- Root-cause Analysis / failure diagnosis
- Routing Anomalies
- Abuse and Misuse
- Distributed Denial of Service

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.

*P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Addison-Wesley, 2006.

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

- 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.

Predictive Modeling

- 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
    - Decision trees, rule-based, nearest-neighbor, Bayesian classifiers, neural networks
  - Regression: for continuous targets --- task of learning a function that maps attributes into a continuous-valued target variable.

Anomaly Detection

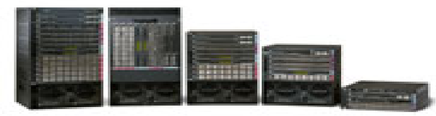
- Anomaly detection is the task of identifying observations whose characteristics are measurably and significantly different form the rest of the data.
- High detection rate and low false positive rate
- Major categories of anomaly detection approaches: statistical, proximity-based, density-based, and cluster-based.

Challenges of Data Mining

- Instrumentation and Measurement
- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Ownership and Distribution
- Privacy Preservation
- …

Raw Traffic

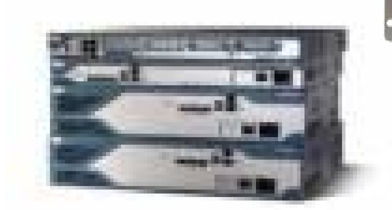
- Getting the traffic
  - Span Port
  - Static Routing
  - NBAR (Cisco)
  - AS-PIC (Juniper)
  - Fiber Tap



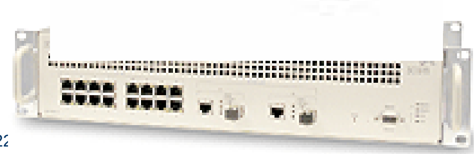
- Reading the traffic
  - Roll your own (hardware) with a network processors like IXP
  - Buy a DAG (e.g. Endace)
  - Roll your own (software) with a PC and NICs

Instrument or Monitor Devices

- Core infrastructure devices
  - Routers
  - SNMP
  - DNS
- Application Servers
  - Web
  - Mail

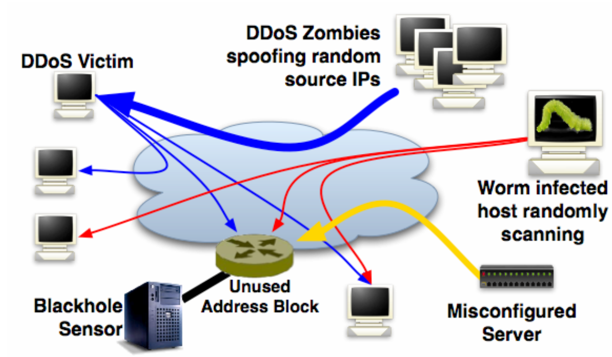


- Security devices
  - Firewalls
  - IDS
  - AV



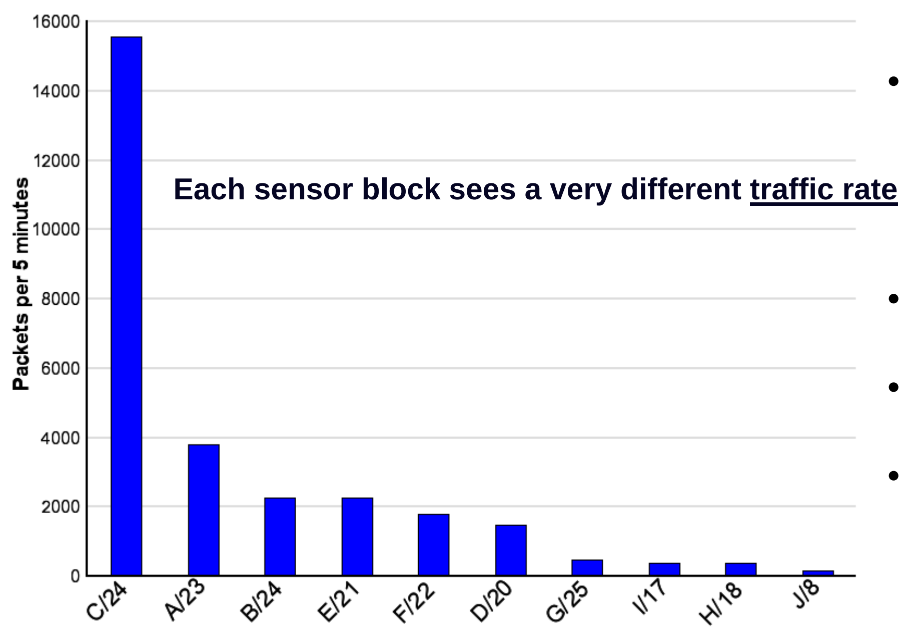
Blackhole Monitoring Sensors

- 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

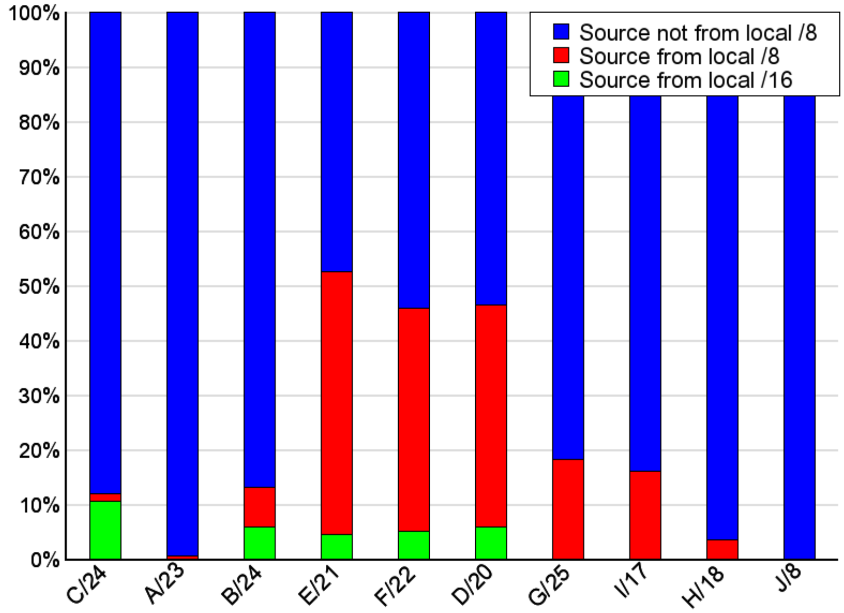
Distribute Sensors
(Not All Blackholes are Created Equal)



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

Cooke, Bailey, Mao, Watson, Jahanian, and McPherson, "Toward Understanding Distributed Blackhole Placement," WORM'04, Washington, DC, October 2004.

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 <u>local preference</u>

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

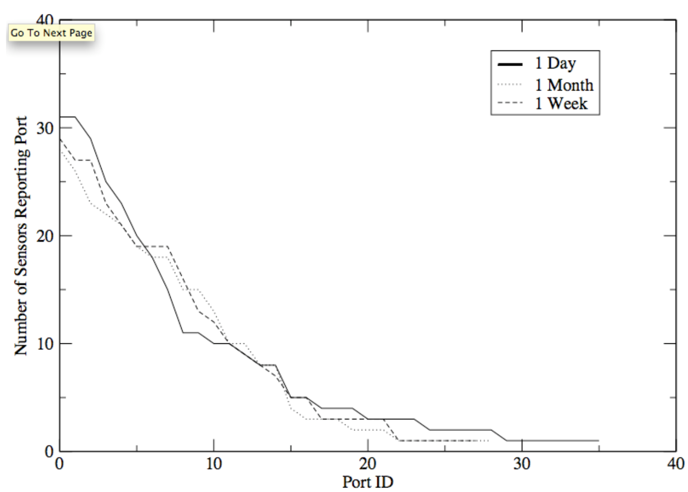
Sources are not Prevalent Across Locations

- 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

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

Destination ports are not equivalent across Locations

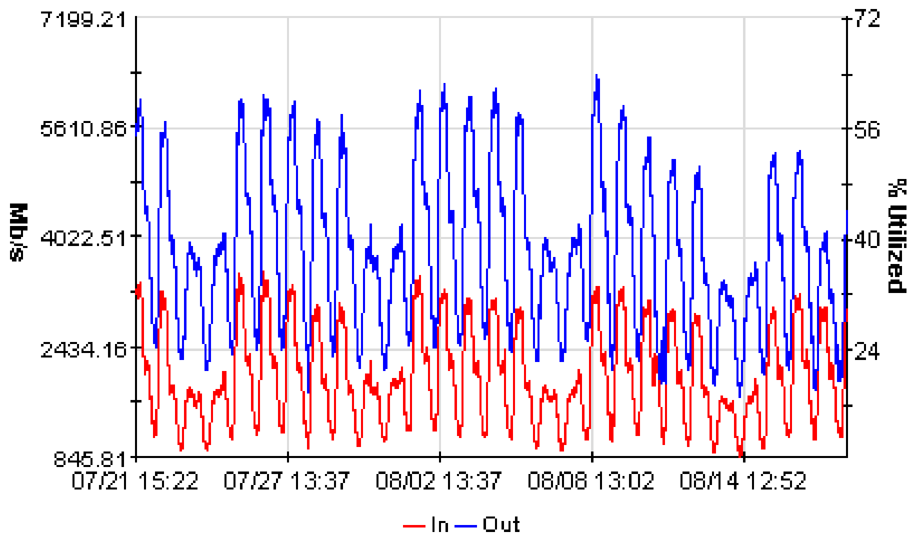
- Examine the top ten ports over a day, week and month time frame.
- Determine how many of those ports appear at each of the 31 blackhole sensors.
- Only a few ports are visible at all sensors. Many are only visible at one.



Bailey et. al., "Data Reduction for the Scalable Automated Analysis of Distributed Darknet Traffic" Internet Measurement Conference (IMC), Oct. 2005.

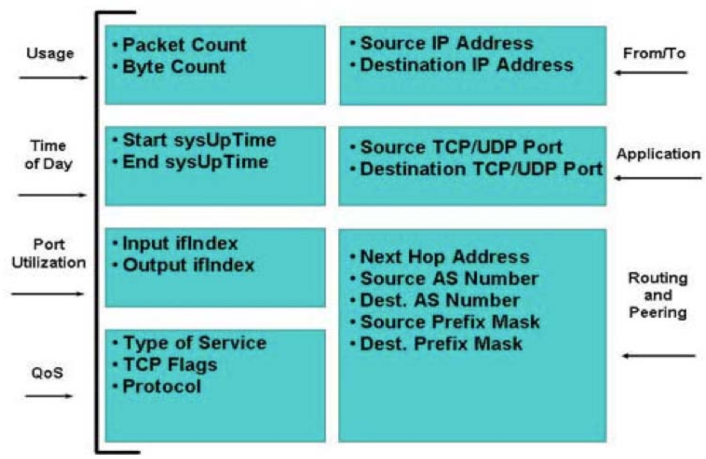
Scalability

- 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

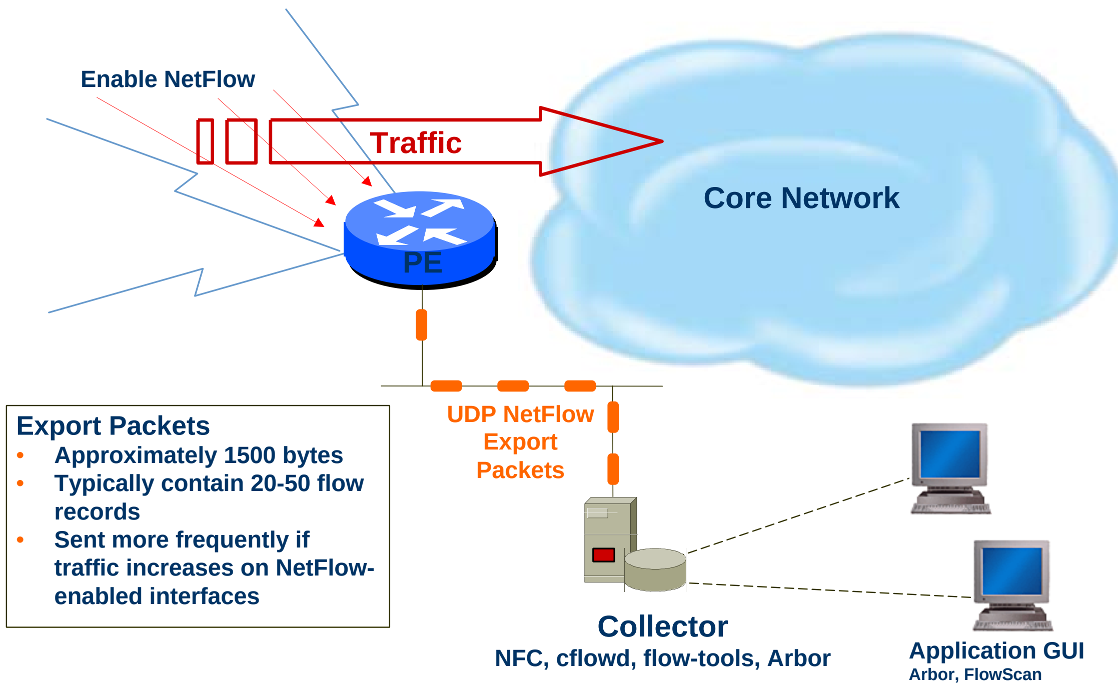


Case Study: Flow Based Abstractions

- 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:



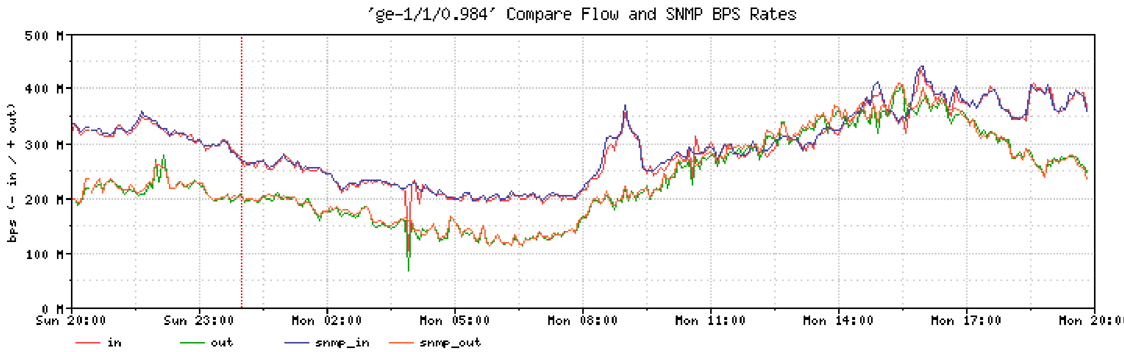
Flow Export



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

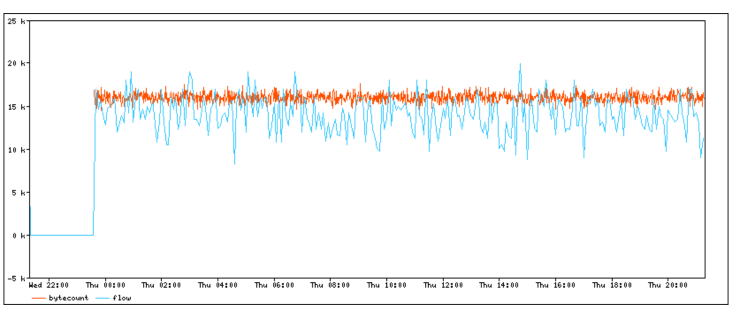
Flow Accuracy versus SNMP Octets



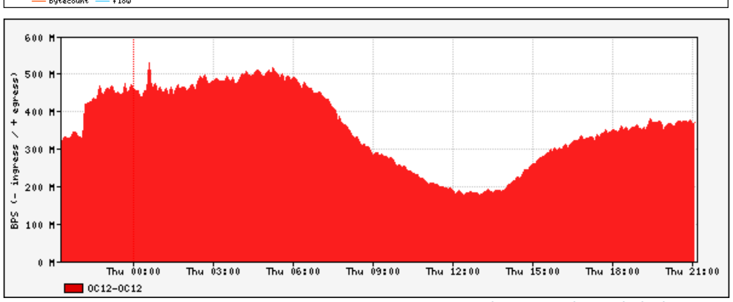
- Flow does not include link layer header (only IP)
- Ground truth sometimes difficult -- significant implementation issues and inaccuracies with SNMP

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

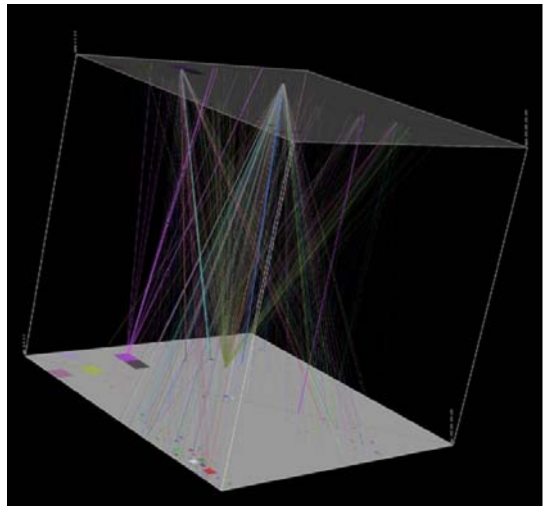


- SNMP ifOctet count for interface during measurement period

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
- …

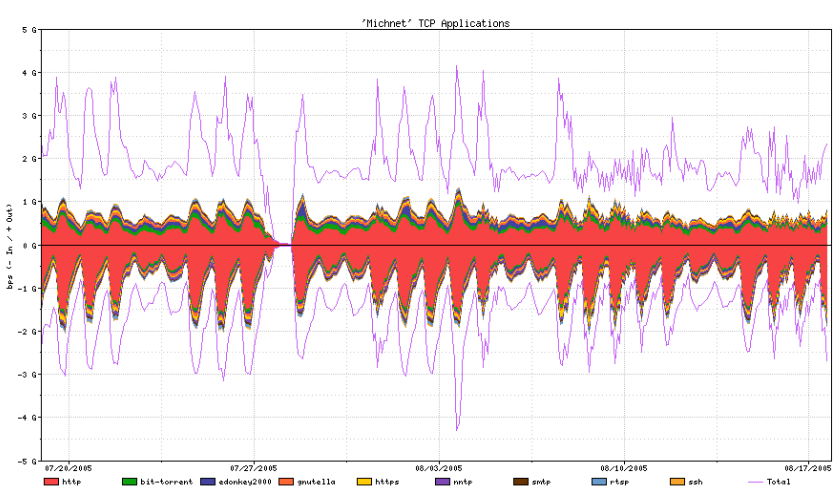
[Patwari, Hero, and Pacholski, MineNet 2005.]
[Lakhina, Crovella, and Diot, SIGCOMM'05]
[Xu, Chandrashekar, and Zhang, MineNet 2005.]

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.



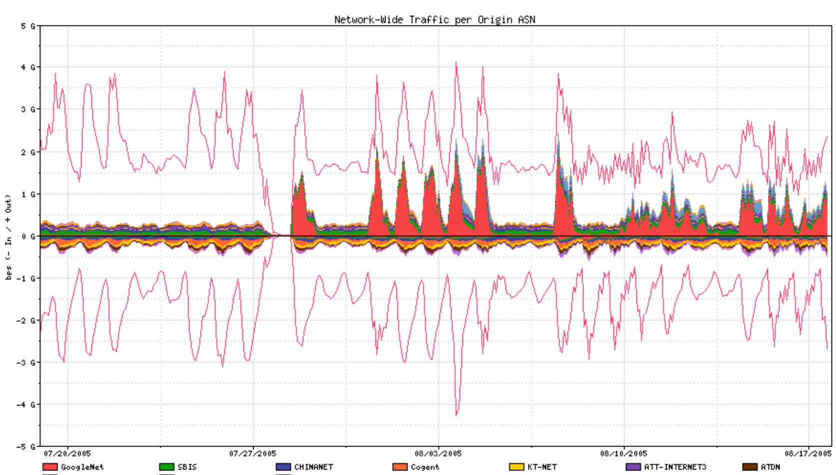
1. HTTP
2. Bit-torrent
3. Edonkey
4. Gnutella
5. HTTPS
6. NNTP
7. SMTP
8. RTSP
9. SSH

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

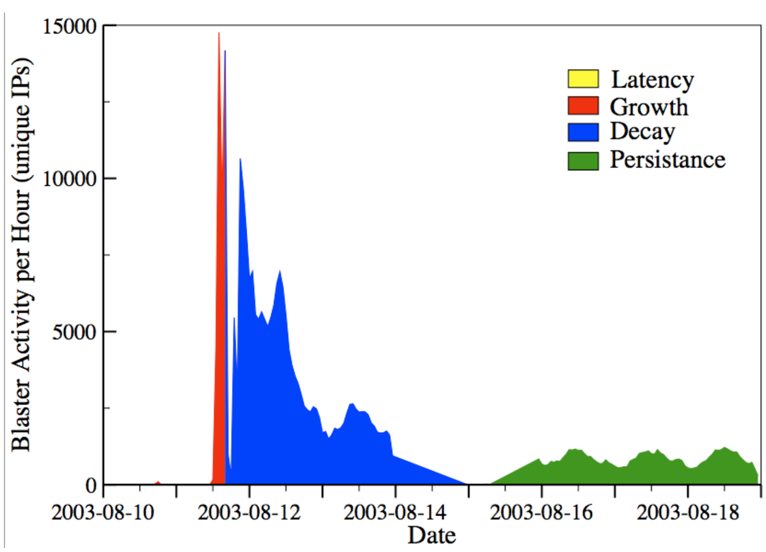


Privacy

- 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”
Bailey et. al. IEEE Privacy & Security, July/August 2005, Vol. 3, No. 4, pp. 26-31.

Worm Growth

- David Moore, Colleen Shannon, Geoffrey M. Voelker, Stefan Savage. Internet Quarantine: Requirements for Containing Self-Propagating Code. IEEE INFOCOM 2003
- The authors describe worm growth using a classic SI epidemic model.

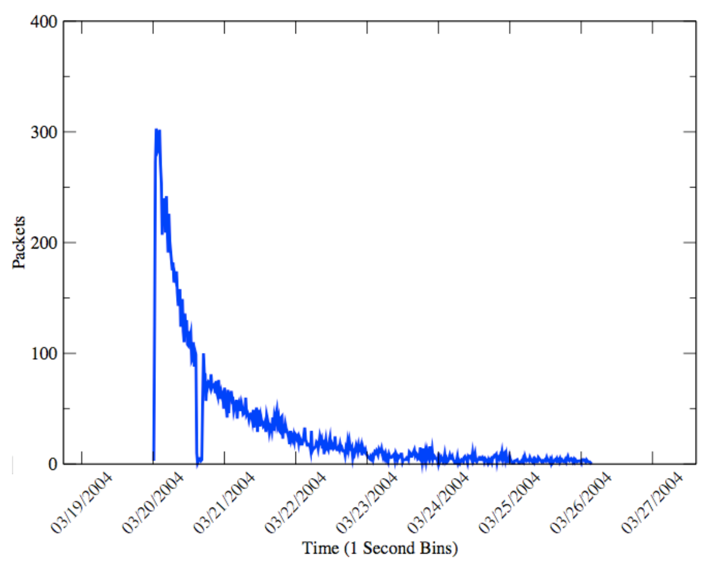
N	size of the total vulnerable population
$S(t)$	susceptibles at time t
$I(t)$	infectives at time t
β	contact rate
$s(t)$	susceptibles ($S(t)$) / population (N) at time t
$i(t)$	infectives ($I(t)$) / population(N) at time t

$$\frac{dI}{dt} = \beta\frac{IS}{N}$$

$$\frac{dS}{dt} = -\beta\frac{IS}{N}$$

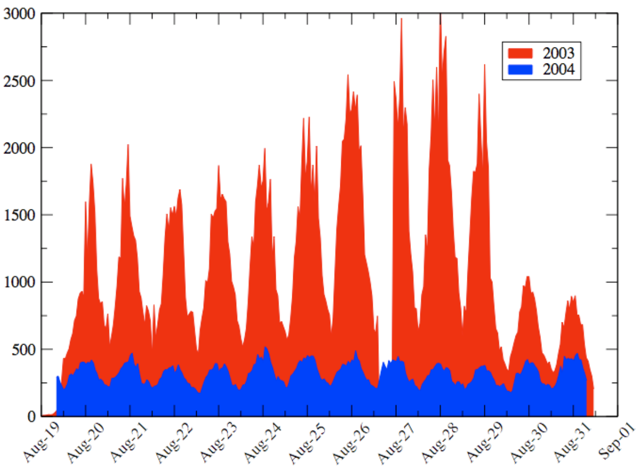
Worm Decay

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



Worm Persistence

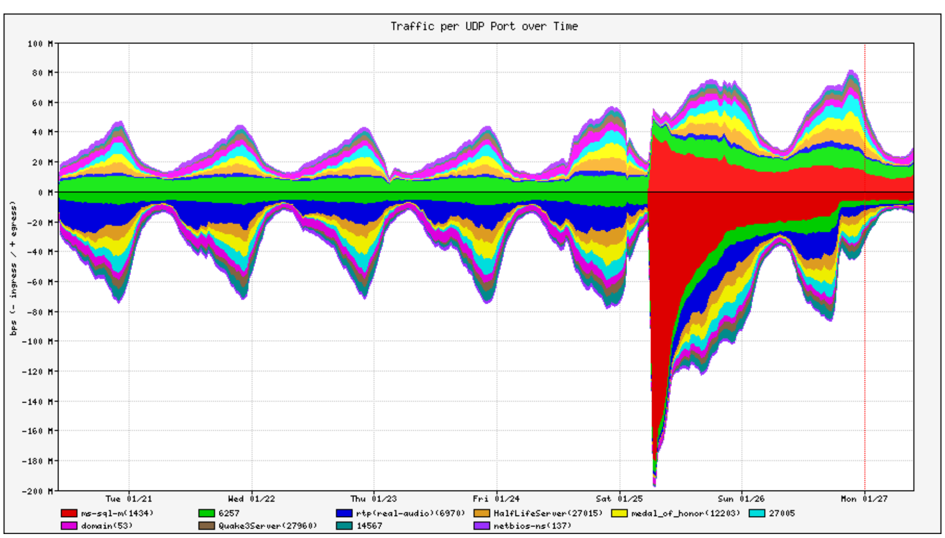
- Decay is a slight misnomer as, unlike witty, most worms persist even years after their release.
- Most persistent populations follow a circadian pattern.



- 45 -

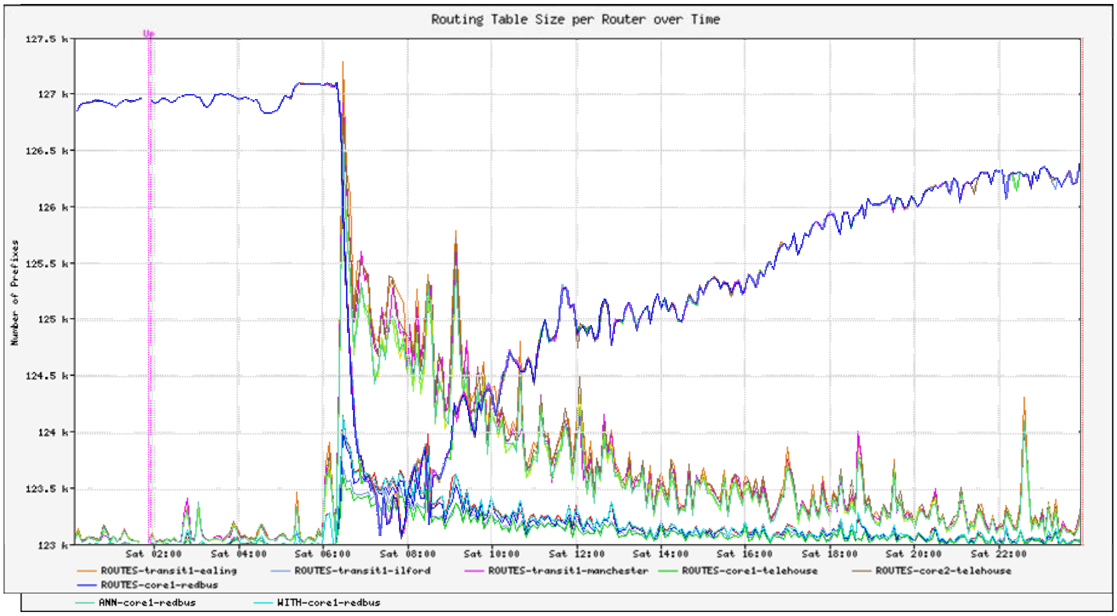
Slammer Data Plane Impact – A European SPs View

- Some DDOS/worms easier to detect than others…



- 46 -

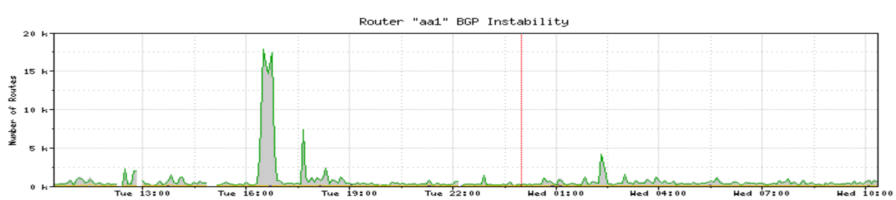
Slammer Control Plane Impact – Temporal Correlation



Root-Cause Analysis?

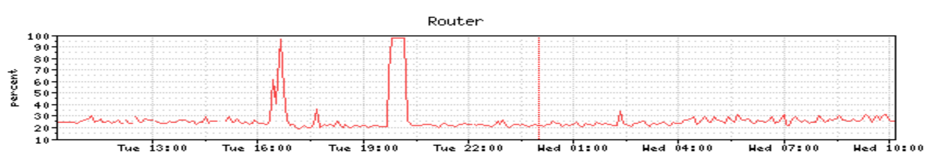
Event Correlation: Is this a DoS attack?

BGP Flaps



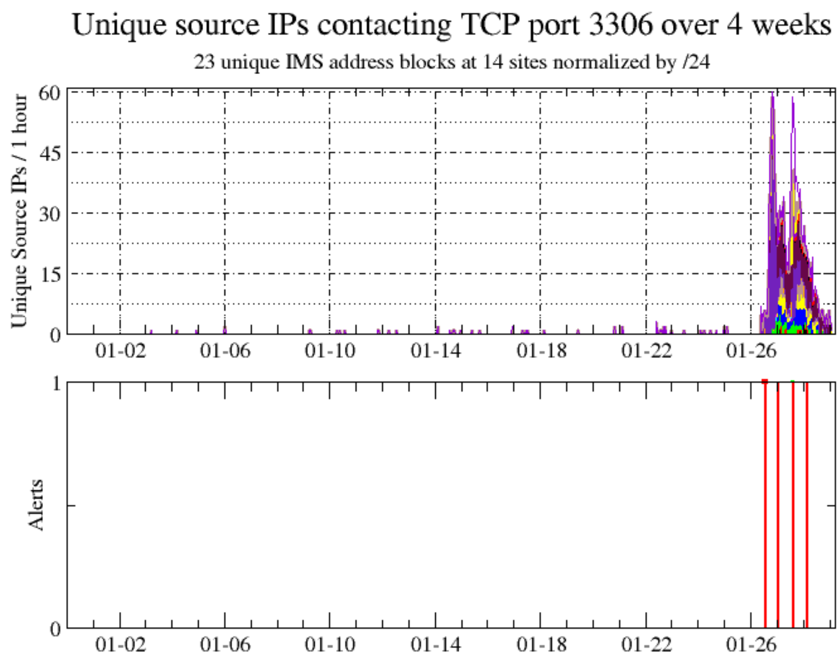
Turns out that configuration/policy change lead to high CPU,
which lead to BGP dropping,
which lead to path change and new traffic on router.

CPU

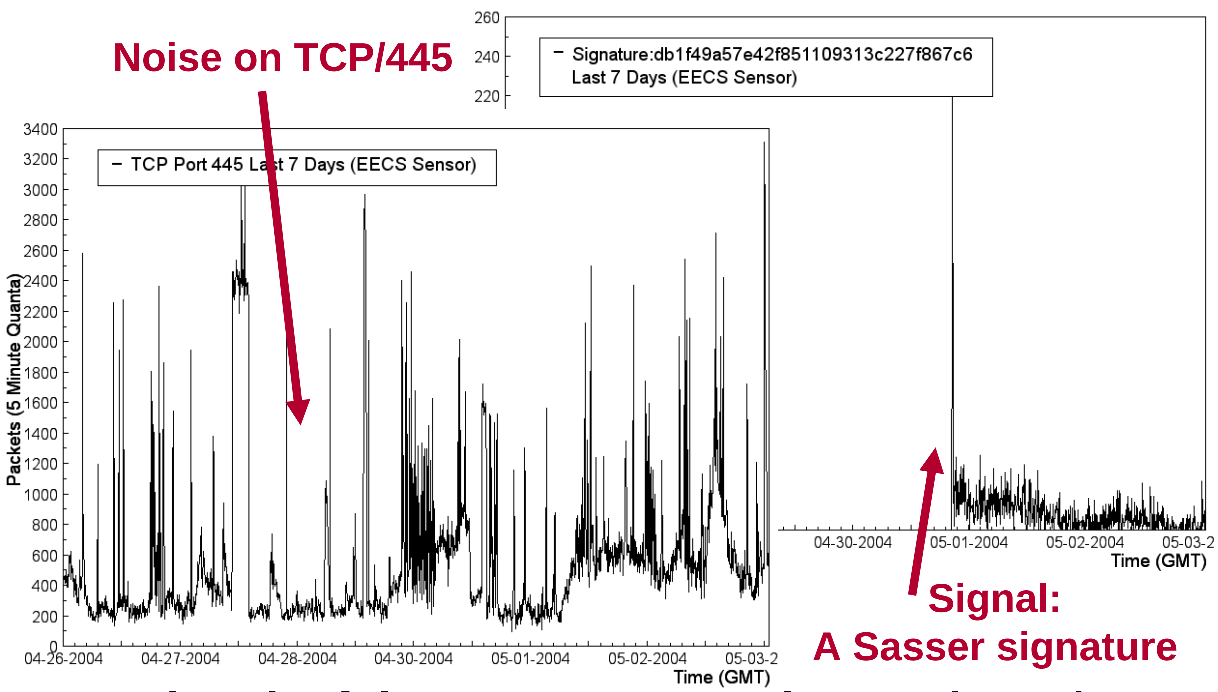


Root-Cause Analysis?

Easy Anomaly Detection: MySQL Worm



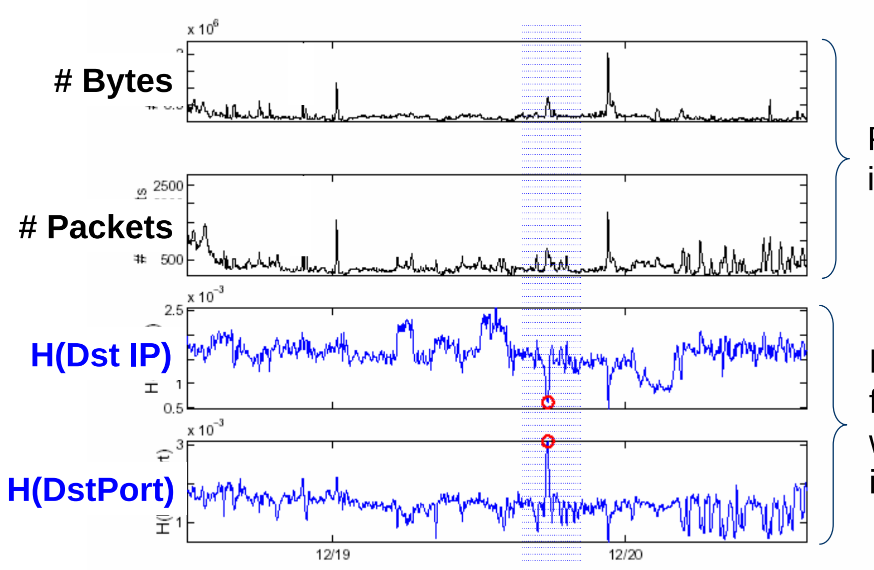
Harder Anomaly Detection: Sasser Worm



Outbreak of the Sasser worm observed at a /24

Feature Entropy Timeseries

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



Port scan dwarfed in volume metrics…

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

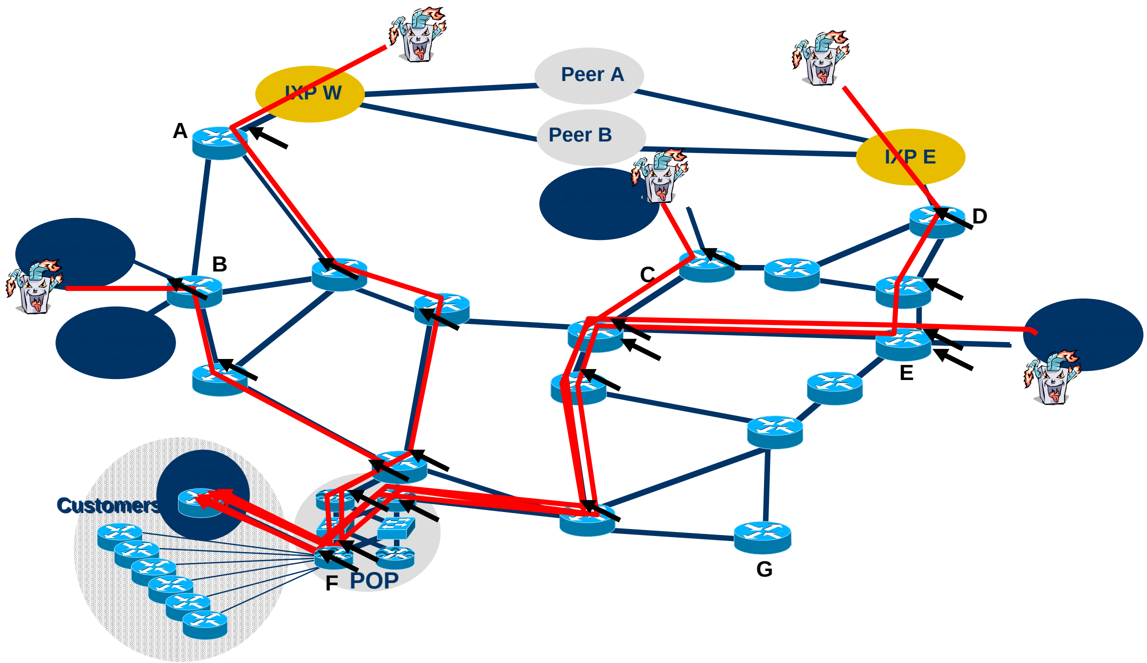
"Mining Anomalies Using Traffic Feature Distributions" by Lakhina, Crovella, and Diot, SIGCOMM 2005.

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