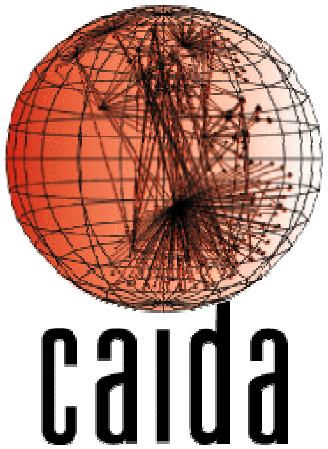
Anomaly Sampling
(bringing diversity to network security)

David Moore



CAIDA
University of California, San Diego

Flocon – October, 2006

Ok, let's get it out of our system

- "Sampling" by itself is a general term, like "aggregation".
  - Sampling: ≈ {things of type X} → smaller {things of type X}.
  - Aggregation: ≈ {things of type X} → smaller {things of type Y}.
  - Both:
    - Turn "too much stuff" into "hopefully enough stuff" to solve problems that *you* care about.
    - Useful at multiple stages of data collection, data management and analysis. Hierarchical approaches are very nice.
- Note: "rotating pcap files, keeping the last 3 days" is **sampling**, with algorithm: sample all packets less than 3 days old.

Ok, let's get it out of our system

- Sampling is not always keep/discard.
  - Sampled items could be given higher priority.
  - Sampled items might be separately kept to allow more efficient initial analyst queries against the dataset. (Same with aggregation)
  - Sampled items may be sufficient for a variety of basic reports which would save processing, etc..

- So…, in this presentation, "sampling" has no direct relationship to the ongoing packet / flow sampling argument ^H^H^H^H^H^H^H^H discussion.
  - You could imagine using this on a stream of packets, or a stream of flows, or syslog entries. Please, imagine that.

Anomaly Sampling

(bringing diversity to network security)

Flagging

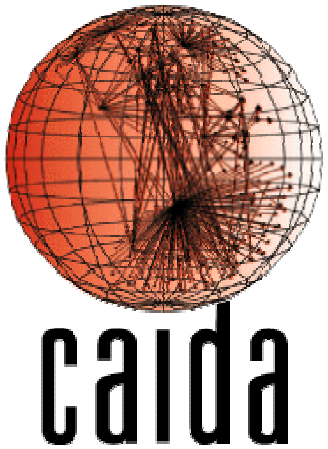
Prioritization

Potential Initial Query

Optimizer

Detection

David Moore



CAIDA
University of California, San Diego

Flocon – October, 2006

Basic Idea

- Existing systems focus on accurate counting of packets (or bytes) for large traffic aggregates
  - e.g., Smart Sampling, Traffic Summaries, Adaptive NetFlow, …

- Instead, focus on **interesting, *new*** information

Why? – Operational Network Security

- Forensics – “Bad guy did something”
  - When did they do it?
  - How did they do it?
  - What other machines did they get?
- Detection
  - Host ABC unexpectedly responded to a probe
  - Host XYZ used a service it never did previously

Living on the Network Edge

- The problem is **ours**, not our customer's.
- We care about **all** of the hosts.
- But each as an **individual**.
  - Some hosts are naturally more important.
  - Each host has its own services, risks, users, threats to other resources, ….
- We care about **small** events, not affecting performance.
- The problem remains **after** the "event" is over.
- Monitored network bandwidths are still high.

Basic Idea

- Existing systems focus on accurate counting of packets (or bytes) for large traffic aggregates
  - e.g., Smart Sampling, Traffic Summaries, Adaptive NetFlow, …
- Instead, focus on **interesting, *new*** information

What is "interesting" and "new"?

- Imagine you are the poor recipient of collected network data. What do you see?
  - Here's a record about our web server. Oh, and here's another record about our web server. And our mail gateway. Oh, here's another packet about our web server, ….
- Please, tell me something ***I don't know***
  - Tell me what is "abnormal", "unusual" or "new".
  - Tell me "just enough" about **everything**.
  - Do not prioritize telling me redundant information.
- These change over time.

System Components

- Diversity Score Assigner
  - Module assigns vague, relative rankings to items (packets, flows, …) based on how similar/different this item is to previously seen items.
  - Many different approaches for this, but there appear to be a decent set of them which cover a wide range of uses when given some parameters.
  - Some approaches are very efficient in memory or CPU requirements.
- Sampling Rate Adjustor
  - The scores produced above are based on the data stream without any knowledge of the desired sampling rate.
  - Variety of algorithms to dynamically keep effective sampling rate near the target sampling rate, while maintaining diversity score information.

Feature Spaces

- Operator chooses sets of fields/etc. over which they want coverage: (e.g.)
  - Source IP address
  - Destination IP address
  - Source & destination IP address pair
  - Protocol, source port, destination port
  - Src. IP addr., protocol, src. port
  - Src. IP addr., protocol, dst. Port
  - …
- Might chose weights to specify relative importance

Controlled Experiment

- Packet trace of live traffic in and out of central computing and network operations building (at university).
- Trace happens to contain some centralized nessus and nmap scanning from network operations. Plus main campus web servers, mail servers, desktops, ….
- Inserted an IRC exchange between 1 server and 3 clients
  - B → X, X → B (message, TCP ACK)
  - X → A, X → B, X → C, X → D (message broadcast)
  - A → X, B → X, C → X, D → X (TCP ACKs)
  - 10 packets for entire exchange, 8 unidirectional flows, 4 bidirectional flows.

Experiment Results

Target Rate	1 / 10
Filter	None
Scheme	Random
Saw at least 1 of IRC test	64%
Avg. # of test packets seen	0.92

Experiment Results

Target Rate	1 / 10	1 / 10
Filter	None	Discard top 85% of traffic
Scheme	Random	Random
Saw at least 1 of IRC test	64%	100%
Avg. # of test packets seen	0.92	7.73

Experiment Results

Target Rate	1 / 10	1 / 10	1 / 1000
Filter	None	Discard top 85% of traffic	None
Scheme	Random	Random	Diversity Counting
Saw at least 1 of IRC test	64%	100%	100%
Avg. # of test packets seen	0.92	7.73	4.66

Experiment Results

Target Rate	1 / 10	1 / 10	1 / 1000	1 / 5000
Filter	None	Discard top 85% of traffic	None	None
Scheme	Random	Random	Diversity Counting	Diversity Counting
Saw at least 1 of IRC test	64%	100%	100%	100%
Avg. # of test packets seen	0.92	7.73	4.66	1.26

Conclusions

- Anomaly detection is radically different for security at the edge compared with performance inside an ISP.

- *Appropriate* sampling techniques can:
  - greatly reduce the amount of data to look at (either by human or software)
  - focus attention on new, interesting events
  - provide good coverage for first-pass forensics analysis

- This approach can be applied to many streams of data: packets, flows, syslog, web logs, …

Internet Measurement Data Catalog http://imdc.datcat.org/

- To facilitate searching for and sharing of data
  - Index as much as possible, including datasets not publicly available
  - DatCat doesn't store any network data itself
- To enhance documentation of datasets via public annotations
  - Easy place for anyone (not just the dataset creator) to provide additional information
- To advance network science by promoting reproducibility
  - Paper X ran their detection algorithm on dataset X and had a false positive rate of 0.2. Using our algorithm on dataset Y, we get a false positive rate of 0.1. Therefore our algorithm is better. …
  - Persistent handles to allow for consistent citing and comparison:

http://imdc.datcat.org/collection/1-003M-5=AOL+500k+User+Session+Collection

DatCat

Internet Measurement Data Catalog http://imdc.datcat.org/