High Level Flow Correlation

Valentino Crespi, California State Los Angeles, CA
Annarita Giani, UC Berkeley, CA
Rajiv Raghunarayan, Cisco Systems, Inc.

FloCon 2008, Savannah GA, January 7-10, 2008.

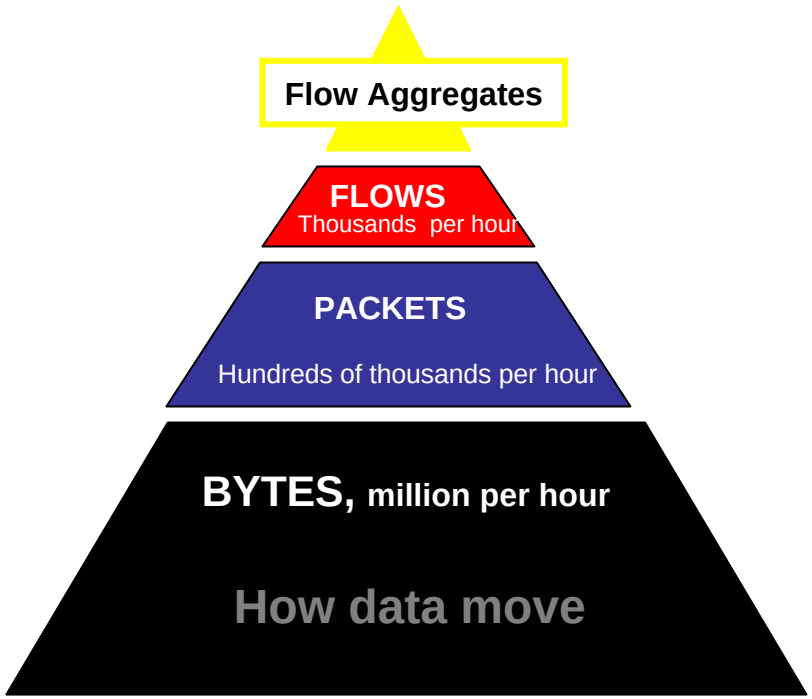
Outline

1. Extension of previous work on Flow Aggregation, (Flocon 2006).
2. Embedding of network traffic in an Euclidian Space.
3. Complex modeling through clustering.
4. Planned work.

Outline

1. Extension of previous work on Flow Aggregation, (Flocon 2006).
2. Embedding of network traffic in an Euclidian Space.
3. Complex modeling through clustering.
4. Planned work.

Behind Flow Aggregation



- Monitoring
- Anomaly detection
- Security analysis
- Traffic profiling
- Debugging
- Traffic engineering
- Usage-based profiling
- Network planning
- Pricing, peering

Data Reduction = Fewer events to be analyzed

Our Previous Work

A. Giani, I. De Souza, V. Berk, G. Cybenko, " Attribution and Aggregation of Network Flows for Security Analysis ," *in Proc. Flocon 2006,* Portland, OR.

We believe that **automated correlation at the raw flow level** is complicated and susceptible to false positives. The world consists of **processes** so our approach to correlation is process-based..

Flow aggregation and correlations between flow data with security events

Implementation of a **PQS based process detection for Cyber Situational Awareness.**

Flow + Snort Alerts

Scenario: several packets in a flow triggered IDS alerts

Snort rule **1560** generates an alert when an attempt is made to exploit a known vulnerability in a web server or a web application.

Snort rule **1852** generates an alert when an attempt is made to access the 'robots.txt' file directly.

Timestamp	Sensor	src IP	dst IP	Proto
Jul 09 16:28:32	S1852	65.54.188.140	208.253.154.195	TCP
Jul 09 16:29:35	S1852	65.54.188.140	208.253.154.195	TCP
Jul 09 16:44:44	S1560	65.54.188.140	208.253.154.195	TCP
Jul 09 18:26:08	S1560	65.54.188.140	208.253.154.195	TCP
Jul 09 21:05:03	S1852	65.54.188.140	208.253.154.195	TCP
Jul 09 22:31:08	S1852	65.54.188.140	208.253.154.195	TCP
Jul 09 22:31:08	S1560	65.54.188.140	208.253.154.195	TCP
Jul 10 02:45:19	S1852	65.54.188.140	208.253.154.195	TCP
Jul 10 02:45:23	S1852	65.54.188.140	208.253.154.195	TCP
Jul 10 09:21:15	S1852	65.54.188.140	208.253.154.195	TCP
Jul 10 14:33:43	S1852	65.54.188.140	208.253.154.195	TCP
Jul 10 17:54:54	S1852	65.54.188.140	208.253.154.195	TCP
Jul 10 22:07:02	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 01:38:09	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 04:05:54	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 04:20:00	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 04:20:00	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 11:07:12	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 11:56:12	S1852	65.54.188.140	208.253.154.195	TCP
Jul 11 17:16:59	S1852	65.54.188.140	208.253.154.195	TCP
S Jul 10 02:30:27	F	65.54.188.140	208.253.154.195	TCP
E Jul 10 23:55:56				

SNORT ALERTS

→ FLOW

Table 2: A sample track of correlated IDS and Flow events

The flow can be characterized as malicious and further investigation must be done.

Outline

1. Extension of previous work on Flow Aggregation, (Flocon 2006).
2. Embedding of network traffic in an Euclidian Space.
3. Complex modeling through clustering.
4. Planned work.

Current aggregators and analyzers

- **POWERFUL TOOLS** to understand the behavior of the network according to certain parameters, e.g. the amount of resources consumed, the variance on the various characteristics of the communication (source ip, destination ip), port.
- **PROBLEM**: They do not provide an analysis and a description of the dynamic evolution of network traffic.
- **NEED** for a structure that summarizes the behavior of the network.

OUR IDEA

Combine flow aggregation techniques with our previous process-based approach:

Use aggregators and flow analyzers to translate traffic into a process to be modeled and estimated.

Build circuits of Aggregating gates

1. Place observing nodes in multiple locations of the network (e.g. on each local router).
2. Each observing nodes dumps traffic flows to a Macro Aggregator (MA).
3. Macro Aggregator: *circuit*. Each gate is a flow aggregator
   - First layer consists of classical aggregators that output flow aggregates. Successive layers process aggregates of flow aggregates
   - Final output: a vector function of the dumped traffic ranging in $\mathbf{R}^n$:

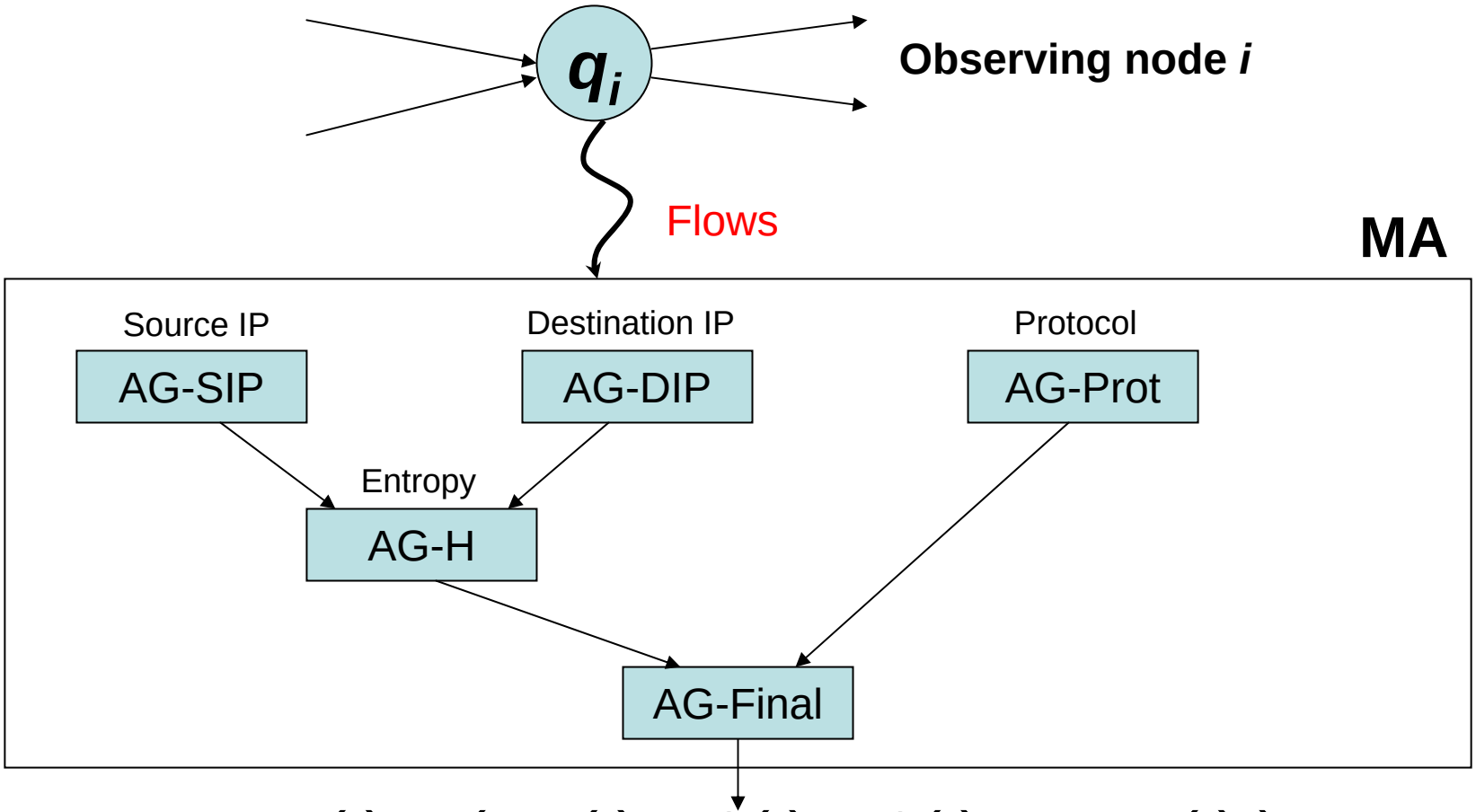
$$\mathbf{X}(t) = (x_1(t), x_2(t), \cdots, x_n(t))$$

   At each time the observing nodes produce a set of vectors:

$$\mathbf{S}(t) = (X_1(t), X_2(t), \cdots, X_n(t))$$

4. Identify and Analyze properties of S(t) over time to characterize/detect anomalies.

Embed Traffic in Euclidean Space



$$\mathbf{X}_i(t) = (x1(t), x2(t), x3(t),\ldots, xn(t))$$

(Entropy S-IP,Entropy D-IP, Average Size,…,%TCP Traffic,%UDP Traffic)

Entropy Based Flow Aggregation (2006)

Yan Hu, Dah-Ming Chiu, and John C.S. Lui
The Chinese University of Hong Kong

Based on Cisco's NetFlow – during flooding attacks the memory and network bandwidth consumed by flow records can increase beyond what is available.
A solution: Adapting sampling rate.

Flows of security attacks usually have common patterns and form conspicuous traffic clusters.
Identifies clusters of attacks flows in real time and aggregated those large number of short attack flows to a few meta flows.

Same sourceIP ~ worm propagation
Same destIP ~ Denial of Service Attack
Same destIP and SourceIP ~ most portscan

Purpose is mostly security.

On the correlation of Internet flow characteristics (2003)

Kun-Chan Lan, JOHN HEIDEMANN
Information Science Institute, University of Southern California

A small percentage of flows consume most of the network bandwidth.

Study of heavy flows in 4 orthogonal dimensions:

- Size
- Duration
- Rate
- Burstiness

and examine their correlations.

Strong correlation between size, rate, burstiness

Automatically Inferring Patterns of Resource Consumption in Network Traffic (2003)

Cristian Estan, Stefan Savage, George Varghese
University of California, San Diego

Method of traffic characterization that automatically groups traffic into minimal clusters of conspicuous consumption.

It is not a static analysis that captures flow characteristics but instead produces hybrid traffic definition that match the underline usage.

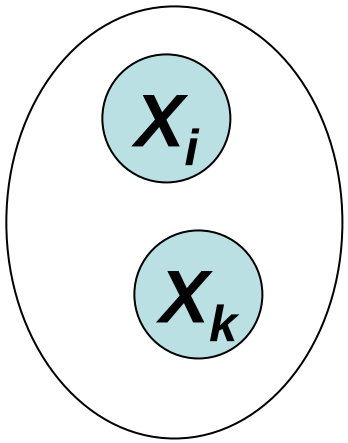
Purpose is mostly resource consumption.

Analyze *S(t)* over time

Approaches:

1. Use clustering techniques (e.g., spectral clustering, k-means based algorithms, etc.) to clusterize the observing nodes and infer correlations between observations and snapshots across the network.
    1. Study how clusters change over time and characterize/detect anomalies.
    2. Use clusters to produce a graphic representation of the traffic.
    3. Define discrete models to describe the evolution of clusters in relation to specific events: coordinated computer attacks, presence of covert channels, bugs in the network software, hardware breakdowns, etc.
2. Define State Space models.
3. Apply learning techniques to learn models.

Spectral Clustering

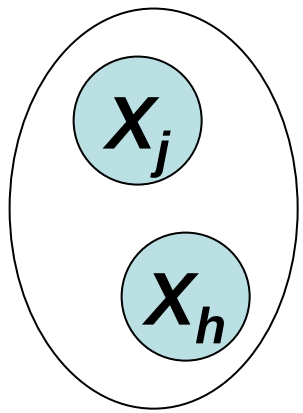


Input: Similarity Matrix M=[aij], , number k>0

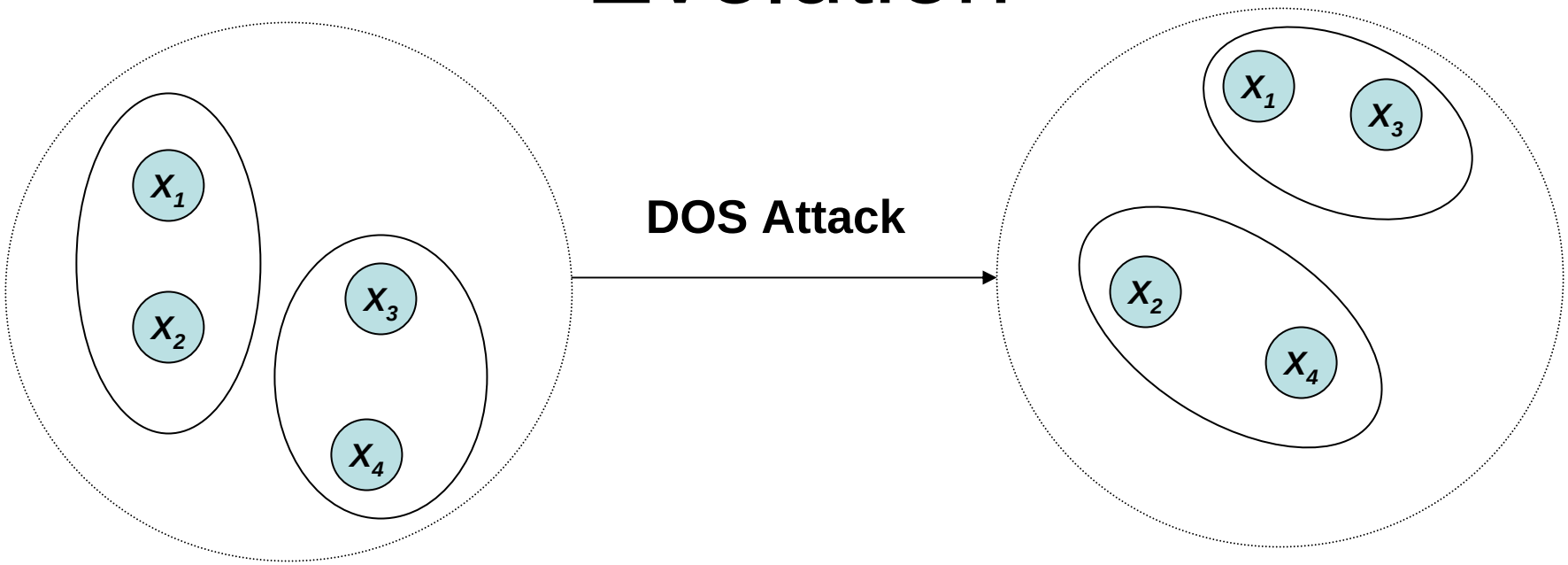
$a_{ij} = s(X_i, X_j)$ e.g $a_{ij} = \exp(-\|X_i - X_j\| / 2\sigma^2)$

- Build similarity graph. For example the Graph whose adjacency matrix AG = M.
- L = Laplacian(AG)
- Compute the k eigenvectors of L associated with the k smallest eigenvalues: v1, v2,…,vk
- V = [v1 v2 … vk], nxk matrix
- Pick the rows of V: y1, y2,…,yn
- Cluster yi's using k-means algorithm into C1,C2,…Ck

Output: clusters C1,C2,…,Ck



Discrete Models of Cluster Evolution



Idea: Build DFA models to identify transitions. In this case we identify anomalies by studying the current clustering in relation to the previous “snapshot” of traffic

Challenges

- Parameter estimation: in our example of clustering k was fixed.
- Apply Bayesian learning techniques to infer k.
- Apply *mixture models* technique to clustering
- Define and learn models of the system’s dynamics.
- Identify relevant attributes of flow aggregators to obtain significant vectors.
- Define appropriate similarity function.

Outline

1. Extension of previous work on Flow Aggregation, (Flocon 2006).
2. Embedding of network traffic in an Euclidian Space.
3. Complex modeling through clustering.
4. Planned work.

Planned Work

- Implement clustering method.
- Develop discrete models.
- Build a software monitor to analyze traffic through clusters and vector representation.
- Experimental analysis of the efficaciousness of our approach.

References

[1] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, Spectral Clustering and Normalized Cuts. In *Proceedings of the KDD'04 Workshop*, Seattle, Washington, August 2004.

[2] C. Estan, S. Savage, and G. Varghese. Automatically Inferring Patterns of Resource Consumption in Network Traffic. In *Proceedings of the 2004 SIGCOMM*.

[3] A. Giani, I. G. D. Souza, V. Berk, and G. Cybenko. Attribution and Aggregation of Network Flows for Security Analysis. In *Proceedings of FloCon 2006*.

[4] Y. Hu, D.-M. Chiu, and J. C. Lui. Adaptive Flow Aggregation - A New Solution for Robust Flow Monitoring under Security Attacks. In *Proceedings of 2006 IEEE/IFIP Network Operations and Management Symposium (IEEE/IFIP NOMS)*.

[5] Y. Hu, D. Chui, and J. C. Lui. Adaptive Flow Aggregation - a New Solution for Robust Flow Monitoring under Security Attacks. In *Proceedings of the 2006 Network Operations and Management Symposium*.

[6] K. Keys, D. Moore, and C. Esten. A Robust System for Accurate Real-time Summaries of Internet Traffic. In *Proceedings of the 2005 SIGMETRICS*, June 2005.

[7] L. Rodrigues and P. R. Guardieiro. A Spatial and Temporal Analysis of Internet Aggregate Traffic at the Flow Level. In *Proceedings of the 2004 Global Telecommunications Conference (GLOBECOM)*, volume 2.

[8] B. Trammell and C. Gates. Naf: The NetSA Aggregated Flow Tool Suite. In *Proceedings of the Large Installation System Administration Conference (LISA 2006)*, 2006.

[9] U. von Luxburg. A Tutorial on Spectral Clustering. Technical Report TR-149, Max-Planck-Institut für biologische Kybernetik, 2006.

Thanks

Annarita Giani <agiani@eecs.berkeley.edu>
Valentino Crespi <vcrespi@calstatela.edu>
Rajiv Raghunarayan <raraghun@cisco.com>