

Lightweight Hierarchical Network Traffic Clustering

Abdulrahman Hijazi, Hajime Inoue, Anil Somayaji

Carleton University

December 8, 2007

Problem Statement

- The complexity of current Internet applications makes the understanding of network traffic a challenging task. New Applications/Protocols/Attacks appear all the time.
- Current solutions have limitations:
 - 1 classifiers based on packet header information are fast but fail with unknown protocols and obfuscated traffic
 - 2 protocol dissectors are more accurate but are very slow
 - 3 machine learning past work identifies traffic as belonging to a small set of pre-defined classes

ADHIC: Our Complementary Solution

- ADHIC (Approximate Divisive Hierarchical Clustering) is a new real-time algorithm that clusters similar network traffic together without prior knowledge of protocol structures.
- Packet similarity is determined through comparisons of substrings within packets at distinguishing offsets.
- ADHIC:
 - 1 finds semantically interesting clusters and appropriately segregates well-known protocols,
 - 2 clusters together traffic of the same protocol running on multiple ports,
 - 3 segregates traffic from applications, such as p2p, that do not use standard ports, and
 - 4 adapts to changing nature of traffic patterns.

Why ADHIC?

- ADHIC is notable in that it
 - ① produces a hierarchical decomposition of network traffic in the form of a cluster-identifying decision tree,
 - ② does not assume prior knowledge of protocols and is unsupervised in every stage in operation,
 - ③ needs only a small fraction of packets (about 3% in our traces) to generate a decision tree, and
 - ④ can be used to cluster packets at wire speeds (250 Mbps in an unoptimized software implementation).

- NetADHICT, our implementation of ADHIC is available at:
<http://www.ccs1.carleton.ca/software>