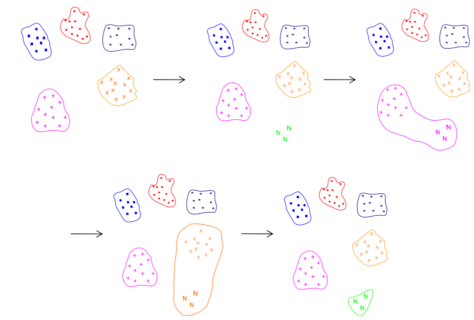
Flow-Based Traffic Clustering Using Unsupervised Machine Learning Techniques

Yanxin Zhang, Pushkar Patankar, George Kesidis, David Miller, Cetin Seren

It is important to detect the applications that create the underlying network traffic in order to keep the network running efficiently and to protect its resources. Machine learning techniques can use measurable statistical properties of traffic flows and make inferences about applications that generate them from these measurements. With the advent of recent applications that encrypt packet payload and scramble port numbers, traffic identification systems cannot rely solely on payload information and packet headers. Further, scalability issues prevent these identification systems from performing deep packet inspection of every packet and thus have to rely on session level aggregation.

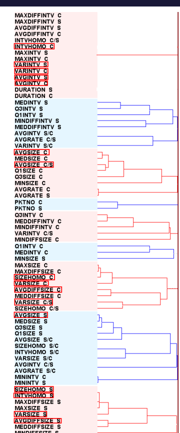
In this work, we look at the clusters the statistical properties of the flows form and observe what new clusters appear through time. The new flow clusters thus become indicators of new applications or threats, *i.e.,* traffic anomalies.



Unsupervised Feature Selection

Key to building a machine learning clustering algorithm is identification of the necessary set of features. We investigated the use of two separate purely unsupervised feature selection approaches (no class labels assumed), i.e., *Hierarchical Clustering Based Feature Selection* and *Iterative KNN Feature Reduction*.

We narrowed down the set of 68 candidate flow-level features to 16 significant features based on the comparison of results from two ML approaches and consideration of computational efficiency.



Feature Name
PktNo_C
PktNo_S
VarSize_C
VarSize_S/C
MinSize_C
MinSize_S
SizeHomo_C/S
SizeHomo_S/C
AvgDiffSize_S
MinDiffSize_C
MinDiffSize_S
VarIntv_C
MinIntv_C
MinIntv_S
IntvHomo_C
IntvHomo_S
IntvHomo_C/S
IntvHomo_S/C

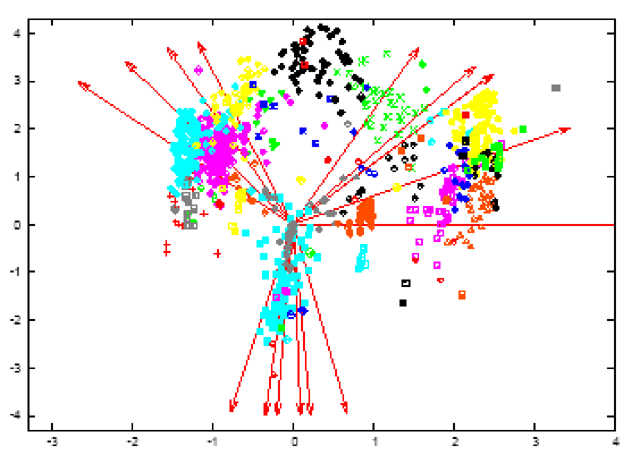
Feature Group	Feature Details
AvgSize	_C, _S, _C/S
VarSize	_C, _S, _C/S
AvgDiffSize	_C, _S
SizeHomo	_C, _S
AvgIntv	_C, _S
VarIntv	_C, _S
IntvHomo	_C, _S

FINAL SET OF 16 FEATURES

Clustering

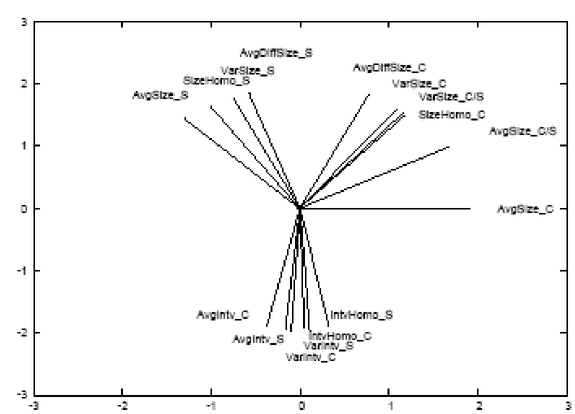
Support Vector Clustering (SVC) is a boundary detection-based clustering algorithm, that gives the unique optimal solution without strong assumption on cluster shapes or number. SVC model saves a small portion of samples to define the cluster boundaries and use them for new application discovery.

Hierarchical Clustering is an iterative clustering technique, that creates clustering solution at different scales or granularities. At higher granularity, the clustering technique shows the relations among the clusters while at the lower level the details within the clusters are revealed.



Visualization and Comparison

We evaluated both the clustering results using the WIDE Internet backbone data, which was a 15 minute IP header file captured in 2006. The test data contained 1845 bi-directional flows (≥ 100 packets). We used star coordinates to visualize clustering results in the 2-dimensional plane.



	A	B	C	D	E
1	100/100	0/0	0/0	0/0	0/0
2	0/0	51.57/99.66	27.97/98.15	0/0	0/0
3	0/0	0/0	0/0	55.76/99.14	0/0
4	0/0	0/0	0/0	0/0	54.54/95.45
5	0/0	0/0	0/0	0/0	0/0

CLUSTER AGREEMENT BETWEEN THE TWO CLUSTERING APPROACHES

This Work is Sponsored by Cisco Systems, Inc.