IST-070
Distributed Mining in Co-evolving Streaming Sensor Data

Christos Faloutsos
Professor, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA

Industry Participants:
Dr. Phil Gibbons
Intel Research Lab, Pittsburgh, PA

Abstract
The goal of this project is to provide tools for automatic discovery of outliers, patterns and correlations on the data streams in a distributed fashion (e.g. measurements from multiple autonomous sensor groups). For example, internet traffic, to and from CMU, can be modeled as multiple co-evolving data streams. Normal traffic patterns are different than the ones when viruses or worms are propagating.

Anomalies in co-evolving streams are useful in numerous settings: water quality sensor measurements, bridge vibration patterns, automobile traffic, to name a few. This problem is crucial to network performance, cyber-security, monitoring of civil infrastructure. The key requirements of the algorithm are:1) fast response (the virus propagation can paralyze the entire network in a few minutes; thus the early detection of the pattern shift is a must); 2) small resource consumption (any algorithm that works in the large network has to be efficient in order to be scalable); and 3) distributed operation (the centralized approach suffers from a single point of failure).

The applications for such a problem are numerous and vital: for example, environmental/weather applications, road traffic monitoring, volcano/earthquake monitoring applications, patient physiological monitoring, as well as the intrusion detection on the internet.

In prior PITA sponsored work, we examined the centralized case, where all sensors send their data to a central site for processing and correlation detection. Here we propose to have *multiple* such sites ('master sites'), each communicating with some of the sensors, and exchanging brief messages with the rest of the master sites. The goal is to have all master sites know which patterns and outliers exist globally, without having a single point of failure. More specifically, we expect to develop distributed on-line algorithms that can automatically identify the hidden variables over multiple streams and also detect the lag correlations among them. We want to investigate the on-line algorithms that work in distributed environments.