Invited Talk
Analyzing Massive Data Streams: Past, Present, and Future

Minos Garofalakis
Bells Labs
600 Mountain Avenue
Murray Hill, NJ 07974, USA
minos@research.bell-labs.com

ABSTRACT

Continuous data streams arise naturally, for example, in the installations of large telecom and Internet service providers where detailed usage information (Call-Detail-Records, SNMP-/RMON packet-flow data, etc.) from different parts of the underlying network needs to be continuously collected and analyzed for interesting trends. Such environments raise a critical need for effective stream-processing algorithms that can provide (typically, approximate) answers to data-analysis queries while utilizing only small space (to maintain concise stream synopses) and small processing time per stream item. In this talk, I will discuss the basic pseudo-random sketching mechanism for building stream synopses and our ongoing work that exploits sketch synopses to build an approximate SQL (multi) query processor. I will also describe our recent results on extending sketching to handle more complex forms of queries and streaming data (e.g., similarity joins over streams of XML trees), and try to identify some challenging open problems in the data-streaming area.