Seminar on Combinatorial Computing
March 5, Wednesday, 6:30 p.m.
Room 6417, Graduate Center
365 Fifth Avenue, New York

Massive data mining

James Abello
DIMACS, Rutgers University

Abstract

A variety of massive data sets exhibit an underlying structure that can be modeled as dynamic weighted multi-digraphs. Their sizes range from tens of gigabytes to petabytes. These include the World Wide Web, Internet Traffic and Telephone Call Detail. These data sets sheer volume brings with it a series of computational and visualization challenges due mainly to I/O and Screen Bottlenecks.

We present extremal memory algorithms for connectivity and minimum spanning trees together with heuristics for quasi-clique finding. We describe how hierarchy trees help us to cope in a unified manner with both, the I/O and screen bottlenecks. This line of research has suggested the need to look for "novel" graph representations in order to provide a user or data mining engine with informed navigation paths.

We will discuss our results with graphs having on the order of 200 million vertices and several billion edges and we will point out some mathematical problems that have surfaced along the way. The overall goal is to extract useful information that can be brought into a user's palm top and to export these techniques to other mining domains.

For further information contact Janos Pach at pachcims.nyu.edu, or visit our website: http://cs.gc.cuny.edu/∼pach/