Workshop on Algorithms for Data Streams

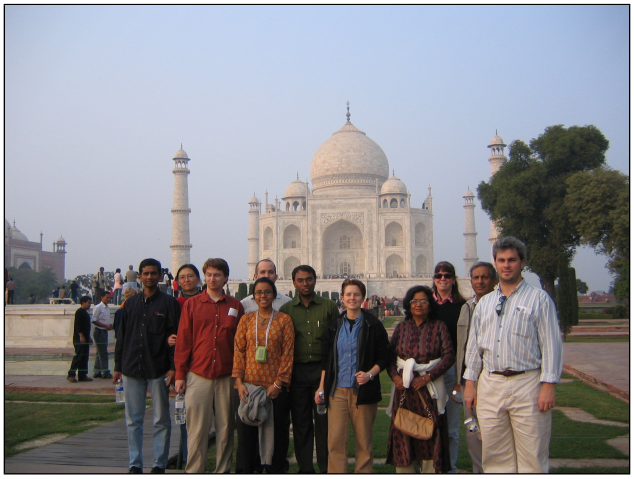
Department of Computer Science and Engineering IIT Kanpur

December 18 - 20, 2006

1 Introduction

A workshop on Algorithms for Data Streams was organized in the Department of Computer Sciences, IIT Kanpur from December 18-20, 2006. The workshop was sponsored by funds from the Research I foundation. The workshop on Algorithms for Data Streams was conceived along the broad charter of Research I foundation of providing impetus to cutting-edge research in active areas in Computer Science in order to boost the current research interests and productivity within the department. In view of current research activity in the area of Data Stream processing, and its wide-spread applicability in multiple disciplines, including, Database Systems, Networking, Data Mining and World-wide-web applications, it was decided to hold a workshop in this area.

Professors Manindra Agrawal and Sumit Ganguly coordinated the efforts for the workshop at IIT Kanpur. The workshop was organized along the lines of established Dagstuhl seminar series in Germany. Professors Sudipto Guha (University of Pennsylvania) and Professor S. Muthukrishnan (Rutgers University) were invited to function as co-organizers of the workshop. The goal was to bring to the IIT Kanpur campus as many of the world's top researchers in this area as we could. Simultaneously, we also wished to invite participation from many notable academic institutions of India, including, IISc, all the IITs, NITs, CMI, Math Sciences, and a few other institutes. In view of this, it was decided that the workshop would bear the cost of economy air-travel of all the international speakers, and bear 2AC train fare for faculty and 3AC train fare for students. Room and board of all guests were covered by the workshop. In view of the limited capacity of lecture halls and guest rooms, attendance of the workshop was restricted by invitation only. Students and faculty of IIT Kanpur were welcome to attend. A post-workshop one day bus-trip to Agra was also held for the international speakers.



2 Participation

We were very fortunate to have in attendance, almost all the top researchers in the area of data streaming. A complete list of the speakers and a brief summary of their talks are given later. There were 22 international speakers from US and Europe and two speakers from IIT Kanpur. [A speaker from IIT Delhi unfortunately could not make it.] There were about 80 delegates, both faculty and students, from colleges all over India, including, NIT Calicut, IIIT Hyderabad, IIIT Allahabad, Engineering College Bareilly, NIT Allahabad, NIT Jaipur, BITS Pilani, IIT Bombay, TIFR Bombay, College of Engineering Pune, College of Engineering Madurai, GSITS Indore. The international delegates were accommodated in the IIT Kanpur Visitors' Hostel and the Visiting Faculty Apartments. Delegates from India were accommodated in the rooms in Halls of Residence 5 and Girls' Hostel. We had a dedicated team of student volunteers, led by Vijaya Saradhi, PhD candidate, who coordinated the task of local arrangements, hospitality and travel reimbursements during the entire period of the workshop.

3 Technical summary

Data Streams processing pertains to very efficient monitoring of fast arriving data, such as network TCP/IP data, sensor data, etc., for anomalies (e.g., fingerprints of denial of service attack), patterns or user-programmed predicates for detecting high-level events. Approximation algorithms and randomized techniques have been invented in the last decade or so to address these problems. The field has gained impetus from a convergence of concerns in multiple application areas, including, approximate query processing and query estimation in Database Systems, telecom network monitoring, financial data monitoring, sensor networks, etc.

The technical topics were divided into several areas. In each area, there was one overview lecture of about an hour, followed by more in-depth talks by each speaker. The principal areas were

1. Data Streams: origins and algorithmic techniques,
2. Lower bounds: results and techniques,
3. Streaming Computational Geometry,
4. Data Stream processing for Networks, databases and the web,
5. Machine learning and streaming,
6. Databases and the Web,
7. Compressed Sensing, and
8. Graphs as Streams.

In addition, a very interesting session on open problems and directions was organized at the close of the second day of the 3-day workshop.

The workshop was inaugurated by Director of IIT Kanpur, Professor S.G. Dhande. An inspiring technical overview of the area was given by Professor Yossi Matias, TelAviv University and Google, Inc., one of the seminal contributors to the area and the recipient of the Gdel Prize for this work. In the area of algorithmic techniques for data streams, Prof. Amit Chakarbarti, Dartmouth University, USA spoke on estimating entropy, Prof. Sudipto Guha, University of Pennsylvania USA, on Order and Information; Prof. Rajeev Raman, University of Leicester UK on algorithms for the reset model; Prof. Srikanta Tirthapura, University of Iowa, USA on estimating the number of distinct elements in a range; Prof. Ravi Kumar, Yahoo! Research, USA, on estimating the number of distinct elements in a stream; and Prof. Sivakumar, Google Inc., USA on multi-pass sketching. In the area of Computational Geometry and Streams, Prof. Piotr Indyk, MIT presented an overview. Prof. Subhash Suri, University of California Santa Barbara, USA spoke on algorithmic techniques for low dimensional geometric streams; Prof. Pankaj Agarwal, Duke University, USA spoke on space-optimal algorithms for computing coreset over stream of two-dimensional points and Prof. Christian Sohler, University of Paderborn, Germany, presented his work on clustering geometric streams.

In the area of applications of data streams to Networking, Database Systems and the World-Wide-Web, Dr. Divesh Srivastava from AT&T Research, USA spoke about challenges in practical processing of streaming data at network line speeds. Prof. S. Muthukrishnan from Rutgers University and Google discussed the applications of stream processing in WWW searching, indexing and clustering. The area of lower bounds for problems in the streaming model is a fascinating subject as it helps prove the minimum computing resources required to solve problems. In this area, we were given an overview talk by Dr. T.S. Jayram, from IBM Research, California, USA. Prof. Nicole Schweikardt, Humboldt-University, Berlin presented her work on lower bounds for query processing on streaming data and external memory data. Dr. Andrew McGregor, University of California, San Diego, USA presented a talk on the intersection of machine learning and data streams.

Graph algorithms over streaming edges have been an active area of research. The overview talk in this area was given by Prof. Sampath Kannan, University of Pennsylvania, Philadelphia, USA. Dr. Surender Baswana gave a lecture on algorithms for computing approximate spanners over streaming graphs. The area of processing matrix data for numerical and non-numerical analysis via data reduction techniques (e.g., sampling) has gathered momentum in recent times. A wonderful lecture on this topic was given by Prof. Ravi Kannan, Yale University, USA. Dr. Michael Mahoney, Yahoo! Research, USA, presented a lecture on Sampling Algorithms and Coresets for L_p regression and applications. Finally, the less structured session on discussions and open problems presented a very fascinating series of presentations, by Piotr Indyk from MIT, Sudipto Guha from University of Pennsylvania, and Sumit Ganguly from IIT Kanpur, Pankaj Mehra from HP Labs, D. Sivakumar from Google Inc., Yossi Matias from Google Inc., Sampath Kannan from University of Pennsylvania along with comments by others.

The workshop ended with a vote of thanks by the organizers to the sponsor, Mr. Narayan Murthy, to the IIT Kanpur student volunteers who helped facilitate the event, and to all the speakers and participants.

4 Workshop archive

A workshop of this quality is best attended multiple times, using the authors slides and lectures to better understand the area, and to obtain further insights. With this view, the proceedings of the workshop were videotaped (courtesy the Media-Television Center at IIT Kanpur) and the digital video is available at the site (`http://www.cse.iitk.ac.in/users/sganguly/revised-schedule.html`). The PowerPoint and other slides used by the speakers are also available at this site.