198:671 Processing Massive Data Sets

S. Muthu Muthukrishnan

Graham Cormode
Details

- Meeting: Core B, Thursday 6—8 PM.
- Graham: x4580, Core 413, Office:

- We meet
  - [1] 01/30
  - [4] 02/06  02/13  02/20  02/27
  - [3] 03/06  03/13  **03/20**  03/27
  - [4] 04/03  04/10  04/17  04/24
  - [1] 05/01  05/08?

- Send us your email addresses today.
More Details

- Course is about massive data sets. What? Where? How?
- HW1: Guess a few data sets that are likely to be large, and determine an estimate of their sizes. What is the largest dataset size you can think of?
- Homeworks should be submitted in latex and ps to muthu and graham by email.
Details....

• Areas:
  – Databases, Networking, Sensors, Hardware, Programming Lang.

• Core problems:
More ....

- 01/30, 02/06: Intro to Streaming.
- 02/13  02/20  02/27: Estimating Distinct Elements, Hot items, Overview of clustering.
- Remainder:
  - Students will make presentations on specific areas of their interest.
  - Involves reading 3—6 papers.
  - Making a ppt based presentation.
  - Writing a survey of the topic, including questions that arise during the discussions. Due at the end of the term. In LaTex.
  - Do it in teams?
Assignments

- Data Stream Management Systems (Wei Zhang)
- Decision trees on data streams. (Yihua Wu)
- Approximation theory results (James Jones).
- Pseudo random variable constructions (Lan Yu)

Statistics:
- Sensors:
Puzzle: Find missing numbers.

- **Paul** permutes numbers $1 \ldots n$, and shows all but one to Carole, in the permuted order, one after the other.
- **Carole** must find the missing number.

Carole can not remember all the numbers she has been shown.
Carole finds the missing number…

- Carole **cumulates** the sum of all the numbers she is being shown. At the end, she can subtract this sum from the total sum of the numbers 1..n.
  - Uses $O(\log n)$ bits to store the partial sum.
  - Performs one $+$ each time Paul shows a number. Takes $O(\log n)$ time per number.
  - At the end, computes the missing number with on subtraction. Takes $O(\log n)$ time for final computation.
Finding two missing numbers…

- What if Paul shows all but two numbers?
- Carole keeps the sum AND product of the numbers Paul shows her.

\[ O(n \log n) \] bits and time.

- Alternatively, Carole keeps the sum AND sum of squares of the numbers Paul shows her.

As before: \[ O(\log n) \] storage, \[ O(\log n) \] process time and \[ O(\log n) \] compute time.
Missing numbers....

- HW2: What is the best algorithm you can design for the problem of finding k missing numbers?
Paul and Carole Games...

• Playing twenty questions (Spencer and Winkler)
  – Paul (for Paul Erdos) asks the questions.
  – Carole (anagram for Oracle) gives answers.

• Pusher, Chooser Games.

• Here, Paul permutes, Carole cumulates.
Data Streams: Motivations and Models
The Data Stream Phenomenon

• Highly detailed, automatic, rapid data feeds.
  – Radar: meteorological observations.
  – Satellite: geodetics, radiation,…
  – Astronomical surveys: optical, IR, radio,…
  – Internet: traffic logs, user queries, email, financial,
  – Sensor nodes: many more “observation points”.

• Need for near-real time analysis of data feeds.
  – Detect outliers, extreme events, fraud, intrusion,
    anomalous activity, complex correlations,
    classification,…
  – Monitoring.
Data Stream Phenomenon

- **Data Stream**: Massive data feeds with very fast updates. The phenomenon is real and recent.

- Normally: *Anecdotics, Numerics,…*

- Let me leave with a mental image:

  We can (and intend to) collect so much data that we have to drop large portions of it to manage it within our space and time resources. This is a new kind of uncertainty. It should jar us one way or the other.
Network management calls for rapid analysis of **MASSIVE** amounts of such data, in particular, summarizing various signals.

- Call detail record,
- ss7 signaling,
- diagnostics, ...

Sample data:

- (202 262 47yx, 800 call att, 01/02/02, 14.12.21, 14.35.00)
- (973 360 7212, 202 262 47yx, 01/02/02, 14.36.00, 14.38.00)
- 111.12.111, 121.25.211, 01/02/02, 14.12.21, 14.35.00, 12412, 100)
- 212.78.123, 121.25.311, 01/02/02, 14.12.21, 14.35.01, 24, 1)
Some queries on network traffic: Fishing in large domains.

- How many distinct IP addresses currently use or used anytime during the day, a given link?
- What are the top $k$ voluminous flows currently in progress in a link?
- How many flows consisted of only on packet?
- How much current traffic pattern is common between two routers?
- What are top $k$ correlated link pairs?

Online statistics
- Paul’s missing numbers.

Rarity
- IPSOFACCTO

Signal analysis: Wavelets, Fourier, etc.
Homework

• HW3: List a few queries you may pose to packet traffic streams.