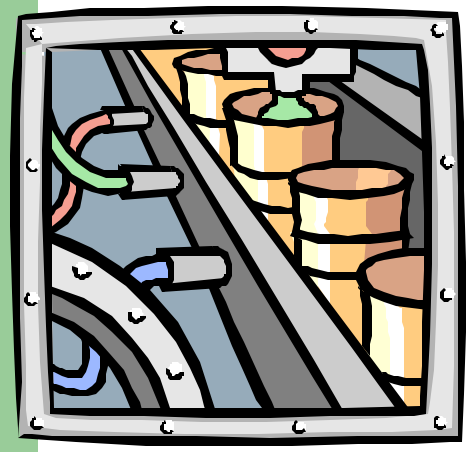
Mining for Patterns and Anomalies in Data Streams

Sampath Kannan
University of Pennsylvania

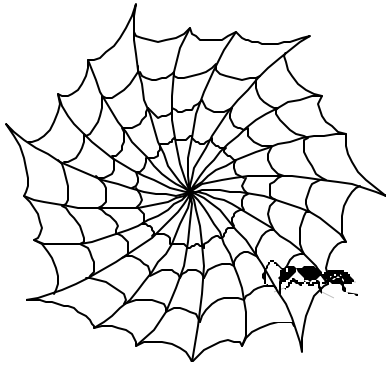
The Problem

- Data sizes too large to fit in primary memory
- Devices with small memory
- Access times to secondary memory are too long for data to be processed in real time

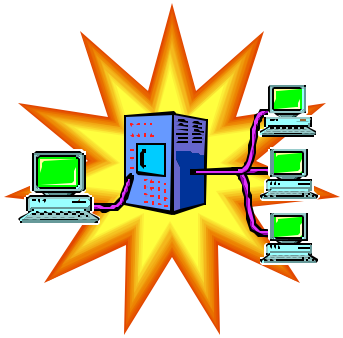
Example scenarios



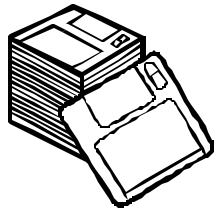
Sensor networks



Web crawlers



Network routers



Databases

Model 1: External Memory Models

- Design and analyze algorithms in terms of
  - Primary memory size
  - Disk block size
  - Page size
  - Number of disk reads and writes
- Efficiency requires new algorithm designs.
- Problems:
  - Algorithms may not be real time.
  - May not have secondary storage.

Sample application: Network of Sensors and Actuators

- Increasingly important in environmental and military applications.
- Sensors monitor physical entities.
- Voluminous monitored data needs to be analyzed in stream fashion.
- Commands to actuators generated based on analysis.

Sensor networks --- cont'd

- Data Characteristics
  - Varying rates of data arrival
  - Many data sources transmitting data
  - Large volume
  - Data produced continuously ... infinite stream, but only "recent" data is relevant.
- Constraints
  - Query and analysis on demand as data arrive
  - Actuators need to process commands

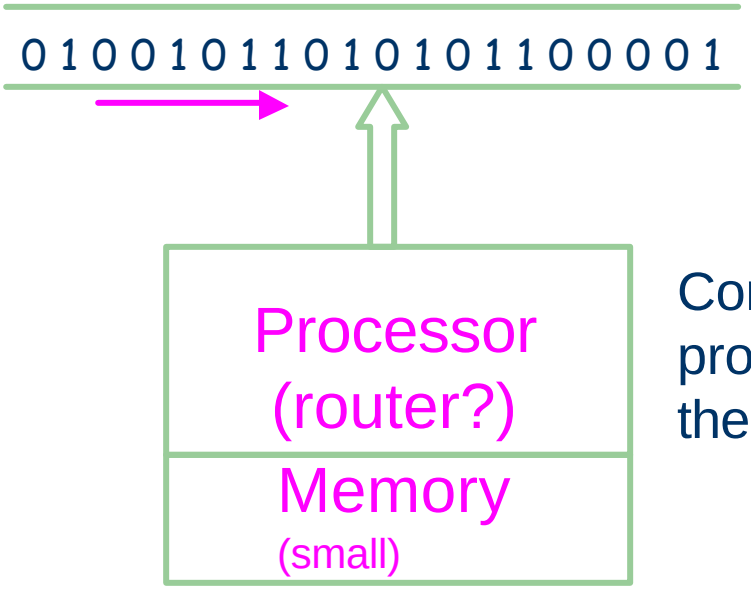
The Size of a Data Stream

- If the data stream is finite: n, the number of items in the stream.
- If the data stream is infinite?

Infinite amount of data does not make computational sense: we will be interested in a window of values which arrived in the last n time steps.

Windowed data stream model.

Model 2: Streaming Algorithms



Compute interesting properties of the data.

- Data processed in real time.
- Need to use randomized strategies and settle for approximate answers

Example 1

Traffic data from successive days:

(source,destination,bytes,day)

... (A,B,20K,1),(A,C,64K,1),(C,D,56K,1) ...

... (A,C,32K,2), (A,D,48K,2), (B,C,10K,2) ...

Are traffic patterns anomalous from one day to the next?

Stream of successive readings from a large number of sensors in a sensor network.

Has there been a “big” change overall?

Example 2

Stream of SYN and ACK packets flowing through a router:

How many SYN packets without corresponding ACKs in last 10,000 packets?

Want to figure out the answer with much fewer than 10,000 words of memory.

This is a question in the sliding window model.

Lower Bounds

For exact computation many simple tasks need lots of space. For example, computing number of 1's in sliding window of size N requires N bits of storage.

Such bounds are proved by using theorems proving lower bounds on communication complexity.

Need to allow approximate and randomized computation.

Algorithms-I

Median: Given stream of *n* integers compute median.

Theorem (Munro & Paterson):

With $O(n^{1/p})$ space, we need p passes and we can do it in p passes.

Frequency moments: Given sequence $s_1, s_2, \ldots, s_n$ where each $s_i \in [1..m]$, let a_i denote number of $i's$ in sequence.

kth frequency moment: $\sum_{i=1}^{m} a_i^k$

Theorem (Alon, Matias, Szegedy) Can approximate the 0^{th} through 6^{th} moments using log n space. However for higher moments the space required is nearly n.

(Input of this form is called cash-register input.)

Algorithms--II

Given two streams representing two vectors compute the norm $(L^1, L^2, \ldots)$ of their differences.

(Feigenbaum, Kannan, Strauss, Viswanathan) show how to do this with polylog space. (Indyk) extends result to cash-register model.

Above results solve the problem posed in Example 1.

Algorithms-III

Database applications:

- Computing histograms --- Given a function represented by a stream approximate it by piecewise constant functions. (Gilbert, Guha, Indyk, Kotidis, Koudas, Muthukrishnan, Srivastava)

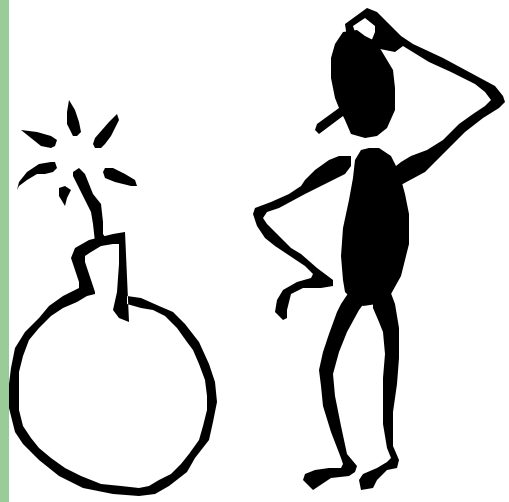
- Quantile summaries --- Compute a summary or sketch of stream of data to answer quantile queries such as `find an element that is in the 30th percentile'. (Greenwald, Khanna)

Algorithms--IV

Transforms: Given a function as a stream, compute the Fourier Transform. Compute a wavelet transform. (Gilbert, Guha, Indyk, Muthukrishnan,Strauss)

Clustering: Given a stream of points in multidimensional space, find a small number of cluster centers that are nearly optimal. (Callaghan, Guha, Meyerson, Mishra, Motwani).

Methods of Attack



1. Sample : Restrict input
2. Compress computation tree
3. Embed
4. All of the above …

A Sampling Approach

- Consider finding the median
- Sample $s = O(d^{-2} \log^2 n)$ values
- Sort and return the median of S
- Error is $\pm dn$ with high prob.
- Uses Hoeffding's Inequality
- [MRL 98, 99]

States of Computation

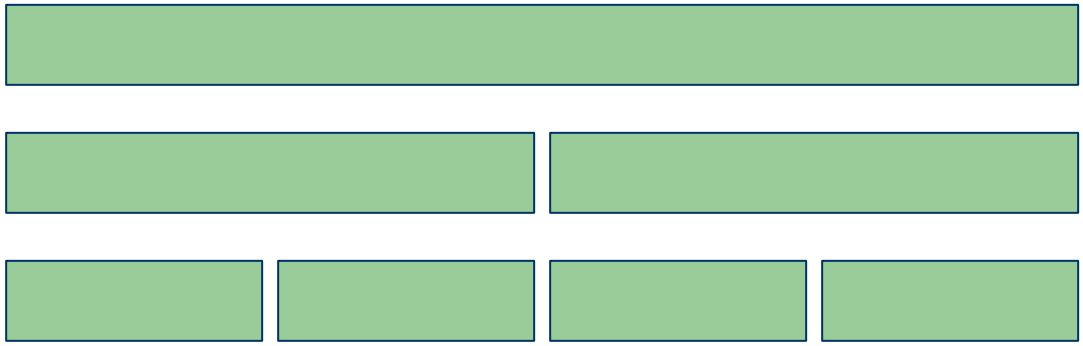
- Consider any offline computation
- Can we store intermediate results in succinct form ?

Targets

1. Divide and Conquer
2. Dynamic Program

Divide and Conquer

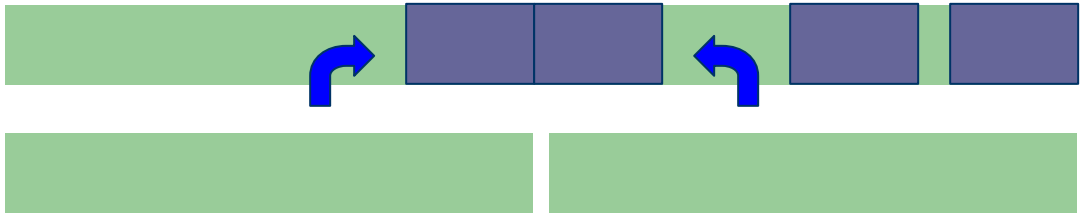
- (GMMO ’00)
- Consider the following clustering algorithm



R-ary Tree. At each node cluster into K clusters and send to parent

Basic Building Block

- Consider a two level process



- Prove that combined new problem has solution close to original cost
- Find (approximate) it

The Dynamic Program

- Store the table in compressed form
- Approximate entries to indicate only the large changes

- For new element, search is reduced since the table is small

Adding points one at a time

- Incremental Algorithms in small space

  [Vitter] Reservoir sampling

  [CCFM '97] Incremental K-center

  [GK 99] Selection in space $O(\frac{1}{d}\log n)$

- Linear Partitioning problems [GK 01, 02]
- Data Structure Questions as well …

Embeddings

Basically Dimensionality reduction

To compute f

- Reduce dimension of input to fit in the memory space available.
- Operate in new space to compute an appropriate function g.
- Lift g back to get f' close to f

Linear Embeddings

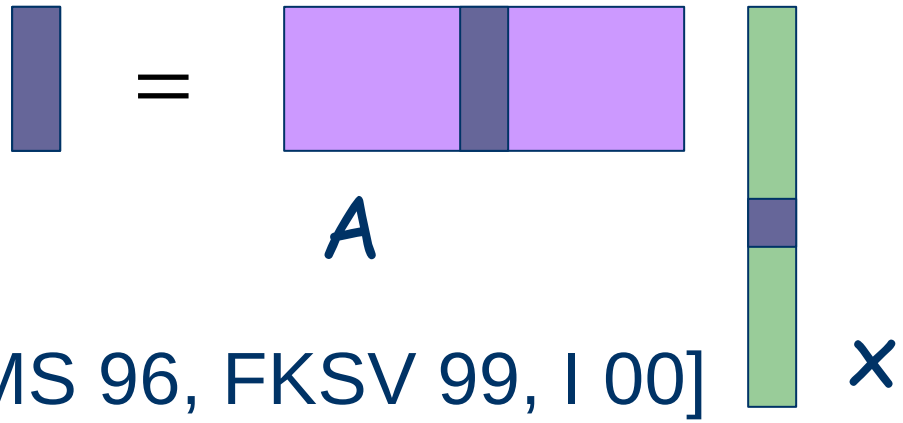
- [JL Lemma] $\|x\|_2 \leq \|Ax\|_2 \leq (1+e)\|x\|_2$
- A is a Random $(e^{-2}\log n)\times n$ Matrix drawn from Gaussian distribution.
- Too many elements in matrix!

Use Pseudorandom Generators [I 00]

P-Stable distribution for ℓ_p **where p** $\in$ **[0,2]**

What it achieves

- Computes Norm when x_i arrive out of order.



- [AMS 96, FKSV 99, I 00]

Our current work

Can we make the data talk?

Often we don’t know what exactly we are looking for, but can tell if an answer is significant...

Example: Intrusion detection --- can enumerate each type of intrusion we can think about and check if the stream of packets at a packet sniffer constitutes this kind of intrusion... but what about the types of intrusions we haven’t listed?

Data Mining

We are doing data mining. What is it?

Searching for answers to questions we cannot formulate?

Looking for statistically significant patterns amidst a background of noise or other patterns?

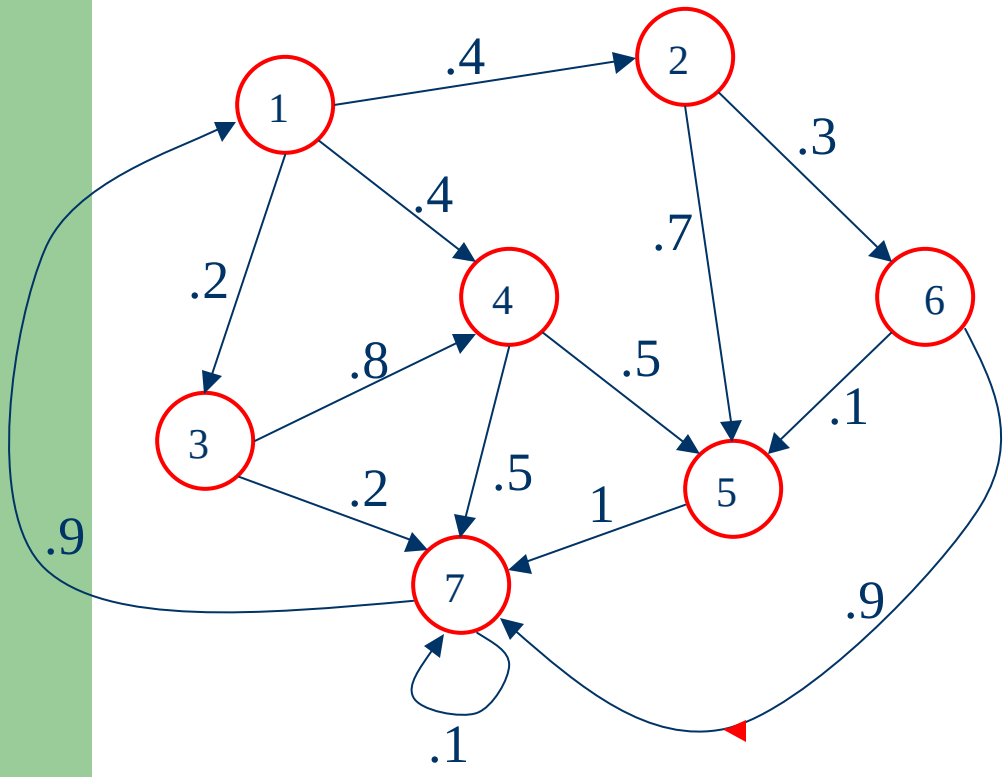
Statistics is key --- otherwise we have no basis for judging significance.

Rough Model Formulation

There are several processes... and we are seeing the interleaved output of these processes. (For intrusion detection some of the processes could represent honest use and others represent intrusion.)

How to model each process? Remember statistics is key. But processes we deal with do have memory. So, natural model is Markov Chains.

Markov Chains



Output sequence:
1 4 7 7 1 2 5 7 ...

Some Basic Facts

- Ergodic Markov chains have a stationary distribution.
- Random walks on *n* node graphs are "within" 2^{-k} from stationary distribution in $O(kn^3)$ steps.
- For a random walk, stationary probability of a vertex is proportional to its degree.

Our Problem

MC1

... 1 3 2 5 1 4

...2 6 1 3 2 7 5 3 1 4 1

MC2

... 2 6 7 3 1

Observe ...2 6 1 3 2 7 5 3 1 4 1 ...

Infer: MC1 & MC2

For our problem we assume:

- Stream is polynomially long in the number of states of each Markov chain (need perhaps $O(n^6)$ long stream).
- "Mixture" probabilities are bounded away from 0.
- Space available is some small polynomial in #states --- possibly $O(n^2)$.
- Assume we have a mixture of two Markov Chains, although results generalize to more.

Our Results

- For Markov chains on disjoint state sets, we can infer a mixture of an arbitrary number of them under a very general interleaving process.

- For Markov chains on overlapping state sets, we can currently deal with inferring a mixture of 2 chains under a technical condition...

Examples

- Computational biology
  - Identifying exons and introns
  - binding sites and other motifs
- Internet intrusion Detection
  - Intruder traffic may have different statistical properties from regular traffic.

Conclusions

Streaming continues to be a source of exciting problems. They are here to stay.

With the interest from the database community and the networks community many streaming algorithms will not remain purely on paper... they will be implemented, empirically tested and improved.

We are planning to implement some of our recent work.