The Big Data Algorithmic Toolkit

Moses Charikar

Stanford University
My Background

• 1991-1995  B.Tech IIT-Bombay
• 1995-2000  Stanford PhD
• 2000-2001  Google Research
• 2001-2015  Princeton
• 2015-      Stanford
My background

• Approximation Algorithms
  – heuristics with provable guarantees for hard optimization problems
  – mathematical programming techniques
    linear programs, etc

• Algorithmic techniques for Massive Data Sets
  – dimension reduction
  – similarity search
Scenarios

• Google search log
Scenarios

- Facebook social network
Scenarios

• Sequencing human genome

• What discoveries can come from more sequencing?
Theorists’ Methodology

• Strip away application specific details
• Focus on core problem

• What can we do?

• What corner cutting needed?

• What is impossible?
Sketching

- Replace original data by compact summary (sketch)
- Problems of interest can be solved using sketches.

Data Reduction

approximated

complex object

sketch

0101100110

0010110010
Streaming Algorithms

• Compute in one pass over data using limited storage
But isn’t this old hat?
Caveats

• Very high level

• A flavor of the questions and results
BASIC PRIMITIVES
Basic Statistics

• Distinct Elements

• Frequency Moments

\[ F_p = \sum f_i^p \]

\[ \sum_i (f_i - g_i)^p \]

• (1+ε) approximation in O(1/ε²) space
Connections

• Derandomization: hash functions

• Impossibility results
Document Similarity: MinHash

- [Broder, ‘97]
- [Broder, Manasse, Glassman, Zweig, ’97]
Locality Sensitive Hashing

- [Indyk, Motwani ‘98]
- Hash functions that distinguish between near and far points
- Building block for Nearest Neighbor Search
Vector representation

• Associate vector with each data point

• Distance between data points = angle between vectors

\[ \cos(\theta) = \frac{u \cdot v}{||u|| \cdot ||v||} \]

• Compact sketch for estimating angular distances?
Hash Function: SimHash

- Hash value is one bit
- Concatenate bits from many hash functions to get compact representation

\[ s(u) = \text{sign}(u \cdot r) \]

\[ \Pr[s(u) \neq s(v)] = \frac{\theta(u, v)}{\pi} \]
Frequent Items

- Most frequent queries in Google query log
Approximate Counting

• Hash table: counter for each bucket
• elements mapped with random signs

\[
\text{estimate} = \text{counter} \times \text{sign}
\]
• Noisy estimator

\[ \sqrt{\frac{\sum_{j>k} f_j^2}{k}} \]

• Many independent copies return median
GRAPH ALGORITHMS
Model

• Stream of edges on n nodes added/deleted
• Run algorithm using $O(n \log(n))$ space
Problems

• Connectivity
• Distances
• Clustering
• Dense subgraphs
• Partitioning
• Random walks
• Matching
Tools

- Spanners
- Sparsifiers
- Sketches
LINEAR ALGEBRA
• Least squares regression
• matrix multiplication
• polynomial fitting
• low rank approximation
• graph sparsification
SAMPLING
Sequencing Human Genome

- What discoveries can come from more sequencing?
Course

• cs369g.stanford.edu