6.5240 Sub-linear Time Algorithms

Prof. Ronitt Rubinfeld
TA: Shyan Akmal
Course Administrator: Joanne Hanley
What is this course about?
Big data?
Really Big data

Impossible to access all of it
Small world phenomenon

Social network graph:

• each “node” is a person

• “edge” between people that know each other
Connectivity properties

• “connected” if every pair can reach each other

• “distance” between two nodes is the minimum number of edges to reach one from another

• “diameter” is the maximum distance between any pair
Small world property

“Six degrees of separation”

In our language:
diameter of the world population is 6
Does earth have the small world property?

• How can we know?
  ▪ data collection problem is immense
  ▪ unknown groups of people found on earth
  ▪ births/deaths

• Stanley Milgram’s 1963 experiment?
The Gold Standard

• linear time algorithms
  – Inadequate...
Approaches when input is too big to view?

• Ignore the problem

• Develop algorithms for dealing with such data
What can we hope to do without viewing most of the data?

- Can’t answer “for all” or “there exists” and other “exactly” type statements:
  - are all individuals connected by at most 6 degrees of separation?
  - exactly how many individuals on earth are left-handed?

- Maybe can answer?
  - is there a large group of individuals connected by at most 6 degrees of separation?
  - is the average pairwise distances of a graph roughly 6?
  - approximately how many individuals on earth are left-handed?
What can we hope to do without viewing most of the data?

• Must compromise:
  ▪ for most interesting problems: algorithm must give *approximate* answer

• we know we can answer *some* questions...
  ▪ e.g., sampling to approximate average, median values
Sublinear time models:

- **Random Access Queries**
  - Can access any word of input in one step
  - How is the input represented?

- **Samples**
  - Can get sample of a distribution in one step,
  - Alternatively, can only get random word of input in one step
    - When computing functions depending on frequencies of data elements
    - When data in random order
Isn’t this just

• Randomized algorithms
• Approximation algorithms
• Statistics
• Learning
• Communication complexity
• Parallel/distributed algorithms?
Course requirements

• Scribing: 25%
  ▪ Signup on web
  ▪ Must be in latex
  ▪ Draft 2 days after lecture

• Problem sets: 35%
• Project: 25%
• Class participation (includes grading): 15%
Course website

• https://people.csail.mit.edu/ronitt/COURSE/F22/
• Announcements
• Pointer to piazza site
• Lecture notes: Posted before lecture
• Homeworks: Check for updates and hints.
• Scribe and grading instructions
• Project ideas
• Probability review
Canvas

- Pset submissions and solutions
- Announcements (with email notification)
Piazza

Please:
help each other without giving too much information!
be nice to each other!

Caution: anonymous to class but NOT to staff
Project Possibilities

- Read a paper or two or three
  - Explain some lemmas
  - Suggest some open problems
  - Even better -- Make some progress on them, or at least explain what you tried and why it didn’t work

- Implement an algorithm or two or three

Can work in groups of 2-3
Plan for this lecture

• Introduce sublinear time algorithms

• Basic algorithms
  ▪ Estimating diameter of a point set
  ▪ Estimating the number of connected components of a graph
Scribe?
I. Classical Approximation Problems
First:

- A very simple example –
  - Deterministic
  - Approximate answer
  - And (of course).... sub-linear time!
Approximate the diameter of a point set

- Given: \( m \) points, described by a distance matrix \( D \), s.t.
  - \( D_{ij} \) is the distance from \( i \) to \( j \).
  - \( D \) satisfies triangle inequality and symmetry.

  (note: input size \( n = m^2 \))

- Let \( i, j \) be indices that maximize \( D_{ij} \) then \( D_{ij} \) is the diameter.

- Output: \( k,l \) such that \( D_{kl} \geq D_{ij}/2 \)

2-multiplicative approximation
Algorithm

- **Algorithm:**
  - Pick $k$ arbitrarily
  - Pick $l$ to maximize $D_{kl}$
  - Output $D_{kl}$

- **Running time?** $O(m) = O(n^{1/2})$

- **Why does it work?**
  
  $D_{ij} \leq D_{ik} + D_{kj}$ (triangle inequality)
  
  $\leq D_{kl} + D_{kl}$ (choice of $l$ + symmetry of $D$)
  
  $\leq 2D_{kl}$ (so $D_{kl}$ is at least diameter/2)