Welcome to CS590 Sublinear Algorithms

Admin.

- Bright space
- Campus Wine; class webs on my page.

- Grading: 4-5 HW to %.
  - Research project 45 %.
  - Scribe notes + peer grading HW 10 %.
  - Class participation 5 %.

- Need scribe for today.
  - Research: max 2 people.

Schedule:
  - See timeline:
    - by Feb 2 should have a topic
    - Feb 16 + 2 paras
    - March 23: 5-10 min present
    - April 29: outcome
    - April 30: written report

Meeting with me before then.
Motivation for sublin. algos:

**Big Data:**
- Internet of Things
- Sales transactions
- Web pages
- Health data
- Genomic data
- Space discovery data
- etc etc

Need algorithm design in \(O(N)\)

1. If data can be stored but no time to read it \(\rightarrow\) sublinear time algos
2. If data is too big to fit in memory \(\rightarrow\) sublinear in space
   - If can throw some away
   \(\rightarrow\) sublinear in communication
   - If can store it on multiple machines that communicate
Sublinear-time algs

- approx algs: e.g., diameter
  # of connected components
  avg degree of a graph

- property testing:
  does an obj have a property or
  is far from having the property?
  (e.g., is G connected or far-
   is G 3-colorable or far)

Model stems from Program checking in PL.

80's
Blum Kannan

80's
Blum Luby Rubinfeld
Formalized by Rubinfeld Sudan 90
Leads to PCP thm

Also many other local models
- Locally decodable / testable codes

- \( \text{poly}(N) \) vs. \( \text{poly} \log(N) \)

Type of research questions:
- Can membership in specific code be tested with \( \text{of} \) many queries?
- Can each entry be corrected with \( \text{whp} \)?
- Best tradeoffs between rate, distance, locality

Leads to a general notion of local computation algos.
Local Computation Algs (LCA-s) 
Rubinfeld, Tamir, Vardy, Xie, '11

queries \[ \uparrow \downarrow \uparrow \downarrow \]
LCA

probe \( \square \) \( \otimes \)

eg: Maximal IS.
probe node \( i \) to see if it is in
the MIS selected or not.

Based on distrib. algs where each node
makes local decision & goal is to have a
consistent MIS
Sublinear-space algo/ streaming

Given a seq of els appearing one at a time, with limited memory, can get eg: statistics about the stream:

- max, min, avg, median
- a random sample?
- estimates of # distinct els
- # heavy hitters?

OR can get an approximate size of:
- max matching
- vertex cover
- # connected components
- avg degree?

Defined by Alon, Matias, Szegedy '96
Sublinear in communication

\[ f(x, y) = x \oplus y \]

\[ f : \mathbb{F}_2 \times \mathbb{F}_2 \rightarrow \mathbb{F}_2 \]

2 bits

- deterministic
- randomized

Many natural problems are hard i.e. require almost full disclosure of inputs
Eq: \( \text{EQUALITY} \)

\( \mathcal{E}_0 \subseteq \mathbb{R} \times \mathbb{R} \)

\( \mathcal{E}_0 \times \mathcal{E}_0 \rightarrow \mathcal{E}_0 \)

\( \mathcal{E}_k (x, y) = \begin{cases} 0 & x = y \\ \text{off} & \text{otherwise} \end{cases} \)

\( \text{Def (Eq)} \geq N \)

\( \text{Indx} \times \mathcal{E}_0 \times \mathbb{N} \rightarrow \mathcal{E}_0 \)

\( \text{Def (Indx)} \geq N \)

\( \text{Rand (Indx)} \geq \Omega (N) \)

\( \text{Set DISJ} \)

\( \text{DISJ} : \mathcal{E}_0 \mathcal{R}_1 \times \mathbb{N} \rightarrow \mathbb{N} \)

\( \text{DISJ} (x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases} \)

\( \text{Rand (DISJ)} = \Omega (\mathcal{N}) \)

\( x, y \text{ are distinct sets of reals} \)
Obs: Data stream edge $\Rightarrow$ communication protocol.

So lb for communication $\Rightarrow$ lbs for streaming.

We'll probably see that lbs for communication $\Rightarrow$ lbs for property testing.

- Other egs: connectivity
  - Bipartiteness
  - Approx weight of spanning tree
- Recent: distributed learning of distributions
Some basic problems

- diameter of a set of pts in $\mathbb{R}^n$
- testing if fuc is constant
- uniform sampling of a stream
- deciding connectivity of a graph in a stream
Deterministic 2-approx of diameter

Def: $D$ dist. metric $\mathbb{R}^n$

1) $D(x, y) \geq 0$; $D(x, y) = 0$ iff $x = y$
2) $D(x, y) = D(y, x)$
3) $D(x, y) \leq D(x, z) + D(y, z)$

Def: $\text{diam}(S) = \max_{x, y \in S} D(x, y)$

Thm (Indyk): Given $S$, def. alg. that outputs $d$ s.t. $\frac{\text{diam}(S)}{2} \leq d \leq \text{diam}(S)$ in time $O(\lvert D \rvert)$

note: $\lvert \text{input} \rvert = O(\lvert D \rvert) = O(\lvert S \rvert^2)$
\[ \forall x, y \in \Omega \quad D(x, y) \leq D(x, t) + D(y, t) \leq 2 \cdot \max \{ D(x, t), D(y, t) \} \leq 2 \max \{ \{ D(r_t) \}_{r \in S} \} \]

Alg: Take any \( t \in S \)

Output \( d = \max \{ \{ D(r_t) \}_{r \in S} \} \)

\[ \text{Diam} \leq 2d \leq \text{diam}(S) \]

\[ \frac{3}{2} \quad \text{approx} \quad O(m n^{1/2}) \]

\[ \text{Diam} \leq d \leq \text{Diam} \]
Property testing

Property = a collection of objects that all have a particular property

\( P_n = \{ \text{all graphs on } n \text{-} \text{vts that are 3-colorable} \} \)

\( P_n = \{ \text{all graphs on } n \text{-} \text{vts that are connected} \} \)

\( P_n = \{ \text{all } \mathbb{F}_2 \text{-} \text{distrib. that are } b_1 \text{-} \text{mod } a \} \)
$\epsilon$-far

Hamming Distance: $f, g : \mathbb{D} \rightarrow \mathbb{R}^k$

$\delta(f, g) = \frac{1}{|\mathbb{D}|} \sum_{x \in \mathbb{D}} 1 \left( f(x) \neq g(x) \right)$

Dist. from $f$ to $P$ is

$\text{dist}(f, P) = \min_{g \in P} \delta(f, g)$
Def: $f$ is $k$-locally testable if

$\mathbf{D}$ A makes $k$ queries to the input.

$\mathbf{D}$ if $f$ is in $P$ then

(one-sided) A accepts $\geq \epsilon$ (completeness)

1. If $f$ is $\epsilon$-far from $f_{c}^{\text{cone}}$

then $\Pr[A \text{ accepts}] < \frac{1}{3}$ (soundness)

$$f(x_{1}, x_{2}, \ldots, x_{n})$$

$$f(y_{1}, y_{2}, \ldots, y_{k}) \neq f(y_{d_{1}}, y_{d_{2}}, \ldots, y_{d_{n}}) \Rightarrow \text{no}.$$
\[ P_n = \{ f : \mathbb{Z}_n \rightarrow \{0,1\} \mid f(x) = 1, \forall x \in [n] \} \]

Claim: \( P_n \) is \( \frac{2}{\varepsilon} \)-locally testable.

Disting bet\( \gamma \) \( f \in P_n \)  
\( \gamma \) \( f \) is \( \varepsilon \)-far from \( P_n \)

for \( n \rightarrow \infty \)  
using only \( \frac{\varepsilon}{(3)} \) many queries.
* Test: pick \( \frac{2}{\epsilon} \) random \( x \)'s

\[
\begin{align*}
&f(x_1) + f(x_2) - \cdots - f(x_{\frac{2}{\epsilon}}) \\
&\quad \\n&\text{if ever see 0 rej: ow acc.}
\end{align*}
\]

Analysis

- queries: \( \frac{2}{\epsilon} = \text{wrt. n} \)
- completeness: \( f \in P_n \) then \( \Pr[\text{acc}] = 1 \)
- soundness: \( f \) is \( \epsilon \)-far from \( P_n \) then

\[
\Pr[\text{acc}] \leq \Pr[\text{no 0 is hit in } \frac{2}{\epsilon} \text{ trials}] \\
\leq 3^{-\frac{2}{\epsilon}} \leq \left( \frac{\epsilon}{3} \right) \leq \left( \frac{\epsilon}{3} \right) = \frac{2}{3} \leq \frac{2}{3} \leq \frac{2}{3} \leq \frac{2}{3} \leq \frac{2}{3}
\]