Welcome to CS590 Sublinear Algorithms

Admin.

- Brightspace ; CampusWire ; class webs. on my page.
- Grading: 4-5 HW 40 %
  Research project 45%
  Scribe notes + peer grading HW 10%
  Class participation 5% ✓

- Need scribe for today ←
- Research : • max 2 people
  • See timeline : • by Feb 2 should have a topic
  Schedule ← meeting with me before then
  • Feb 16 + 2 paras Proposal
  • March 23 5-10 min present
  • April 29 outcomes
  • April 30 written report.

Motivation for sublin. algs :

BIG DATA :

internet of things
sales transactions
web pages
health data
genomic data
space discovery data

etc etc

Need algorithm design in $o(N)$

- If data can be stored but no time to read it $\longrightarrow$ sublinear time algs
- If data is too big to fit in memory $\rightarrow$ sublinear in space
  if can throw some away

  $\longrightarrow$ sublinear in communication
  if can store it on multiple machines that communicate

$\boxed{\text{Sublin - time algs}}$

- approx algs : eg : diameter ✓
  - \# of connected compts.
  - avg degree of a graph
- property testing :
  does an obj have a property or
  is far from having the
  property ?
  (eg. is G connected or far..
  is G 3-colorable or far)

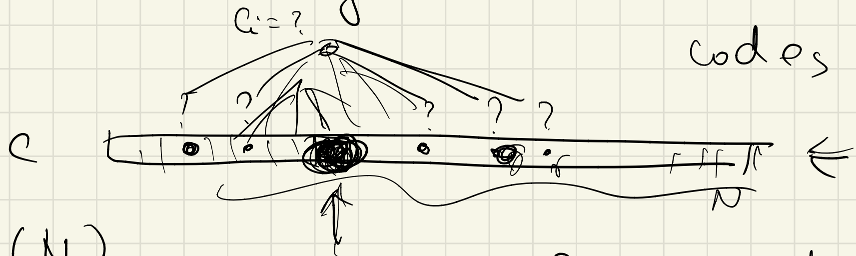
Model stems from Program checking
in PL.

80's { Blum Kannan
Blum Luby Rubinfeld

formalized by Rubinfeld Sudan 90
Leads to PCP thm

Also many other local models

• Locally decodable / testable codes



poly(N)

polylog(N)

$c_i = ?$ if can only query few posns into word?

Type of research questions:

Can membership in specific code be tested with cst. many queries?

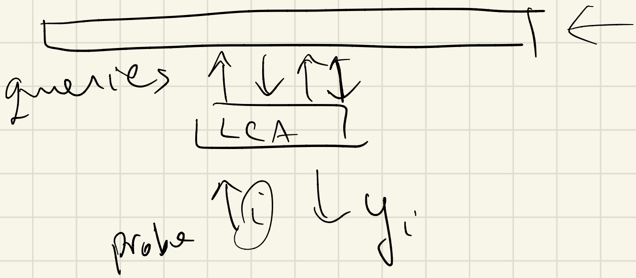
Can each entry be corrected whp?

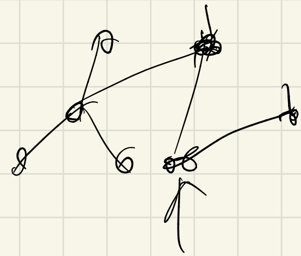
Best tradeoffs bet: rate, distance, locality

Leads to a general notion of Local Computation Algs.

Local Comput Algs (LCAs)

Rubinfeld Tamir Vardy Xie '11





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 algs / streaming

Given a seq of elts appearing one at a time, with limited memory

can get eg: statistics abt the stream

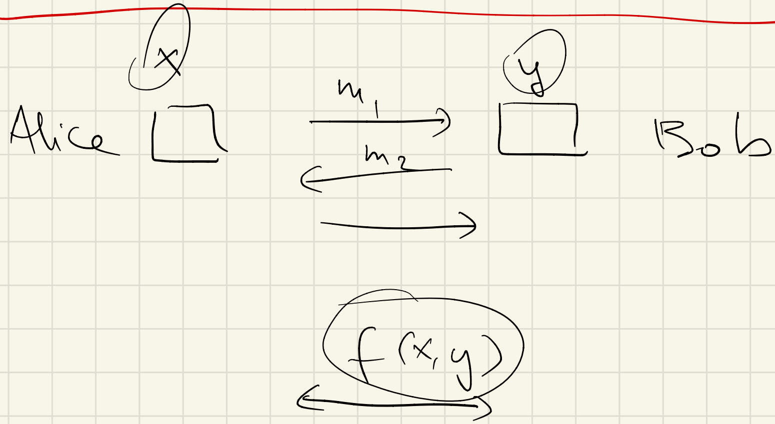
- max, min, avg, median
- a rand sample?
- estimates of # distinct elts
- \# heavy hitters?

OR can get an approx size of

- max matching!
- vertex cover!
- \# connect compts!
- avg degree?!

Defined by Alon Matias Szegedy '96

Sublinear in communication



Eg $\quad f(x,y) = x \oplus y$

$f : \mathbb{F}_2 \times \mathbb{F}_2 \to \mathbb{F}_2$

2 bits

- deterministic
- randomized

Many natural problems are hard

i.e. require almost full disclosure of inputs

Eg!

EQUALITY $\quad EQ : \{0,1\}^N \times \{0,1\}^{N} \to \{0,1\}$

$$EQ(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{ow.} \end{cases}$$

$Det(EQ) \geq N$

INDEX $\quad iDX : \underbrace{\{0,1\}^N}_{\text{Bob}} \times \underbrace{[N]}_{\text{Alice}} \to \{0,1\}$

$iDX(\bar{x}, y) = x_y$

$Det^{\text{one-way}}(iDX) \geq N$

$Rand^{\text{one-way}}(iDX) \geq \Omega(N)$

SET DISJ. $\quad DISJ : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$

$$DISJ(\bar{x}, \bar{y}) = \begin{cases} 1 & \text{if } x \cap y = \phi \\ 0 & \end{cases}$$

x, y are char vects of sets.

$Rand(DISJ) = \Omega(N)$

Obs Data stream algs $\Rightarrow$ communic 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

- diam. of a set of pts in $\mathbb{R}^n$ ←
- testing if fnc is constant
- uniform sampling of a stream
- deciding connectivity of a graph in a stream.

Deterministic 2-approx of diameter

<u>Def</u>: 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) $D(x,y) \leq D(x,z) + D(y,z)$

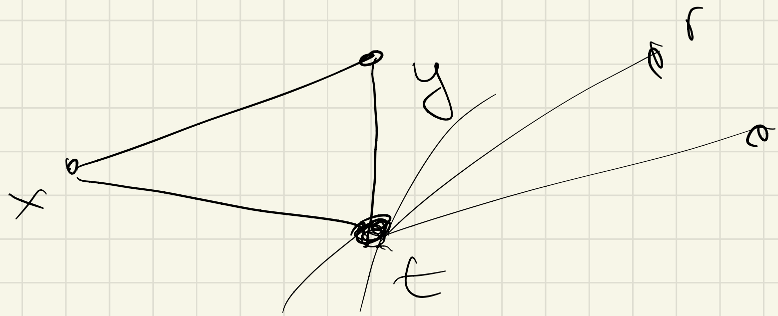
Def: $\text{diam}(S) = \max\limits_{x,y \in S} D(x,y)$

Thm (Indyk): Given S, D: $\exists$ det. alg that outputs d s.t. $\frac{\text{diam}(S)}{2} \leq d \leq \text{diam}(S)$

in time $\underset{\sim}{o(|D|)}$

note: $|\text{input}| = O(\underline{|D|}) = O(\underline{|S|}^2)$

$|S|$



$\underline{\forall x, y}$ $\forall t$ $D(x,y) \le D(x,t) + D(y,t)$

$\le 2 \cdot \max\{D(x,t), D(y,t)\}$

$\le 2 \max_{r \in S} \{D(r,t)\}$

$\boxed{\frac{3}{2}\text{-approx} \quad O(m n^{1/2})}$

Alg: Take any $t \in S$

output $\underline{d} = \max_{r \in S} \{D(r,t)\}$

$\text{Diam} \le 2d \le \text{diam}(S)$

$\frac{\text{Diam}}{2} \le d \le \text{Diam}$ ✓

Property testing

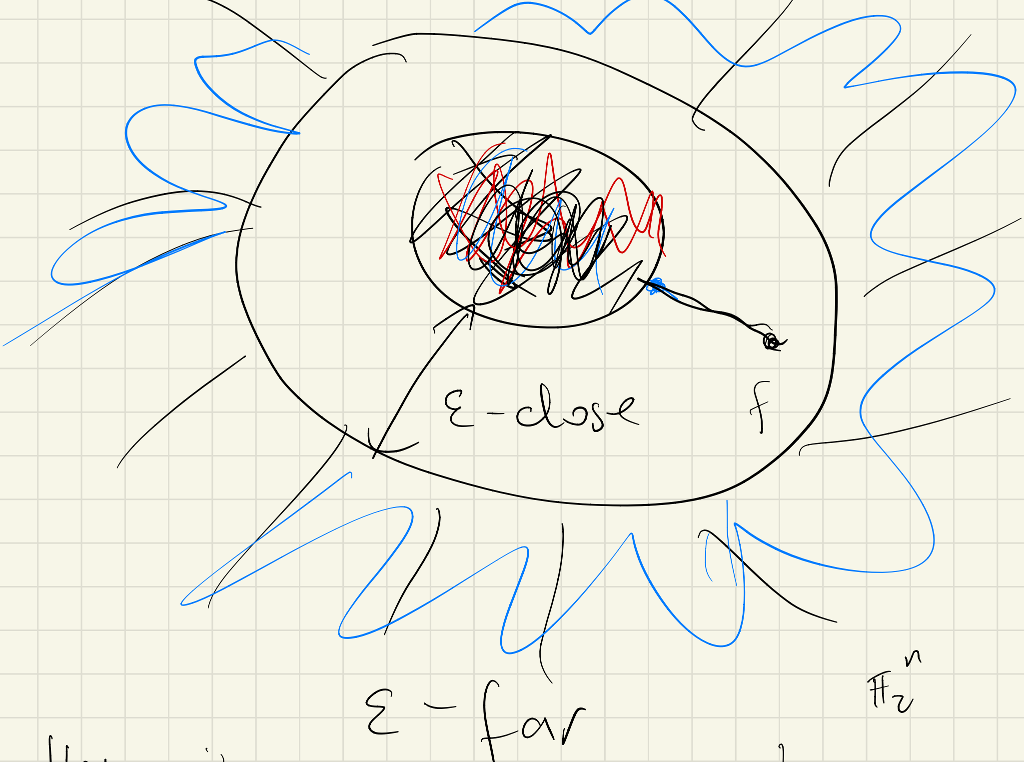
Property = a collection of obj that all have a particular property

$\mathcal{J}_n = \{$ all graph on n-vts that are 3-colorable $\}$

$P_n = \{$ all graphs on n vts that are connected $\}$

$\{$ ~~[illegible]~~ that are Δ-free $\}$

$P_n = \{$ distrib. that are bimodal $\}$



$\mathbb{F}_2^n$

Rel Hamming Distance : $f, g : D \to R$

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

Dist. from f to $\mathcal{P}$ is

$$\text{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \delta(f,g)$$

Def: $\mathcal{P}$ is k-locally testable if $\exists$ rand alg A with black-box access to input s.t.

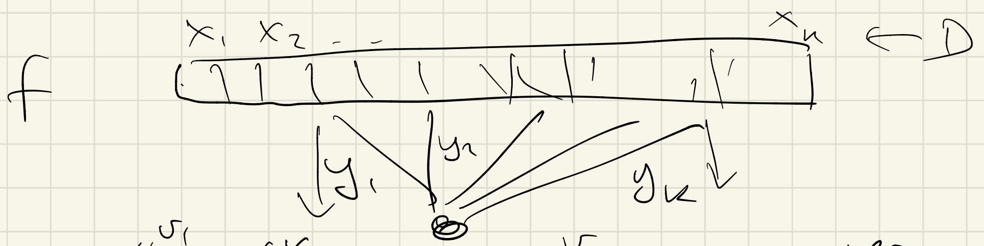
1) A makes k queries to the input.

2) ◦ if f is in $\mathcal{P}$ then

(one-sided) A accepts (completeness)

◦ If f is ε-far from $\mathcal{P}$

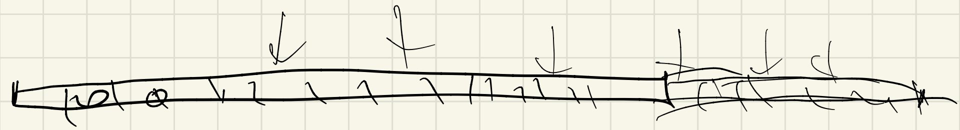
then $\Pr[A \text{ accepts}] < \frac{1}{3}$ (soundness)



$f(y_1)$ "v_1" $f(y_2)$ "v_2" - - - $f(y_k)$ "v_k" $\longrightarrow$ yes / no.

$$P_n = \{ f \mid f(x) = 1,\ x \in [n],\ f : [n] \to \{0,1\} \}$$

f



if $f \in P_n$ or

f is ε-far from P_n.

(i.e. f has at least an ε-fract of 0's)

$> \varepsilon n$ values of f are 0's.

claim: P_n is $\frac{2}{\varepsilon}$-locally testable.

Disting bet $\begin{cases} f \in P_n \\ f \text{ is } \varepsilon\text{-far from } P_n \end{cases}$

using only $\frac{2}{(\varepsilon)}$ many queries.

for $n \to \infty$

- Test: pick $\frac{2}{\varepsilon}$ random x's

$$f(x_1) \quad f(x_2) \quad \ldots \quad f(x_{\frac{2}{\varepsilon}})$$
$$\;\;\| \qquad\quad \| \qquad\qquad\quad \|$$
$$\cancel{0} \qquad 1 \quad 1 \qquad\quad \cancel{1}$$

if ever see 0 rej.

o.w. acc.

<u>Analysis</u>

queries: $\frac{2}{\varepsilon}$ = ct. w.r.t. n

<u>completeness</u>: $f \in P_n$ then $\Pr[\text{acc}] = 1$

<u>soundness</u>: f is ε-far from P_n

then

$\Pr[\text{acc}] = \Pr[\text{no } 0 \text{ is hit in } \frac{2}{\varepsilon} \text{ trials}]$

$$\leq (1-\varepsilon)^{2/\varepsilon} \leq (e^{-\varepsilon})^{2/\varepsilon} = e^{-2} < \frac{1}{3}$$