Streaming Algorithms for Set Cover

Sariel Har-Peled (UIUC)

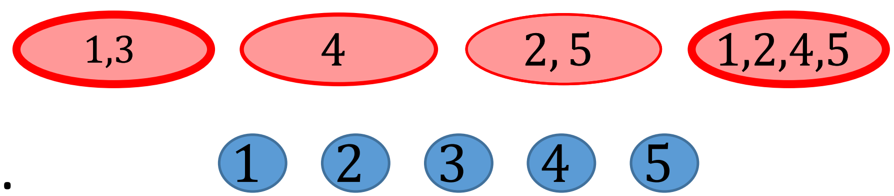
Piotr Indyk (MIT)

Sepideh Mahabadi (MIT)

Ali Vakilian (MIT)

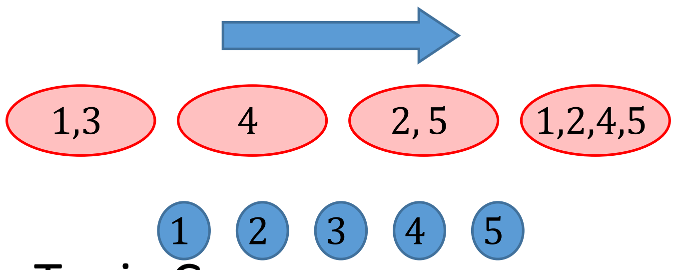
Set Cover Problem

- Input: a collection $\mathcal{S}$ of sets $S_1 \ldots S_m$ that covers $U = \{1 \ldots n\}$
  - i.e., $S_1 \cup S_2 \cup \cdots \cup S_m = U$
- Output: a subset $\mathcal{I}$ of $\mathcal{S}$ such that:
  - $\mathcal{I}$ covers U
  - $|\mathcal{I}|$ is minimized
- Classic optimization problem:
  - NP-hard
  - Greedy $(\ln n)$-approximation algorithm
  - Can't do better unless P=NP [Feige 98][Alon, Moshkovitz, Safra 06][Dinur, Steurer 14]



Streaming Set Cover [SG09]

- Model
  - Sequential access to $S_1, S_2, \dots, S_m$
  - One (or few) passes, sublinear (i.e., $o(mn)$) storage
  - (Hopefully) decent approximation factor



- Why?
  - A classic optimization problem
  - Application in “Big Data”: Clustering, Topic Coverage
  - One of few NP-hard problems studied in streaming
    - Other examples: Clustering, Max-Cut, Sub-Modular Optimization, FPT

Previous and Our Results: Algorithms

Algorithms	Approximation	Passes	Space	Type
Greedy Alg	$O(\log n)$ $O(\log n)$	1 n	$O(mn)$ $O(n)$	Deterministic Deterministic
[Getoor and Saha 09]	$O(\log n)$	$O(\log n)$	$O(n \log n)$	Deterministic
[Emek and Rósen 14]	$O(\sqrt{n})$	1	$\tilde{O}(n)$	Deterministic
[Demaine, Indyk, **M**, Vakilian 14]	$O(\rho 4^{1/\delta})$	$O(4^{1/\delta})$	$\tilde{O}(mn^{\delta})$	Randomized
[Chakrabarti, Wirth 16]	$O(n^{\delta}/\delta)$	$1/\delta - 1$	$\tilde{O}(n)$	Deterministic

This Work	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	**Randomized**

ρ = approximation guarantee for offline **Set Cover**

n = number of *elements*
m = *number of sets.*

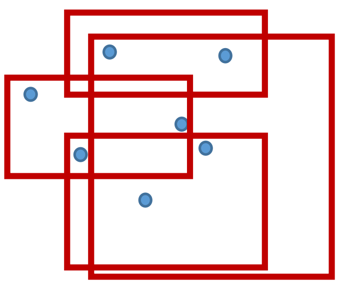
Previous and Our Results: Lower-bounds

[Guha and McGregor] A p-pass *streaming* algorithm of problem **P** using s bits of storage yields a $(2p-1)$ rounds protocol with $(2p-1)s$ bit of communication for **P** in *2-party communication complexity* model.

Lower bounds	**Approximation**	**Passes**	**Space**	**Type**
[Nisan 02]	$(\log n)/2$	Any	$\Omega(m)$	Randomized
[Emek, Rosen 14]	$\sqrt{n}$	1	$\Omega(n)$	Randomized
[Demaine, Indyk, **M**, Vakilian 14]	Constant	Any	$\Omega(mn)$	Deterministic
[Chakrabarti, Wirth 14]	$\delta^2 n^{1/\delta}$	$1/\delta$	$\tilde{\Omega}(n)$	Randomized
This Work	3/2	1	$\Omega(mn)$	**Randomized**
This Work	1	$1/\delta$	$\Omega(mn^{\delta})$	**Randomized**

Our Results

Our Results	Approximation	Passes	Space	Type
Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	Randomized
Geometric Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(n)$	Randomized
Lower-bound	3/2	1	$\Omega(mn)$	Randomized
Lower-bound	1	$1/\delta$	$\Omega(mn^{\delta})$	Randomized
Sparse Case Lower-bound	1	$1/\delta$	$\Omega(ms)$	Randomized



$s =$ sparsity of the sets
$(s \leq n^{\delta})$

Outline of the Algorithm

Approach: "dimensionality reduction"

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses $\mathrm{O}(\rho k)$ sets (k = min cover size)
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

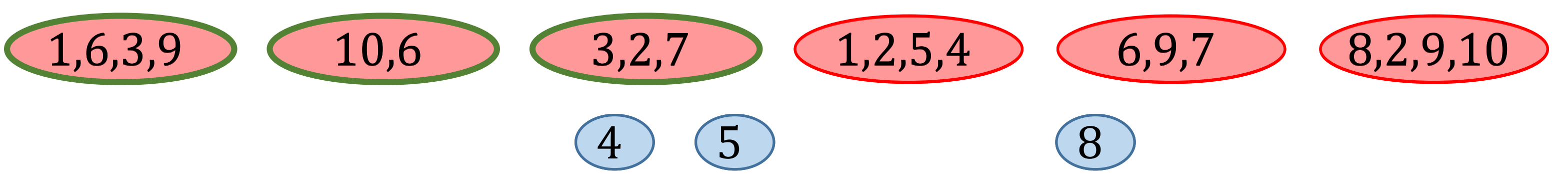
Repeat $O(1/\delta)$ times:

- Covers all the elements
- $O(\rho/\delta)$ approximation
- Uses $\tilde{O}(mn^{\delta})$ space
- $O(1/\delta)$ passes

Dimensionality reduction:

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses $O(\rho k)$ sets
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

- Suppose we know k = min cover size
- Select a set R of $kn^{\delta} \log m \log n$ random elements from U
- Pass 1:
  - For each set S_i , select S_i if it covers $\Omega(|R|/k)$ uncovered elements of R
  - Otherwise, store projection of S_i over R
- Compute a ρ-approximate set cover I' over R
- Pass 2:
  - Update the set of uncovered elements
- Report sets found in Pass 1



Dimensionality reduction:

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses $O(\rho k)$ sets
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

- Suppose we know k = min cover size
- Select a set R of $kn^{\delta} \log m \log n$ random elements from U
- Pass 1:
  - For each set S_i, select S_i if it covers $\Omega(|R|/k)$ uncovered elements of R
  - Otherwise, store projection of S_i over R
- Compute a ρ-approximate set cover I' over R
- Pass 2:
  - Update the set of uncovered elements
- Report sets found in Pass 1

Dimensionality reduction:

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses $O(\rho k)$ sets
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

- Suppose we know k = min cover size
- Select a set R of $kn^{\delta}\log m \log n$ random elements from U
- Pass 1:
  - For each set S_i , select S_i if it covers $\Omega(|R|/k)$ uncovered elements of R
  - Otherwise, store projection of S_i over R

k sets

- Compute a ρ-approximate set cover I' over R
- Pass 2:
  - Update the set of uncovered elements
- Report sets found in Pass 1

ρk sets

Dimensionality reduction:

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses $O(\rho k)$ sets
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

- Suppose we know k = min cover size
- Select a set R of $kn^{\delta} \log m \log n$ random elements from U
- Pass 1:
  - For each set S_i , select S_i if it covers $\Omega(|R|/k)$ uncovered elements of R
  - Otherwise, store projection of S_i over R
- Compute a ρ-approximate set cover I' over R
- Pass 2:
  - Update the set of uncovered elements
- Report sets found in Pass 1

Increases space by $\log n$

$k \leq n$ sets : $n \log m$

$$m \frac{|R|}{k}$$
$$= m \cdot kn^{\delta} \log m \log n / k$$
$$= \tilde{O}(mn^{\delta})$$

Dimensionality reduction:

- Covers all but $1/n^\delta$ fraction of elements
- Uses $O(\rho k)$ sets
- Uses $\tilde{O}(mn^\delta)$ space
- Two passes

- Suppose we know k = min cover size
- Select a set R of $kn^\delta \log m \log n$ random elements from U
- Pass 1:
  - For each set S_i , select S_i if it covers $\Omega(|R|/k)$ uncovered elements of R
  - Otherwise, store projection of S_i over R
- Compute a ρ-approximate set cover I' over R
- Pass 2:
  - Update the set of uncovered elements
- Report sets found in Pass 1

Relative (p, ϵ)-approximation

- Let U be a set of elements
- Let $\mathcal{H} \subseteq 2^V$ be a collection of subsets of the ground set U

Then a subset Z is a relative (p, ϵ)-approximation for $(U, \mathcal{H})$ if for each $S \in \mathcal{H}$

- $(1-\epsilon)\frac{|S|}{|U|} \leq \frac{|S \cap Z|}{|Z|} \leq (1+\epsilon)\frac{|S|}{|U|}$ if $|S| \geq p|U|$
- $\frac{|S|}{|U|} - \epsilon p \leq \frac{|S \cap Z|}{|Z|} \leq \frac{|S|}{|U|} + \epsilon p$ if $|S| < p|U|$

$(1 \pm \epsilon)$-multiplicative estimator

(ϵp)-additive estimator

[Har-Peled and Sharir] For any p, ϵ and q, a random sample of U of size $O(\frac{1}{\epsilon^2 p}(\log|\mathcal{H}| \log\frac{1}{p} + \log\frac{1}{q}))$ is a relative (p, ϵ)-approximation of $(U, \mathcal{H})$ with probability at least $(1-q)$.

Dimensionality reduction:

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses ρk sets
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

- Suppose we know k = min cover size
- Select a set R of $\boldsymbol{kn^{\delta} \log m \log n}$ random elements from U
- Pass 1:
  Relative $(1/n^{\delta}, 1/2)$-approximation
  - For each set S_i, select S_i if it covers $\Omega(|R|/k)$ uncovered elements of R
  - Otherwise, store projection of S_i over R
- Compute a ρ-approximate set cover I' over R
- Pass 2:
  - Update the set of uncovered elements
- Report sets found in Pass 1

Algorithm

- Repeat $1/\delta$ times
  - Dimensionality Reduction component

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses ρk sets
- Uses $\tilde{O}(mn^{\delta})$ space
- Two passes

Our Results	Approximation	Passes	Space	Type
Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	Randomized
Geometric Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(n)$	Randomized
Lower-bound	3/2	1	$\Omega(mn)$	Randomized
Lower-bound	1	$1/\delta$	$\Omega(mn^{\delta})$	Randomized
Sparse Case Lower-bound	1	$1/\delta$	$\Omega(ms)$	Randomized

Lower bound: single pass

- Have seen that $O(1)$ passes can reduce space requirements
- What can(not) be done in one pass?
- We show that distinguishing between $k = 2$ and $k = 3$ requires $\widetilde{\Omega}\,(mn)$ space

Many vs One Set-Disjointness

- Two sets cover U iff their complements are disjoint

- Consider the following one-way communication complexity problem:
  - Alice: sets $S_1, \dots, S_m$
  - Bob: set S_B
  - Question: is S_B disjoint from one of S_i's ?

[Our Result] The randomized one way communication complexity of Many vs. One Set-disjointness is $\Omega(mn)$ if error probability is 1/poly(m).

Many vs One Set-Disjointness

[Our Result] The randomized one way communication complexity of Many vs. One Set-disjointness is $\Omega(mn)$ if error probability is 1/poly(m).

- Alice's sets are selected *uniformly* at random

- There exist poly(m) sets S_B such that if Bob learns answers to all of them, he can recover all S_i's with high probability

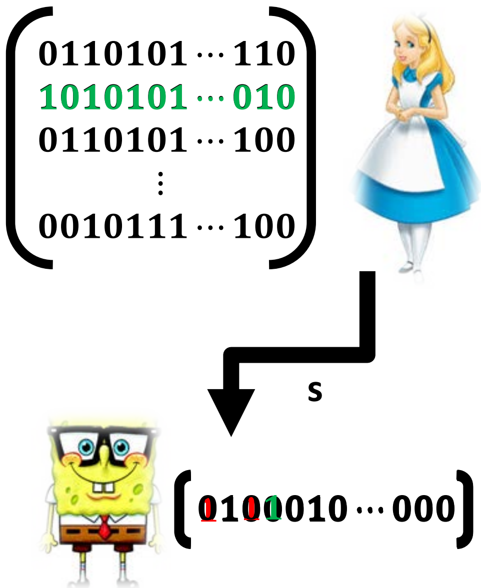
$$\begin{array}{cc} & n \\ m & \begin{pmatrix} \mathbf{0110101 \cdots 110} \\ \mathbf{1010101 \cdots 010} \\ \mathbf{0110101 \cdots 100} \\ \vdots \\ \mathbf{0010111 \cdots 100} \end{pmatrix} \end{array}$$

s

- Bob can recover mn random bits from $o(mn)$ bits of communication -> contradiction

Recovering Alice's Collection

- Recovery procedure
  - Suppose that Bob has a set S_B that is disjoint from *exactly* one S_i (we do not know which one)
    - Call it a "good seed" for S_i
  - Then Bob queries all extensions $S_B \cup \{e\}$ to recover S_i



- Bob's queries:
  - A random "seed" of size $c \log m$ is disjoint from exactly one S_i w.p. $m^{-O(c)}$
  - Try $m^{O(c)}$ times
- Recover all S_i

Result

[Our Result] The randomized one way communication complexity of Many vs. One Set-disjointness is $\Omega(mn)$ if error probability is 1/poly(m).

Our Results	Approximation	Passes	Space	Type
Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	Randomized
Geometric Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(n)$	Randomized
Lower-bound	3/2	1	$\Omega(mn)$	Randomized
Lower-bound	1	$1/\delta$	$\Omega(mn^{\delta})$	Randomized
Sparse Case Lower-bound	1	$1/\delta$	$\Omega(ms)$	Randomized

Lower bound: Multipass

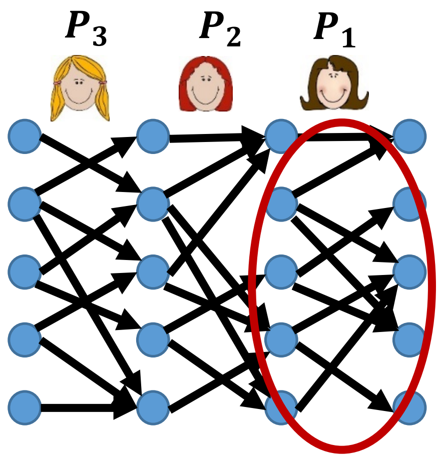
- Reduction from Intersection Set Chasing [Guruswami, Onak 13]
- Very "fragile", works only for the exact problem

[Our Result] Any $1/\delta$ pass *exact* algorithm of Set Cover requires $\widetilde{\Omega}(mn^{\delta})$ space

In **s-Sparse Set Cover**, each input set is of size at most s.

[Our Result] Any $1/\delta$ pass *exact* algorithm of s-Sparse Set Cover requires $\widetilde{\Omega}(ms)$ space (for $s \leq n^{\delta}$)

Set Chasing(n, p) Problem

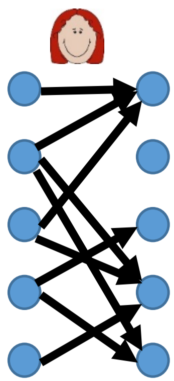


Set Chasing (5,3)

- p players,
- Each knows an $n * n$ bipartite directed graph

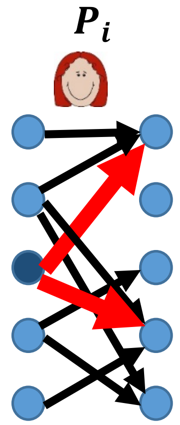
Set Chasing(n, p) Problem

$\boldsymbol{P_i}$



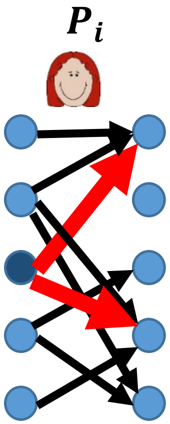
$f_i\colon [n] \to 2^{[n]}$

Set Chasing(n, p) Problem



$f_i\colon [n] \to 2^{[n]}$

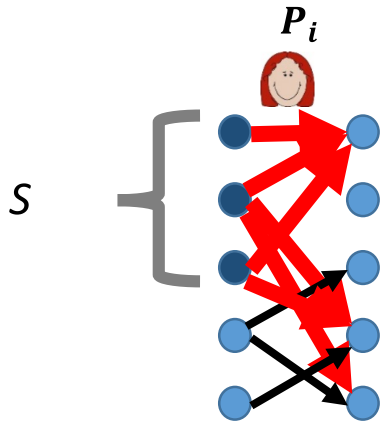
Set Chasing(n, p) Problem



$f_i: [n] \to 2^{[n]}$

$\overline{f_i}(S) = \cup_{a \in S} f_i(a)$

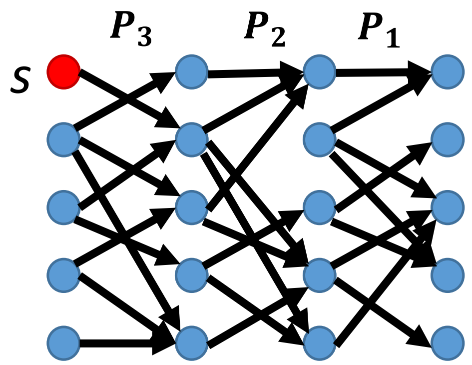
Set Chasing(n, p) Problem



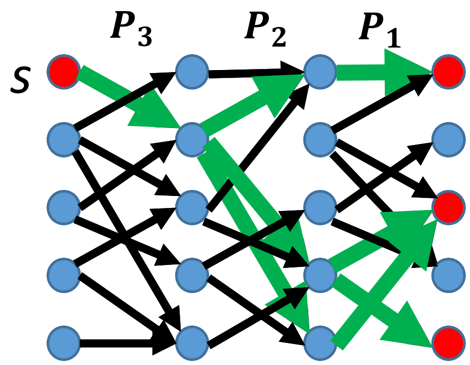
$f_i : [n] \to 2^{[n]}$

$\overline{f_i}(S) = \bigcup_{a \in S} f_i(a)$

Set Chasing(n, p) Problem



Set Chasing(n, p) Problem



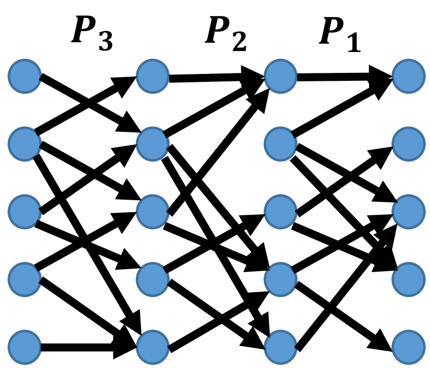
- p players,
- r rounds; in each round starting from P_1 a player speaks (to all)

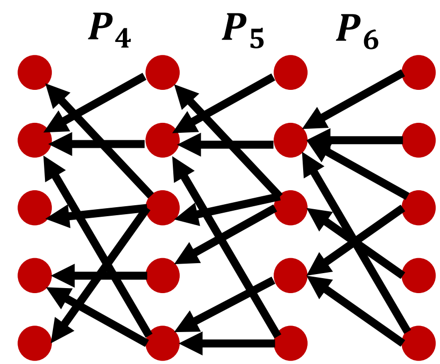
Goal: P_p computes $\overline{f_1}\left(\overline{f_2}\left(\cdots\left(\overline{f_p}(s)\right)\cdots\right)\right)$ at the end of the last round.

Interesting instance: $r = p - 1$

CC(SC(n, p)) = $n^{1+\Omega(1/p)}$ [Feigenbaum et al. 08]

Intersection Set Chasing(n, p) Problem

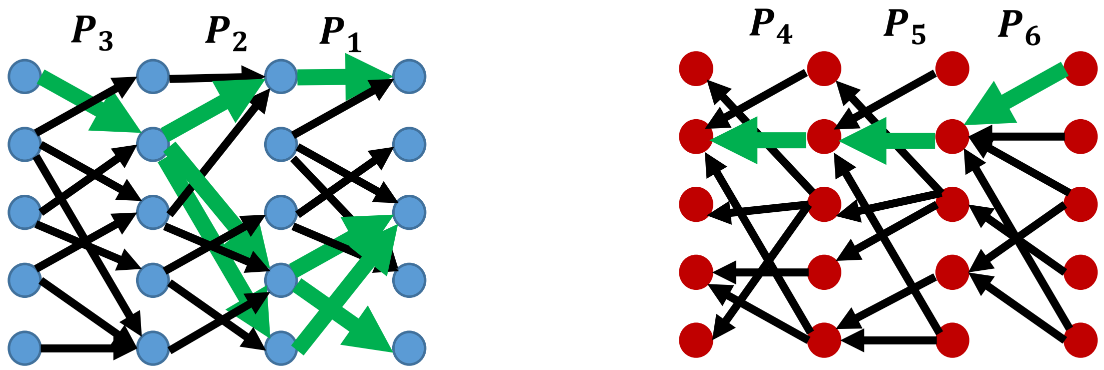




- Two instances of Set Chasing

Goal: Whether $\overline{f_1}\left(\overline{f_2}\left(\cdots\left(\overline{f_p}(s)\right)\cdots\right)\right)$ and $\overline{f_{p+1}}\left(\overline{f_{p+2}}\left(\cdots\left(\overline{f_{2p}}(s)\right)\cdots\right)\right)$ intersect?

Intersection Set Chasing(n, p) Problem

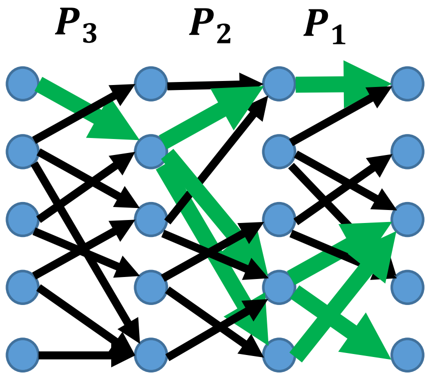


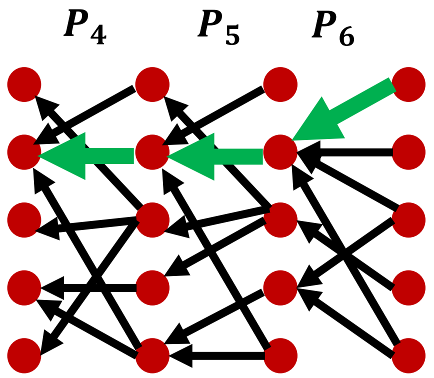
- Two instances of Set Chasing

Goal: Whether $\overline{f_1}\left(\overline{f_2}\left(\cdots\left(\overline{f_p}(s)\right)\cdots\right)\right)$ and $\overline{f_{p+1}}\left(\overline{f_{p+2}}\left(\cdots\left(\overline{f_{2p}}(s)\right)\cdots\right)\right)$ intersect?

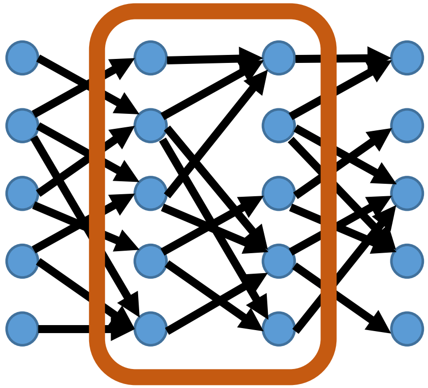
Intersection Set Chasing(n, p) Problem

[Guruswami and Onak] Any randomized protocol that solves Intersection Set Chasing(n, p) with error probability less than $1/10$, requires $\widetilde{\Omega}(\frac{n^{1+1/(2p)}}{p^{16}})$ bits of communication where n is sufficiently large and $p \leq \frac{\log n}{\log \log n}$.

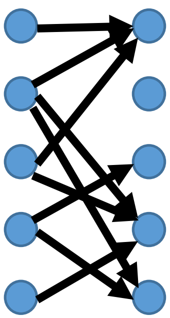




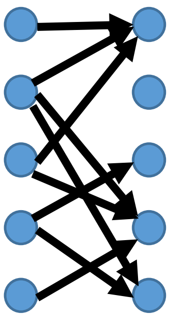
Reduction

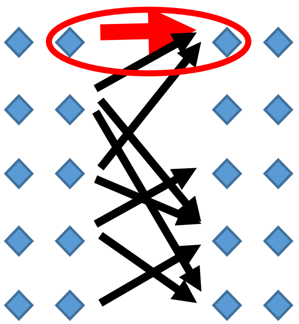


Reduction



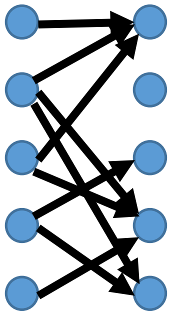
Reduction

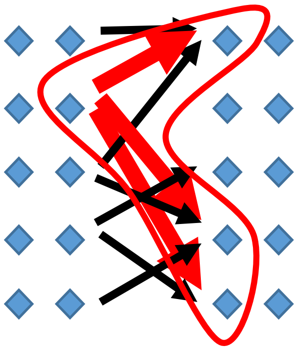




Function set

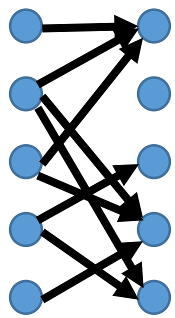
Reduction

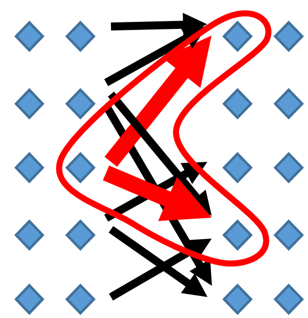




Function set

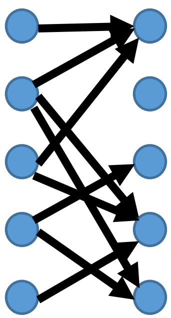
Reduction

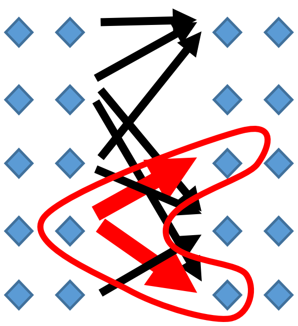




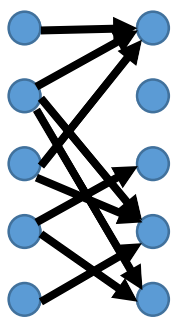
Function set

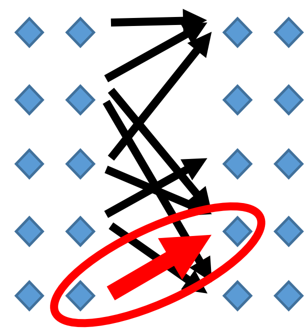
Reduction





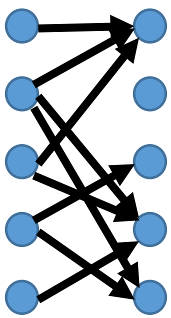
Reduction

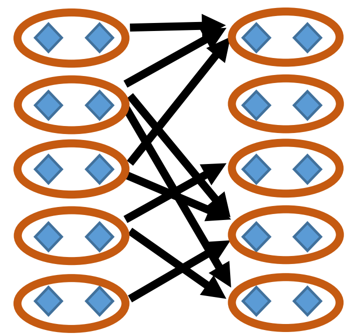




Function set

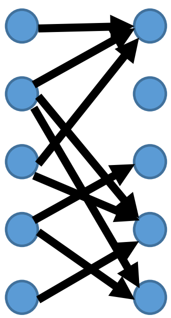
Reduction

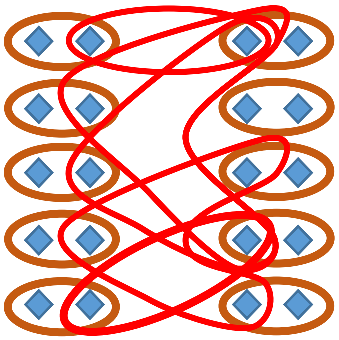




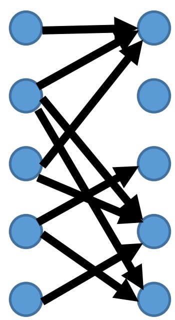
Border sets

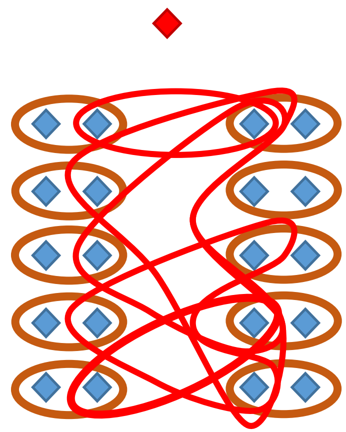
Reduction





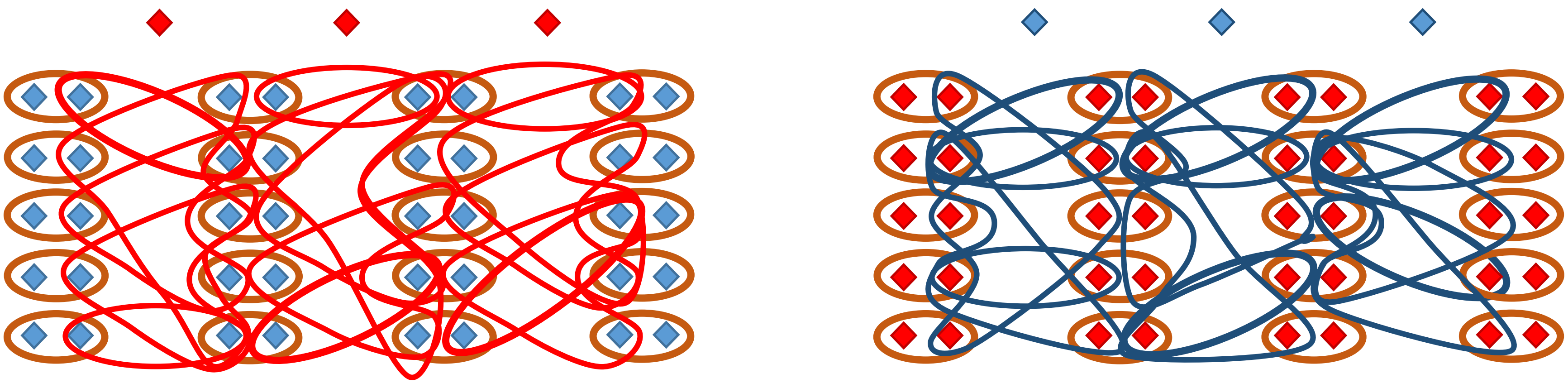
Reduction





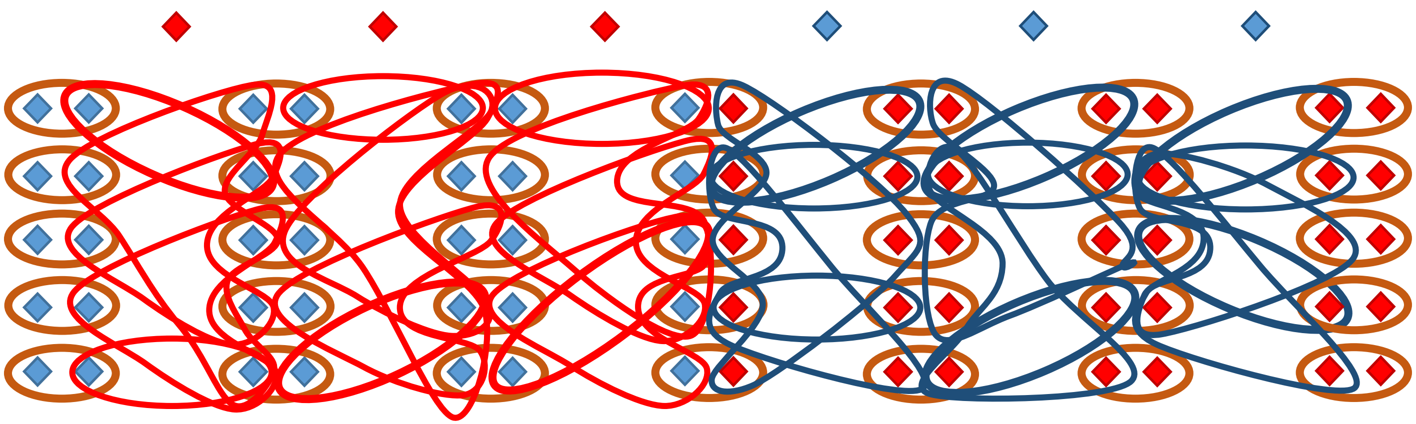
Enforce to pick one of the function sets.

Reduction



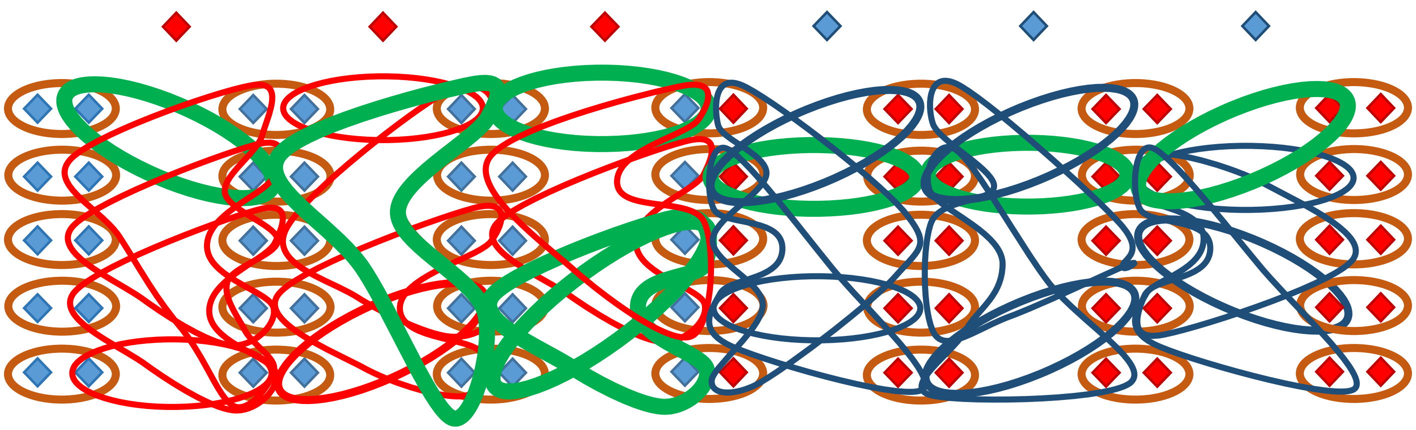
Reduction

- Size of Set Cover in such an instance is at least $(2p+1)n+1$
- There exists an intersection between the corresponding nodes iff size of the set cover is exactly $(2p+1)n+1$



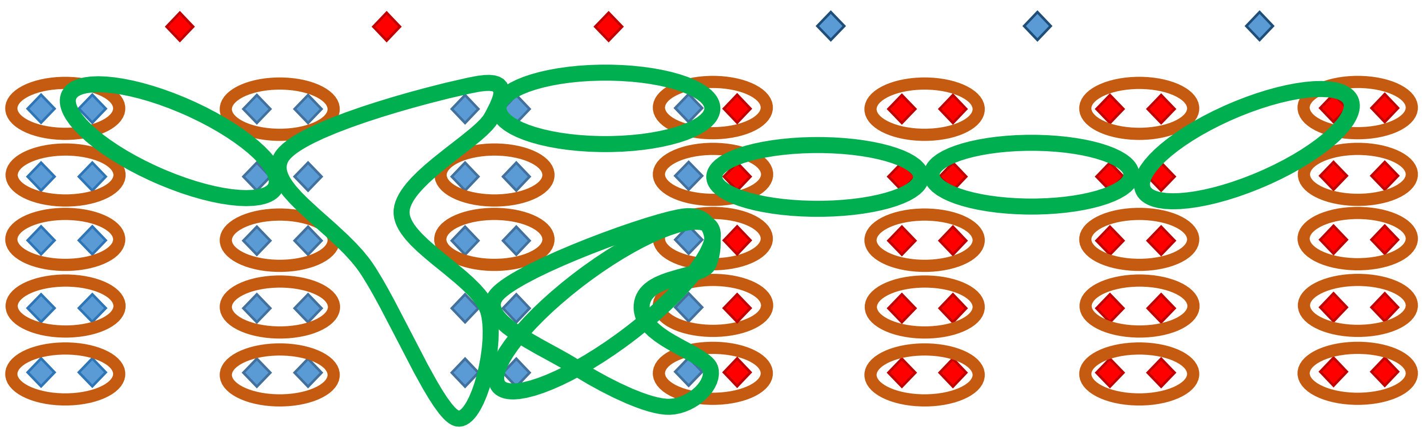
Reduction

- Size of Set Cover in such an instance is at least $(2p+1)n+1$
- There exists an intersection between the corresponding nodes iff size of the set cover is exactly $(2p+1)n+1$



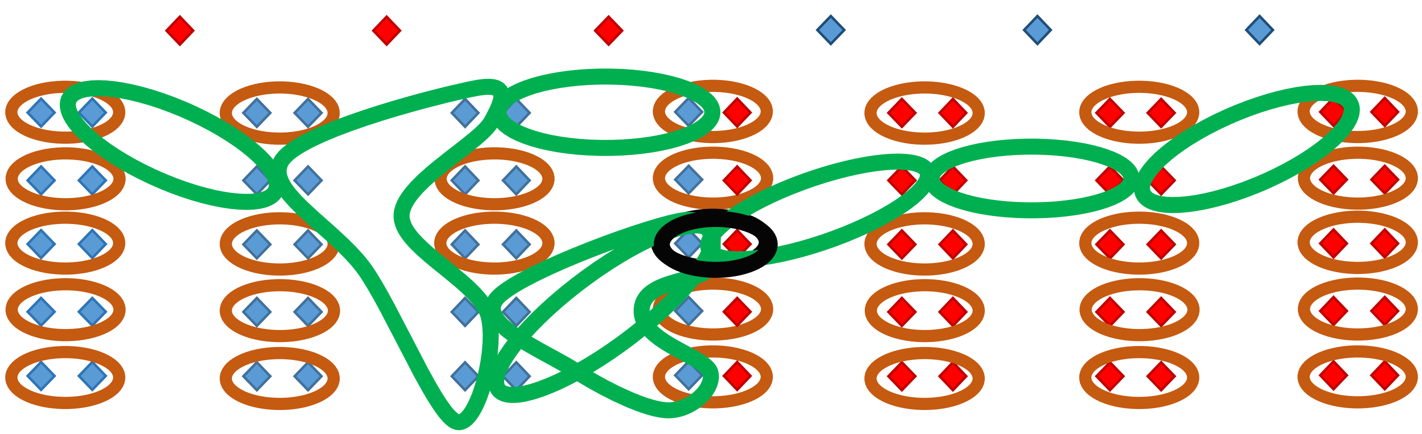
Reduction

- Size of Set Cover in such an instance is at least $(2p+1)n+1$
- There exists an intersection between the corresponding nodes iff size of the set cover is exactly $(2p+1)n+1$



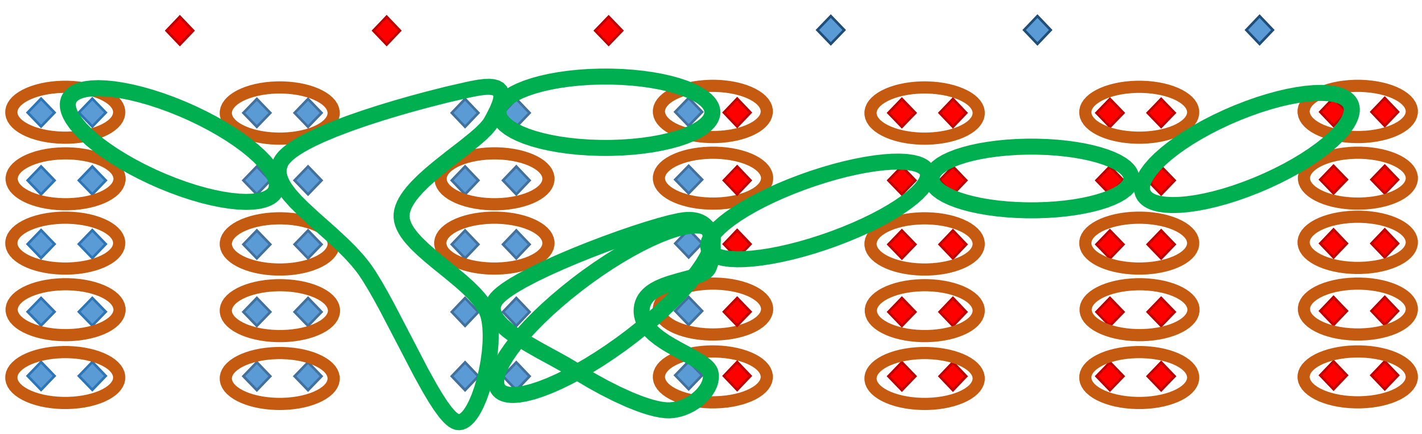
Reduction

- Size of Set Cover in such an instance is at least $(2p+1)n+1$
- There exists an intersection between the corresponding nodes iff size of the set cover is exactly $(2p+1)n+1$



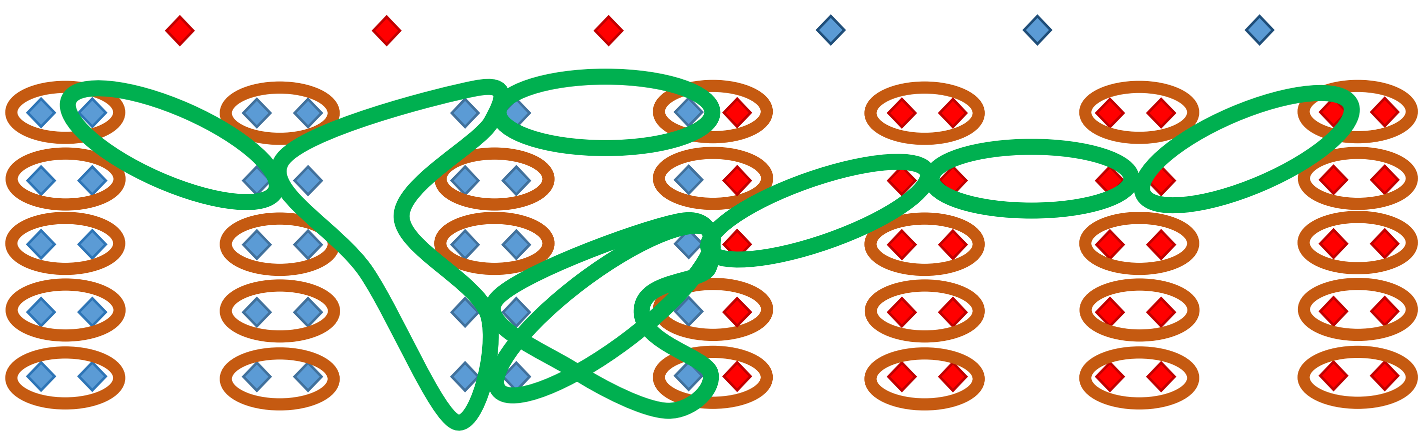
Reduction

- Size of Set Cover in such an instance is at least $(2p+1)n+1$
- There exists an intersection between the corresponding nodes iff size of the set cover is exactly $(2p+1)n+1$



Reduction

- $M_{SC} = O(np)$, $N_{SC} = O(np)$, $1/\delta = O(p)$
- Lower bound of $\widetilde{\Omega}(n^{1+1/2p}) = \widetilde{\Omega}(M_{SC} N_{SC}^{O(\delta)})$



Result

Any Streaming Algorithm that solves the set cover problem with constant probability of error in $\frac{1}{2\delta} - 1$ passes , requires $\widetilde{\Omega}\left(mn^{\delta}\right)$ memory space where $\delta \geq \frac{\log\log n}{\log n}$.

Our Results	Approximation	Passes	Space	Type
Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	Randomized
Geometric Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(n)$	Randomized
Lower-bound	3/2	1	$\Omega(mn)$	Randomized
Lower-bound	1	$1/\delta$	$\Omega(mn^{\delta})$	Randomized
Sparse Case Lower-bound	1	$1/\delta$	$\Omega(ms)$	Randomized

Future Directions

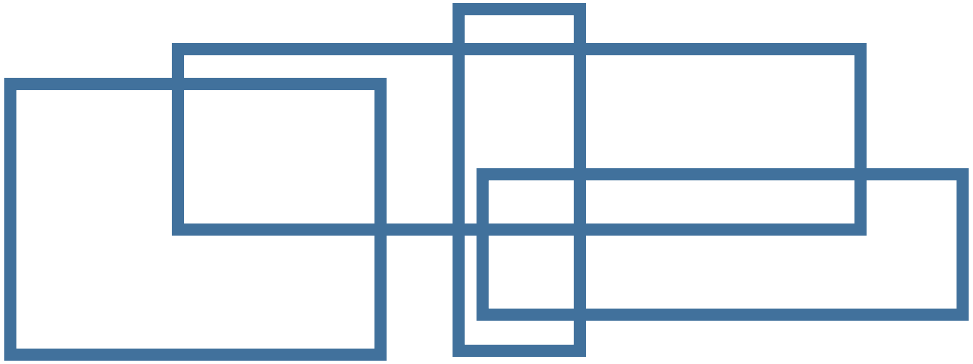
Our Results	Approximation	Passes	Space	Type
Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	Randomized
Geometric Algorithm	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(n)$	Randomized
Lower-bound	3/2	1	$\Omega(mn)$	Randomized
Lower-bound	1	$1/\delta$	$\Omega(mn^{\delta})$	Randomized
Sparse Case Lower-bound	1	$1/\delta$	$\Omega(ms)$	Randomized

- Weighted Set Cover Problem
- Improving lower bound for single pass protocols
- Improving Lower bound for multiple pass protocols: for approximate algorithms
- Geometric set cover in higher dimensions

Thank You!

Geometric Set Cover

- Elements are points in R^2.
- Sets are *discs*, *axis-parallel rectangles* and *fat triangles* (*shapes*).
- **Main Observation**: Transform the sets $\mathcal{F}$ to *canonical representation* $\mathcal{F}'$
  1. Each set in $\mathcal{F}'$ is contained by a set in $\mathcal{F}$.
  2. Each set in $\mathcal{F}$ is union of at most c sets in $\mathcal{F}'$.
  3. The size of $\mathcal{F}'$ is small, given that each of them has few points in them



Geometric Set Cover

- Elements are points in R^2.
- Sets are *discs*, *axis-parallel rectangles* and *fat triangles* (*shapes*).
- **Main Observation**: Transform the sets $\mathcal{F}$ to *canonical representation* $\mathcal{F}'$
  1. Each set in $\mathcal{F}'$ is contained by a set in $\mathcal{F}$.
  2. Each set in $\mathcal{F}$ is union of at most c sets in $\mathcal{F}'$.
  3. The size of $\mathcal{F}'$ is small, given that each of them has few points in them