Fair Set Cover

Mohsen Dehghankar[†] R. Raychaudhury[‡] S. Sintos[†] A. Asudeh [†]

† University of Illinois Chicago
{mdehgh2, stavros, asudeh}@uic.edu

‡ Duke University
rahul.raychaudhury@duke.edu

KDD'25: 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 3–7, 2025 – Toronto, ON, Canada

Outline

Motivation

Set Cover Problem

- **Given**:
  - A universe of n elements $U = \{e_1, e_2, \ldots, e_n\}$
  - A family of μ sets $\mathcal{S} = \{S_1, S_2, \ldots, S_\mu\}$, where $\cup_{i=1}^{\mu} S_i = U$
- **Goal**: Find the smallest sub collection $X \subseteq S$ such that $\bigcup_{s_i \in X} = U$.

Set Cover Applications

- *classic applications*: airline crew scheduling, facility location, computational biology, network security, etc.
- *problems with societal impact*: business license distribution, team formation, fair clustering, etc.
  - Prevent biased selection.

Motivation Example

Team of Experts Assembly

The HR of a company wants to form a team of data scientists.

- Form a minimal-size team that collectively satisfies a set of skills (e.g., {`python`, `sql`, `data-visualization`, `statistics`, `deep-learning`, $\cdots$}).
- Historical Biases + solely optimizing for the team size $\Rightarrow$ selection bias: mostly selecting from privileged groups.
- *Societal Requirement*: Equal (or proportionate) representation of various demographic groups.

Outline

(Group) Fairness Notion: Demographic Parity

General Definition

- Given: non-negative coefficients $\{f_1, \cdots, f_k\}$, where $\sum_{h=1}^{k} f_h = 1$
- For all groups $\mathbf{g}_h \in \mathcal{G}$: $\left|X \cap \mathcal{S}_h\right| = f_h \left|X\right|$.

Customized Definitions

- *Count-parity* – when $f_h = \frac{1}{k}$, $\forall \mathbf{g}_h \in \mathcal{G}$: equal number from each group.
- *Ratio-parity* – when $f_h = \frac{m_h}{\mu}$, $\forall \mathbf{g}_h \in \mathcal{G}$: maintains the original group ratios.

ε-unfairness

If, for each group $\mathbf{g}_h \in \mathcal{G}$, it holds that $1 - \varepsilon \leq \frac{|X \cap \mathcal{S}_h|}{f_h |X|} \leq 1 + \varepsilon$.

Problem Definition

Generalize Fair Set Cover

- **Given**:
  - A universe of n elements $U = \{e_1, e_2, \ldots, e_n\}$
  - A family of m sets $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$, where $\cup_{i=1}^{m} S_i = U$.
  - A set of k groups $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_k\}$. Each $S \in \mathcal{S}$ is associated with a group $\mathbf{g}(S) \in \mathcal{G}$
  - A fraction f_h for each $\mathbf{g}_h \in \mathcal{G}$ such that their sum is equal to 1
  - A weighted function $w : \mathcal{S} \to \mathbb{R}^+$.
- **Goal**: Find a fair cover X_w such that $\sum_{S \in X_w} w(S)$ is minimized.
- Under count parity, we call the problem Fair Set Cover (FSC).
-

Hardness

- FSC is NP-complete.
- FSC cannot be approximated with a sublinear approximation factor unless P = NP.

Key Aspects of Proposed Algorithms

Table: Summary of Algorithms for Zero-unfairness

Fairness	**Setting**	**Algorithm**	**Approx. Factor**	**Runtime**
Count Parity	Unweighted	Baseline	$k(\ln n + 1)$	$\mathcal{O}(mkn)$
		GreedyAlg	$\ln n + 1$	$\mathcal{O}(\text{poly}(n, m, k))$
		FasterAlg	$\frac{e}{e-1}(\ln n+1)$	$\mathcal{O}(\lvert X^*\rvert nm \log n)$
	Weighted	Baseline	$k\Delta(\ln n + 1)$	$\mathcal{O}(mkn)$
		GreedyAlg	$\Delta(\ln n + 1)$	$\mathcal{O}(m^k n)$
		FasterAlg	$\frac{e}{e+1}\Delta(\ln n+1)$	$\mathcal{O}(mn\,\mathcal{L} + mkn^3)$
Ratio Parity	Unweighted	GreedyAlg	$\ln n + 1$	$\mathcal{O}(m^p n)$
		FasterAlg	$\frac{e}{e+1}(\ln n+1)$	$\mathcal{O}(m(p+\mathcal{L}))$[1]

[1] $\mathcal{L}$ = time to solve an LP with $n + k$ variables and $2n + 2mk$ constraints.

Key Aspects of Proposed Algorithms

Table: Summary of Algorithms for Zero-unfairness

Fairness	**Setting**	**Algorithm**	**Approx. Factor**	**Runtime**
Count Parity	Unweighted	Baseline	$k(\ln n + 1)$	$\mathcal{O}(mkn)$
Count Parity	Unweighted	☞ GreedyAlg	$\ln n + 1$	$\mathcal{O}(\text{poly}(n, m, k))$
Count Parity	Unweighted	☞ FasterAlg	$\frac{e}{e-1}(\ln n+1)$	$\mathcal{O}(\lvert X^*\rvert nm \log n)$
Count Parity	Weighted	Baseline	$k\Delta(\ln n + 1)$	$\mathcal{O}(mkn)$
Count Parity	Weighted	GreedyAlg	$\Delta(\ln n + 1)$	$\mathcal{O}(m^k n)$
Count Parity	Weighted	FasterAlg	$\frac{e}{e+1}\Delta(\ln n+1)$	$\mathcal{O}(mn\,\mathcal{L} + mkn^3)$
Ratio Parity	Unweighted	GreedyAlg	$\ln n + 1$	$\mathcal{O}(m^p n)$
Ratio Parity	Unweighted	FasterAlg	$\frac{e}{e+1}(\ln n+1)$	$\mathcal{O}(m(p+\mathcal{L}))$[2]

[2] $\mathcal{L}$ = time to solve an LP with $n + k$ variables and $2n + 2mk$ constraints.

Outline

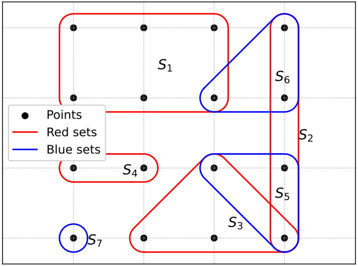
FSC Greedy (binary groups)

Let $U^- = U$ be the set of uncovered elements. Repeat until $U^- = \emptyset$:

- find the *pair* of sets (S_{Red}, S_{Blue}) from the unselected sets that covers the max # uncovered elements.
- move S_{Red} and S_{Blue} to the selected set X.
- update the uncovered elements: $U^- \leftarrow U^- \setminus (S_A \cup S_B)$

Analysis

- Always Fair (0-unfairness)
- Approximation Ratio: $\log n$
- Time Complexity: $O(m^2 n)$



Standard Greedy:

- selects $\{S_1, S_2, S_3, S_4, S_7\}$
- unfair: 4 red sets and 1 blue set

FSC Greedy selects $\langle (S_1, S_5), (S_3, S_6), (S_4, S_1) \rangle$.

Unweighted Fair Set Cover – General Grouping

Extension of Greedy to non-binary groups

- Extending beyond binary groups: at every iteration select k sets, one from each group that maximally cover the uncovered elements.
- Time complexity: $O(m^k n)$ (exponential to the number of groups)

Instead, at every iteration, the Faster alg. finds the k sets approximately:

Max k-color Cover Problem

- Given the uncovered U^- and non-selected sets $\mathcal{S}^-$, and k colors
- select k sets $X \subseteq \mathcal{S}^-$, one from each color, such that $\left|X \cap U^-\right|$ is maximized.

Max k-color Cover – A $(1 - \frac{1}{e})$-Approximation Algorithm

The LP-Relaxtion Algorithm

1. Model the problem as IP
2. Relax to LP and Solve
3. Rounding: For every group $\mathbf{g}_h \in \mathcal{G}$, sample exactly one set from $\mathcal{S}_h^-$, using the probabilities $\{x_i^* \mid S_i \in \mathcal{S}_h^-\}$.

Analysis

- Approximation factor: $(1 - \frac{1}{e})$.
- Time Complexity: $O(\mathcal{L}(n + k, 2(n + m)))$.

IP Formulation

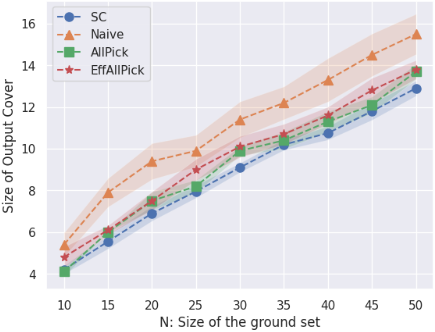
$$
\begin{aligned}
\max \quad & \sum_j y_j \\
\text{s.t.} \quad & \sum_{i:S_i \in \mathcal{S}_h^-} x_i = 1, && \forall \mathbf{g}_h \in \mathcal{G} \\
& \sum_{i:e_j \in S_i} x_i \geq y_j, && \forall e_j \in U^- \\
& x_i \in \{0, 1\}, && \forall S_i \in \mathcal{S}^- \\
& y_j \in \{0, 1\}, && \forall e_j \in U^-
\end{aligned}
$$

Outline

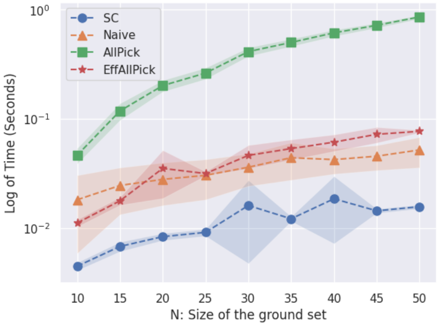
Highlighted Experiments – Resume Skills

Algorithm	Avg. Fairness Ratio	Avg. Cover Size
Opt-SC	0.48	3.32
Greedy-SC	0.55	3.42
Opt-FSC	**1.00**	3.75
EffAllPick	**1.00**	3.90

Output size



Time



Thank you!

Question?