UIC

Maximum Coverage with Fairness for Social Equity

Abolfazl Asudeh, Tanya Berger-Wolf, Bhaskar DasGupta, Anastasios Sidiropoulos

InDeX Lab
Innovative
Data
eXploration
Laboratory

Outline

- Reminder: max cover problem
- Application Demonstration
- Problem Formulation
- Experiment results
- Solution/Analysis
- Discussions

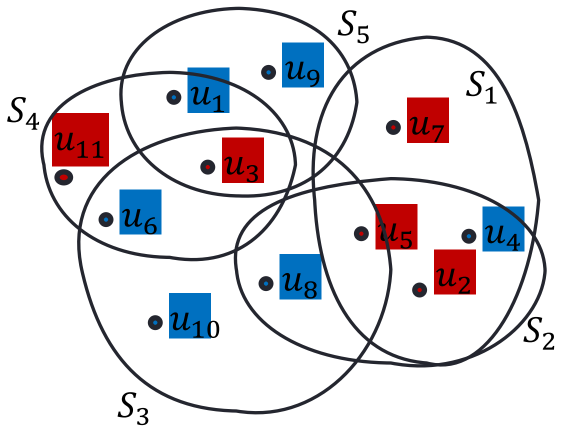
Reminder

Reminder: Max Cover problem

- Given:
  - A universe $\boldsymbol{U} = \{u_1, \ldots, u_n\}$ of *n* elements
  - A collection of sets $\mathbf{S} = \{S_1, \ldots, S_n\}$ where each set belongs to the powerset of *U*: $S_i \in 2^U$
  - A value *k*
- Objective:
  - Find *k* sets that cover maximum number of elements:

$$\underset{|\boldsymbol{S}'|=k}{\operatorname{argmax}} \bigcup_{S_i \in \boldsymbol{S}'} S_i$$

- $\boldsymbol{U} = \{u_1, \dots, u_{10}\}$
- $\boldsymbol{S} = \{S_1, \dots, S_5\}$
- $S_1 = \{u_2, u_4, u_5, u_7\}$
- $k = 3$



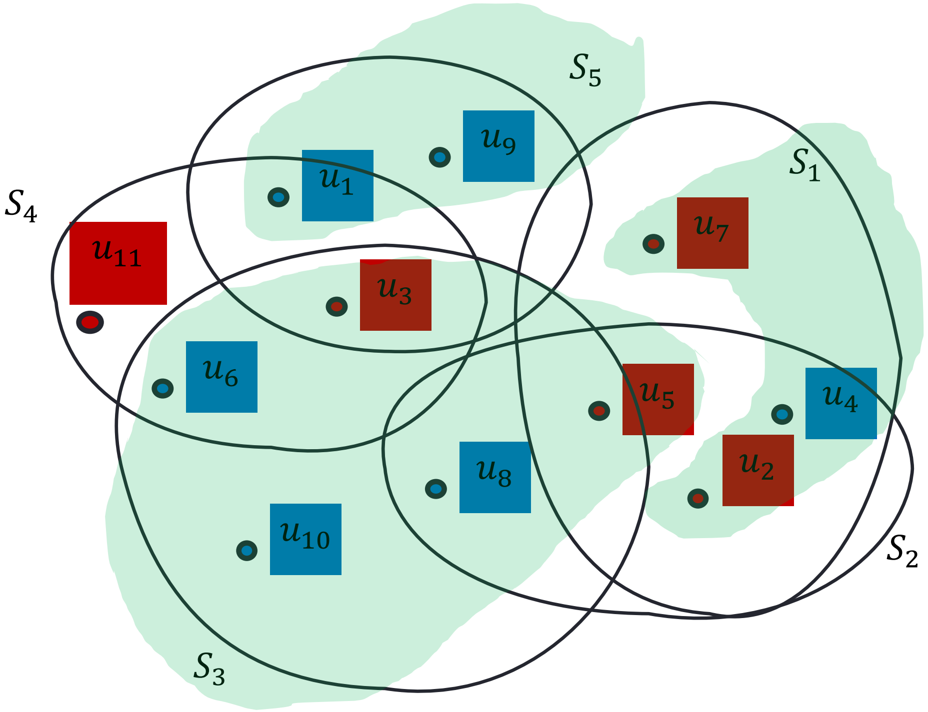
- Opt. answer:
  - S_1, S_3, S_5
  - Coverage=10

Max. cover is NP-complete

- A polynomial solution for max-cover would be a polynomial solution for **all** NP-complete problems

- The problem is (i) monotonic, (ii) submodular
  - The greedy solution achieves $\left(1 - \frac{1}{e}\right)$ approximation

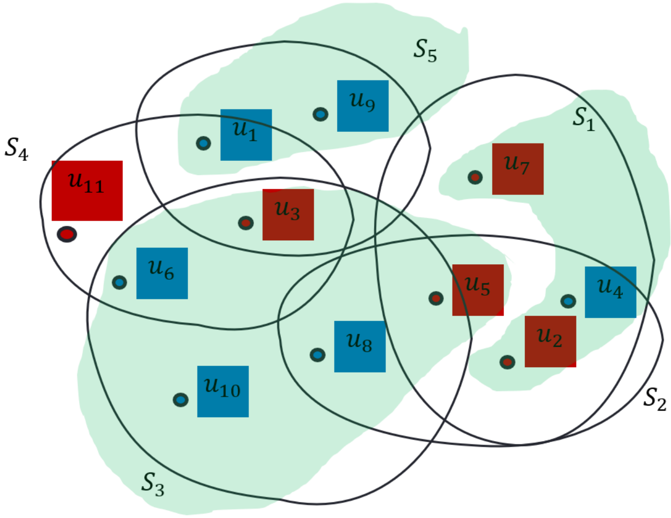
Greedy



Unfair coverage

- Coverage of blue = 6
- Coverage of red = 4

* each color shows a demographic group



(Diverse) Application Demonstration

Application1: Service/Facility Allocation

- One of the most common policy decisions made based on data is assigning services/facilities across different places.
    - Bus stops
    - School closing
    - Fire Station
    - Selecting hospitals for special treatments
- Objective: to maximize the number of people covered.
- Issue (historical biases, e.g. redlining): unfair coverage.

Service/Facility Allocation: Real examples from news

- NYC Bike sharing:
  - U: individuals
  - S: each bike station identifies a set, covering the individuals that are "close" to them
- Amazon same-day delivery:
  - U: individuals
  - S: each delivery location covers (the individuals in) a neighborhood

Application2: Data Integration

- Idea: combine multiple data sources to augment the power of any individual data source.
  - U: data points
  - S: data sources
- Objective: to collect (cover) the maximum number of data points.
- Issue: failing to include an adequate number of instances from minorities (biased datasets)

Application 3: Targeted Ad.

- Targeted advertising is popular in social media.
- Scenario:
  - A company wants to target its “potential customers”
    Needs to select a set of features (such as “single” or “college student”) that specify the groups of users to be targeted.
  - **U**: customers
  - **S**: relevant features, each showing a group of customers, having those features
- Objective: select k features that hit max customers
- Issue: racism in the advertisement

Running Example 1: COVID-19 Testing Facilities

- Providing proper testing facilities that effectively identify infected cases is critical for minimizing the spread of the Coronavirus.
    - Limited number of testing facilities
    - Goal: maximize coverage of ppl "close" to the facilities
- Issue:
    - The coronavirus testing locations heatmap in *city of Memphis* reveal that most screening happens in predominantly white and well-off suburbs, not in black-majority, lower-income neighborhoods

Running Example 2: Targeted Job Advertisement

- Employer in Linkedin:
  - Select k keywords (resume skills) to highlight in job advertisement
  - Goal: to attract the maximum number of applicants.

- Due to the underlying biases and false stereotypes, the company may end up with, for example, sexism in the advertisement
- A major concern in employment, the company would like to attract a diverse group of applicants.

Problem formulation

Fair Max Cover (FMC) problem

- Given:
  - A universe $\boldsymbol{U} = \{u_1, \dots, u_n\}$ of n elements. Each element belongs to a demographic group, identified by its color (it can be more than 2)
  - A collection of sets $\mathbf{S} = \{S_1, \dots, S_n\}$ where each set belongs to the powerset of U: $S_i \in 2^U$
  - A value k
- Objective:
  - Find k sets that cover maximum (weighted) number of elements such that the number of elements covered from each group is equal*

UIC InDeX Lab Innovative Data eXploration Laboratory

* A more general definition has been proposed in the paper

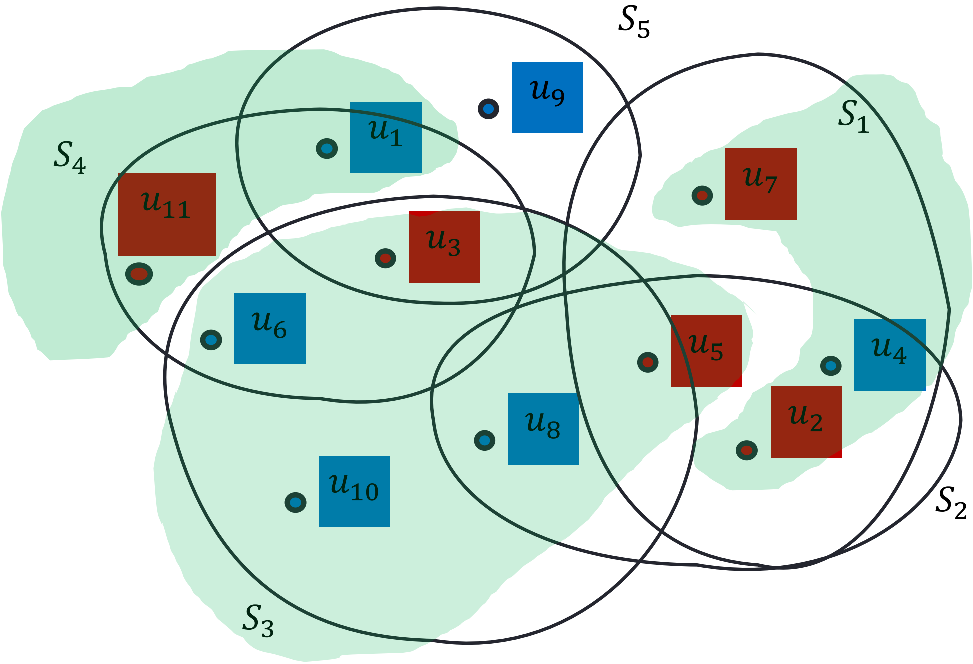
Approximation

$$\frac{1}{\epsilon} \leq \frac{E[cov1]}{E[cov2]} \leq \epsilon$$

- **disparity:** $\epsilon - 1$

FMC

- Coverage of blue = 5
- Coverage of red = 5



Experiments

Proof of Concept (Job Advertisement)

- n=1,986 individual resumes
- m=218 resume skills (keywords)
- k=5 keywords
- S: gender

- MC

Auditing, Job Scanning, Partnerships, Graphic Design, Drawing

Coverage: 1046

Disparity: 0.1

- FMC

Auditing, Coaching, Interviewing, Integrated Marketing, Organizational Development

Coverage: 1012

Disparity: 0

Proof of Concept (Covid-19 testing facilities)

- Individuals in City of Chicago $n \cong 2M$
- Each zip-code is a set, covering the ppl within 2 miles travel distance
- k=5 zip-codes
- S: race (black and white)

- MC (Greedy)

60613, 60625, 60636, 60642, 60651

Coverage: 790K

Disparity: 0.78

- FMC

60613, 60620, 60636, 60642, 60651, 60653

Coverage: 817K

Disparity: 0.01

Performance Evaluation

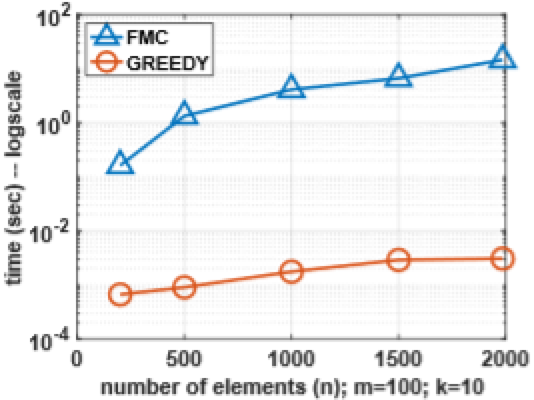


Figure 2: Varying number of items; time

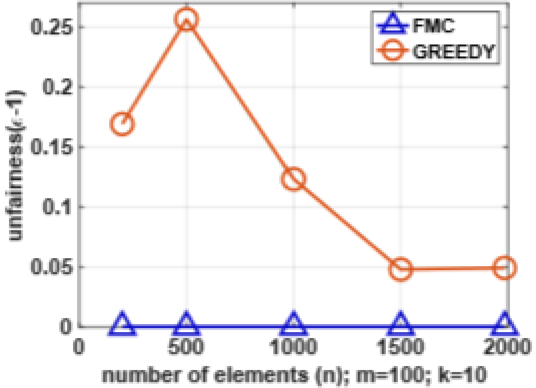


Figure 3: Varying number of items; fairness

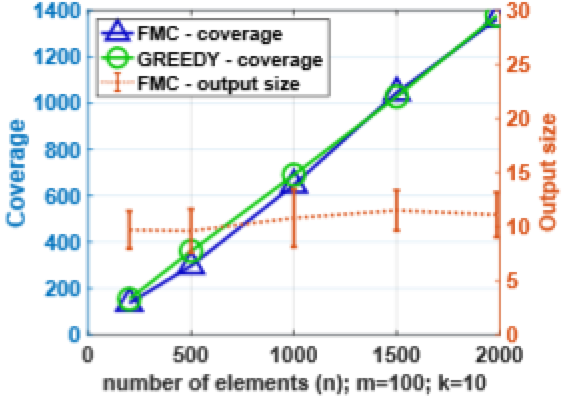


Figure 4: Varying number of items; coverage

Solution:
IP formulation, Randomized Algorithm, Approximation

High-level idea (LP-relaxation)

1. Formulate the problem as Integer Programming (IP)
   - $y_i = 1$ if set i is selected, 0 otherwise
   - $x_i = 1$ if element i is covered, 0 otherwise
2. Relax the integer (binary) variables to real values ([0,1])
   - $y_i \in [0,1]$ and $x_i \in [0,1]$
   - Let $y_1^*, \ldots, y_m^*$ and $x_1^*, \ldots, x_n^*$ be the output of the LP problem
3. $disp_{min} = \infty$
4. for *n* iterations do: //do multiple roundings and choose the best
   1. for i=1 to m do: set $y_i^+ = 0$ with probability y_i^* (, 1 otherwise)
   2. for i=1 to n do: set $x_i^+ = 1$ if at least one of the sets containing is selected ($y_j^+ = 1$)
   3. Compute disparity for x^+, and update $disp_{min}$ if a better solution has been discovered

UIC InDeX Lab Innovative Data eXploration Laboratory

LP Relaxation of IP formulation

maximize	$\sum_{i=1}^{n} w(u_i)x_i$	
subject to	$x_j \leq \sum_{u_j \in S_\ell} y_\ell$	for $j = 1, \ldots, n$
	$\sum_{\ell=1}^{m} y_\ell = k$	
	$\sum_{u_\ell \in C_i} x_\ell \geq k/\chi$	for $i \in \{1, \ldots, \chi\}$
	$\sum_{u_\ell \in C_i} x_\ell = \sum_{u_\ell \in C_j} x_\ell$	for $i, j \in \{1, \ldots, \chi\}, i < j$.[3]
	$0 \leq x_j \leq 1$	for $j = 1, \ldots, n$
	$0 \leq y_\ell \leq 1$	for $\ell = 1, \ldots, m$

Switch to notes

InDeX Lab
Innovative
Data
eXploration
Laboratory

Thank you, Discussions