Section 20.2: Maximum Coverage

Problem Definition

Input: Subsets $T_1, T_2, \ldots, T_m$ of a ground set U, and a positive integer k.

Output: set $K \subseteq \{1, 2, \ldots, m\}$ of k indices to maximize the coverage:

$$\underbrace{f_{cov}(K) := \left| \bigcup_{i \in K} T_i \right|}_{\text{\# of distinct elements covered by the } T_i\text{'s}}$$

Quiz #1

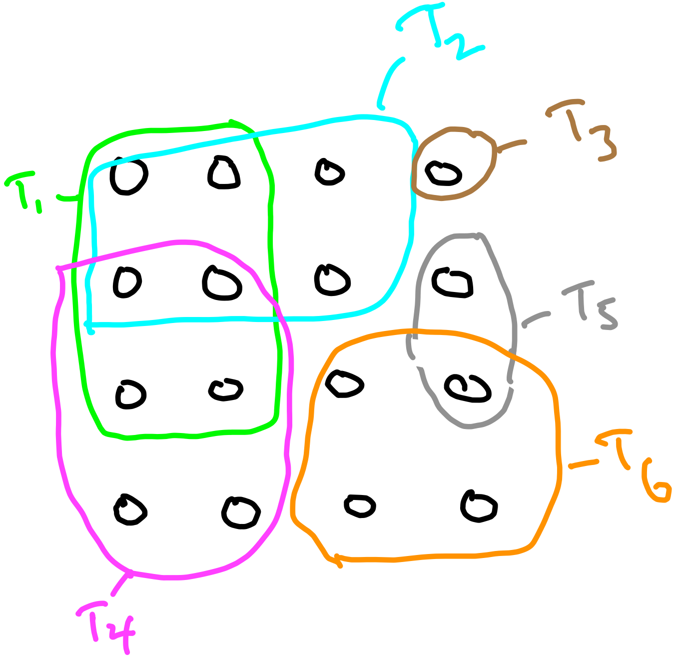
What is the largest coverage achieved by 4 of the subsets?

(a) 13

(b) 14

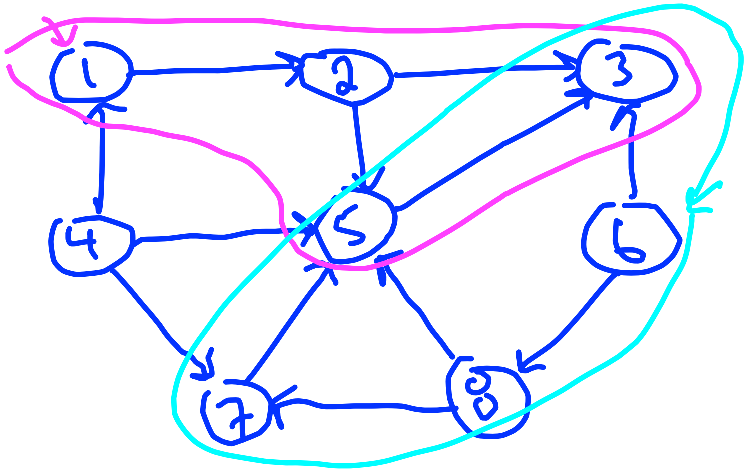
(c) 15

(d) 16



Further Applications

- choosing locations for firehouses, schools, etc. to maximize # of people close to one of them
- recruitment to maximize participation



A Greedy Algorithm

$f_{cov}(K) = |\bigcup_{i \in K} T_i|$

GreedyCoverage

$K := \emptyset$ [indices of chosen subsets]

for $j = 1$ to k:

$i^* = \underset{i=1}{\overset{m}{\operatorname{argmax}}} \left[\underbrace{f_{cov}(K \cup \{i\}) - f_{cov}(K)}_{\text{increase in coverage from } T_i} \right]$ [greedily increase coverage]

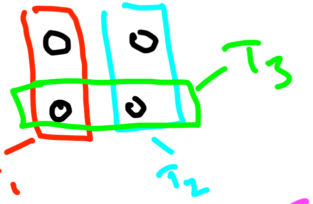
$K := K \cup \{i^*\}$

return K

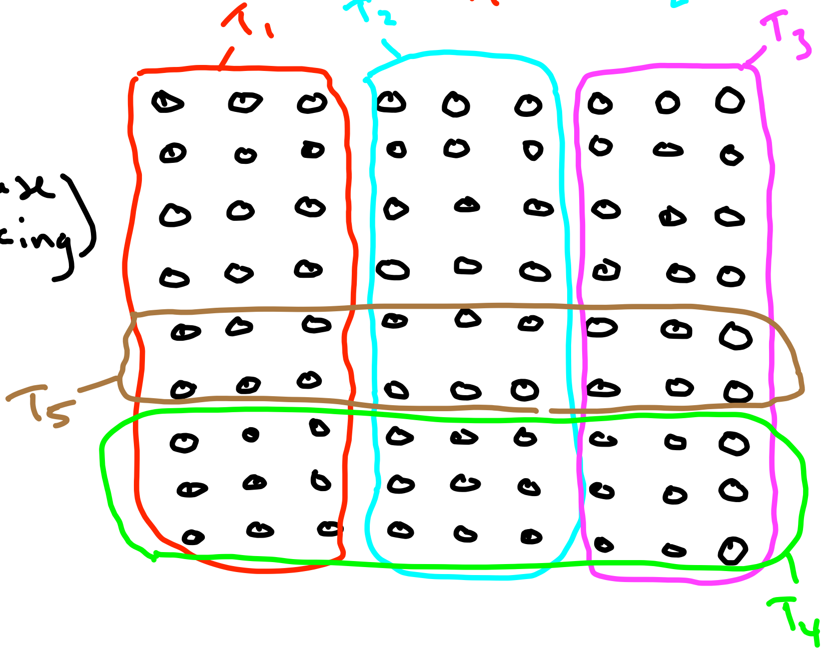
$s = \max_{i=1}^{m} |T_i|$

Running time: $O(kms)$

Quiz #2



For $k=3$, what are (i) the maximum-possible coverage; and (ii) smallest coverage that might be achieved by GreedyCoverage (w/ worst-case tie-breaking)



a) 72 and 60

b) 81 and 57

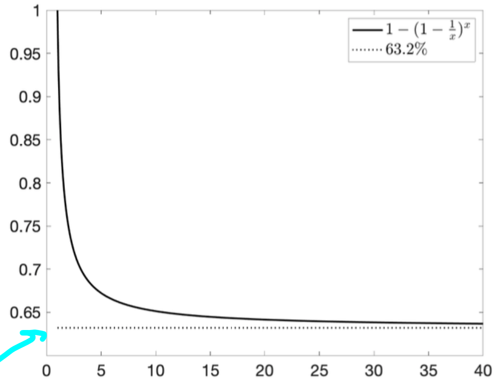
c) 81 and 60

d) 81 and 64

Bad Examples for GreedyCoverage

So: $k=2 \Rightarrow$ GreedyCoverage might only get 75%
$k=3 \Rightarrow$ " " " " $\approx 70.4\%$

In general: GreedyCoverage might achieve coverage only $1-(1-1/k)^k$ times the maximum possible.



$\approx 63.2\%$

$$\left[1-\left(1-\frac{1}{k}\right)^k \approx 1-\left(e^{-\frac{1}{k}}\right)^k = 1-\frac{1}{e} \approx 0.632\right]$$

Approximate Correctness

Theorem: coverage of GreedyCoverage $\geq \left(1 - \left(1 - \frac{1}{k}\right)^k\right) \cdot$ max-possible coverage

Key Lemma: Let C^* = max-possible coverage.
Each subset chosen by GreedyCoverage covers at

$$\frac{1}{k}\left(C^* - \begin{array}{c}\text{\# of elements}\\ \text{already covered}\end{array}\right)$$

Proof of Key Lemma

Notation: $\hat{K}$ = set of k indices (e.g., of an optimal solution). coverage $= \hat{C}$

Fix: Iteration j, first $j-1$ subsets chosen by Greedy Coverage

Claim: $\sum_{i \in \hat{K}} \left(\text{coverage increase from } T_i\right) \geq \hat{C} - \#\text{ of elements already covered}.$

↖ area of green region wedged in between

$\Rightarrow \max_{i \in \hat{K}} \left(\text{coverage increase from } T_i\right) \geq \frac{1}{k}\left(\hat{C} - \#\text{ of elements already covered}\right)$

[maximum is at least average]

$\Rightarrow$ Greedy Coverage does at least as well

QED!

Proof of Approximate Correctness

Notation: C^* = max-possible coverage, C_j = coverage of GreedyCoverage's first j subsets.

[Key Lemma: $\forall j$, $C_j - C_{j-1} \geq \frac{1}{k}(C^* - C_{j-1})$.]

[j=k] $C_k - C_{k-1} \geq \frac{1}{k}(C^* - C_{k-1}) \Leftrightarrow C_k \geq \frac{C^*}{k} + \left(1-\frac{1}{k}\right)C_{k-1}$

[j=k-1] $C_{k-1} \geq \frac{C^*}{k} + \left(1-\frac{1}{k}\right)C_{k-2}$

Geometric series ($r \neq 1$)

$1 + r + r^2 + \dots + r^\ell = \frac{1-r^{\ell+1}}{1-r}$

$\Rightarrow C_k \geq \frac{C^*}{k}\left(1 + \left(1-\frac{1}{k}\right)\right) + \left(1-\frac{1}{k}\right)^2 C_{k-2}$

[j=k-2] $\Rightarrow C_k \geq \frac{C^*}{k}\left(1 + \left(1-\frac{1}{k}\right) + \left(1-\frac{1}{k}\right)^2\right) + \left(1-\frac{1}{k}\right)^3 C_{k-3}$

[j=1] $\Rightarrow C_k \geq \frac{C^*}{k}\left(1 + \left(1-\frac{1}{k}\right) + \left(1-\frac{1}{k}\right)^2 + \dots + \left(1-\frac{1}{k}\right)^{k-1}\right) + \cancel{\left(1-\frac{1}{k}\right)^k C_0}$ $= 0$

$= k\,\frac{1-\left(1-1/k\right)^k}{\cancel{1-\left(1-1/k\right)}}$

$\geq C^*\left(1 - \left(1-\frac{1}{k}\right)^k\right)$. QED!