# Set Cover

Tao Xiao

# Mass Mailing

- Say you'd like to send some message to a large list of people (e.g. all campus)
- There are some available mailing-lists, however, the moderator of each list charges $1 for each message sent
- You'd like to find the smallest set of lists that covers all recipients

# SET-COVER

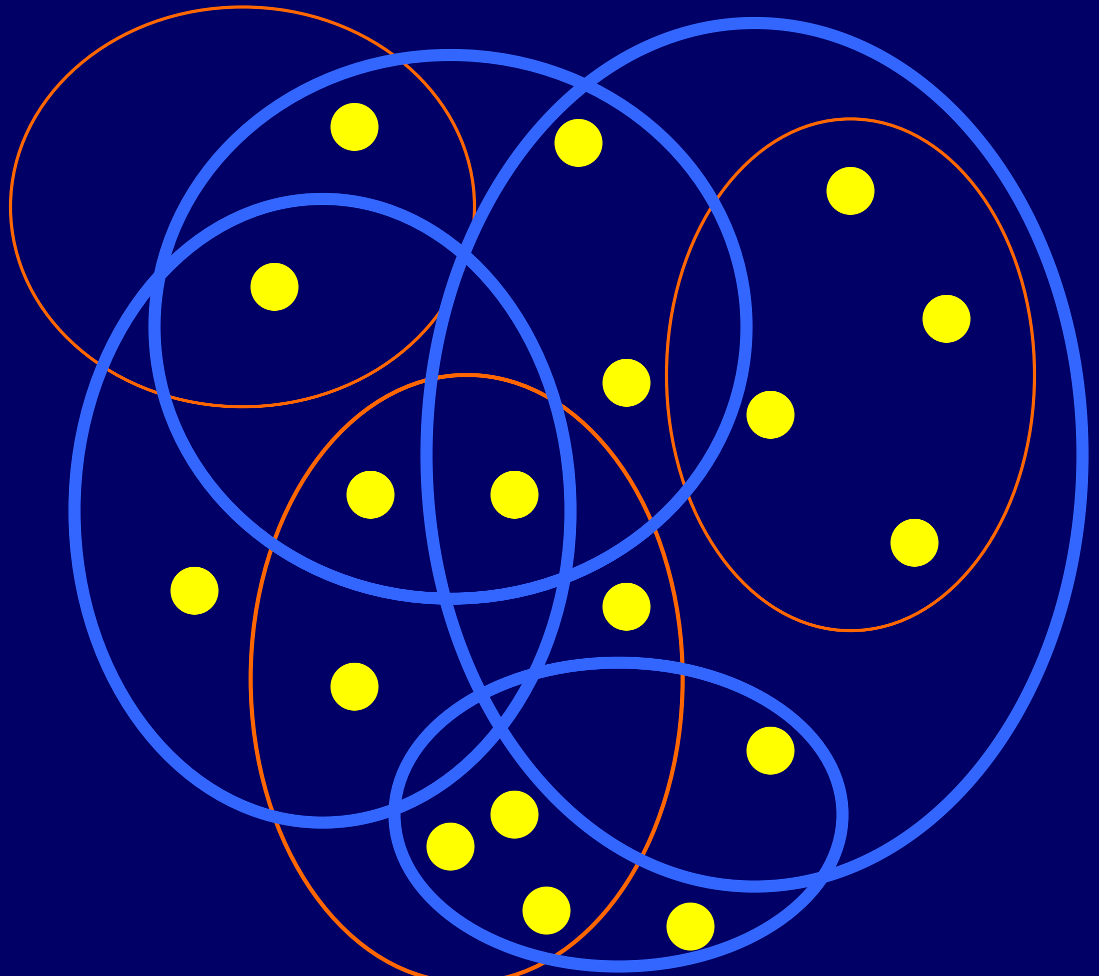
- Instance: a finite set X and a family F of subsets of X, such that

$$X = \bigcup_{S \in F} S$$

- Problem: to find a set $C \subseteq F$ of minimal size which *covers* X, i.e -

$$X = \bigcup_{S \in C} S$$

# SET-COVER: Example

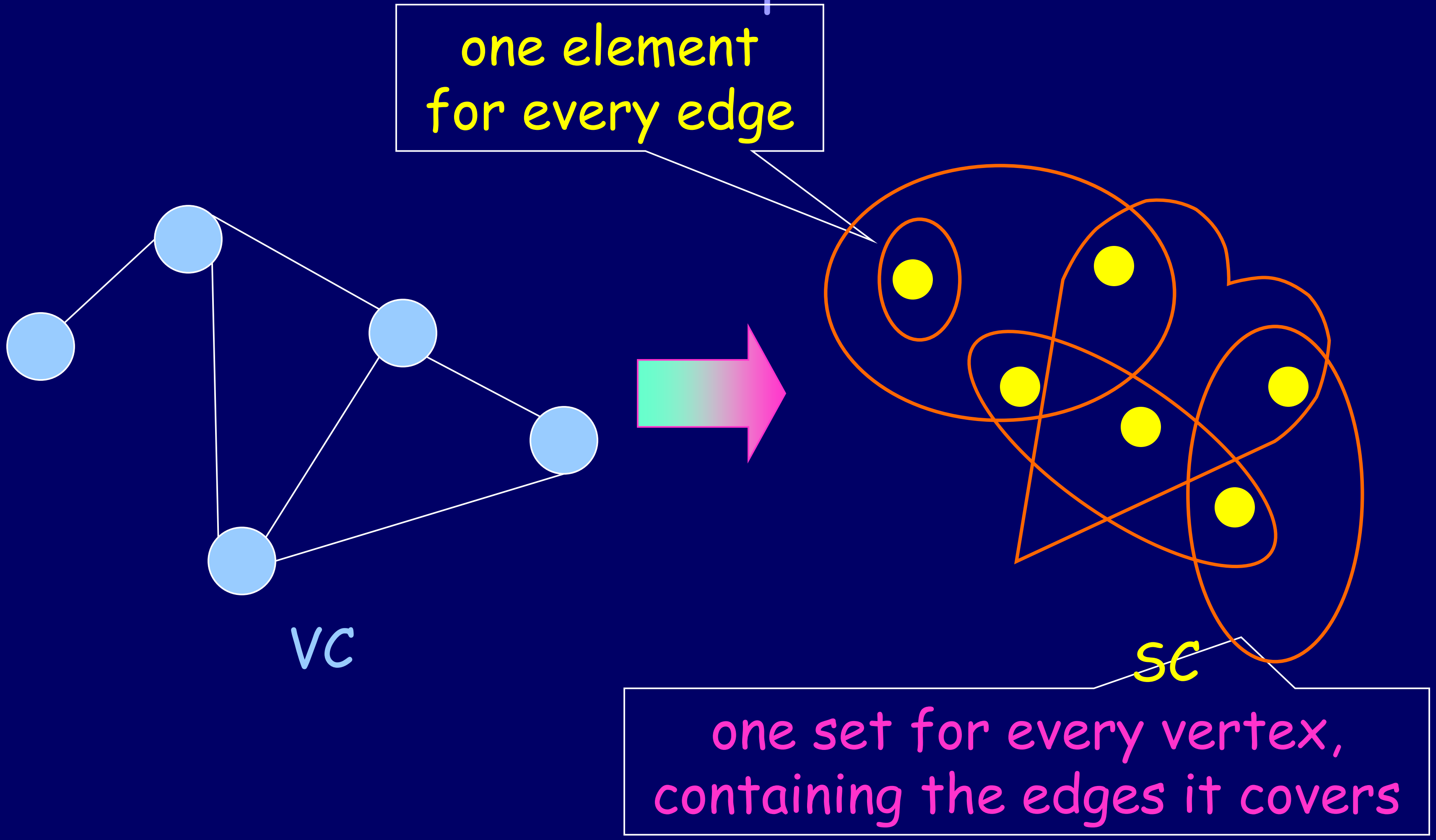


# SET-COVER is NP-Hard

Proof: Observe the corresponding decision problem.

- Clearly, it's in NP (Check!).
- We'll sketch a reduction from (decision) VERTEX-COVER to it:

# VERTEX-COVER $\leq_p$ SET-COVER



# The Greedy Algorithm

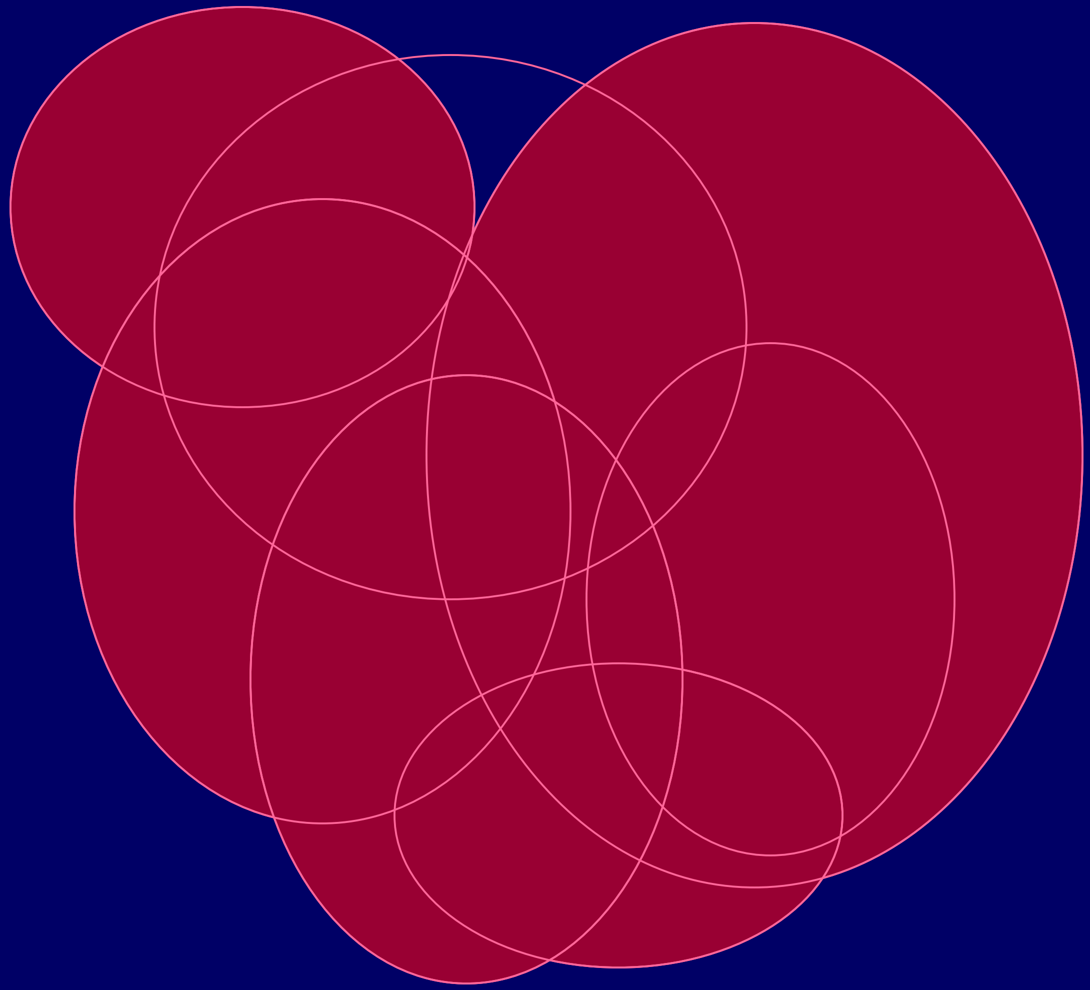
- $C \leftarrow \phi$
- $U \leftarrow X$
- **while** $U \neq \phi$ **do**
  - select $S \in F$ that maximizes $|S \cap U|$ $O(|F| \cdot |X|)$
  - $C \leftarrow C \cup \{S\}$
  - $U \leftarrow U - S$
- return $C$

$\min\{|X|, |F|\}$

# Demonstration

compare to the optimal cover

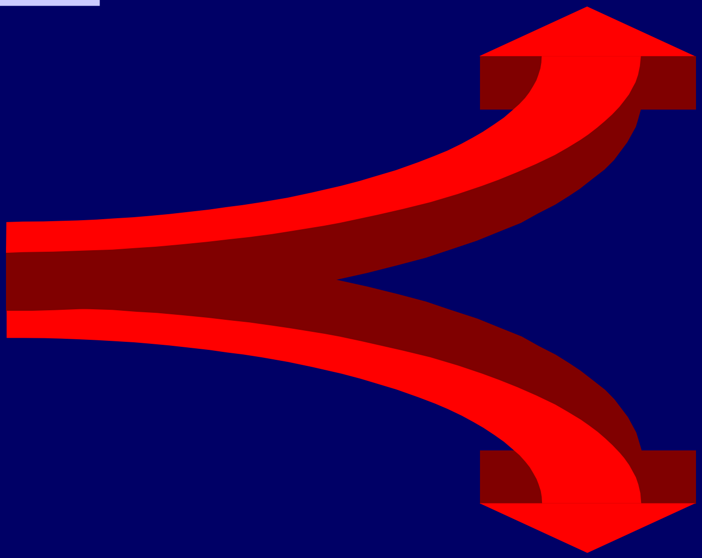
5



# Is Being Greedy Worthwhile?

## How Do We Proceed From Here?

- We can easily bound the approximation ratio by log n.
- A more careful analysis yields a tight bound of ln n.



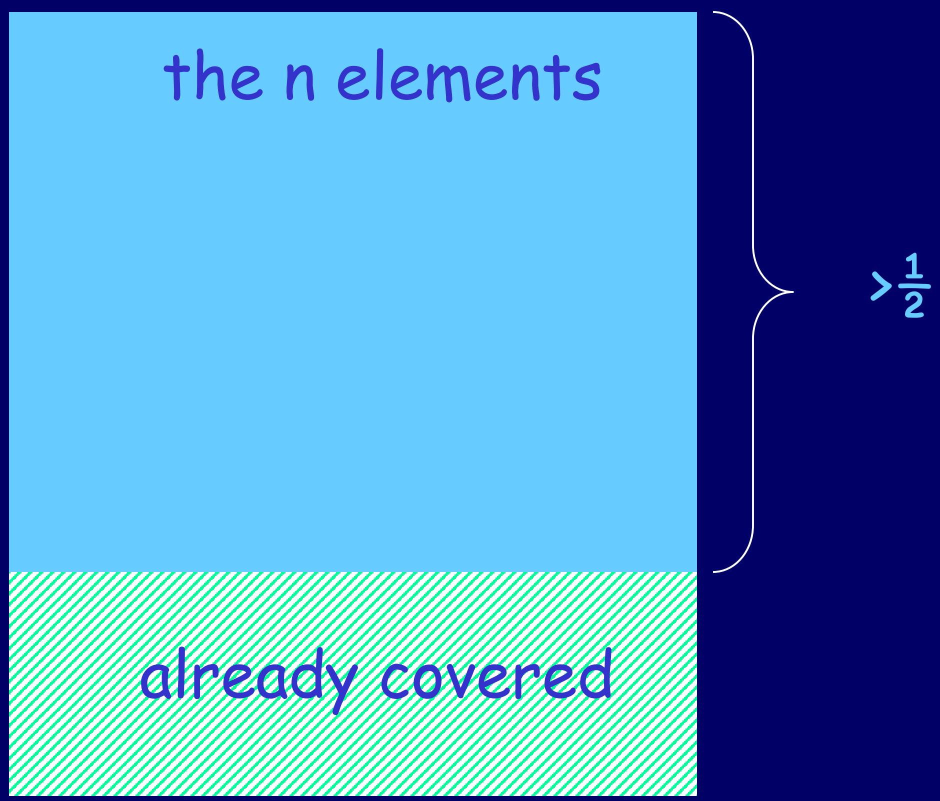
# The Trick

- We'd like to compare the number of subsets returned by the greedy algorithm to the optimal
- The optimal is unknown, however, if it consists of k subsets, then any part of the universe can be covered by k subsets!
- Which is exactly what the next 3 distinct arguments take advantage of

# Loose Ratio-Bound

Claim: If ∃ cover of size k, then after k iterations the algorithm have covered at least ½ of the elements

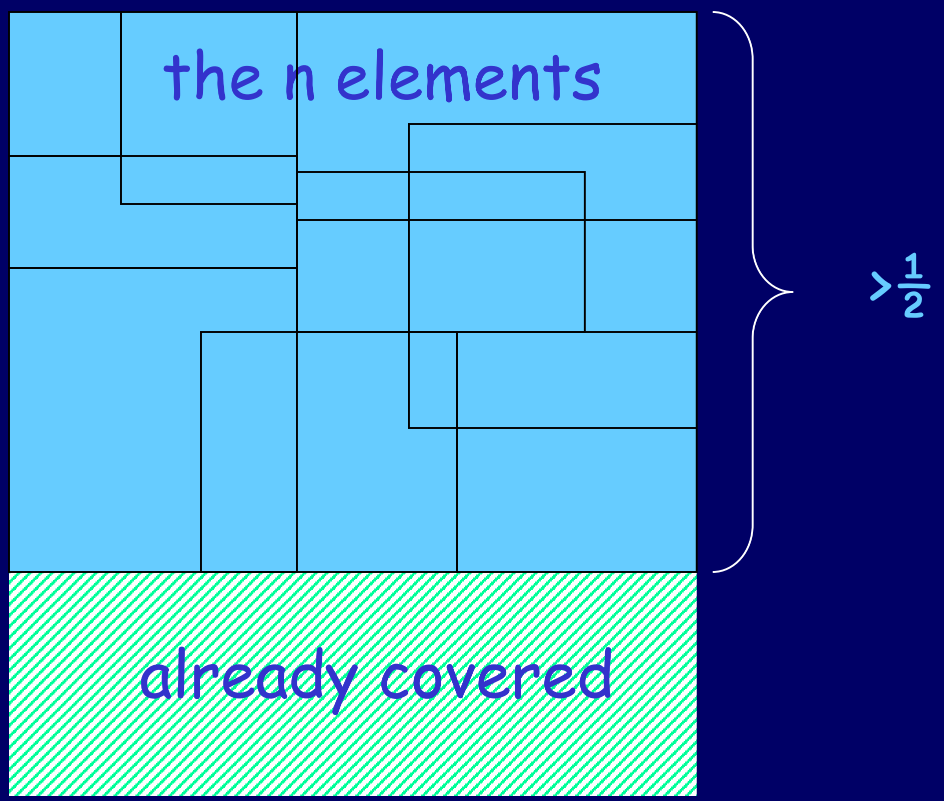
Suppose it doesn't and observe the situation after k iterations:



# Loose Ratio-Bound

<u>Claim</u>: If ∃ cover of size k, then after k iterations the algorithm have covered at least ½ of the elements

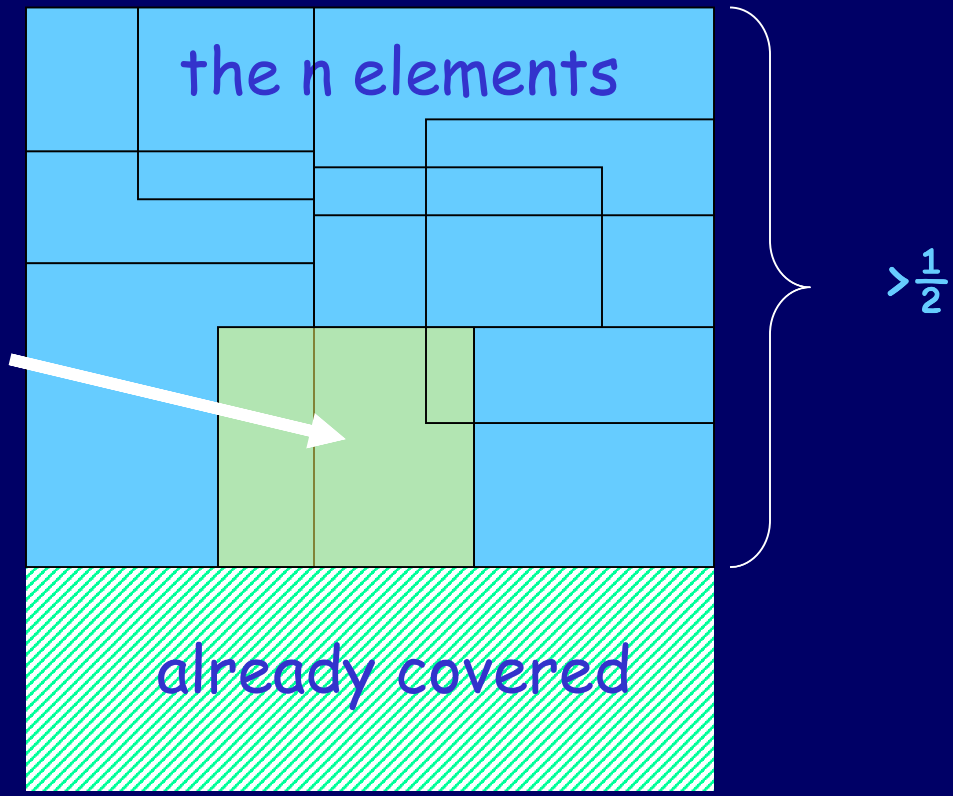
Since this part → can also be covered by k sets...



# Loose Ratio-Bound

Claim: If ∃ cover of size k, then after k iterations the algorithm have covered at least ½ of the elements

there must be a set not chosen yet, whose size is at least ½n·1/k

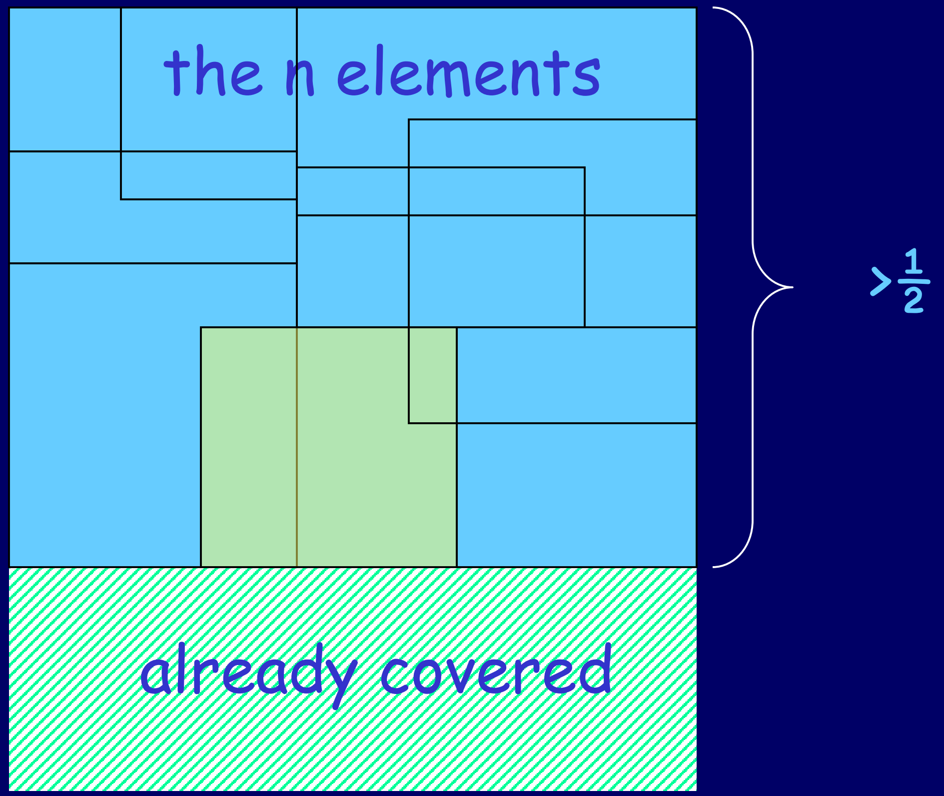


# Loose Ratio-Bound

Claim: If $\exists$ cover of size k, then after k iterations the algorithm have covered at least ½ of the elements

and the claim is proven!

Thus in each of the k iterations we've covered at least $\frac{1}{2}n \cdot 1/k$ new elements



# Loose Ratio-Bound

Claim: If $\exists$ cover of size $k$, then after $k$ iterations the algorithm covered at least $\frac{1}{2}$ of the elements.

Therefore after klogn iterations (i.e - after choosing klogn sets) all the n elements must be covered, and the bound is proved. $\blacksquare$

# Better Ratio Bound

Let $S_1, \ldots, S_t$ be the sequence of sets outputted by the greedy algorithm. Let, for $0 \le i \le t$

$$U_i \equiv X - \bigcup_{j=1}^{i} S_j$$

Since, for every $i$, $U_i$ can be covered by $k$ sets, it follows

$$|U_{i+1}| = |U_i - S_{i+1}| \le |U_i| \frac{k-1}{k}$$

# Better Ratio Bound

$$|U_{i+1}| = |U_i - S_{i+1}| \leq |U_i| \frac{k-1}{k}$$

Hence, for any $0 \leq i < j \leq t$

$$|U_j| \leq |U_i| \cdot \left(\frac{k-1}{k}\right)^{j-i}$$

Which implies that for every $i$

$$|U_{i \cdot k}| \leq |U_0| \cdot \left(\frac{k-1}{k}\right)^{i \cdot k} \leq |X| \cdot \frac{1}{e^i}$$

Therefore, $t \leq k \ln(n) + 1$

# Tight Ratio-Bound

Claim: The greedy algorithm approximates the optimal set-cover to within a factor

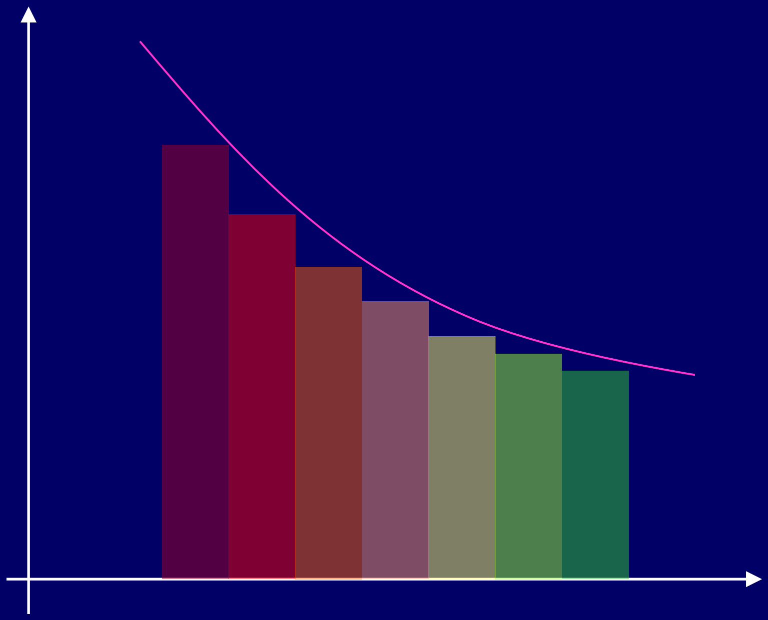
$$H(\max\{ |S| : S \in F \})$$

Where $H(d)$ is the $d$-th harmonic number:

$$H(d) \stackrel{\text{def}}{=} \sum_{i=1}^{d} \frac{1}{i}$$

# Tight Ratio-Bound

$$\sum_{k=1}^{n}\frac{1}{k}=\sum_{k=2}^{n}\frac{1}{k}+1\leq\int_{1}^{n}\frac{1}{x}dx+1=\ln n+1$$

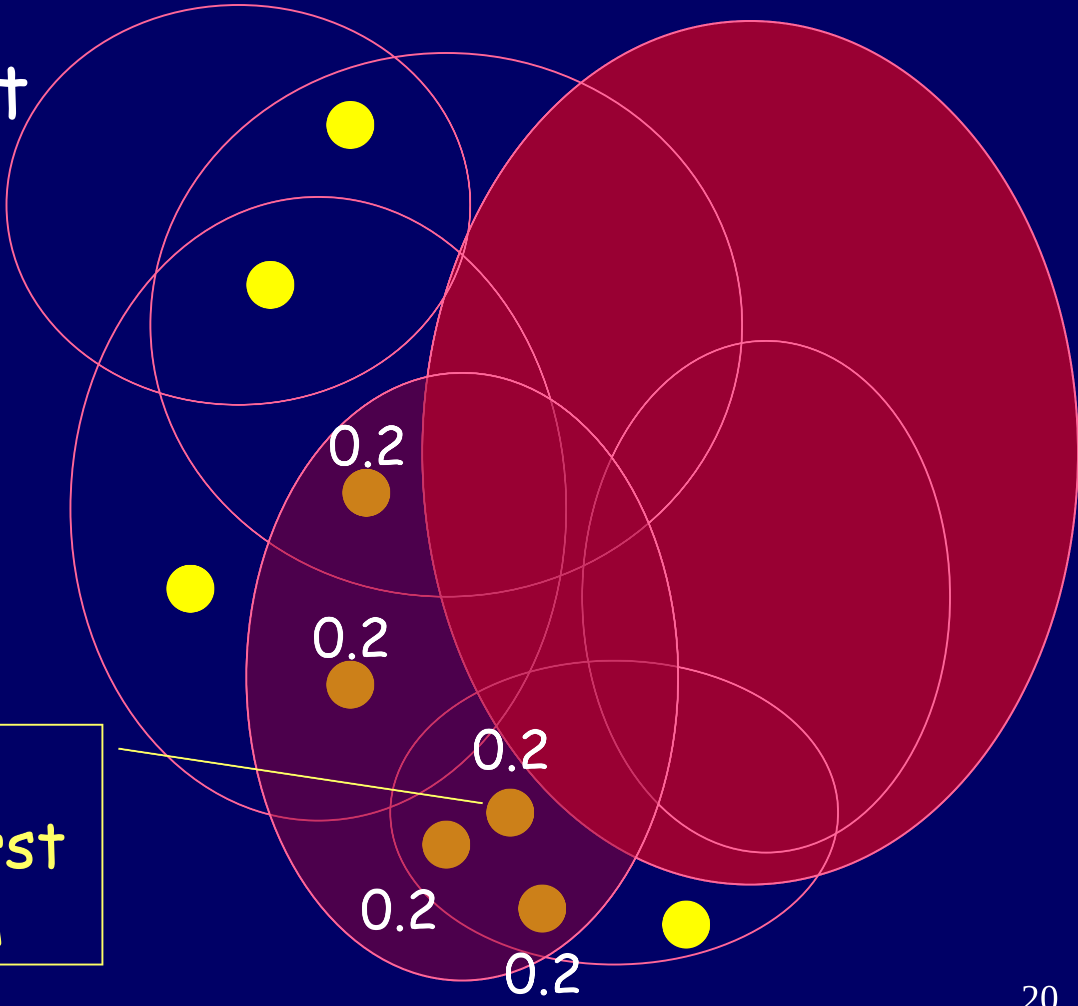


# Claim's Proof

Charge $1 for each set

Split cost between covered elements

Bound from above the total fees paid

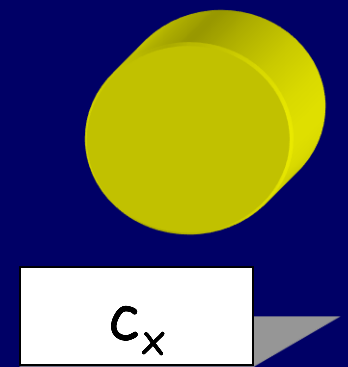



each recipient pays the fractional cost for the first mailing-list it appears in

# Analysis

- Thus, every element $x \in X$ is charged

$$c_x \overset{\text{def}}{=} \frac{1}{|S_i - (S_1 \cup \ldots \cup S_{i-1})|}$$




- Where $S_i$ is the first set that covers $x$.

# Lemma

number of members of $S$ still uncovered after $i$ iterations

Lemma: For every $S \in F$

$$\sum_{x \in S} c_x \le H(|S|)$$

Proof: Fix an $S \in F$. For any $i$, let

$$u_i \overset{\text{def}}{=} |S - (S_1 \cup \ldots \cup S_i)|$$

$\forall 1 \le i \le k$ : $S_i$ covers $u_{i-1} - u_i$ elements of $S$

Let $k$ be the smallest index, s.t. $u_k = 0$

# Lemma

$$\sum_{x\in S} c_x = \sum_{i=1}^{k} \frac{u_{i-1}-u_i}{|S_i-(S_1\cup\ldots\cup S_{i-1})|} \le \sum_{i=1}^{k} \frac{u_{i-1}-u_i}{|S-(S_1\cup\ldots\cup S_{i-1})|} =$$

sum charges

else greedy strategy would have taken $S$ instead of $S_i$

definition of $u_{i-1}$

$$\sum_{i=1}^{k} \frac{u_{i-1}-u_i}{u_{i-1}} \le \sum_{i=1}^{k} H(u_{i-1})-H(u_i) = H(u_0)-H(u_k) = H(|S|)$$

$\forall a<b$

$H(b)-H(a) = \frac{1}{a+1}+\ldots+\frac{1}{b} \ge \frac{b-a}{b}$

Telescopic sum

$H(u_k)=H(0)=0$

$H(u_0)=H(|S|)$

# Analysis

Now we can finally complete our analysis:

$$|C| = \sum_{x \in X} c_x \le \sum_{S \in C^*} \sum_{x \in S} c_x \le |C^*| \cdot H(\max\{|S| : S \in F\})$$

■