# New and Improved Bounds for the Minimum Set Cover Problem

Rishi Saket (IBM)

Maxim Sviridenko (University of Warwick)

# Minimum Set Cover Problem

- We are given the ground set $\{1, \ldots, n\} = [n]$ and $m$ subsets $S_j \subseteq [n]$ for $j = 1, \ldots, m$.
- Each set $S_j$ has an associated weight $w_j \geq 0$.
- The goal is to choose a collection of sets indexed by $\mathcal{C} \subseteq \{1, \ldots, m\} = [m]$ such that $[n] = \cup_{j \in \mathcal{C}} S_j$ and minimize $\sum_{j \in \mathcal{C}} w_j$.
- Let $\Delta = \max_{j \in [m]} |S_j|$ be the maximal cardinality of a set in the instance. For each element $i \in [n]$, let $k_i = |\{S_j : i \in S_j, j \in [m]\}|$ be the number of sets in the instance containing the element $i \in [n]$ and let $k = \max_{i \in [n]} k_i$.

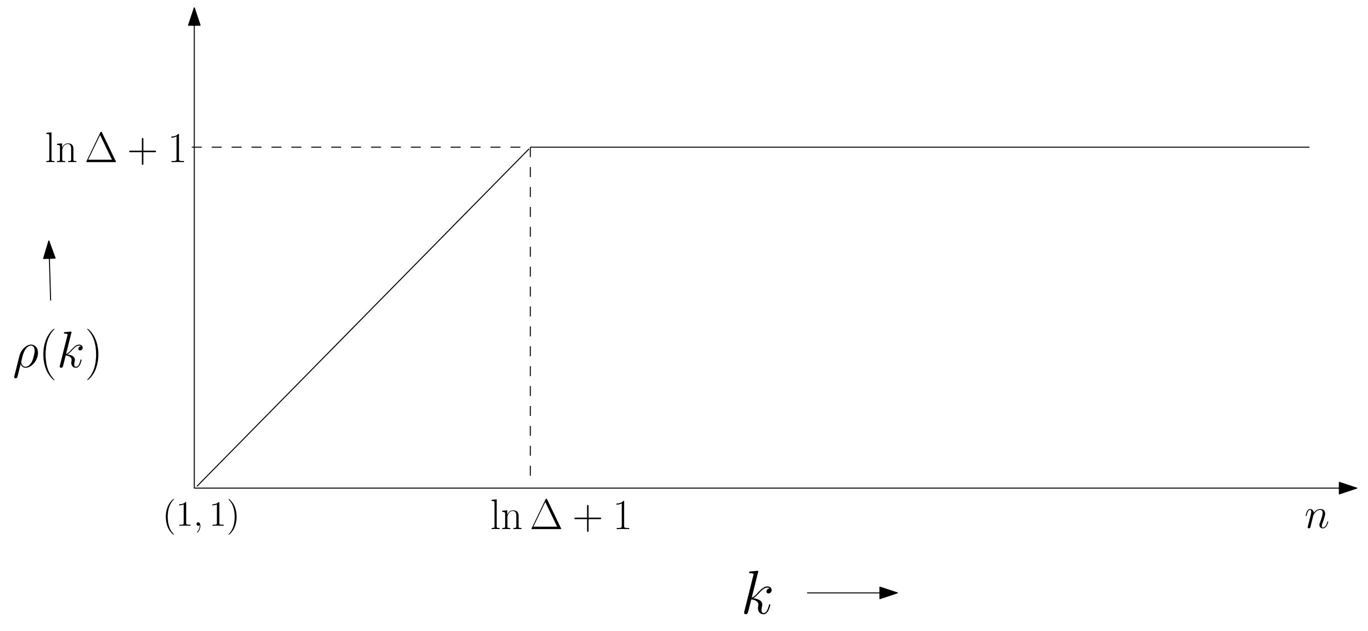
# Very Short History

- The natural greedy algorithm has performance guarantee $\ln \Delta + 1$ due to Johnson [1974], Lovasz [1975], Chvatal [1979].
- For any $\epsilon > 0$, we cannot do better than $(1 - \varepsilon) \ln n$ unless $NP \subseteq DTIME(n^{\log n})$ due to Feige (1998) ($\Delta, k \approx n$ in the reduction).

# Very Short History

- Another well-known type of algorithms (primal-dual or local ratio) has performance guarantee $k$ due to Hochbaum [1982], Bar-Yehuda and Even [1981].
- Khot and Regev [2003] showed that there is no $k - \varepsilon$ approximation algorithm under UGC for constant $k$.

# Nevertheless

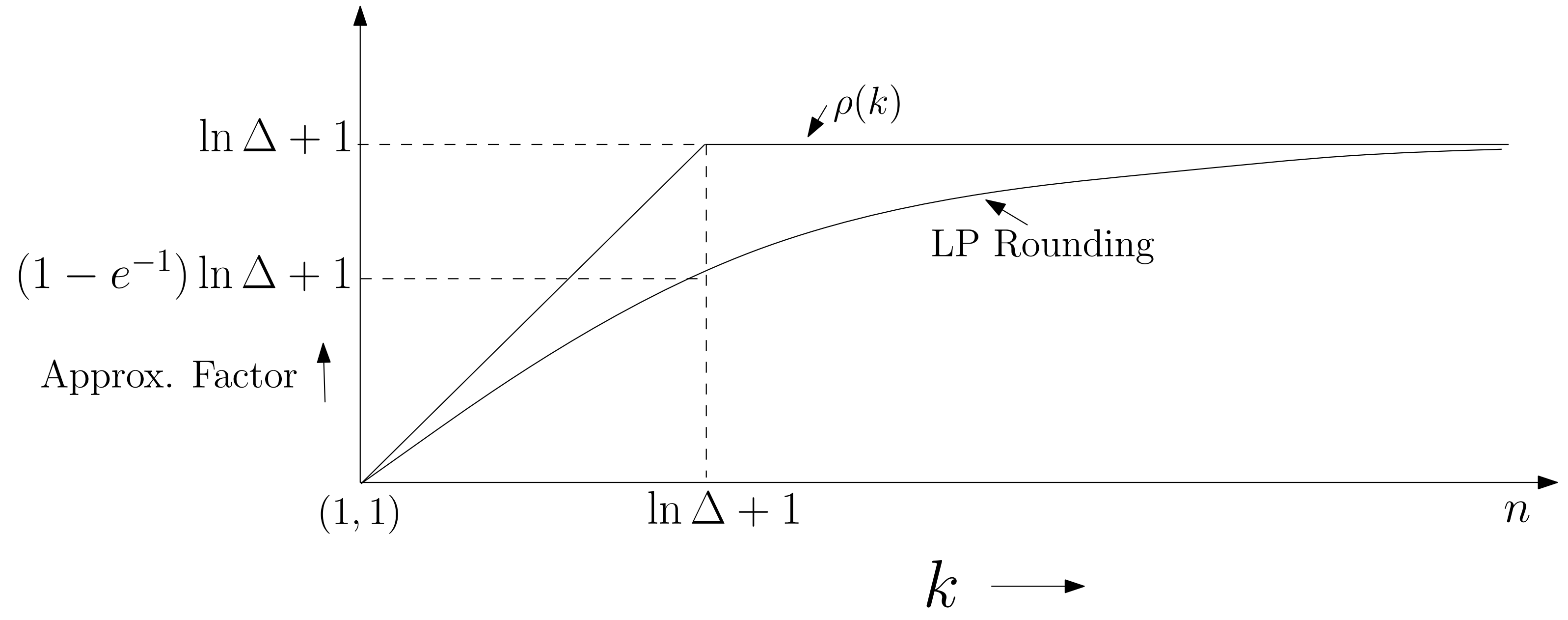



- Consider $\rho(k) = \min\{k, \ln\Delta + 1\}$.
- That is the performance guarantee of classical algorithms as a function of $k$.

# Our Results

- Randomized LP Rounding approximation algorithm with performance guarantee $R(k) = (k-1)(1 - e^{-\frac{\ln \Delta}{k-1}}) + 1$.
- Note, that $R(k) < \rho(k) = \min\{k, \ln \Delta + 1\}$ for all $k$ and $R(k) \approx \rho(k)$ when $k << \ln \Delta$ or $k >> \ln \Delta$.
- In particular, when $k = \ln \Delta + 1$, our algorithm has performance guarantee $(1 - e^{-1}) \ln \Delta + 1$.
- For $k = \theta(\log \Delta)$, we show an LP integrality gap of $k(1 - e^{-\frac{\ln \Delta}{k}} - \delta)$ for any constant $\delta > 0$.

# Our Results



$\rho(k)$
$\ln\Delta + 1$
LP Rounding
$(1 - e^{-1})\ln\Delta + 1$
Approx. Factor
$(1, 1)$
$\ln\Delta + 1$
$n$
$k$

# Summary of known hardness factors

| Range of $k$ | Hard. Factor | Assumption | Reference |
|---|---|---|---|
| $k$: arb. large const. | $k - \varepsilon$ | UGC | KR[2003] |
| $O((\log\log\Delta)^{1/c})$ | $k - 1 - \varepsilon$ | $n^{O(\log\log n)}$ | DGKR[2005] |
| $O((\log\Delta)^{1/c})$ | $k/2 - \varepsilon$ | $n^{O(\log\log n)}$ | DGKR[2005] |
| $k = \theta(\log\Delta)$ | $\Omega(\log^{1-\varepsilon}\Delta)$ | $n^{\mathrm{poly}(\log n)}$ | KS[2008] |
| $k = \theta(\log\Delta)$ | $\Omega\left(\frac{\log\Delta}{(\log\log\Delta)^2}\right)$ | $n^{\mathrm{poly}(\log n)}$ | This work. |
| $k = \Omega((\log\Delta)^c)$ | $\Omega(\log\Delta)$ | $n^{O(\log\log n)}$ | LY[1994] |
| $k = \Omega(2^{\log^{1-\varepsilon}\Delta})$ | $\Omega(\log\Delta)$ | $P \neq NP$ | RS[2007] |
| $k = \Omega(\Delta^\gamma)$ | $(1-\varepsilon)\ln\Delta$ | $n^{O(\log\log n)}$ | Feige[1998] |

# Standard LP

$$\min \sum_{j \in [m]} w_j x_j, \tag{1}$$

$$\sum_{j : i \in S_j} x_j \geq 1, \qquad \forall i \in [n], \tag{2}$$

$$x_j \geq 0, \qquad \forall j \in [m]. \tag{3}$$

Let $LP^*$ be the optimal value of the linear programming relaxation and $x_j^*, j \in [m]$ be the optimal fractional solution found by the LP solver.

# Randomized Rounding

- Let $\alpha = 1 - e^{-\frac{\ln \Delta}{k-1}}$. Define $p_j = \min\{1, \alpha k \cdot x_j^*\}$ for each set $S_j$, $j \in [m]$.
- Choose the set $S_j$ with probability $p_j$ independently at random. Let $R_1$ be the indices of sets chosen by our random procedure.
- Let $I^r$ be the set of the elements that are still not covered.
- Each element in $I^r$ chooses the cheapest set in our instance that covers it. Let $R_2$ be the set of indices of such sets covering $I_r$.
- Our algorithm outputs $R_1 \cup R_2$ as the final solution.

# Analysis

- $E[\sum_{j\in R_1} w_j] = \sum_{j\in[m]} w_j p_j \leq k(1-e^{-\frac{\ln\Delta}{k-1}})LP^*$.
- For each $i \in [n]$, let $j_i$ be the index such that $w_{j_i} = \min_{j\in[m]:i\in S_j} w_j$ and $W = \sum_{i\in[n]} w_{j_i}$. Then

$$
\begin{aligned}
W &= \sum_{i\in[n]} w_{j_i} \leq \sum_{i\in[n]} w_{j_i} \sum_{j:i\in S_j} x_j^* \\
&\leq \sum_{i\in[n]} \sum_{j:i\in S_j} w_j x_j^* \leq \Delta \cdot LP^*.
\end{aligned}
$$

- We estimate $Pr[i \in I^r]$. If $p_j = 1$ for at least one set such that $i \in S_j$ then $Pr[i \in I^r] = 0$.

# Analysis

Otherwise, $p_j = \alpha k \cdot x_j^*$ for all sets $S_j$ such that $i \in S_j$ and

$$
\begin{aligned}
Pr[i \in I^r] &= \prod_{j|i \in S_j} (1 - p_j) \leq \left(1 - \frac{\sum_{j|i \in S_j} p_j}{k_i}\right)^{k_i} \\
&\leq \left(1 - \frac{\sum_{j|i \in S_j} p_j}{k}\right)^{k} \\
&= \left(1 - \frac{\sum_{j|i \in S_j} \alpha k \cdot x_j^*}{k}\right)^{k} \\
&\leq (1 - \alpha)^k = \frac{1}{\Delta^{k/(k-1)}}.
\end{aligned}
$$

# Analysis

- Therefore, by linearity of expectation, the expected weight of the sets in $R_2$ can be estimated above by $\sum_{i=1}^{n} w_{j_i} Pr[i \in I^r] \leq W/\Delta^{k/(k-1)} \leq LP^*/\Delta^{1/(k-1)}$.
- Overall, the expected cost of the approximate solution is upper bounded above by

$$\left( k(1 - e^{-\frac{\ln \Delta}{k-1}}) + \frac{1}{\Delta^{1/(k-1)}} \right) LP^*$$
$$= \left( (k-1)(1 - e^{-\frac{\ln \Delta}{k-1}}) + 1 \right) LP^*$$

# Integrality Gap Instance

- Given a ground set of $n$ elements and $m = n^{\varepsilon}$ sets.
- Fix an arbitrary constant $c > 0$ and $k = c \cdot \ln n$.
- Each element $i \in [m]$ independently at random chooses $k$ sets out of possible $m$ sets.
- Each set $S_j$ for $j \in [m]$ consists of elements that chose that set.
- Let $\mathcal{I}_{\varepsilon}$ be the resulting random instance of the minimum set cover problem.

# Analysis

- The fractional solution $x'_j = 1/k$ for all $j \in [m]$ is feasible. Therefore, $LP^* \leq m/k$.
- Fix an arbitrary collection of sets indexed by $\mathcal{C} \subseteq [m]$ such that $|\mathcal{C}| = (1 - e^{-1/c} - \delta)m$.
- Probability that any $i \in [n]$ is not covered by $\mathcal{C}$ is

$$\begin{aligned}
\frac{\binom{(e^{-1/c}+\delta)m}{k}}{\binom{m}{k}} &= \prod_{i=0}^{k-1} \frac{(e^{-1/c}+\delta)m - i}{m-i} \\
&\geq (e^{-1/c} + \delta/2)^k \\
&= \frac{(1 + e^{1/c}\delta/2)^{c \ln n}}{n} = n^{-(1-F_{c,\delta})},
\end{aligned}$$

# Analysis

- where $F_{c,\delta} = c \ln(1 + \delta e^{1/c}/2)$ is a constant depending on $c$ and $\delta$. We assume that $\delta$ is small enough that $F_{c,\delta} \in (0, 1)$.
- Probability that all $n$ elements are covered by $\mathcal{C}$ is at most

$$\left(1 - n^{-(1-F_{c,\delta})}\right)^n \leq e^{-n^{F_{c,\delta}}}.$$

- The total number of choices for the index set $\mathcal{C}$ is at most $2^m = 2^{n^\varepsilon}$.

# Analysis

- Therefore, by the union bound, the probability that there exists a feasible index set $\mathcal{C}$ is at most

$$e^{-n^{F_{c,\delta}}} 2^{n^{\varepsilon}} \leq e^{n^{\varepsilon} - n^{F_{c,\delta}}}.$$

- Choose $\varepsilon = F_{c,\delta}/2$. With probability at least $1 - e^{n^{\varepsilon} - n^{F_{c,\delta}}}$ one needs to choose at least $(1 - e^{-1/c} - \delta)m$ sets into any feasible integral solution.

# Open Problems

- In the case when $k \approx \log \Delta$, can we design an approximation algorithm for the Minimum Set Cover Problem with performance guarantee $o(\log \Delta)$?
- The best known upper bound is from our algorithm. The best known complexity lower bound is $\Omega\left(\frac{\log \Delta}{(\log \log \Delta)^2}\right)$.
- One must use stronger math. programming lower bounds: SDP?, Lasserre hierarchy?

# Submodular Set Cover 1

- Find a set cover indexed by the set $\mathcal{C} \subseteq [m]$ such that $f(\mathcal{C})$ is minimized.
- Iwata, Nagano [2009] introduced the problem and designed a $k$-approximation algorithm. And showed that there is no polynomial time approximation algorithm with performance guarantee better than $o(n/\log^2 \log n)$.

# Submodular Set Cover 2

- Wolsey [1982] introduced the following problem.
- We are given a set $[m]$ and a monotone submodular function $f : 2^{[m]} \to R_+$ find a collection of indices $\mathcal{C} \subseteq [m]$ minimizing $\sum_{j \in \mathcal{C}} c_j$ such that $f(\mathcal{C}) = f([m])$.
- He proved that the greedy algorithm has performance guarantee $\ln\left(\sum_{j=1}^{m} f(\{j\})\right) + 1$.

# Submodular Set Cover 3

- Kamiyama [2011] combined two models, i.e. we have an arbitrary submodular objective $g$ and monotone submodular function $f$.
- He designed an algorithm with performance guarantee

$$\max_{S \subseteq [m] : f(S) < f([m])} \frac{\sum_{j \in [m] \setminus S} f_S(\{j\})}{f_S([m] \setminus S)}$$

where $f_S(X) = f(S \cup X) - f(S)$.

- Fujito [2000] had an analogous algorithm and guarantee for the Wolsey's special case.

# Submodular Set Cover 4

- Hayrapetyan, Swamy, and Tardos [2005] introduced the following generalization.
- We are given a set cover instance and a monotone submodular function $h_j(T)$ for each set $S_j$. Find a set cover $\mathcal{C}$ and subsets $T_j \subseteq S_j$ covering the ground set, minimizing $\sum_{j \in \mathcal{C}} h_j(T_j)$.
- Hayrapetyan, Swamy, and Tardos claim that a variant of greedy and primal-dual gives performance ratios similar to the classical ones.
- Chudak and Nagano [2007] designed algorithms to solve continuous relaxation of this problem.

# Submodular Set Cover 5

$$\min \sum_{j\in[m]} w_j x_j, \tag{4}$$

$$\sum_{j:i\in S_j} x_j \geq z_i, \qquad \forall i \in [n], \tag{5}$$

$$\sum_{i\in S} z_i \geq f(S), \qquad \forall S \subseteq [n], \tag{6}$$

$$0 \leq z_i \leq 1, \qquad \forall i \in [n], \tag{7}$$

$$x_j \geq 0, \qquad \forall j \in [m]. \tag{8}$$

The classical problem is when $f(S) = |S|$. The problem is interesting with $f(S)$ is submodular or supermodular.