K-Medians, Facilities Location, and the Chernoff-Wald Bound

Neal Young

Dartmouth College

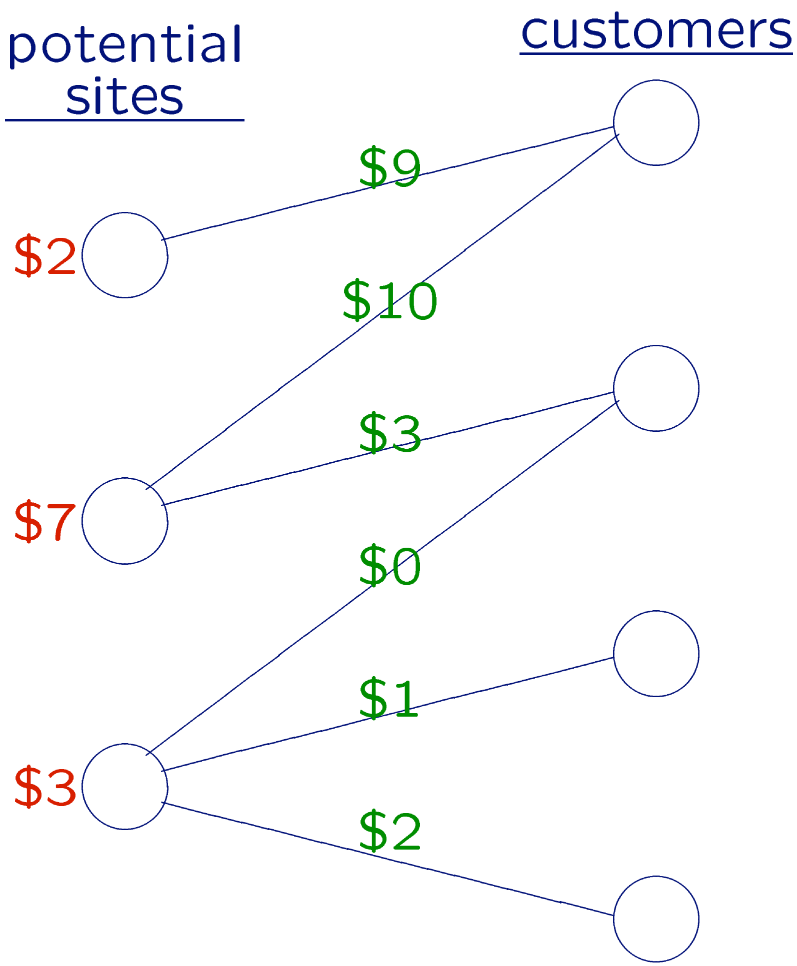
Results

1. Improved performance guarantees for two NP-hard problems: *facilities location* and k-*medians* (no triangle inequality).

2. Fast Lagrangian-relaxation algorithms for *fractional* variants.

3. A new probabilistic inequality: *Chernoff-Wald* bound.

facilities location and weighted k-medians



Both generalize set cover.

No Δ-inequality assumed.

Facilities location: Minimize site cost + total distance.

Wtd k-medians: Minimize total distance s.t. site cost $\leq k$.

Improved performance guarantees

Let $k =$ site cost of opt, $d =$ total distance of opt.

Facilities location:

site cost + total distance $\leq H_\Delta(k + d)$. [Hochbaum '82]

site cost + total distance $\leq H_\Delta \cdot k + d$. NEW

K-medians:

site cost $\leq \ln(n)\, k/\epsilon$, total distance $\leq (1 + \epsilon)\, d$. [Lin,Vitter '92]

site cost $\leq \ln(n/\epsilon)\, k$, total distance $\leq (1 + \epsilon)\, d$. NEW

Faster algorithm for fractional k-medians

site cost $\leq (1+\epsilon)k$, total distance $\leq (1+\epsilon)\, d$.

Algorithm: Lagrangian relaxation (via randomized rounding).

Time: $O(k \ln(n)/\epsilon^2)$ linear-time passes.

Chernoff-Wald bound.

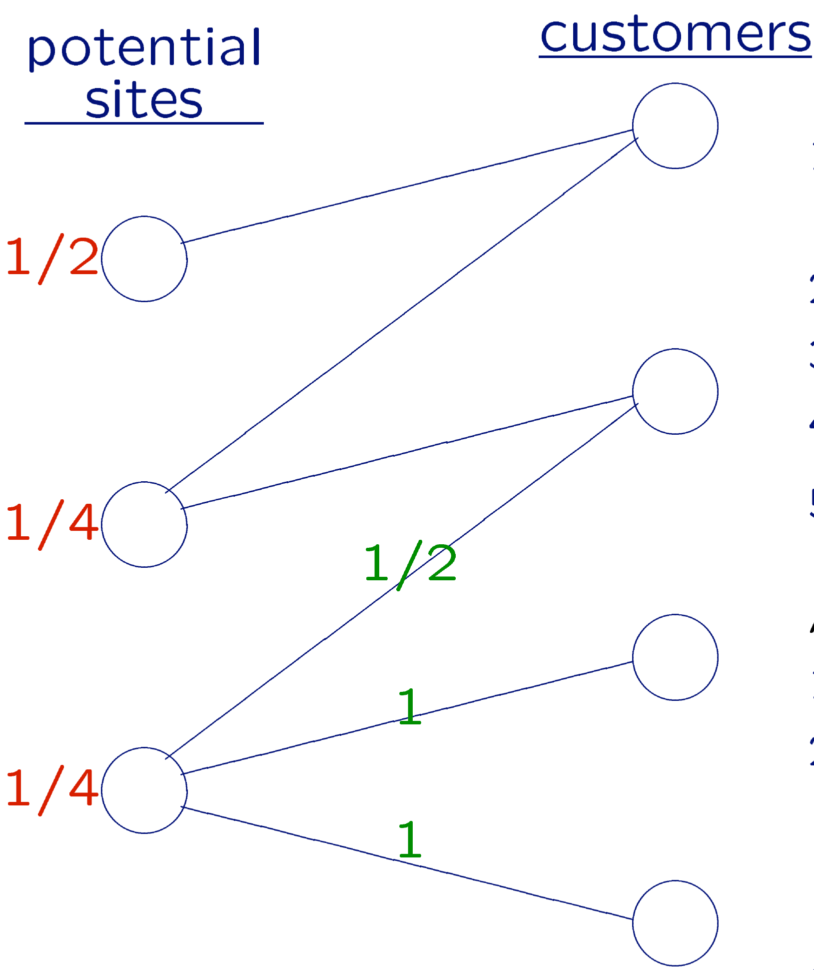
Previous: [Garg, 1998, unpublished]

Lagrangian-relaxation algorithm for fractional set cover.

New:

1. Recast Garg's result in *randomized rounding* framework.
2. *Abstract* the analysis to get a new probabilistic bound.

randomized rounding scheme (facilities location)



1. Each site and edge gets probability proportional to wt. in fractional sol'n x.
2. Repeat until all customers assigned:
3. choose one site s randomly w/pr. Pr(s)
4. assign each cust. c to s w/pr. Pr(s, c)
5. Return chosen sites, assignment.

Analysis:

1. site cost — bound E[#iterations].
2. distance — Pr[s gets c] $= x(s, c)$.

Other problems use same inner loop.

Chernoff-Wald bound:

$$\begin{array}{|c|c|c|c|}
x_{00} & x_{01} & \cdots & x_{0m} \\
x_{10} & x_{11} & \cdots & x_{1m} \\
x_{20} & x_{21} & \cdots & x_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
x_{T0} & x_{T1} & \cdots & x_{Tm} \\
\hline
\end{array}$$

Let $T = \#$ rows, a random *stopping time*.

Let $M =$ min. column sum.

x_{ij}'s are independent random 0-1 variables, $E[x_{ij}] \geq \mu$.

Then

$$E[M] \geq \mu E[T] - \sqrt{2\mu E[T] \ln(m)}.$$

proof: Combines proofs of Wald's equation and Chernoff bound.

Analogous bound for $E[\max_i S_i]$.

Chernoff-Wald bound, example

Throw balls randomly in 100 bins until some condition is met.

Let $N = \#$balls thrown.

Let $B_i = \#$balls in ith bin. $E[B_i] = \frac{1}{100}E[N]$.

Let $\texttt{lo} = \min_i B_i$.

Chernoff-Wald:

$$E[\texttt{lo}] \;>\; \frac{1}{100}E[N] - \sqrt{\frac{2\ln(100)}{100}E[N]}$$

$$\approx\; \frac{1}{100}E[N] - \frac{1}{3.3}\sqrt{E[N]}$$

Chernoff-Wald bound, example 1

Stop when 10000 balls are thrown.

$N = 10000$.

Chernoff-Wald:

$$E[\texttt{lo}] > \frac{1}{100}E[N] - \frac{1}{3.3}\sqrt{E[N]}$$

$$\approx 100 - 30 = 70$$

Chernoff: $\Pr[\texttt{lo} > 70] < 1$.

Chernoff-Wald bound, example 2

Stop when bin 1 gets two balls in a row.

$E[N] \approx 10000.$

Chernoff-Wald:

$$E[\mathbf{1o}] \;>\; \frac{1}{100}E[N] - \frac{1}{3.3}\sqrt{E[N]}$$

$$\approx\; 100 - 30 = 70$$

Chernoff: *doesn't apply.*

Chernoff-Wald bound, example 3

Stop when $\texttt{lo} = 70$ (every bin has at least 70 balls).

Know $\texttt{lo}$, what is $E[N]$?

Chernoff-Wald:

$$\begin{aligned} E[\texttt{lo}] &> \frac{1}{100}E[N] - \frac{1}{3.3}\sqrt{E[N]} \\ 70 &> \frac{1}{100}E[N] - \frac{1}{3.3}\sqrt{E[N]} \\ \text{implies} \qquad E[N] &< 10000 \end{aligned}$$

Chernoff: *doesn't apply.*

Summary

1. Improved performance guarantees for two NP-hard problems:

   Facilities location: site cost + total distance $\le H_\Delta \cdot k + d$.

   K-medians: site cost $\le \ln(n/\epsilon)\,k$, total distance $\le (1+\epsilon)\,d$.

2. Fast Lagrangian-relaxation algorithms for fractional variants.

3. New *Chernoff-Wald* bound:

   Bounds $E[\max]$ or $E[\min]$ of a collection of sums of r.v.'s.

   Applies even when number of trials is a random variables.