Distributed k-Means and k-Median Clustering on General Topologies

Maria-Florina Balcan, Steven Ehrlich, and Yingyu Liang

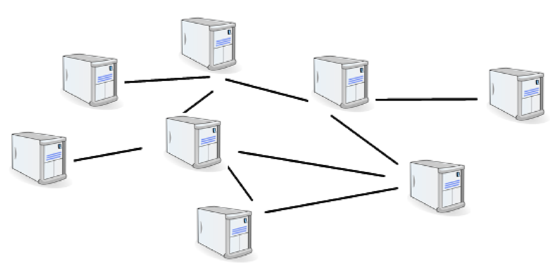
Georgia Tech College of Computing

Problem Setup

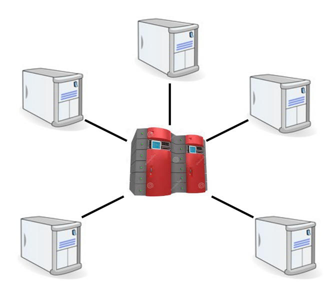
▷ **k-Clustering:** Given a set P of N points in $\mathbf{R}^d$, find centers $\mathbf{x} = \{x_1, \dots, x_k\}$ to minimize $\sum_{p \in P} \mathrm{cost}(p, \mathbf{x})$.
Widely studied cost functions in ML & TCS

- k-median: $\mathrm{cost}(p, \mathbf{x}) = \min_{x \in \mathbf{x}} d(p, x)$
- k-means: $\mathrm{cost}(p, \mathbf{x}) = \min_{x \in \mathbf{x}} d^2(p, x)$

▷ **Modern Challenge:** data distributed over different sites, e.g. distributed databases, images and videos over networks, ...



general communication network



star network

▷ **Distributed Clustering:**

- Communication graph: undirected graph G on n nodes with m edges, where an edge indicates that the two nodes can communicate
- Global data: P is divided into local data sets $P_1, \dots, P_n$
- Goal: efficient distributed algorithm with low communication

Our Results

▷ Efficient algorithm that

- outputs $(1+\epsilon)\alpha$-approx, given any non-distributed α-approx algo
- has low communication independent of #points in global data set
  - communication on a star network: $\tilde{O}(kd + nk)$ points
- has good experimental performance

▷ Two stages of our distributed algorithm

1. Each node constructs a local portion of a global summary
2. Communicate the local portions, and compute approximation solution on the summary

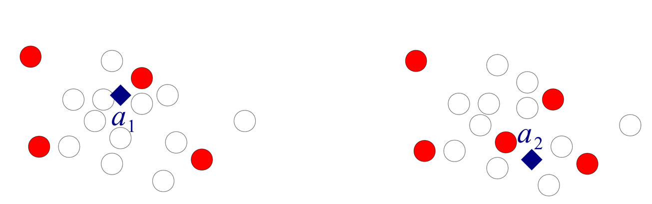
Coreset

▷ **Coreset** [Har-Peled-Mazumdar, STOC04]: short summaries capturing relevant info w.r.t. all clusterings

Definition. *An ϵ-coreset for P is a set of points D and weights w on D s.t.* $\forall \mathbf{x}, (1-\epsilon)\mathrm{cost}(P, \mathbf{x}) \le \sum_{q \in D} w_q \mathrm{cost}(q, \mathbf{x}) \le (1+\epsilon)\mathrm{cost}(P, \mathbf{x})$.

▷ **Non-distributed coreset construction** [Feldman-Langberg, STOC11]

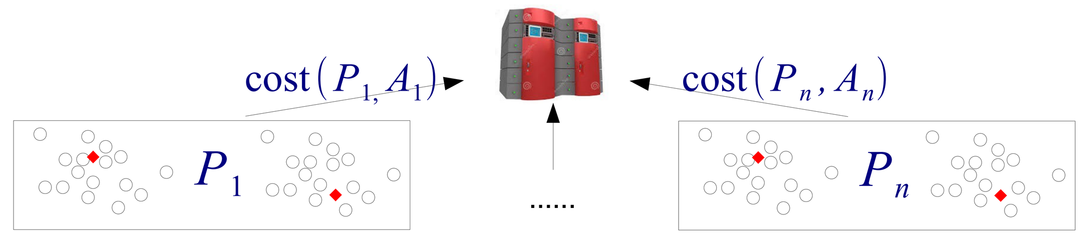
1. Compute a constant approximation solution A
2. Sample points S with probability proportional to $\mathrm{cost}(p, A)$; $|S| = \tilde{O}(kd)$ for constant ϵ



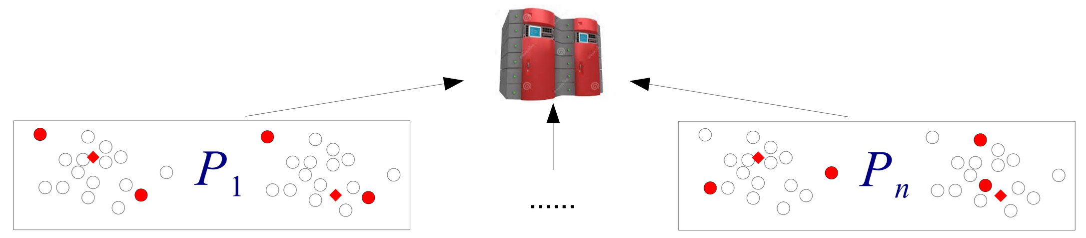
Distributed Coreset Construction

Algorithm (two rounds, interactive)

1. Compute a constant approximation solution A_i for P_i; Communicate the costs $\mathrm{cost}(P_i, A_i)$

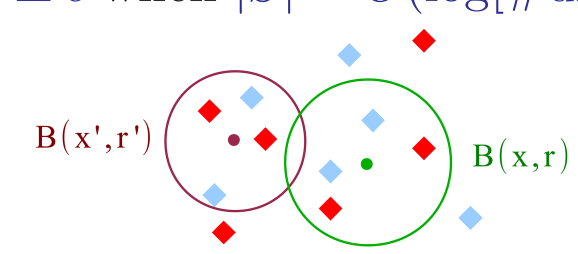


2. Sample points from P_i according to the multinomial distribution given by $\mathrm{cost}(P_i, A_i)$; #sampled points$=\tilde{O}(kd)$ for constant ϵ



Analysis

▷ **Uniform sampling for metric balls:** $\forall B(x,r) = \{p : d(p,x) \le r\}$, $\frac{|B(x,r) \cap S|}{|S|} = \frac{|B(x,r) \cap P|}{|P|} \pm \epsilon$ when $|S| = \tilde{O}(\log[\#\text{distinct } B(x,r) \cap P]/\epsilon^2)$



▷ **Sampling for general function space:** [Feldman-Langberg, STOC11]
Let $B(f,r) = \{p : f(p) \le r\}$ for $f : P \mapsto \mathbf{R}_{\ge 0}, f \in F$.

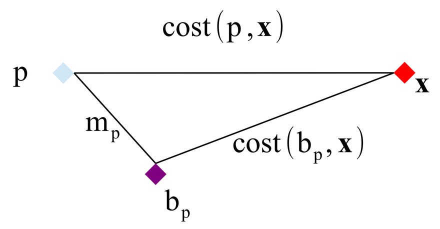
Lemma. *Sample S from P with prob. prop. to m_p, and let $w_p = \frac{\sum_q m_q}{m_p |S|}$. If $|S| = \tilde{O}(\dim(F,P)/\epsilon^2)$, then w.h.p.*

$$\forall f \in F, \left|\sum_{p \in P} f(p) - \sum_{q \in S} w_q f(q)\right| \le \epsilon \left(\sum_{p \in P} m_p\right)\left(\max_{p \in P} \frac{f(p)}{m_p}\right).$$

Proof idea: replace p with m_p copies p'; let $f(p') = f(p)/m_p$

▷ **Intuition for distributed k-median:**

- Let a_p be an anchor point for $p \in P_i$, and use sampling to approximate $f_{\mathbf{x}}(p) = \mathrm{cost}(p, \mathbf{x}) - \mathrm{cost}(a_p, \mathbf{x})$.
  - Set $m_p = \mathrm{cost}(p, a_p) \ge |f_{\mathbf{x}}(p)|$.
  - Error $\le \epsilon \sum_{p \in P} \mathrm{cost}(p, a_p)$.



- Keypoints for low communication:
  - sufficient to choose a_p to be the nearest center in the local approximation solution A_i so that error $\le O(\epsilon)\mathrm{OPT}$;
  - sufficient to do the sampling locally.

▷ **Intuition for distributed k-means:** similar as k-median except

- Upper bounds not available for $f_{\mathbf{x}}(p) = \mathrm{cost}(p, \mathbf{x}) - \mathrm{cost}(a_p, \mathbf{x})$
- Bound separately the errors of bad points $P \backslash G(\mathbf{x})$ and good points $G(\mathbf{x}) = \{p \in P : |\mathrm{cost}(p, \mathbf{x}) - \mathrm{cost}(a_p, \mathbf{x})| \le \mathrm{cost}(a_p, \mathbf{x})/\epsilon\}$

Distributed Clustering

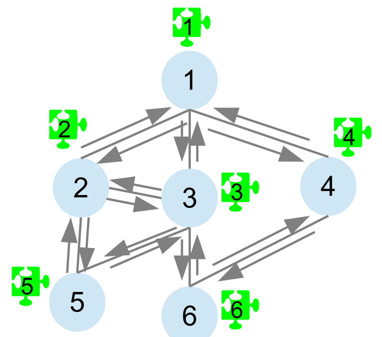
▷ **Algorithm**

1. Distributed coreset construction
2. Communicate the local portions of the coreset
3. Compute approximation solution on the coreset

Theorem. *Given any non-distributed α-approx algo as a subroutine, our algo computes a $(1+\epsilon)\alpha$-approx solution. The total communication cost is $\tilde{O}(m(kd+nk))$ points for constant ϵ.*

▷ **Total Communication on Different Networks** (for constant ϵ):

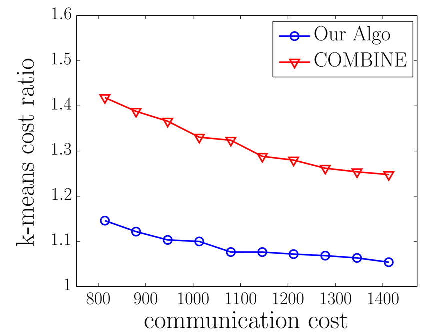
1. **Star graph:** $\tilde{O}(kd + nk)$ points
   by sending the local portions of the coreset to the coordinator
2. **Rooted Tree:** $\tilde{O}(h(kd + nk))$ points
   by sending the local portions of the coreset to the root
3. **General Topologies:** $\tilde{O}(m(kd + nk))$ points
   Message Passing: on each node do
   - Communicate its local message to all its neighbors
   - When the node receives new message, communicate to all its neighbors



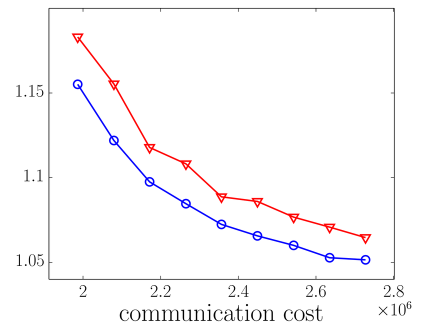
Experiments

▷ Data set: ColorHistogram ($\approx$ 68k points in $\mathbf{R}^{32}, k = 10, n = 25$); YearPredictionMSD ($\approx$ 0.5m points in $\mathbf{R}^{90}, k = 50, n = 100$)

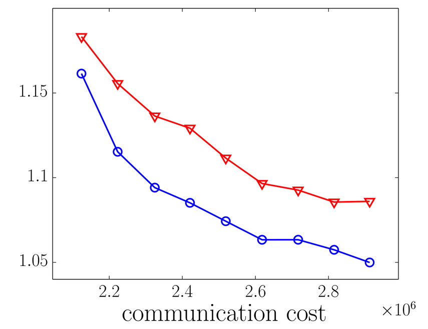
▷ Results on ColorHistogram:



star, exponential

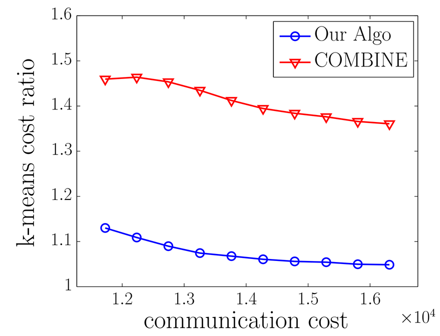


grid, weighted

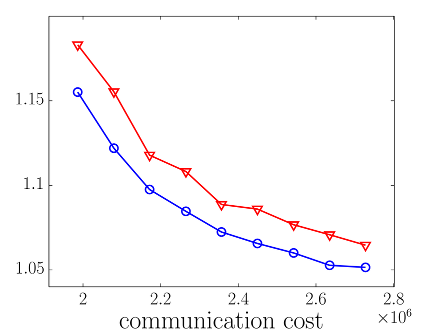


preferential, degree-based

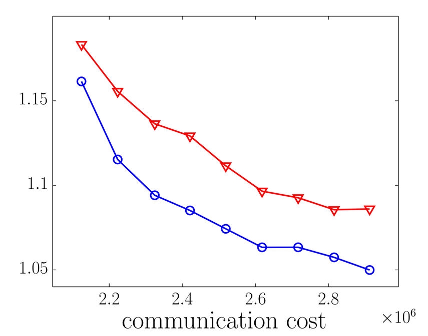
▷ Results on YearPredictionMSD:



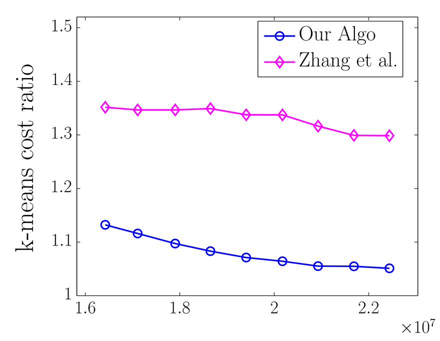
star, exponential



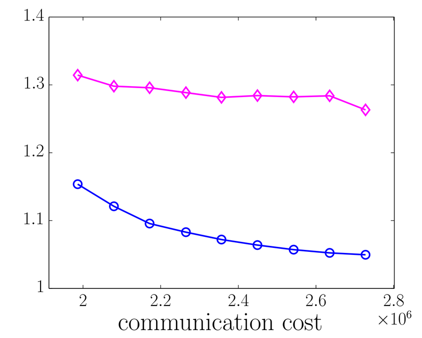
grid, weighted



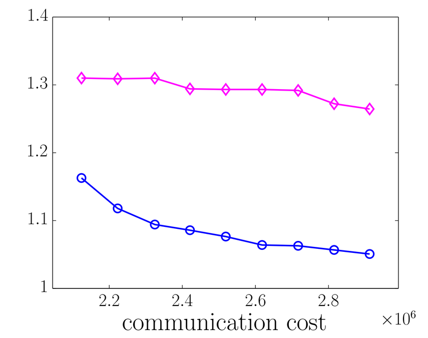
preferential, degree-based



tree, uniform



tree, weighted



tree, degree-based