Distributed \textit{k-Clustering} with Heavy Noise
\textit{(NeurIPS’18)}

\textbf{Xiangyu Guo} and \textbf{Shi Li}
State University of New York at Buffalo
Distributed \((k,z)\)-clustering

MapReduce model

\(m\) servers

The coordinator

\(P_1\)

\(P_2\)

\(P_m\)

\(P_3\)

Dataset \(P\)

\(n: \text{size of } P\)

\(m: \#\text{servers}\)
Distributed \((k,z)\)-clustering

**MapReduce model**

- The coordinator
- \(m\) servers

**Dataset** \(P\)

- \(n\): size of \(P\)
- \(m\): #servers
- \(k\): #clusters
- \(z\): #outliers

**Task:** discarding \(z\) outliers & clustering non-outliers into \(k\) clusters
Major concerns

**Clustering quality**
$O(1)$-approximation: objective $\leq O(1) \cdot \text{OPT}$

**Communication cost**
Focus on the case when data is heavily noisy: $z \gg k, m$
Can we achieve $O(1)$-approx with communication cost $\ll \Theta(z)$?
Can we achieve $O(1)$-approx with communication cost $\ll \Theta(z)$?

NO. 😱

**Theorem:** Any $O(1)$-approx algorithm needs communication cost $\Omega(z)$
Can we achieve $O(1)$-approx with communication cost $\ll \Theta(z)$?

**NO. 😳**

**Theorem:** Any $O(1)$-approx algorithm needs communication cost $\Omega(z)$

**Yes! 😊**

If allow removing slightly more than $z$ outliers
## Distributed \((k,z)\)-center

<table>
<thead>
<tr>
<th></th>
<th>approx. ratio</th>
<th>comm. cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>[\text{MKCWM15}]</td>
<td>((O(1),1))</td>
<td>(O(m(k+z)))</td>
</tr>
<tr>
<td>[\text{GLZ17}]</td>
<td>((O(1),2+\epsilon))</td>
<td>(\tilde{O}(m(1/\epsilon+k)))</td>
</tr>
<tr>
<td>Ours</td>
<td>((O(1),1+\epsilon))</td>
<td>(\tilde{O}(mk/\epsilon))</td>
</tr>
</tbody>
</table>

- objective \(\leq O(1) \cdot \text{OPT}\)
- #outliers \(\leq (1+\epsilon)z\)
<table>
<thead>
<tr>
<th></th>
<th>problem</th>
<th>approx. ratio</th>
<th>comm. cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>[GLZ17]</td>
<td>(k,z)-median</td>
<td>(O(1), 2 + \epsilon)</td>
<td>\tilde{O}(m/\epsilon + mk)</td>
</tr>
<tr>
<td></td>
<td>(k,z)-means</td>
<td>(O(1), 2 + \epsilon)</td>
<td>(O(1), 2 + \epsilon)</td>
</tr>
<tr>
<td>[CAZ18]</td>
<td>(k,z)-median/means</td>
<td>(O(1), 1)</td>
<td>O(k \log n + z)</td>
</tr>
<tr>
<td>Ours</td>
<td>(k,z)-median</td>
<td>(1 + \epsilon, 1 + \epsilon)</td>
<td>\tilde{O} (k \epsilon^{-3} + mk \epsilon^{-1})</td>
</tr>
<tr>
<td></td>
<td>(k,z)-means</td>
<td>(1 + \epsilon, 1 + \epsilon)</td>
<td>\tilde{O} (k \epsilon^{-5} + mk \epsilon^{-1})</td>
</tr>
</tbody>
</table>

(Note: To achieve (1 + \epsilon)-approx in the objective, we need exponential (in m, k, \epsilon^{-1}) running time)
Thank you!