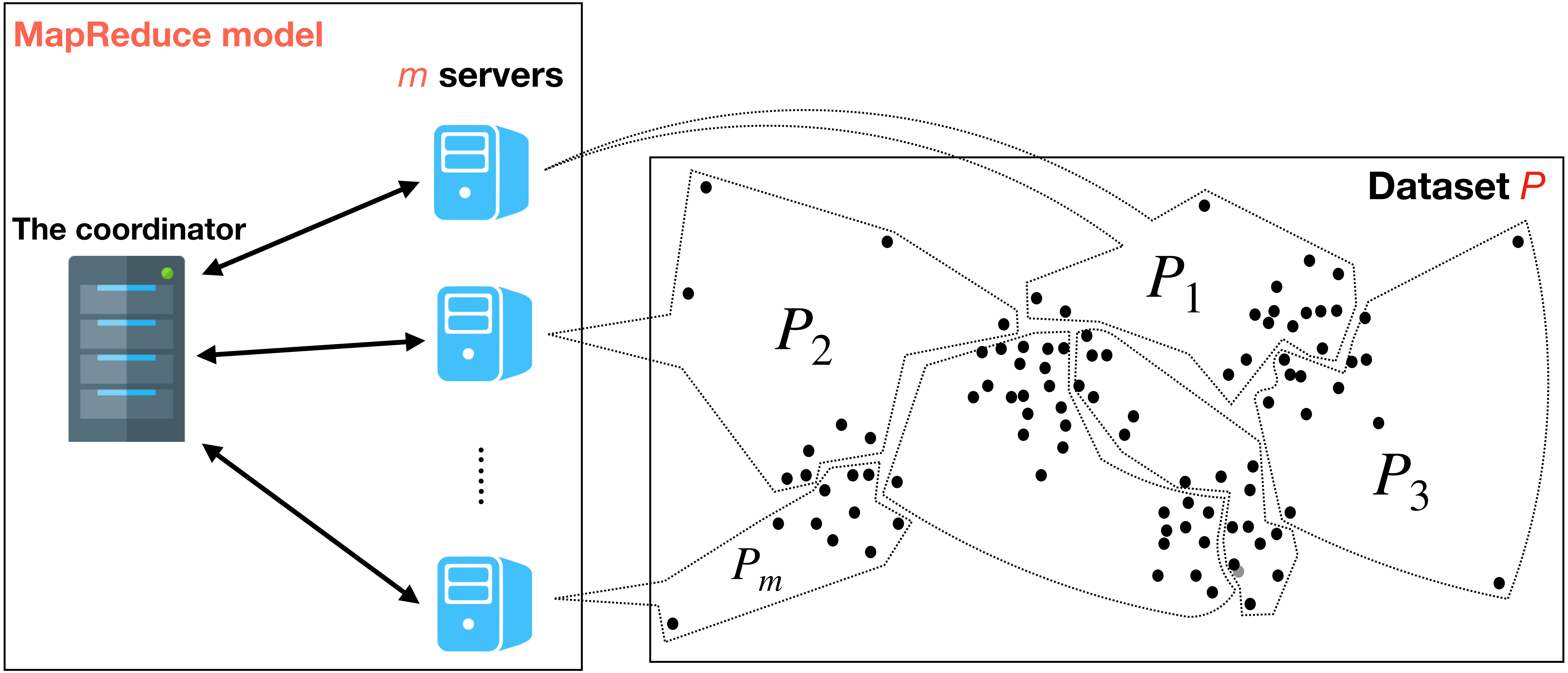
Distributed *k*-Clustering with Heavy Noise

(*NeurIPS'18*)

Xiangyu Guo and **Shi Li**
State University of New York at Buffalo

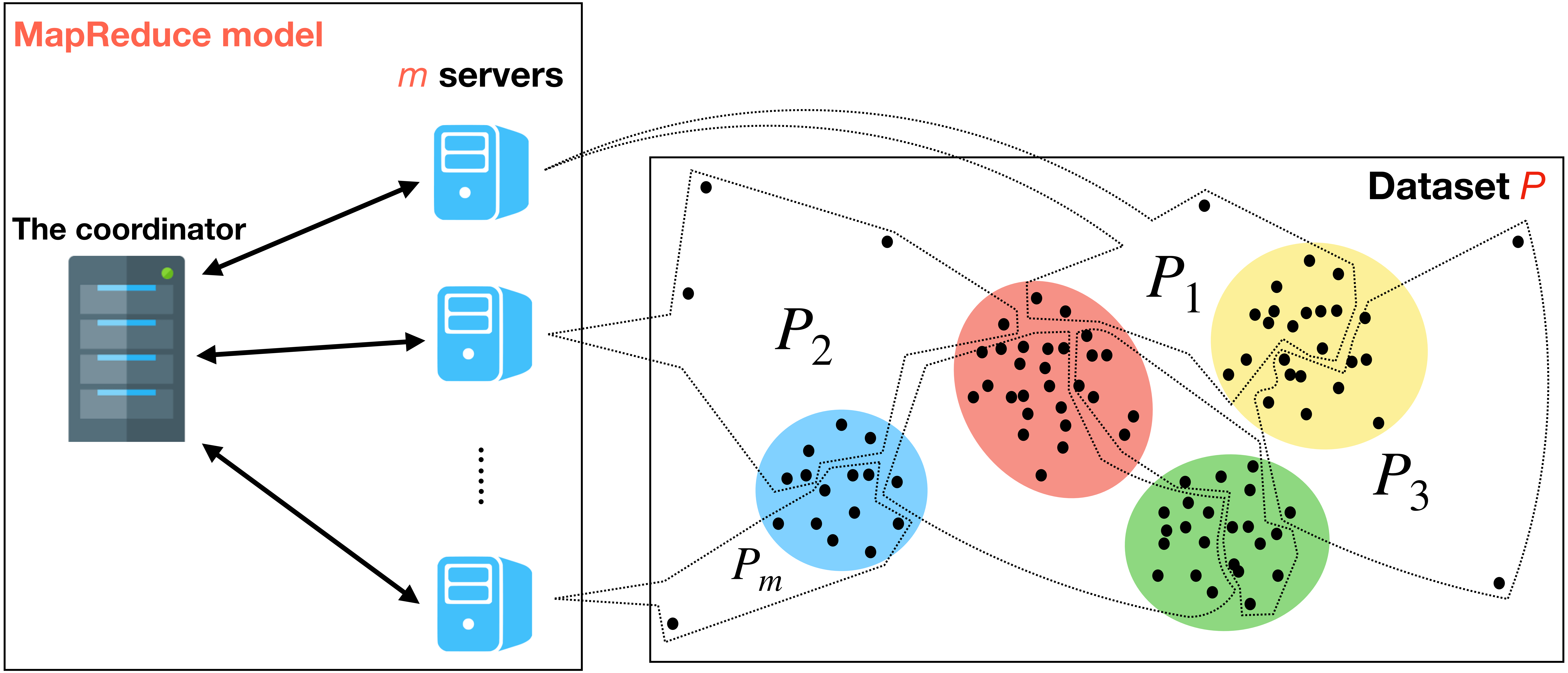
Distributed (*k*,*z*)-clustering



n: size of *P*

m: **#servers**

Distributed (*k*,*z*)-clustering



n: size of *P*

m: **#servers**

k: **#clusters**

z: **#outliers**

Task: discarding *z* outliers & clustering non-outliers in to *k* clusters

Major concerns

Clustering quality

$O(1)$-approximation: objective $\leq O(1) \cdot \mathrm{OPT}$

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

	approx. ratio	comm. cost
[MKCWM15]	$(O(1), 1)$	$O(m(k+z))$
[GLZ17]	$(O(1), 2+\epsilon)$	$\tilde{O}(m(1/\epsilon + k))$
Ours	$(O(1), 1+\epsilon)$	$\tilde{O}\left(mk/\epsilon\right)$

objective $\leq O(1) \cdot \mathrm{OPT}$

#outliers $\leq (1+\epsilon)z$

(k,z)-median/means

	problem	approx. ratio	comm. cost
[GLZ17]	(k,z)-median	$(O(1), 2+\epsilon)$	$\tilde{O}(m/\epsilon + mk)$
	(k,z)-means	$(O(1), 2+\epsilon)$	$(O(1), 2+\epsilon)$
[CAZ18]	(k,z)-median/means	$(O(1), 1)$	$O(k \log n + z)$
Ours	(k,z)-median	$(1+\epsilon, 1+\epsilon)$	$\tilde{O}\left(k\epsilon^{-3} + mk\epsilon^{-1}\right)$
	(k,z)-means	$(1+\epsilon, 1+\epsilon)$	$\tilde{O}\left(k\epsilon^{-5} + mk\epsilon^{-1}\right)$

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

Thank you!