Scalable constant k-means approximation on *well-clusterable* data

Cheng Tang and Claire Monteleoni

George Washington University

Clustering algorithms theory vs practice

- Gap between worst-case analysis and performance in practice
- Bridge the gap via easy-data assumption

Heuristic (*Buckshot* algorithm [Cutting et al., 92])

1. Take a uniform random sample S from X
2. Run single-linkage (Kruskal's) on S until k components left.
3. Take the mean $\boldsymbol{c_i}$ of the i-th component, $\forall i \in [k]$

Output $C = \{c_i, i \in [k]\}$

Result: constant k-means appx, with time independent of data size

Observation: a uniform random sample can amplify the easiness of the original dataset

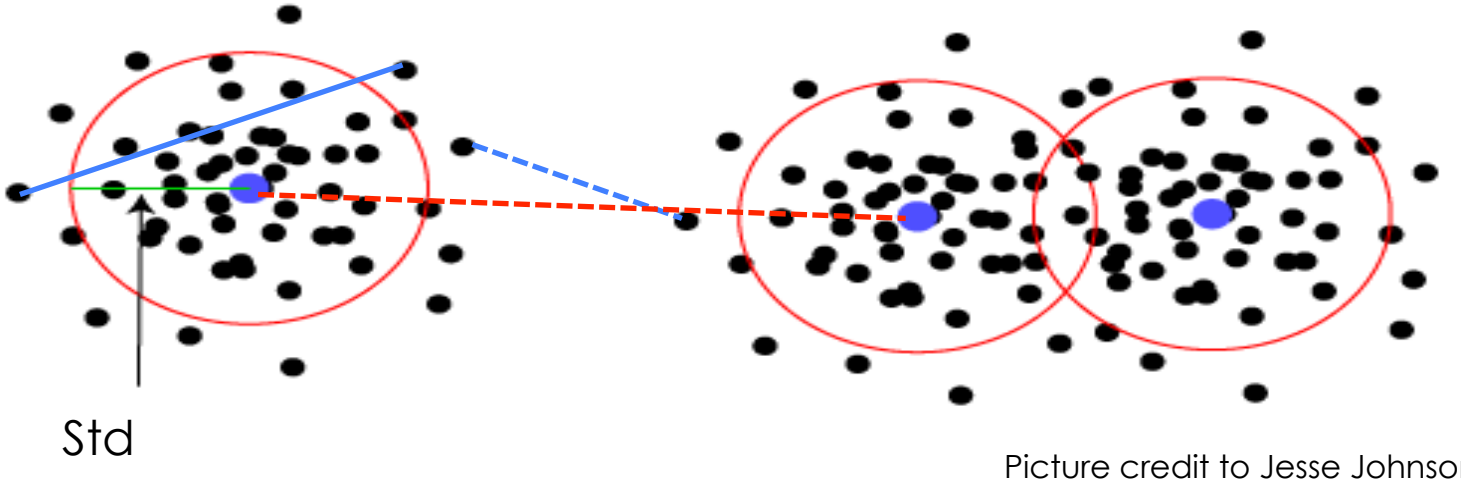
Acceleration techniques for clustering

Random projection preserves clustering structure	**Spectral projection amplifies clustering structure**	**Weighted/uniform sampling preserves clustering structure**	**Uniform sampling amplifies clustering structure**
Dasgupta-Gupta03 Boutsidis-Zouzias-Drineas10	Vempala-Wang02 Kannan-Vempala09	Boutsidis-Mahoney-Drineas09 **Daniely-Linial-Saks12**	** **This work**

**Results with the "k-means algorithm" at *Nonconvex Opt.* workshop tomorrow

Main ideas

1. We show single-linkage can correctly identify k components in a dataset if it satisfies a *strong separability* *assumption* **
2. We introduce center-separability of a dataset, and show w.h.p., a random sample of the dataset satisfies the *strong separability*



Picture credit to Jesse Johnson

** This property of single-linkage was also studied in Ackerman-Ben-David09

Constant k-means approximation with no dependence on N

k-means objective $F_X(C) = \sum_{x \in X} \|x - C(x)\|^2, C(x) = \arg\min_{c \in C} \|x - c\|$

- $C = \{c_i, i \in [k]\}$ achieves constant approximation guarantees w.h.p. with sample size

$$O(\frac{N}{N_{min}} \log k) \quad \text{where } \frac{N}{N_{min}} \geq \Omega(k)$$

Running time poly(k), if $\frac{N}{N_{min}} = o(e^k)$