k-variates++: more pluses in the *k*-means++

R. Nock[1,2], R. Canyasse[3,4], R. Boreli[1,5], F. Nielsen[3,6]

Data61[1], ANU[2], Ecole Polytechnique[3], Technion[4], UNSW[5], Sony CS Labs[6]

contact: richard.nock@data61.csiro.au

Key contribution: **(a)** generalise *k*-means++ (applications+) & **(b)** approximation bounds, **(c)** likelihood ratio thms

(b) **Objective**: minimise the *k*-means potential for $\mathcal{C}$

$$\phi(\mathcal{A};\mathcal{C}) \doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}-\boldsymbol{c}(\boldsymbol{a})\|_2^2$$

with

$$\boldsymbol{c}(\boldsymbol{a}) \doteq \arg\min_{\boldsymbol{c}\in\mathcal{C}} \|\boldsymbol{a}-\boldsymbol{c}\|_2^2$$

Suppose $\wp_t$ is η-stretching: for any *optimal* cluster A (size > 1) and $\boldsymbol{a}_0 \in A$, we have:

$$\frac{\phi(A;\mathcal{C})}{\phi(A;\{\boldsymbol{a}_0\})} \leq (1+\eta)\cdot\frac{\phi(\wp_t(A);\mathcal{C})}{\phi(\wp_t(A);\{\wp_t(\boldsymbol{a}_0)\})}, \forall t$$

Input: data $\mathcal{A} \subset \mathbb{R}^d$ with $|\mathcal{A}| = m$, $k \in \mathbb{N}_*$, random variables $\{X_{\boldsymbol{a}}, \boldsymbol{a}\in\mathcal{A}\}$, probe functions $\wp_t : \mathcal{A} \to \mathbb{R}^d$ $(t \geq 1)$;
Step 1: Initialise centers $\mathcal{C} \leftarrow \emptyset$;
Step 2: **for** $t = 1, 2, ..., k$
2.1: randomly sample $\boldsymbol{a} \sim_{q_t} \mathcal{A}$, with $q_1 \doteq u_m$ and, for $t > 1$,

(a) $$q_t(\boldsymbol{a}) \doteq D_t(\boldsymbol{a})\left(\sum_{\boldsymbol{a}'\in\mathcal{A}} D_t(\boldsymbol{a}')\right)^{-1}, \text{ where } D_t(\boldsymbol{a}) \doteq \min_{\boldsymbol{x}\in\mathcal{C}} \|\wp_t(\boldsymbol{a})-\boldsymbol{x}\|_2^2 \ ;$$

2.2: randomly sample $\boldsymbol{x} \sim X_{\boldsymbol{a}}$;
2.3: $\mathcal{C} \leftarrow \mathcal{C} \cup \{\boldsymbol{x}\}$;
Output: $\mathcal{C}$;

k-means++: $X_{\boldsymbol{a}} = \mathrm{Dirac}(\boldsymbol{a})$, $\wp_t = \mathrm{Id}(.)$

(c) **Principle**: define parameters that quantify *spreading* (δ_w) & *monotonicity* (δ_s) of neighborhood $\mathrm{NN}(\mathcal{N},\mathcal{B})$ in $\mathcal{A}$ with

$$\mathrm{NN}(\mathcal{N},\mathcal{B}) \doteq \sum_{\boldsymbol{a}\in\mathcal{B}} \|\boldsymbol{a}-\mathrm{NN}_{\mathcal{N}}(\boldsymbol{a})\|_2^2$$

and $\mathcal{N},\mathcal{B} \subseteq \mathcal{A}$

$$|\mathcal{N}| \in \{1,2,...,k-1\}$$

$$|\mathcal{B}| \in \{m-1, m\}$$

Suppose data lives in $\Omega \doteq \mathrm{Ball}(L_2, R)$ and $X_{\cdot}$ satisfies the pointwise likelihood constraint:

$$(dp_{X_{a'}}/dp_{X_a})(\boldsymbol{x}) \leq \varrho(R) \ , \forall \boldsymbol{a},\boldsymbol{a}' \in \mathcal{A}, \forall \boldsymbol{x} \in \Omega$$

Then

$$\mathbb{E}_{\mathcal{C}\sim k-\mathrm{variates++}}[\phi(\mathcal{A};\mathcal{C})] \leq (2+\log k)\cdot\Phi$$

, with $\Phi \doteq (6+4\eta)\phi_{\mathrm{opt}} + 2\phi_{\mathrm{bias}} + 2\phi_{\mathrm{var}}$

k-means++: $\phi_{\mathrm{bias}} = \phi_{\mathrm{opt}}$, $\phi_{\mathrm{var}} = 0$, $\eta = 0$, $\Rightarrow \Phi = 8\phi_{\mathrm{opt}}$

$$\phi_{\mathrm{opt}} \doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}-\boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2$$

$$\phi_{\mathrm{bias}} \doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\mathbb{E}[X_{\boldsymbol{a}}]-\boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2$$

$$\phi_{\mathrm{var}} \doteq \sum_{a\in\mathcal{A}} \mathrm{tr}\left(\mathrm{cov}[X_{\boldsymbol{a}}]\right)$$

(Can be **better** than Arthur and Vassilvitskii's bound (by constant factor) if X_a well chosen ($\phi_{\mathrm{bias}} < \phi_{\mathrm{opt}}$))

general approximation bound for *k*-variates++

Then, for any neighbour $\mathcal{A}' \approx \mathcal{A}$ (differ by one arbitrary element) and with $\wp_t = \mathrm{Id}(.)$,

$$\frac{\mathbb{P}_{\mathcal{C}\sim k-\mathrm{variates++}}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}_{\mathcal{C}\sim k-\mathrm{variates++}}[\mathcal{C}|\mathcal{A}]} \leq (1+\delta_w)^{k-1} + f(k)\cdot\delta_w\cdot(1+\delta_s)^{k-1}\cdot\varrho(R)$$

In particular, if densities for all X_a lie in $[\epsilon_m, \epsilon_M] \not\ni 0$, then letting $\rho_{\mathcal{D}} \doteq \epsilon_M/\epsilon_m$,
for any $0 < \delta < 1/2$ and $k \leq \frac{\delta^2}{4\rho_{\mathcal{D}}}\cdot\sqrt{m}$, there is prob. $\geq 1-\delta$ (sampling of $\mathcal{A}$) that

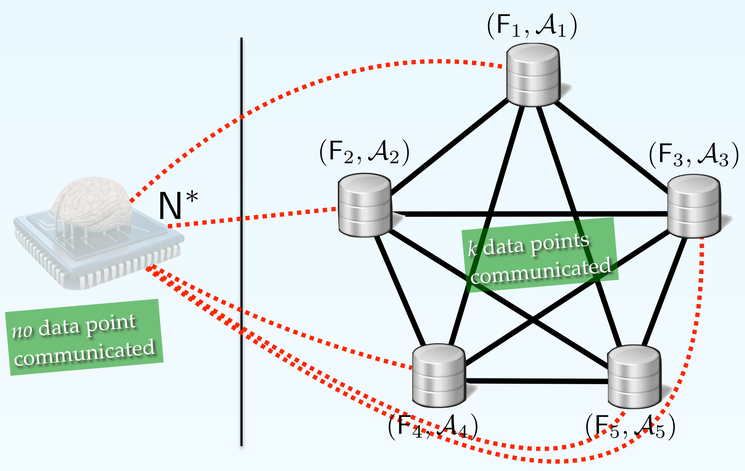
$$\mathbb{P}[\mathcal{C}|\mathcal{A}']/\mathbb{P}[\mathcal{C}|\mathcal{A}] \leq 1 + \rho_{\mathcal{D}}^k \cdot \underbrace{\left(\frac{4}{m^{\frac{1}{4}+\frac{1}{d+1}}} + \left(\frac{64}{k^{\frac{2}{d}}}\right)^k \cdot \frac{\varrho(2R)}{m}\right)}_{o(1)} \quad \forall \mathcal{A}' \approx \mathcal{A}$$

no δ_w, δ_s in bound (proof exhibits small δ_w, δ_s whp)

general likelihood ratio for *k*-variates++

applications: reductions from *k*-variates++ to get approximability ratios

Principle: show algorithm is equivalent to *k*-variates++ for specific X_a and $\wp_t$, and get approximability ratio !
Example: **distributed clustering**
(see paper for *online* or *streaming clustering*)
n "Forgy" nodes (*uniform* seeding) $(\mathsf{F}_i, \mathcal{A}_i)$, handling subset of data
1 "special node" (*non-uniform* seeding, can be "Forgy")



Algorithm 1 D-*k*-means++

Input: Forgy nodes $(\mathsf{F}_i, \mathcal{A}_i), i \in [n]$, (see paper)
for $t = 1, 2, ..., k$
(Round 0 : elicit N^*)
Round 1 : N^* picks $i^* \sim_{\mathrm{biased}} [n]$ and asks F_{i^*} for a center;
Round 2 : F_{i^*} picks $\boldsymbol{a} \sim_{\mathrm{uniform}} \mathcal{A}_{i^*}$ and sends $\boldsymbol{a}$ to $\mathsf{F}_i, \forall i$;
Round 3 : $\forall i, \mathsf{F}_i$ updates $D_t(\mathcal{A}_i)$ and sends it to N^*;
Output: $\mathcal{C}$ = set of broadcasted $\boldsymbol{a}$s;

Thm: let $\phi_s^F \doteq \sum_{i\in[n]}\sum_{\boldsymbol{a}\in\mathcal{A}_i} \|\boldsymbol{c}(\mathcal{A}_i)-\boldsymbol{a}\|_2^2$, with
Then Thm 1 holds for D-*k*-means++, with

$$\Phi \doteq 10\phi_{\mathrm{opt}} + 6\phi_s^F$$

as n increases, ϕ_s^F may decrease: bound gets better. Competitive for highly distributed and "dense" nodes (e.g. wrt Balcan et al., proc. NIPS 2013)

applications: (differential) privacy with k-variates++

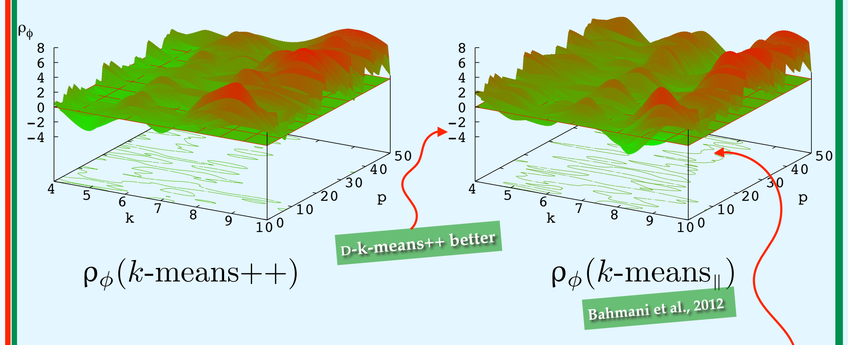
Differential private release of $\mathcal{C}$ ($\mathbb{P}[\mathcal{C}|\mathcal{A}']/\mathbb{P}[\mathcal{C}|\mathcal{A}] \leq \exp\epsilon$) with Laplace mechanism (X_a) can be ensured while guaranteeing

$$\Phi = \tilde{O}\left(\phi_{\mathrm{opt}} + \frac{mR^2}{(\epsilon+\log m)^2}\right)$$

sublinear noise

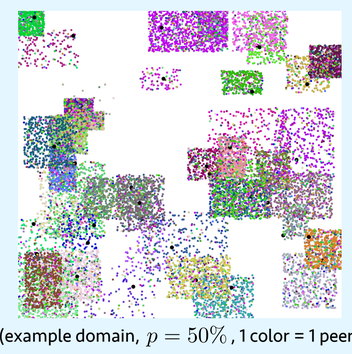
experiments

distributed clustering:
Simulated data, ***d***=50, sample peers with $\mathbb{E}[|\mathcal{A}_i|] = 500$ until $\sum_i |\mathcal{A}_i| \approx 20000$
For each peer, (a) data uniformly sampled in an hyper rectangle + (b) ***p***% of points given to a random peer (increases ϕ_s^F, problem more difficult)

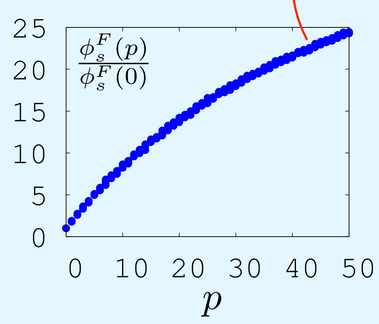


$$\rho_\phi(\mathcal{H}) \doteq \frac{\phi(\text{D-}k\text{-means++})-\phi(\mathcal{H})}{\phi(\mathcal{H})}\cdot 100$$

getting hard but D-k-means++ still competitive



(example domain, $p = 50\%$, 1 color = 1 peer)



D-*k*-means++ may compete with non distributed approach (k-means++) and optimised approach (Bahmani et al., proc. VLDB 2012), **even for large *p*, all the more if *k* not too large**

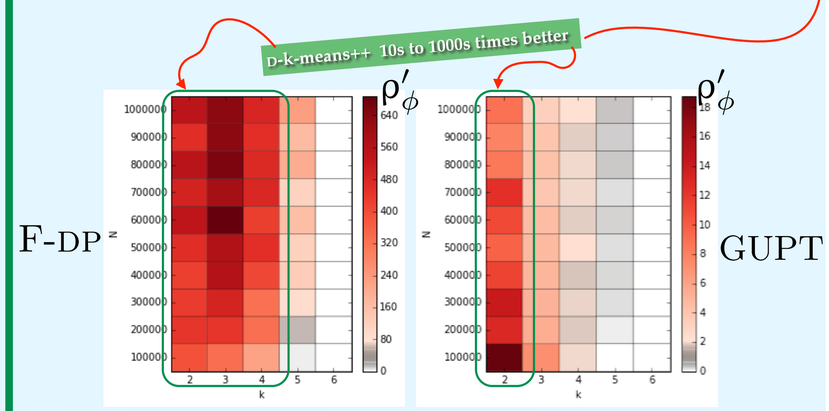
differential privacy:
Real world datasets: comparison with Forgy differentially private (F-DP) and GUPT

Dataset	m	d	k	ρ'_ϕ(Forgy-DP)	ρ'_ϕ(GUPT)
LifeSci	26733	10	2	311	1.6
			3	172	0.4
			4	6	0.02
Image	34112	3	2	300	4.8
			3	77	0.9
EuropeDiff	169308	2	2	1200	46.1
			3	3120	66.5
			4	3750	55.0
			5	4000	51.0
			6	5000	36.0
			7	2600	26.0
			8	350	2.0

$$\rho'_\phi(\mathcal{H}) \doteq \frac{\phi(\mathcal{H})}{\phi(k\text{-variates++})}$$

Simulated datasets: uniform ball sampling, ***d***=15

D-k-means++ 10s to 1000s times better



k-variates++ (predictably) beats Forgy-DP and competes with GUPT (Mohan et al., proc. SIGMOD 2012), **all the more if *k* not too large and *m* large (and in general, *d* large)**