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

**Input:** data \( A \subset \mathbb{R}^d \) with \( |A| = m, k \in \mathbb{N} \), random variables \( \{X_a, a \in A\} \), probe functions \( \psi_i : A \rightarrow \mathbb{R}^d \) \((t \geq 1)\);  
Step 1: Initialise centers \( \mathcal{C}' = \emptyset \);  
Step 2: for \( t = 1, 2, \ldots \) \( k \)  
2.1: randomly sample \( a \sim_{un} A \), with \( q_i = u_{a_{un}} \) and, for \( t > 1 \),  
2.2: randomly sample \( s \sim_{un} X_a \);  
2.3: \( e \leftarrow e \cup \{s\} \);  
Output: \( \mathcal{C}' \).

**Suppose** \( \psi_i \) is \( n \)-stretching: for any optimal cluster \( A \) (size \( > 1 \)) and \( a_0 \in A \), we have:

\[
\phi_{\mathcal{C}(\psi_i)}(A; a_0) = \arg \min_{a \in A} \phi(\psi_i(A); \{\psi_i(a_0)\}) \leq 1 + \eta \cdot \frac{\phi(\psi_i(A); e)}{\phi(\psi_i(A); \{\psi_i(a_0)\})} \quad \forall \mathcal{C}(\psi_i) \in \mathcal{C}.
\]

Then

\[
\mathbb{E} e \sim_{k \text{-variates}} [\phi(A; \mathcal{C})] \leq (2 + \log k) \cdot \Phi,
\]

with \( \Phi = \{ (6 + 4 \eta) \phi_{\text{opt}} + 2 \phi_{\text{bias}} + 2 \phi_{\text{var}} \}

\[= \sum_{a \in A} (a - c_{\text{opt}}(a))^2 \]

\[= \sum_{a \in A} \left[ \mathbb{E}[X_a] - c_{\text{opt}}(a) \right]^2 \]

\[= \sum_{a \in A} \text{tr} (\text{cov}[X_a]) \]

**Then, for any neighbour \( A' \approx A \) (differ by one arbitrary element) and with \( \psi_i = \text{Id}(A) \),

\[
\mathbb{E} e \sim_{k \text{-variates}} [\phi(A; \mathcal{C})] \leq (1 + \delta w)^{-k-1} f(k) \cdot \delta w \cdot (1 + \delta)^{-k-1} \cdot \varrho(R).
\]

**In particular,** if densities for all \( X_a \) lie in \( \{e_m, \mathcal{E}_A \} \), then there is \( \geq 1 - \delta \) (sampling of \( A \))

\[
P[\mathcal{C}[A'] / \mathcal{C}[A]] \leq 1 + \rho_{p_{k}} \cdot \left( \frac{4}{m + \frac{1}{2}} \right) + \left( \frac{64 k^4}{k^4} \right) \cdot \frac{\varrho(R)}{m}
\]

**Differential private release of \( \varrho(\mathcal{C}[A']) / \varrho(\mathcal{C}[A]) \leq \exp(\epsilon) \) with Laplace mechanism \( (\mathcal{M}) \) can be ensured while guaranteeing

\[
\phi = O \left( \phi_{\text{opt}} + \frac{m R^2}{(\epsilon + \log m)^2} \right)
\]

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

**Algorithm 1:** k-means++

**Input:** Forgotten nodes \( (F_i, A_i) \in \mathcal{B} \)

**Output:** \( \mathcal{C} \) set of centres

1. for \( t = 1, 2, \ldots, k \) \( (\text{Round } t \text{ of } \mathcal{B}) \)
2. Round 1: \( N \) peers \( \sim \mathbb{N} \) and asks \( F_i \) for a center;
3. Round 2: \( F_i \) peers \( \sim \mathbb{N} \), \( A_i \), and sends \( a \) to \( F_i \); \( \psi_i \);
4. Round 3 \( \sim \psi_i \), \( F_i \) updates \( c_{\text{opt}}(a) \), and sends it to \( N \).
5. \( (\text{Output: } C) \)

**Applications:**  
**Optimality:** show algorithm is equivalent to k-means++ for specific \( X_a \) and \( \psi_i \), and get approximability ratio +  
**Example:** distributed clustering  
(see paper for online or streaming clustering)

**n** "Forgy" nodes (uniform sampling) \( F_i, A_i \), handling subset of data
1 "special node" (non-uniform sampling, can be "Forgy")

**Distributed clustering:**  
Simulated data, \( d=50 \), sample peers with \( X_a | A \sim \mathcal{E}_A \) \( \mathcal{E}_A \) \( \approx 50000 \)  
For each peer, (a) data uniformly sampled in an hyper rectangle + (b) \( p \% \) of points given to a random peer (increases \( \phi_{k} \), problem more difficult)

**Differential privacy:**  
Real world datasets: comparison with Forgy differentially private (F-DP) and G UPT

<table>
<thead>
<tr>
<th>Dataset</th>
<th>m</th>
<th>n</th>
<th>( \phi_{k} ) (Forgy-avg)</th>
<th>( \phi_{k} ) (G UPT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LifeSci</td>
<td>10</td>
<td>2</td>
<td>120</td>
<td>10</td>
</tr>
<tr>
<td>Image</td>
<td>4</td>
<td>2</td>
<td>200</td>
<td>20</td>
</tr>
<tr>
<td>Europe</td>
<td>4</td>
<td>2</td>
<td>600</td>
<td>60</td>
</tr>
<tr>
<td>Diff</td>
<td>6</td>
<td>2</td>
<td>5000</td>
<td>5000</td>
</tr>
</tbody>
</table>

**Simulated datasets:** uniform ball sampling, \( d=15 \)

**Differential privacy:**

\[
\phi_{\text{GUP}}(X) = \phi_{\text{GUP}}(X|\mathcal{K}(X) = \emptyset)
\]

\[
\phi_{\text{GUP}}(X) = \phi_{\text{GUP}}(X|\mathcal{K}(X) = \emptyset)
\]