SoK: Efficient Privacy-preserving Clustering

Aditya Hegde, Helen Möllering, Thomas Schneider, Hossein Yalame

Partitioning-based Clustering

Distribution-based Clustering

Density-based Clustering

Hierarchical Clustering
<table>
<thead>
<tr>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Motivation and Preliminaries</strong></td>
</tr>
<tr>
<td><strong>2. Survey of Private Clustering</strong></td>
</tr>
<tr>
<td><strong>3. Evaluation of State-of-the-Art Protocols</strong></td>
</tr>
<tr>
<td><strong>4. Challenges to Real-life Application</strong></td>
</tr>
</tbody>
</table>
Clustering is applied on highly sensitive information
Our Contributions

- First comprehensive review and analysis of private clustering protocols
- Guideline on how to choose an appropriate private clustering protocol for concrete applications
- Open-source implementation and benchmark of four most efficient, fully private clustering schemes: [CKP19], [MPO+19], [MRT20], [BCE+21]
59 works were analyzed
Fully private clustering does not leak anything beyond the output

Ideal Functionality

\[ \emptyset | \text{output(a)} \]

\[ \emptyset | \text{output(b)} \]
Fully private clustering does not leak anything beyond the output

Ideal Functionality

- **TTP**
- $\emptyset | output(a)$
- $\emptyset | output(b)$

Requirements

- Privacy
- Efficiency
- Clustering Quality
- Flexibility
## Agenda

1. **Motivation and Preliminaries**
2. **Survey of Private Clustering**
3. **Evaluation of State-of-the-Art Protocols**
4. **Challenges to Real-life Application**
Multiple aspects influence the choice for a private clustering scheme

Plaintext Algorithm

K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift
Multiple aspects influence the choice for a private clustering scheme

<table>
<thead>
<tr>
<th>Plaintext Algorithm</th>
<th>K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Model</td>
<td>Semi-honest, Malicious</td>
</tr>
</tbody>
</table>
Multiple aspects influence the choice for a private clustering scheme

| Plaintext Algorithm                           | K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift |
| Security Model                               | Semi-honest, Malicious |
| Scenarios                                    | 2PC/MPC, Outsourcing |
Multiple aspects influence the choice for a private clustering scheme

<table>
<thead>
<tr>
<th>Plaintext Algorithm</th>
<th>K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Model</td>
<td>Semi-honest, Malicious</td>
</tr>
<tr>
<td>Scenarios</td>
<td>2PC/MPC, Outsourcing</td>
</tr>
<tr>
<td>Data Partition</td>
<td>horizontal (h), vertical (v), arbitrary (a)</td>
</tr>
</tbody>
</table>
Multiple aspects influence the choice for a private clustering scheme

<table>
<thead>
<tr>
<th>Plaintext Algorithm</th>
<th>K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Model</td>
<td>Semi-honest, Malicious</td>
</tr>
<tr>
<td>Scenarios</td>
<td>2PC/MPC, Outsourcing</td>
</tr>
<tr>
<td>Data Partition</td>
<td>horizontal (h), vertical (v), arbitrary (a)</td>
</tr>
<tr>
<td>PETs</td>
<td>Homomorphic Encryption (HE, [GB09]), Public Key Cryptography, Garbled Circuits (GC, [Yao86]), Arithmetic Secret-Sharing (ASS, [GMW87])</td>
</tr>
</tbody>
</table>
Multiple aspects influence the choice for a private clustering scheme

<table>
<thead>
<tr>
<th>Plaintext Algorithm</th>
<th>K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Model</td>
<td>Semi-honest, Malicious</td>
</tr>
<tr>
<td>Scenarios</td>
<td>2PC/MPC, Outsourcing</td>
</tr>
<tr>
<td>Data Partition</td>
<td>horizontal (h), vertical (v), arbitrary (a)</td>
</tr>
<tr>
<td>PETs</td>
<td>Homomorphic Encryption (HE, [GB09]), Public Key Cryptography, Garbled Circuits (GC, [Yao86]), Arithmetic Secret-Sharing (ASS, [GMW87])</td>
</tr>
<tr>
<td>Privacy</td>
<td>Fully privacy-preserving, Leakage</td>
</tr>
</tbody>
</table>
Multiple aspects influence the choice for a private clustering scheme

<table>
<thead>
<tr>
<th>Plaintext Algorithm</th>
<th>K-means, K-medoid, Mean-shift, Gaussian Mixture Models Clustering (GMM), DBSCAN, hierarchical clustering (HC), Affinity Propagation, Mean-shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Model</td>
<td>Semi-honest, Malicious</td>
</tr>
<tr>
<td>Scenarios</td>
<td>2PC/MPC, Outsourcing</td>
</tr>
<tr>
<td>Data Partition</td>
<td>horizontal (h), vertical (v), arbitrary (a)</td>
</tr>
<tr>
<td>PETs</td>
<td>Homomorphic Encryption (HE, [GB09]), Public Key Cryptography, Garbled Circuits (GC, [Yao86]), Arithmetic Secret-Sharing (ASS, [GMW87])</td>
</tr>
<tr>
<td>Privacy</td>
<td>Fully privacy-preserving, Leakage</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Computation, Communication, Memory</td>
</tr>
</tbody>
</table>
There are only 10 fully private clustering schemes

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Paper</th>
<th>PETs</th>
<th>Scenario</th>
<th>Data</th>
<th>Output</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HE</td>
<td>GC</td>
<td>MIX</td>
<td>MPC</td>
<td>Out h a</td>
</tr>
<tr>
<td>K-means</td>
<td>[BO07]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[RSB+16]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[JA18]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[KC18]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[MRT20]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mean-shift</td>
<td>[CPK19]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Affinity Prop.</td>
<td>[KMS+21]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>[ZE13]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[BCE+21]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HC</td>
<td>[MPO+19]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
There are only 10 fully private clustering schemes

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Paper</th>
<th>PETs</th>
<th>Scenario</th>
<th>Data</th>
<th>Output</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>[BO07]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>final centroids</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[RSB+16]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>final centroids</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[JA18]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>final centroids</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[KC18]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>cluster sizes</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[MRT20]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>final centroids</td>
<td>✓</td>
</tr>
<tr>
<td>Mean-shift</td>
<td>[CKP19]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>final centroids</td>
<td>✓</td>
</tr>
<tr>
<td>Affinity Prop.</td>
<td>[KMS+21]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>final clusters</td>
<td>✓</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>[ZE13]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Cluster labels</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>[BCE+21]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Cluster labels</td>
<td>✓</td>
</tr>
<tr>
<td>HC</td>
<td>[MPO+19]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Final dendrogram</td>
<td>✓</td>
</tr>
</tbody>
</table>
Agenda

1. Motivation and Preliminaries

2. Survey of Private Clustering

3. Evaluation of State-of-the-Art Protocols

4. Challenges to Real-life Application
Performance is the decisive metric

Small Datasets:
- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

- HE-Meanshift [CKP19]
- PCA/OPT [MPO+19]
- ppDBSCAN [BCE+21]
- MPC-KMeans [MRT20]
Performance is the decisive metric

Small Datasets:
- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

- **LAN**
  - HE-Meanshift [CKP19]
    - 1.7 hours
    - 1.7x
  - PCA/OPT [MPO+19]
    - 1 hour
    - 20x
  - ppDBSCAN [BCE+21]
    - 3 min
    - 7x
  - MPC-KMeans [MRT20]
    - 25 s
    - 7x
### Performance is the decisive metric

<table>
<thead>
<tr>
<th>Method</th>
<th>LAN</th>
<th>WAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE-Meanshift [CKP19]</td>
<td>1.7 hours</td>
<td>1.7 hours</td>
</tr>
<tr>
<td></td>
<td>1.7x</td>
<td>1.7x</td>
</tr>
<tr>
<td>PCA/OPT [MPO+19]</td>
<td>1 hour</td>
<td>3 hours</td>
</tr>
<tr>
<td></td>
<td>20x</td>
<td>10x</td>
</tr>
<tr>
<td>ppDBSCAN [BCE+21]</td>
<td>3 min</td>
<td>20 min</td>
</tr>
<tr>
<td></td>
<td>7x</td>
<td>1.2x</td>
</tr>
<tr>
<td>MPC-KMeans [MRT20]</td>
<td>25 s</td>
<td>17 min</td>
</tr>
</tbody>
</table>

**Small Datasets:**
- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$
Performance is the decisive metric

<table>
<thead>
<tr>
<th>Method</th>
<th>Small Datasets</th>
<th>Large Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE-Meanshift [CKP19]</td>
<td>1.7 hours, 1.7x</td>
<td>1.7 hours, 1.7x</td>
</tr>
<tr>
<td>PCA/OPT [MPO+19]</td>
<td>1 hour, 20x</td>
<td>3 hours, 10x</td>
</tr>
<tr>
<td>ppDBSCAN [BCE+21]</td>
<td>3 min, 7x</td>
<td>20 min, 1.2x</td>
</tr>
<tr>
<td>MPC-KMeans [MRT20]</td>
<td>25 s, 1.7 hours</td>
<td>23 hours, 5x</td>
</tr>
</tbody>
</table>

**Small Datasets:**
- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

**Large Datasets:**
- Number of points: $2^{13} \leq N \leq 2^{16}$
- Dimension: $1 \leq d \leq 16$
- Number of clusters: $2 \leq K \leq 20$
Performance is the decisive metric

### Small Datasets
- **HE-Meanshift** [CKP19]
  - LAN: 25 s, 7x
  - WAN: 17 min, 1.2x
  - 1.7 hours

- **PCA/OPT** [MPO+19]
  - LAN: 3 min, 20x
  - WAN: 20 min, 10x
  - 1 hour

- **ppDBSCAN** [BCE+21]
  - LAN: 1.7 hours
  - WAN: 1.7 hours
  - 1.7x

- **MPC-KMeans** [MRT20]

### Large Datasets
- **HE-Meanshift** [CKP19]
  - LAN: 23 hours

- **PCA/OPT** [MPO+19]
  - LAN: 5 hours

- **ppDBSCAN** [BCE+21]
  - LAN: 5 hours

- **MPC-KMeans** [MRT20]

**Performance strongly affects choice of protocol.**

**Small Datasets:**
- Number of points: $50 \leq N \leq 200$
- Dimension: $1 \leq d \leq 8$
- Number of clusters: $2 \leq K \leq 10$

**Large Datasets:**
- Number of points: $2^{13} \leq N \leq 2^{16}$
- Dimension: $1 \leq d \leq 16$
- Number of clusters: $2 \leq K \leq 20$
Several factors affect clustering quality

- Protocol/Algorithm
- Parameters
- Randomness
### Agenda

1. **Motivation and Preliminaries**

2. **Survey of Private Clustering**

3. **Evaluation of State-of-the-Art Protocols**

4. **Challenges to Real-life Application**
Plaintext clustering eases parameter selection
Plaintext clustering eases parameter selection

Choose Clustering Algorithm and its Parameters
Plaintext clustering eases parameter selection

Choose Clustering Algorithm and its Parameters
Plaintext clustering eases parameter selection

Choose Clustering Algorithm and its Parameters

Evaluate
- Visual
- Clustering indices
Plaintext clustering eases parameter selection

Choose Clustering Algorithm and its Parameters

Different Random Seed
Plaintext clustering eases parameter selection

Choose Clustering Algorithm and its Parameters

Evaluate
- Visual
- Clustering indices

Different Random Seed
Distributed data and protocol efficiency are the main challenges
Distributed data and protocol efficiency are the main challenges

Choose Clustering Protocol and Parameters
- Preliminary analysis of dataset
- Parameters depend on input data
- Efficiency determines protocol choice
Distributed data and protocol efficiency are the main challenges

Choose Clustering Protocol and Parameters
- Preliminary analysis of dataset
- Parameters depend on input data
- Efficiency determines protocol choice

Overhead of secure computation
Distributed data and protocol efficiency are the main challenges

Choose Clustering Protocol and Parameters
- Preliminary analysis of dataset
- Parameters depend on input data
- Efficiency determines protocol choice

Evaluate
- Securely computing clustering indices
- Handling outliers and noise
Future research directions for private clustering

- **Efficiency**: runtime, communication, and memory
- Parameters that can be set *independent* of input data
- Protocols that handle *outliers* and *noise*
- Techniques to securely *evaluate* clustering output
THANKS FOR YOUR ATTENTION!

Contact: [https://encyrpto.de/moellering](https://encyrpto.de/moellering)
Code: [https://encyrpto.de/code/SoK_ppClustering](https://encyrpto.de/code/SoK_ppClustering)
References (1)


References (2)


