Federated Spectral Clustering via Secure Similarity Reconstruction

Dong Qiao[1,2], Chris Ding[1], Jicong Fan[1,2*]

[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Research Institute of Bigdata, Shenzhen

December 9, 2023

Motivation

- Spectral Clustering (SC)
  - Step 1: Construct a similarity matrix;
  - Step 2: Perform normalized cut and Kmeans [Shi and Malik, 2000]
- Limitations of SC
  - SC heavily relies on the construction of similarity matrix;
  - SC cannot be directly performed on distributed dataset;
  - Vanilla SC cannot preserve the user's privacy.

$$K_{xx} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & k(\boldsymbol{x}_n, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}$$

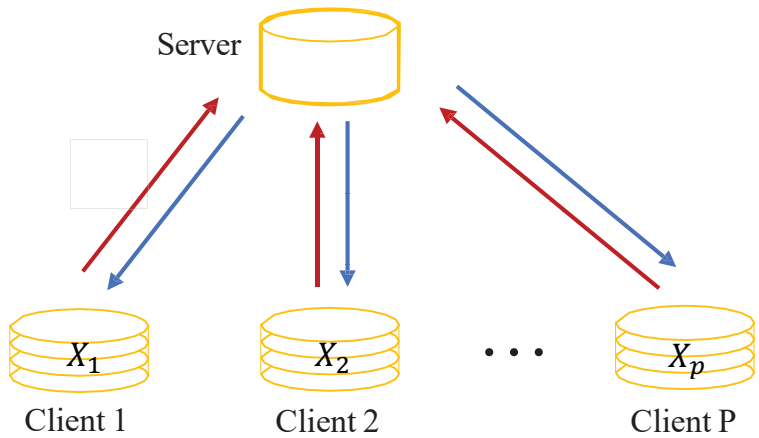


Figure: Distributed data

Similarity Reconstruction via Feature Space Factorization

- Gaussian kernel: $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \exp(-\frac{1}{r^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2)$
- Nonlinear approximation: $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \simeq \widehat{\phi(\boldsymbol{x}_i)}^T \widehat{\phi(\boldsymbol{x}_j)}$ for $\widehat{\phi(\boldsymbol{x}_i)} = \phi(\boldsymbol{Z})\boldsymbol{c}_i$
- For a client p with $\boldsymbol{X}_p \in \mathbb{R}^{m \times N_p}$, the nonlinear approximation has a matrix form of

$$\widehat{\phi(\boldsymbol{X}_p)} \simeq \phi(\boldsymbol{Z})\boldsymbol{C}_p$$

- To do the nonlinear approximation for P clients, we consider the distributed optimization problem:

$$\min_{\boldsymbol{Z},\boldsymbol{C}} \triangleq \sum_{p=1}^{P} \omega_p f_p(\boldsymbol{Z}, \boldsymbol{C}_p)$$

where $f_p(\boldsymbol{Z}, \boldsymbol{C}_p) = \frac{1}{2}\|\phi(\boldsymbol{X}_p) - \phi(Z)\boldsymbol{C}_p\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{C}_p\|_F^2$

Our design: FedSC

- **Stage I**: Federated similarity reconstruction (See left part in Fig 1)
- (1) Local update scheme:

$$\boldsymbol{C}_p^s = \arg\min_{\boldsymbol{C}_p} f_p(\boldsymbol{Z}^{s-1}, \boldsymbol{C}_p)$$

$$\boldsymbol{Z}_p^s = \arg\min_{\boldsymbol{Z}_p} f_p(\boldsymbol{Z}_p, \boldsymbol{C}_p^s)$$

- (2) Aggregation scheme:

$$\boldsymbol{Z}^s = \frac{1}{|\mathcal{A}^{s-1}|} \sum_{p \in \mathcal{A}^{s-1}} \boldsymbol{Z}_p^s$$

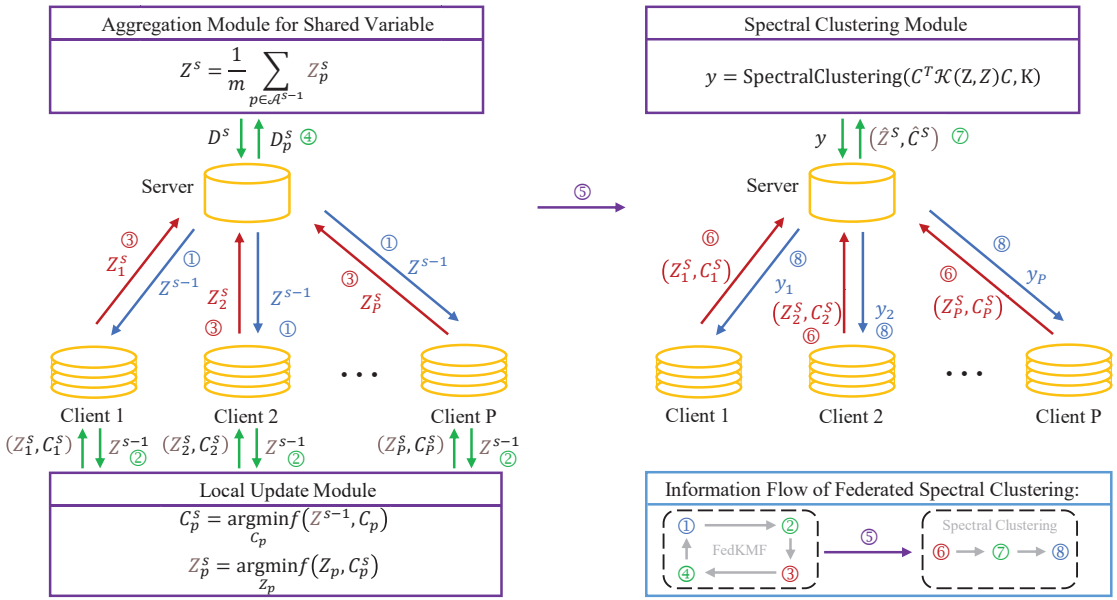


Figure: FedSC

- **Stage II**: Spectral Clustering (See right part in Fig 1)

$$y = \text{SpectralClustering}(\boldsymbol{C}^T \mathcal{K}(\boldsymbol{Z}, \boldsymbol{Z})\boldsymbol{C}, K)$$

Our design: Secure-Enhanced FedSC

- **Alternative 1**: FedSC with perturbed data
  Raw data $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is perturbed by a noise matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ to form the noisy data matrix:

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}$$

- **Alternative 2**: FedSC with perturbed factors
  We added Gaussian noise to $\boldsymbol{Z}$ in every round but added to Gaussian noise to $\boldsymbol{C}$ in the last round. That is,

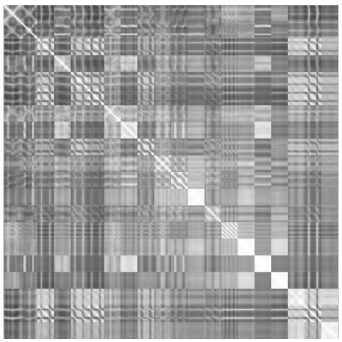
$$\begin{cases} \tilde{\boldsymbol{Z}}_p^s = \boldsymbol{Z}_p^s + \boldsymbol{E}_{Z_p}^s & s = 1, \ldots, S; p = 1, \ldots, P \\ \tilde{\boldsymbol{C}}_p^S = \boldsymbol{C}_p^S + \boldsymbol{E}_{C_p} & p = 1, \ldots, P \end{cases}$$

  where $\boldsymbol{E}_{Z_p}^s$ and $\boldsymbol{E}_{C_p}$ are drawn from $\mathcal{N}(0, \sigma_Z^2)$ and $\mathcal{N}(0, \sigma_C^2)$. Then, we perform SC on

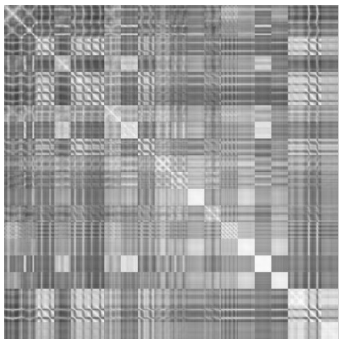
$$\tilde{K}_{xx} = \tilde{\boldsymbol{C}}^T \mathcal{K}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{Z}}) \tilde{\boldsymbol{C}}$$

Numerical results

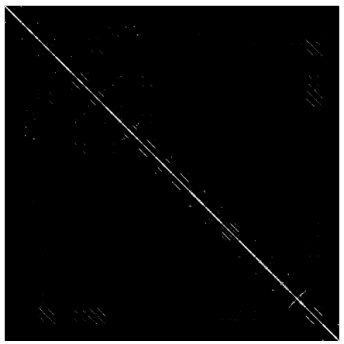
- Part I: Similarity Reconstruction



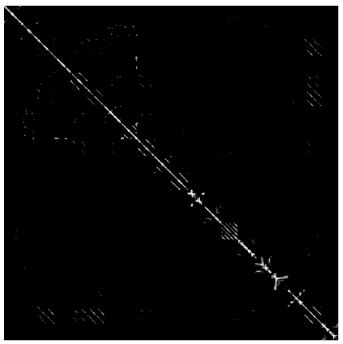
(a) Vanilla SC



(b) FedSC



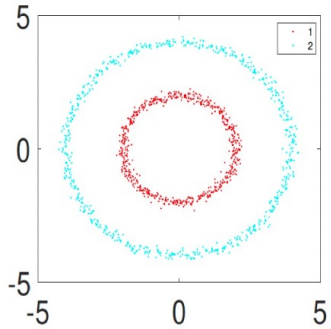
(c) Vanilla SC



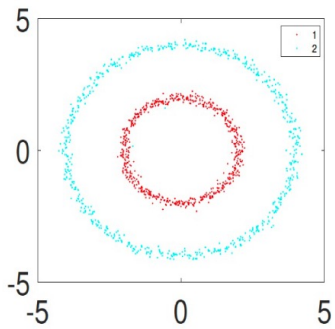
(d) FedSC

Figure: COIL20

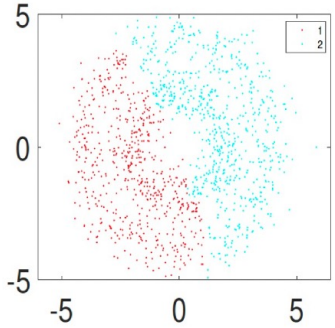
- Part II: Clustering performance of FedSC



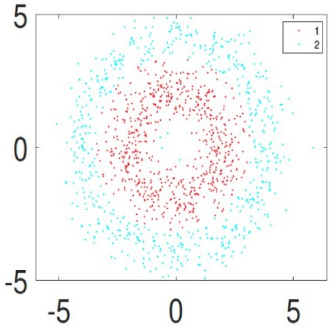
(a) Vanilla SC



(b) FedSC



(c) Vanilla SC



(d) FedSC

Figure: Concentric circles

Numerical results

- Part III: Comparison with baselines

Table: Comparison of clustering accuracy

		Kmeans	SC	DSC	FedSC
X	Iris	0.8933 ± 0.0000	0.9000 ± 0.0000	0.5480 ± 0.0679	0.9000 ± 0.0031
	COIL20	0.6113 ± 0.0534	0.8025 ± 0.0009	0.1009 ± 0.0100	0.7828 ± 0.0231
	MNIST	0.5110 ± 0.0424	0.5872 ± 0.0468	0.1340 ± 0.0113	0.6002 ± 0.0410
	CIFAR10	0.2171 ± 0.0132	0.2182 ± 0.0133	0.1235 ± 0.0062	0.2134 ± 0.0131
X with 0.3σ	Iris	0.8420 ± 0.0274	0.8327 ± 0.0267	0.4533 ± 0.0674	0.8427 ± 0.0404
	COIL20	0.6422 ± 0.0366	0.7997 ± 0.0029	0.0981 ± 0.0084	0.7793 ± 0.0240
	MNIST	0.5530 ± 0.0364	0.6246 ± 0.0492	0.1355 ± 0.0114	0.5908 ± 0.0510
	CIFAR10	0.2209 ± 0.0154	0.2198 ± 0.0172	0.1194 ± 0.0051	0.2187 ± 0.0125
X with 0.7σ	Iris	0.6500 ± 0.0420	0.6087 ± 0.0468	0.3927 ± 0.0349	0.6120 ± 0.0455
	COIL20	0.6220 ± 0.0627	0.7662 ± 0.0172	0.0893 ± 0.0055	0.6803 ± 0.0198
	MNIST	0.5317 ± 0.0506	0.5763 ± 0.0385	0.1377 ± 0.0148	0.5181 ± 0.0462
	CIFAR10	0.2202 ± 0.0147	0.2205 ± 0.0084	0.1209 ± 0.0045	0.2134 ± 0.0152

Table: Clustering accuracy (average over 10 trials) of FedSC with perturbed factors on COIL20.

		α_c			
		0	0.05	0.1	0.15
		Mean±Std	Mean±Std	Mean±Std	Mean±Std
α_z	0	0.7831±0.0268	0.6573±0.0291	0.6012±0.0391	0.4835±0.0526
	0.05	0.7824±0.0126	0.6573±0.0243	0.6165± 0.0286	0.4998±0.0567
	0.1	0.7817±0.0299	0.6625±0.0258	0.6453±0.0186	0.5508±0.0415
	0.15	0.7881±0.0223	0.6526±0.0258	0.6219±0.0467	0.5901±0.0326

Conclusion

- Secure kernelized factorization method for federated spectral clustering on distributed data
- Theoretical guarantees for optimization convergence, correct clustering, and differential privacy
- Numerical experiments on synthetic and real image datasets

THANK YOU