Professorship of Data Mining and Analytics
Department of Informatics
Technical University of Munich

Robust Spectral Clustering for Noisy Data

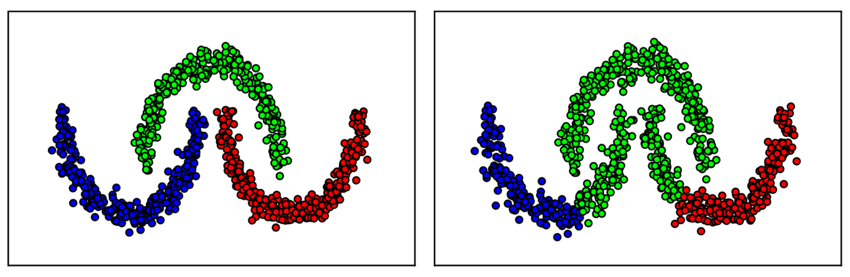
Modeling Sparse Corruptions Improves Latent Embeddings

Aleksandar Bojchevski, Yves Matkovic, Stephan Günnemann

MOTIVATION

- Spectral clustering (SC) **widely** used, but highly **sensitive** to noisy data
- Noise distorts the embedding space and obfuscates the clustering structure

Noise 0.07 | Noise 0.0875



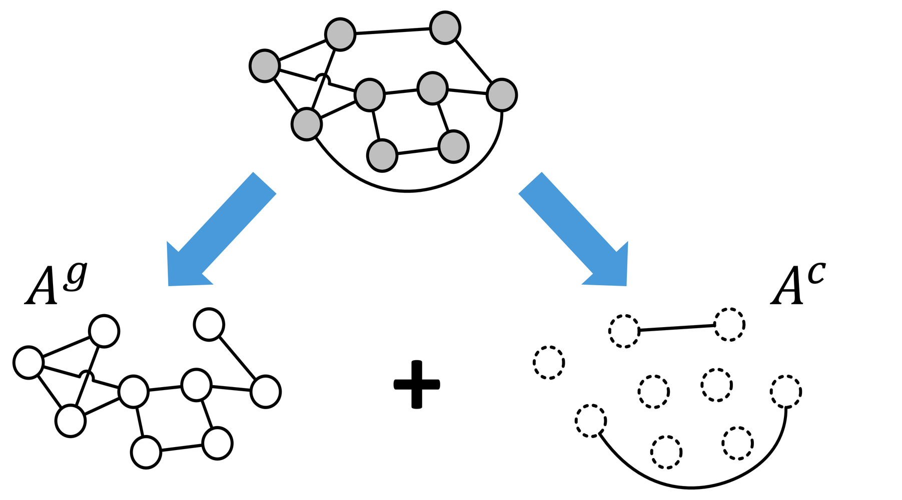
SC ✓ RSC ✓ | SC ✗ RSC ✓

- We propose a robust version: **RSC**

PROBLEM FORMULATION

❑ Core idea: **Latent Decomposition**

$$A = A^g + A^c$$



❑ Jointly learn decomposition and embedding

- Decomposition steered by the underlying clustering

$$A^*, H^* = \underset{H\in\mathbb{R}^{n\times d},\ A^g\in(\mathbb{R}_{\geq 0})^{n\times n}}{\operatorname{argmin}} \ Tr(H^T \cdot L(A^g)\cdot H) \qquad (1)$$

subject to: $H^T \cdot D(A^g)\cdot H = I$ and

$A = A^g + A^c, \|A^c\|_0 \leq 2\theta, \|a_i^c\|_0 \leq \omega_i$

- Robust formulation for all SC versions
- Result → improved embedding

ALGORITHMIC SOLUTION

Alternating Optimization

❑ Update H, Given A^g/A^c → ***Easy***

- Trace minimization problem
- Solution for H are the k first generalized eigenvectors of $L(A^g)$

❑ Update A^g/A^c, Given H → ***(NP) Hard***

- Express eigenvalues of A^g_{new} in closed form
- A^g_{new} that minimizes (1) equivalent to maximizing:

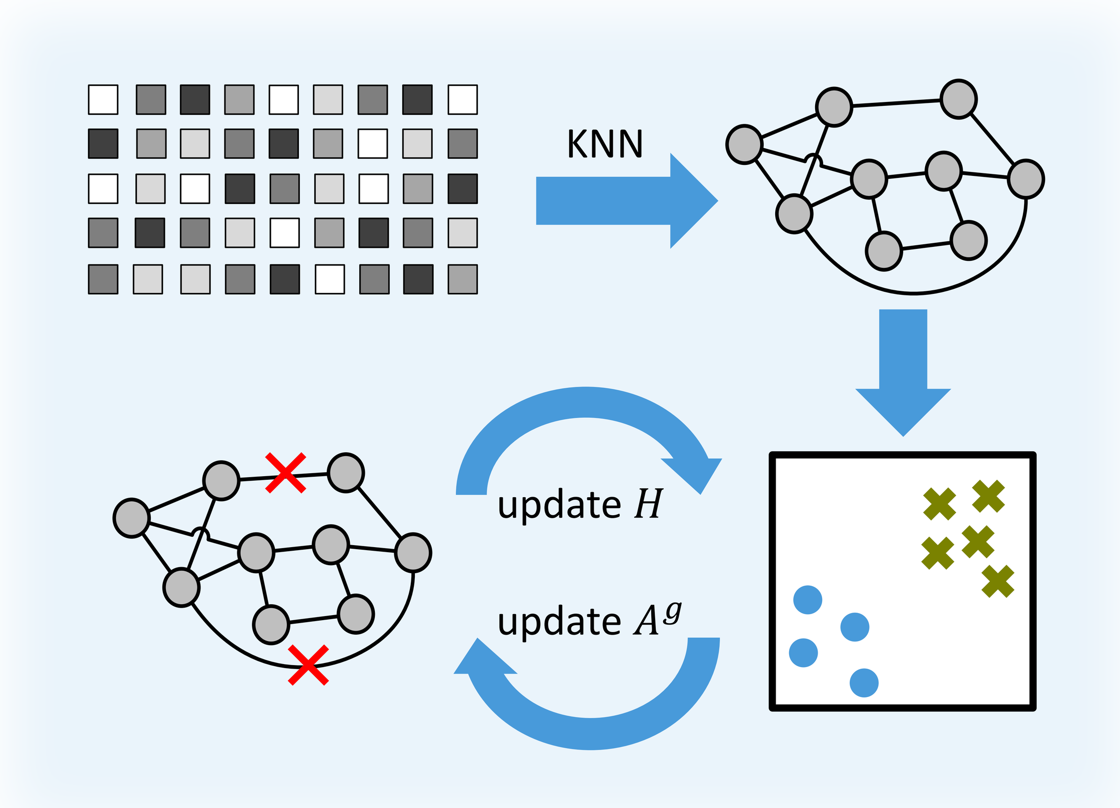
} Step 1: Eigenvalue Perturbation

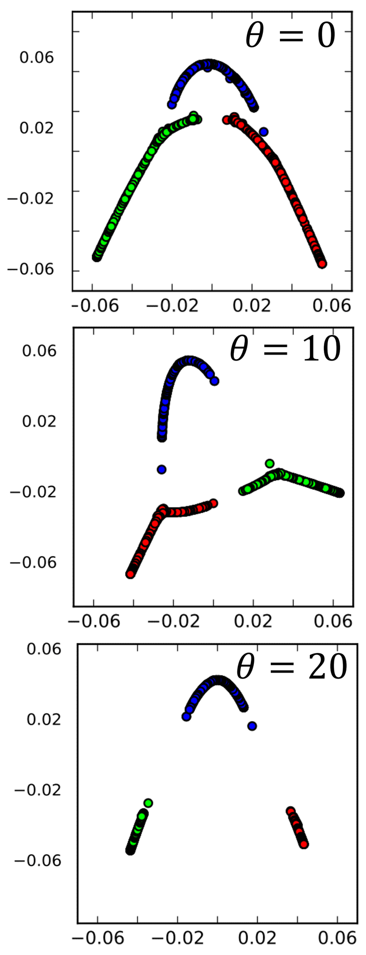
$$f([a^c_{uv}]_{u,v\in E}) = \sum_{u,v\in E} a^c_{uv}\left(\underbrace{\|\boldsymbol{h}_u - \boldsymbol{h}_v\|_2^2}_{nodes\ far\ away\ in\ the\ embedding\ space} - \underbrace{\left\|\sqrt{\lambda}\circ \boldsymbol{h}_u\right\|_2 - \left\|\sqrt{\lambda}\circ \boldsymbol{h}_v\right\|_2}_{prefers\ edges\ close\ to\ critical\ region}\right)$$

subject to $\|\cdot\|_0$ constraints

- Observation: Above problem is equivalent to **Multidimensional Knapsack** problem
- Greedy approximation scheme
- Best possible approximation ratio of $1/\sqrt{N+1}$

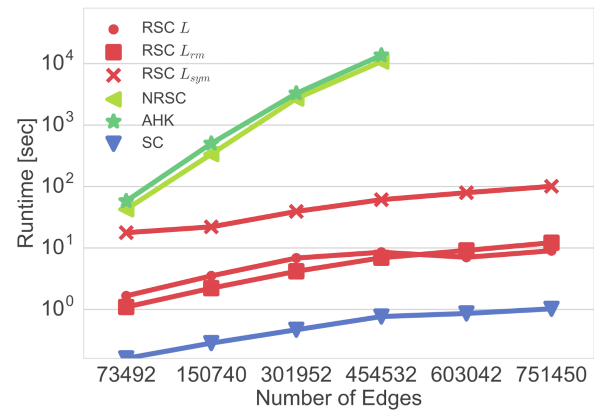
} Step 2: Efficient Solution





❑ Complexity

- Derivations for all 3 established versions of SC (different Laplacians)
- Each version has linear runtime in the number of edges $O(E)$



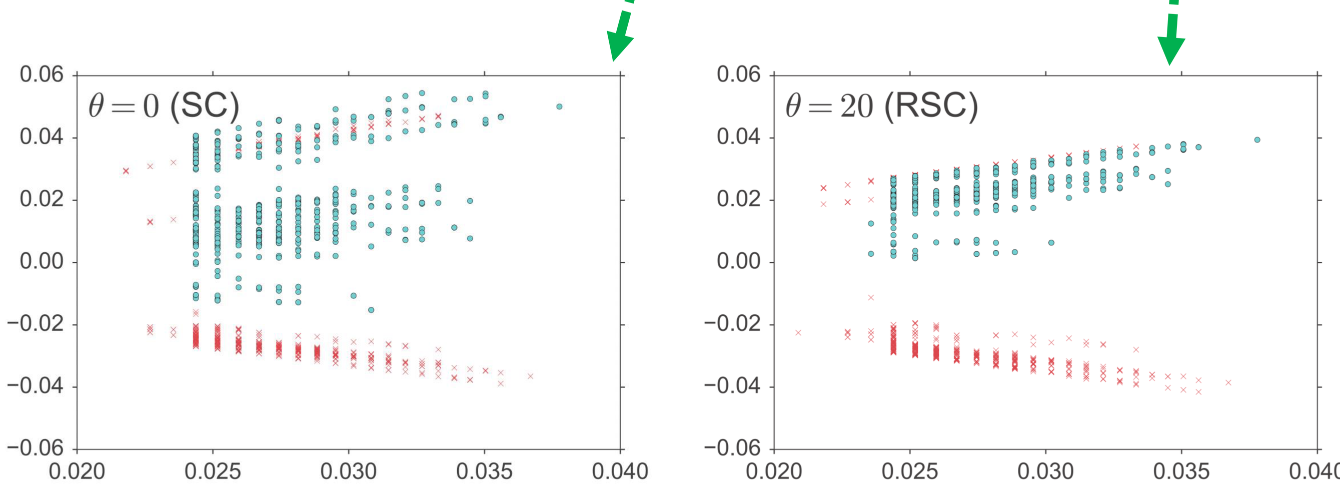
RESULTS

How can we evaluate the quality of the **clustering**?

❑ RSC improves the clustering as measured by the NMI

Dataset	AHK	NRSC	SC		RSC
Moons	0.53	0.99	0.47	+112.8 %	1.00
Banknote	0.53	0.47	0.46	+ 32.6 %	0.61
USPS	0.77	0.83	0.78	+ 8.9 %	0.85
MNIST	0.71	0.76	0.70	+ 11.4 %	0.78
Pendigits	0.94	0.94	0.93	+ 3.2 %	0.96

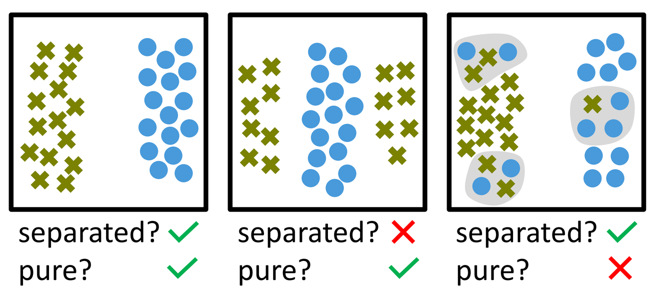
❑ RSC improves discrimination in the embeddding space



How can we evaluate the quality of the **embeddings**?

❑ Global Separation

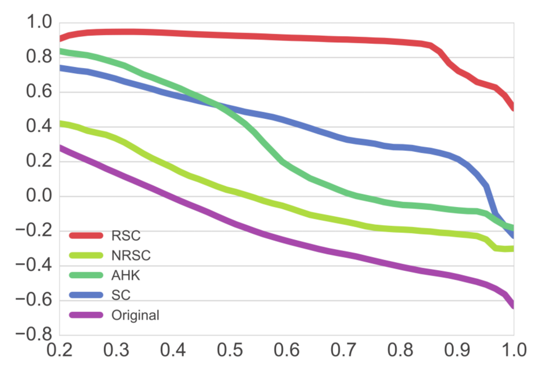
- Degree of separability between clusters
- Robust silhouette coefficient



$$P_{c,c'}(x) = \underset{x\,\%\,smallest}{\text{average}}\left[dist(\boldsymbol{h}_i,\boldsymbol{h}_j)\right]_{i\in C_c, j\in C_{c'}}$$

$$GS_c(x) = \frac{P_{c,c'}(x) - P_{c,c}(x)}{\max\{P_{c,c'}(x), P_{c,c}(x)\}}$$

$$c' = \underset{c\neq c'}{\operatorname{argmin}}\, P_{c,c'}(x)$$



(worst) $-1 \leq GS_c(x) \leq 1$ (best)

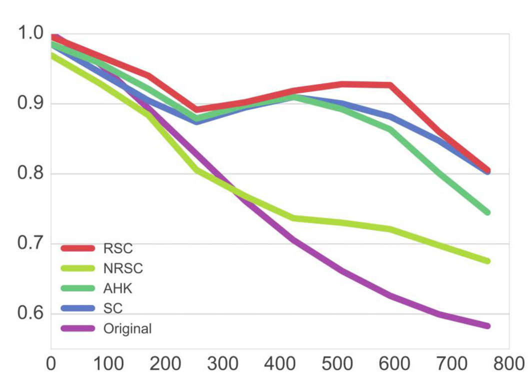
❑ Local Purity

- Homogeneous local neighborhood

$$occ_x(c,i) = \left|\{ j \in NN_x(i) \cup \{i\} \mid c_j = c\}\right|$$

$$pur_x(i) = \frac{1}{x+1}\max_{c\in C} occ_x(c,i)$$

$$\text{PUR}(x) = \frac{1}{N}\sum_{i=1}^{N} pur_x(i)$$



(worst) $\frac{1}{|C|} \leq PUR(x) \leq 1$ (best)