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
- We propose a robust version: RSC

PROBLEM FORMULATION
- Core idea: Latent Decomposition
  \[ A = A^\theta + A^c \]
  clean graph + sparse corruptions
- Jointly learn decomposition and embedding
  - Decomposition steered by the underlying clustering
    \[ A^*, H^* = \arg\min_{A^\theta, H} \quad \text{Tr}(H^T \cdot L(A^\theta) \cdot H) \]
    subject to: \[ H^T \cdot D(A^\theta) \cdot H = I \]
    and \[ A = A^\theta + A^c, \|A^c\|_0 \leq 2\beta, \|a_i^\theta\|_0 \leq \omega_i \]
  - Robust formulation for all SC versions
  - Result \( \rightarrow \) improved embedding

ALGORITHMIC SOLUTION
- Update \( H \), Given \( A^\theta / A^c \) \( \rightarrow \) Easy
  - Trace minimization problem
  - Solution for \( H \) are the \( k \) first generalized eigenvectors of \( L(A^\theta) \)
- Update \( A^\theta / A^c \), Given \( H \) \( \rightarrow \) (NP) Hard
  - Express eigenvalues of \( A^\theta_{\text{new}} \), in closed form
  - \( A^\theta_{\text{new}} \) that minimizes (1) equivalent to maximizing:
    \[ f(\{a_i^\theta\}_{i=1}^N) = \sum_{i \in E} \alpha_{ij}^\theta \left( \|h_i - h_j\|^2 \right)^{0.46} \left( \|h_i + h_j\|^2 \right)^{0.46} \]
    subject to \( \|h_i\| \) constraints
  - Observation: Above problem is equivalent to Multidimensional Knapsack problem
  - Greedy approximation scheme
  - Best possible approximation ratio of \( 1/\sqrt{N+1} \)
- 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AHK</th>
<th>NRSC</th>
<th>SC</th>
<th>RSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moons</td>
<td>0.53</td>
<td>0.99</td>
<td>0.47</td>
<td>+112.8</td>
</tr>
<tr>
<td>Banknote</td>
<td>0.53</td>
<td>0.47</td>
<td>0.78</td>
<td>+32.6</td>
</tr>
<tr>
<td>USPS</td>
<td>0.77</td>
<td>0.83</td>
<td>0.78</td>
<td>+8.9</td>
</tr>
<tr>
<td>MNIST</td>
<td>0.71</td>
<td>0.76</td>
<td>0.70</td>
<td>+11.4</td>
</tr>
<tr>
<td>Pendigits</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>+3.25</td>
</tr>
</tbody>
</table>

How can we evaluate the quality of the embeddings?
- Global Separation
  - Degree of separability between clusters
  - Robust silhouette coefficient
    \[ P_{c,c'}(x) = \text{average}_{a \in a_{c,c'}} \text{dist}(h_i, h_j)_{a_{c,c'}} \]
    \[ G_{c,c'}(x) = P_{c,c'}(x) - P_{c,c'}(c) \] \[ \max(P_{x,x'}, P_{x,x'}) \]
    \[ c' = \arg\min_{c \neq c'} P_{c,c'}(x) \]
- Local Purity
  - Homogeneous local neighborhood
    \[ a_{c,c'}(c) = |\{j \in N_{c'}(c) \cup \{i\}) | c = c \} \]
    \[ \text{pur}_{c,c'}(i) = \frac{1}{x+1} \max_{c \neq c'} a_{c,c'}(c,i) \]
    \[ \text{PUR}(c) = \frac{1}{N} \sum_{i=1}^{N} \text{pur}_{c,c'}(i) \]

August 13 - 17, 2017
Halifax, Nova Scotia - Canada