Clustering with Multi-Layer Graphs: A Spectral Perspective [1]

Souheib Ben Mabrouk - Imen Mahdi

Supervisor: Jean Feydey

MATHÉMATIQUES
VISION
APPRENTISSAGE

December 18, 2023

Contents

Current Section

Context: Multi-Layer Graphs

A multilayer graph is a collection of interconnected graphs that share a common set of vertices while featuring distinct sets of edges.
$\Longrightarrow$ inherent representation for **multi-model data**.

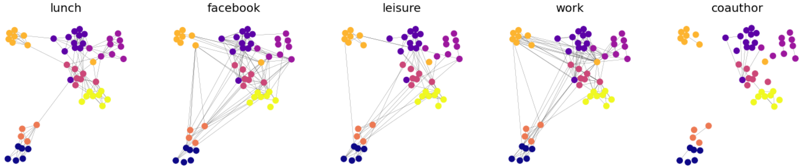

Figure 1: A multi-layer graph representing AUCS social network interactions.

A conventional clustering algorithm :

- Identifies k clusters in a **single-layer graph**.
- Selects the k smallest eigenvalues of graph Laplacian matrix (L) and corresponding eigenvectors $(u_1, ..., u_k)$, and applies K-means on rows of the spectral embedding matrix U composed from the eigenvectors as its columns.
- **Theoretical guarantees** of the effectiveness of vertex projection into a lower-dimensional spectral domain.

$\Longrightarrow$ **Extension to Multi-layer Graphs using a common spectrum of the Laplacian matrices of the different layers.**

Current Section

Methodology: Clustering with Generalized Eigen-Decomposition

- **Core Idea:**
  - **Aggregate** information from multiple layers : Find matrix P such that $L_{\mathrm{rw}}^{(i)} \approx P\Lambda^{(i)}P^{-1}$ for all i.
  - Prioritize information from the most relevant layer for clustering.
  - Compensation for sparse layers through the integration of information from remaining layers through **optimization**.
- **Optimization Problem:**

$$\min_{P,Q} \frac{1}{2}\sum_{i=1}^{M} \|L_{\mathrm{rw}}^{(i)} - P\Lambda^{(i)}Q\|_F^2 + \frac{\alpha}{2}(\|P\|_F^2 + \|Q\|_F^2) + \frac{\beta}{2}\|PQ - I_n\|_F^2$$

Methodology: Clustering with Spectral Regularization (1/2)

- **Core Idea:** Map the graph into 1-dimensional space such that projections of connected vertices are preserved as close as possible.
- Find a **smoothing** function $f : V \rightarrow \mathbb{R}$ that preserves the proximity between the vertices for each layer:

$$\arg\min_{f} f^T L f \;\; s.t \; \|f\| = 1, f \perp 1$$

- If the eigenvalues are small, the eigenvectors of L are **good candidates**.

Methodology: Clustering with Spectral Regularization (2/2)

- **For 2 layers:** we consider the first layer as **the most informative** and we compute the eigenvectors of $L_{rw}^{(0)}$ denoted $(u_1, ..., u_k)$

$$\textit{The smooth spectrum:} \quad \arg\min_f \underbrace{\frac{1}{2}\|f_i - u_i\|^2}_{f_i \text{ close to } u_i} + \underbrace{\lambda f_i^T L_{rw}^{(1)} f_i}_{\text{f smooth to the next layer}}$$

- **For m layers:** we can repeat the process by considering the current combination and the next informative layer.

$\Longrightarrow$ The final function is the **join spectrum** for the clustering.

Current Section

Cora[3]

- Extensive bibliographic repository with diverse research papers.
- Selected 900 papers within three ground truth clusters.

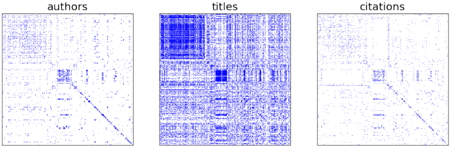

Figure 2: Spy plots of the adjacency matrices for the layers of the CORA dataset.

MIT Reality Mining [4]

- Social network of 94 mobile users on the MIT campus.
- After preprocessing, narrowed down to 90 subjects organized into 7 clusters.

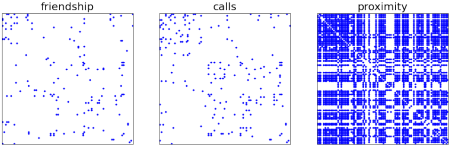

Figure 3: Spy plots of the adjacency matrices for the layers of the MIT dataset.

AUCS[5]

- Extension to the original work with relationships among 61 university employees.
- Each subject is affiliated with one of 8 research groups, designated as the ground truth cluster.

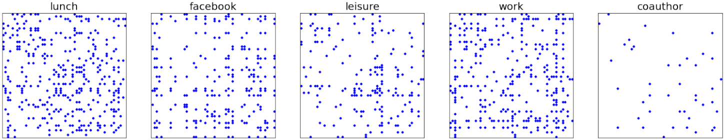

Figure 4: Spy plots of the adjacency matrices for the layers of the Aucs dataset.

Experiments: Algorithms and Evaluation Metrics

- **Implementation** of Spectral Clustering method for single-layer graphs, Generalized Eigen-Decomposition (**SC-GED**), and Spectral Regularization (**SC-SR**).
- **Baseline methods :** Kernel K-means (**K-Kmeans**) algorithm, the Spectral Clustering applied to **Single Layers**, the averaged Laplacian matrix algorithm (**SC-AL**), and the summation of adjacency matrices (**SC-SUM**)
- **Evaluation** of the performance using the **Purity**, the Normalized Mutual Information (**NMI**), and Rand Index (**RI**).

Results

Cora Dataset

	Single layer graph		Multi-layer graph				
	Titles	Citations	SC-SUM	SC-AL	K-Kmeans	SC-GED	SC-SR
Purity	0.609	0.521	0.523	0.526	**0.705**	0.560	0.584
NMI	0.237	0.005	0.009	0.018	**0.427**	0.122	0.238
RI	0.524	0.401	0.396	0.396	**0.640**	0.567	0.600

MIT Dataset

	Single layer graph		Multi-layer graph				
	Friendship	Calls	SC-SUM	SC-AL	K-Kmeans	SC-GED	SC-SR
Purity	0.578	0.500	0.478	0.478	**0.589**	0.578	0.433
NMI	**0.361**	0.181	0.222	0.220	0.302	0.313	0.202
RI	**0.682**	0.414	0.530	0.529	0.571	0.653	0.625

AUCS Dataset

	Single layer graph		Multi-layer graph				
	Lunch	Facebook	SC-SUM	SC-AL	K-Kmeans	SC-GED	SC-SR
Purity	0.870	0.500	0.833	0.833	**0.944**	0.889	**0.944**
NMI	0.849	0.413	0.857	0.857	**0.908**	0.880	0.905
RI	0.937	0.725	0.925	0.925	**0.966**	0.950	0.962

Current Section

Sensitivity to Layer Importance Order: SC-SR

How to properly sort layers according to their importance?

- Ideally, rely on **knowledge of expertise**
- **Exhaustive** search along all different permutations (d! complexity)
- Selection criteria evaluated **for each layer**

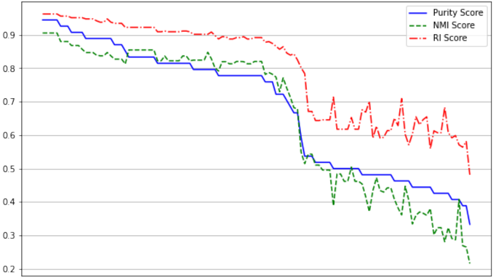

Figure 5: Influence of Informative Layer Order on SC-SR Method Performance evaluated on AUCS Dataset

Modularity Analysis

- **Modularity:** Measures the quality of a partition. Higher modularity indicates more cohesive clusters, making a layer more meaningful for clustering.
- Evaluate layer ranking from most to least informative.
- AUCS dataset layer analysis shows Co-author and Lunch with the highest modularity.

Table 1: Modularity of Layers Based on Clustering Results

Lunch	**Facebook**	**Leisure**	**Work**	**Coauthor**
0.599	0.272	0.474	0.400	0.662

Scalability

- **Limitations:**
  - SC-GED method complexity: $O(mn^3)$.
  - Complexity arises from eigen-decomposition and matrix inversions.
- **Appplication:**
  - Applicable to small networks (approximately 100 nodes, 3 to 5 layers).
  - Impractical and non scalable for real-life datasets are larger and more complex.

Reproducibility

- **Challenges:**
  - **Non-existence** of the preprocessing details for Cora and MIT datasets.
  - Unclear criteria for paper selection in Cora and node/cluster selection in MIT.

  $\Longrightarrow$ Variability in results due to different data engineering strategies for adjacency matrix extraction.
- **Solutions:**
  - Introduced a new dataset, AUCS, for reproducibility.
  - Carefully constructed adjacency matrices.
  - Precise ground truth labels for benchmarking.

Current Section

Conclusion

- Re-implemented two spectral-based multi-layer graph clustering methods: **SC-GED** and **SC-SR**.
- Extended the evaluation on a new **well-constructed** dataset.
- Demonstrated a slight **improvement** over existing methods and **competition** with certain baselines.
- Emphasized the **critical** impact of the order of informative layers.
- Encountered challenges such as **hyperparameter sensitivity**, requiring expert domain knowledge, and facing computational intensity issues with **limited scalability**.

Thank You for your attention !

References

[1] Xiaowen Dong, Pascal Frossard, Pierre Vandergheynst, and Nikolai Nefedov.
Clustering with multi-layer graphs: A spectral perspective.
IEEE Transactions on Signal Processing, 60(11):5820â5831, 2012.

[2] B. Luo, R. C. Wilson, and E. R. Hancock.
Spectral clustering of graphs.
In Nicolai Petkov and Michel A. Westenberg, editors, *Computer Analysis of Images and Patterns*, pages 540–548, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[3] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore.
Automating the construction of internet portals with machine learning.
Information Retrieval Journal, 3:127–163, 2000.
www.research.whizbang.com/data.

References

[4] Nathan Eagle, Alex Pentland, and David Lazer.
Inferring social network structure using mobile phone data.
Proceedings of the National Academy of Sciences (PNAS), 106(36):15274–15278, 2009.

[5] Luca Rossi and Matteo Magnani.
Towards effective visual analytics on multiplex and multilayer networks.
Chaos, Solitons and Fractals, 72:68–76, 2015.

[6] Ulrike von Luxburg.
A tutorial on spectral clustering.
Statistics and Computing, 17(4):395–416, 12 2007.

[7] Liang Liu, Zhao Kang, Jiajia Ruan, and Xixu He.
Multilayer graph contrastive clustering network.
Information Sciences, 613:256–267, 2022.