The Power Mean Laplacian for Multilayer Graph Clustering

Pedro Mercado[1], Antoine Gautier[1], Francesco Tudisco[2], Matthias Hein[1]
[1]Saarland University [2]University of Strathclyde

Introduction

GOAL: Extend Spectral Clustering to networks that have multiple kind of relationships, by defining a new Laplacian that blends all sources of information.

CONTRIBUTIONS:

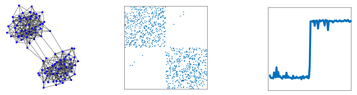
1. We introduce the *power mean Laplacian* as an alternative way to blend the information of different kinds of relations.
2. We show that in expectation under the Stochastic Block Model our method *outperforms* current approaches.
3. We show that eigenvectors of the matrix power mean can be computed efficiently *without ever computing the matrix itself*.

Clustering as Graph Partitioning

1 Get eigenvectors $\{\mathbf{u}_i\}_{i=1}^k$ corresponding to the k **smallest** eigenvalues of L.
2 Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$.
3 Cluster the rows of U with k-means into clusters $C_1, \ldots, C_k$.

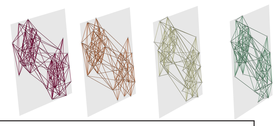
Single Layer Case

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$
$$\mathbf{L}_{\text{sym}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$$



CASE OF MULTILAYER GRAPHS:

A multilayer graph is the set $\mathbb{G} = \{G^{(1)}, \ldots, G^{(T)}\}$ where each graph $G^{(t)} = (V, W^{(t)})$ encodes a particular kind of relationship.



How to merge the information coming from different layers?

Scalar Power Mean

The power mean of a set of non-negative scalars $x_1, \ldots, x_T$ is defined as

$$m_p(x_1, \ldots, x_T) = \left(\tfrac{1}{T}\sum_{i=1}^T x_i^p\right)^{1/p}$$

	$p \to -\infty$	$p = -1$	$p \to 0$	$p = 1$	$p \to \infty$
$m_p(x_1, \ldots, x_T)$	$\min\{x_1, \ldots, x_T\}$	$T(\sum_{i=1}^T x_i^{-1})^{-1}$	$(\prod_{i=1}^T x_i)^{1/T}$	$\frac{1}{T}\sum_{i=1}^T x_i$	$\max\{x_1, \ldots, x_T\}$
name	minimum	harmonic mean	geometric mean	arithmetic mean	maximum

Matrix Power Mean

The matrix power mean of positive definite matrices $\mathbf{A}_1, \ldots, \mathbf{A}_T$ is defined by [1]

$$M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T) = \left(\tfrac{1}{T}\sum_{i=1}^T \mathbf{A}_i^p\right)^{1/p}$$

OBSERVATION: Let $\mathbf{A}_t\mathbf{u} = \lambda_t\mathbf{u}$, for $t = 1, \ldots, T$. Then,

$$M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T)\mathbf{u} = m_p(\lambda_1, \ldots, \lambda_T)\mathbf{u}$$

Relative ordering of eigenvalues is different among matrix means.

Power Mean Laplacian

We define the power geometric mean Laplacian of $\mathbb{G}$ as

$$\mathbf{L}_\mathbf{p} = M_p(\mathbf{L}_{\text{sym}}^{(1)}, \ldots, \mathbf{L}_{\text{sym}}^{(T)})$$

For $p \leq 0$ we add a small diagonal shift to enforce Laplacians to be p.d.

Stochastic Block Model Analysis

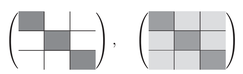
In the Stochastic Block Model (SBM), the edge W_{ij} exists with probability p_{in} if v_i and v_j are in the **same** cluster and p_{out} if they are in **different** clusters.

For multilayer graphs we consider a SBM for each graph $G^{(t)}$, with $(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)})$.

ALL LAYERS CONTAIN FULL INFORMATION

Each layer contains information regarding all clusters.

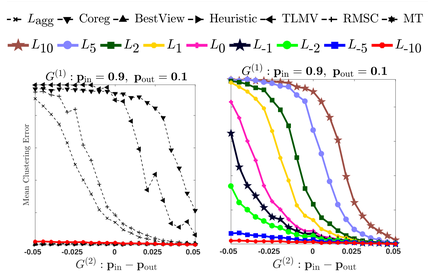
$$P(W_{ij}^{(t)} = 1) = \begin{cases} p_{\text{in}}^{(t)} & \text{if } v_i, v_j \text{ are in the same cluster} \\ p_{\text{out}}^{(t)} & \text{if } v_i, v_j \text{ are in the different clusters} \end{cases}$$



THEOREM: Let $\boldsymbol{\chi}_1 = \mathbf{1}$ and $\boldsymbol{\chi}_i = (k-1)\mathbf{1}_{c_i} - \mathbf{1}_{\overline{c_i}}$. Then, $\{\boldsymbol{\chi}_i\}_{i=1}^k$ correspond to the k-smallest eigenvalues of $\mathcal{L}_\mathbf{p}$ if and only if $m_p(\boldsymbol{\mu}) < 1$, where $\mu_t = 1 - \rho_t$, and $\rho_t = (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)})$.
In particular, for $p \to \pm\infty$, $\{\boldsymbol{\chi}_i\}_{i=1}^k$ correspond to the k-smallest eigenvalues of:

- $\mathcal{L}_\infty$ if and only if $p_{\text{in}}^{(t)} > p_{\text{out}}^{(t)}$ holds for **all** $t \in \{1, \ldots, T\}$,
- $\mathcal{L}_{-\infty}$ if and only if there **exists** a $t \in \{1, \ldots, T\}$ such that $p_{\text{in}}^{(t)} > p_{\text{out}}^{(t)}$.

$\mathcal{L}_\infty$ and $\mathcal{L}_{-\infty}$ are related to the logical operators AND and OR.



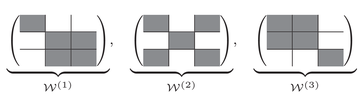
Setting: 50 runs on graphs with two clusters of 100 nodes.

- The larger the value of p, the larger the clustering error of $\mathbf{L}_\mathbf{p}$.
- $\mathbf{L}_{-10}$ presents the smallest clustering error.
- For small values of p, the more robust is $\mathbf{L}_\mathbf{p}$ against noise.

The smaller the value of p, the larger the coverage of $\mathcal{L}_\mathbf{p}$ under the SBM.

NO LAYER CONTAINS FULL INFORMATION

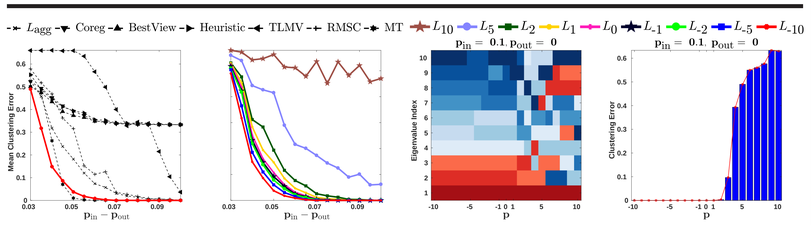
Each layer contains information regarding one particular cluster. Single layers do not recover ground truth clusters.



THEOREM: Let $p_{\text{in}} > p_{\text{out}}$. For any nonzero integer p the eigenvectors of $\mathcal{L}_\mathbf{p}$ corresponding to its three smallest eigenvalues are

$$\boldsymbol{\chi}_1 = \mathbf{1}, \quad \boldsymbol{\chi}_2 = \mathbf{1}_{c_2} - \mathbf{1}_{c_1}, \quad \text{and} \quad \boldsymbol{\chi}_3 = \mathbf{1}_{c_3} - \mathbf{1}_{c_1}$$

In expectation $\mathcal{L}_\mathbf{p}$ recovers ground truth clusters for all p



Variance of L_p is smaller for $p < 0$.

SBM: NON-CONSISTENT LAYER PARTITIONS

Layers follow the same node partition, but partitions fluctuate from layer to layer with a certain probability. We take the Degree Corrected SBM from [2].

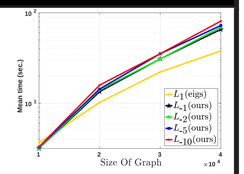
	$\tilde{p}$ = 0.5	0.6	0.7	0.8	0.9	1.0
L_{agg}	0.3	1.3	3.0	8.0	22.3	100.0
Coreg	0.3	0.0	0.3	0.0	0.0	64.7
BestView	9.7	1.0	0.3	0.0	0.7	77.3
Heuristic	0.0	0.0	0.0	0.0	0.3	59.3
TLMV	0.7	0.7	4.0	6.0	24.7	100.0
RMSC	1.0	1.7	4.0	7.0	19.7	100.0
MT	1.3	0.3	0.7	3.0	17.0	100.0
L_{10}	0.0	0.0	0.0	0.0	1.0	100.0
L_5	0.0	0.0	0.0	0.0	5.0	100.0
L_2	0.0	0.0	0.3	2.3	18.3	100.0
L_1	1.0	1.0	3.0	7.0	30.3	100.0
L_0	4.3	4.3	9.7	15.3	38.3	100.0
L_{-1}	6.7	7.7	15.7	16.3	42.3	100.0
L_{-2}	8.0	13.0	20.3	20.7	42.7	100.0
L_{-5}	22.3	23.0	36.3	37.7	50.0	100.0
L_{-10}	**69.0**	**76.3**	**68.0**	**67.3**	**59.7**	100.0

- $\tilde{p} = 0$: the partitions between layers are independent,
- $\tilde{p} = 1$: the partitions between layers are identical,
- $\mu = 0$: all edges are within communities, and
- $\mu = 1$: edges are assigned ignoring clustering structure.

The smaller the value of p, the smaller the clustering error of $\mathbf{L}_\mathbf{p}$.

Power Method for Power Mean Laplacian

- Computation of $M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T)$ is expensive.
- $M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T)$ is in general a dense matrix, even if $\mathbf{A}_1, \ldots, \mathbf{A}_T$ are sparse.



Observations for the Inverse Power Method (IPM) with $p < 0$:

- $M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T)\mathbf{u} = \lambda\mathbf{u} \iff M_p^p(\mathbf{A}_1, \ldots, \mathbf{A}_T)\mathbf{u} = \lambda^p\mathbf{u}$
- The function $f(x) = x^p$ is order reversing for $p < 0$: if $\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T)$ then $\lambda_1^p \geq \cdots \geq \lambda_n^p$ are eigenvalues of $M_p^p(\mathbf{A}_1, \ldots, \mathbf{A}_T)$

$$M_p^p(\mathbf{A}_1, \ldots, \mathbf{A}_T)\mathbf{u} = \frac{1}{T}\sum_{i=1}^T (\mathbf{A}_i^p)\mathbf{u}$$

For $(\mathbf{A}^p)\mathbf{u}$ we use the Polynomial Krylov Subspace Method (PKSM)[3]:

- Project $\mathbf{A}$ onto the subspace $\mathbb{K}^s(\mathbf{A}, \mathbf{u}) = \text{span}\{\mathbf{u}, \mathbf{A}\mathbf{u}, \ldots, \mathbf{A}^{s-1}\mathbf{u}\}$,
- At each step we have $AV_s = V_sH_s + \mathbf{v}_{s+1}\mathbf{e}_s^T$ where H_s is $s \times s$ symmetric tridiagonal.

The matrix vector product $\mathbf{x} = (\mathbf{A}^p)\mathbf{u}$ is then approximated by $\mathbf{x}_s = V_s(H_s)^p\mathbf{e}_1\|\mathbf{u}\|$.

We compute eigenvectors of $M_p(\mathbf{A}_1, \ldots, \mathbf{A}_T)$ without computing the matrix itself.

Experiments

For each layer we build $W^{(t)}$ from the k-nearest neighbour graph based on the Pearson linear correlation between nodes with $k \in \{20, 40, 60, 80\}$.

	3Sources	BBC	BBCS	Wiki	UCI	Citeseer	Cora	WebKB
# vertices	169	685	544	693	2000	3312	2708	187
# layers	3	4	2	2	6	2	2	2
# classes	6	5	5	10	10	6	7	5
L_{agg}	0.194	0.156	0.152	0.371	0.162	0.373	0.452	**0.277**
Coreg [4]	0.215	0.196	0.164	0.784	0.248	0.395	0.659	0.444
Heuristic	**0.192**	0.218	0.198	0.697	0.280	0.474	0.515	0.400
TLMV [5]	0.284	0.259	0.317	0.412	0.154	0.363	0.533	0.430
RMSC [6]	0.254	0.255	0.194	0.407	0.173	0.422	0.507	0.279
MT [7]	0.249	**0.133**	0.158	0.544	0.103	0.371	0.436	0.298
L_1	0.194	0.154	0.148	0.373	0.163	0.285	**0.367**	0.440
$\boldsymbol{L_{-10}}$ (ours)	0.200	0.159	**0.144**	**0.368**	**0.095**	**0.283**	0.374	0.439

- $\boldsymbol{L_{-10}}$ reaches the smallest clustering error in 50% of datasets.
- The largest error gap happens in UCI dataset.
- The largest clustering error happens in WebKB.

References

[1] K. V. Bhagwat and R. Subramanian. Inequalities between means of positive operators. *Math. Proc. of the Cambridge Phil. Soc.*, 1978.
[2] Marya Bazzi, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Model. Simul.*, 14(1):1–41, 2016.
[3] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, PA, USA, 2008.
[4] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *NIPS*, 2011.
[5] Dengyong Zhou and Christopher JC Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, 2007.
[6] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, 2014.
[7] Caterina De Bacco, Eleanor A. Power, Daniel B. Larremore, and Cristopher Moore. Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E*, 95:042317, Apr 2017.