GENERALIZED MATRIX MEANS FOR SEMI-SUPERVISED LEARNING WITH MULTILAYER GRAPHS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

PEDRO MERCADO[1], FRANCESCO TUDISCO[2], MATTHIAS HEIN[1]

[1]UNIVERSITY OF TÜBINGEN [2]GRAN SASSO SCIENCE INSTITUTE

GSSI

INTRODUCTION

GOAL: Extend Graph-Based Semi-Supervised Learning to graphs that have multiple kinds of relations, by defining a new graph regularizer that blends information encoded by different kinds of interactions.

CONTRIBUTIONS:

1. We introduce the *power mean Laplacian* as a regularizer for semi-supervised learning with multilayer graphs.
2. We show that in expectation under the Stochastic Block Model our method *outperforms* current approaches.
3. We show that solutions to our approach can be computed efficiently *without ever computing the power mean Laplacian matrix itself*.

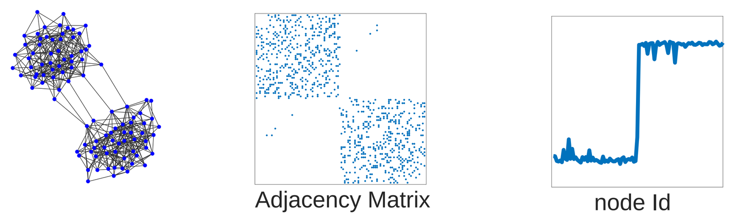
GRAPH-BASED SEMI-SUPERVISED LEARNING

1. Let $Y_i^{(r)} = 1$ if v_i is a labeled node of class $\mathcal{C}_r$, and $Y_i^{(r)} = 0$ else.
2. Let $f^{(r)} = \arg\min_{f\in\mathbb{R}^n} \|f - Y^{(r)}\|^2 + \lambda f^T \mathbf{L} f$.
3. Assign node v_i to class $y_i = \arg\max\{f_i^{(1)}, \ldots, f_i^{(k)}\}$.

Single Layer Case

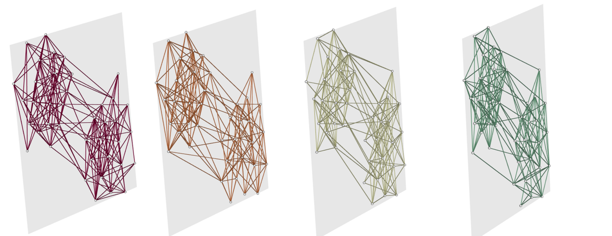
$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

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



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 extend SSL when different kinds of interactions are available?

SCALAR POWER MEAN

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

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

	$p\to-\infty$	$p=-1$	$p\to 0$	$p=1$	$p\to\infty$
$m_p(x_1,\ldots,x_T)$	$\min\{x_1,\ldots,x_T\}$	$T\left(\sum_{i=1}^T x_i^{-1}\right)^{-1}$	$\left(\prod_{i=1}^T x_i\right)^{1/T}$	$\frac{1}{T}\sum_{i=1}^T x_i$	$\max\{x_1,\ldots,x_T\}$
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]

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

POWER MEAN LAPLACIAN

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

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

SSL WITH MULTILAYER GRAPHS

We extend SSL to multilayer graphs via the regularizer $\mathbf{L_p}$ with parameter p:

$$f^{(r)} = \arg\min_{f\in\mathbb{R}^n} \|f - Y^{(r)}\|^2 + \lambda f^T \mathbf{L_p} f \qquad (1)$$

STOCHASTIC BLOCK MODEL ANALYSIS

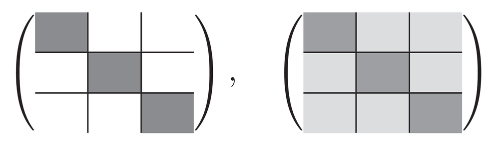
In the Stochastic Block Model (SBM), the edge W_{ij} exists with probability p_{in} if v_i and v_j are in the **same** class and p_{out} if they are in **different** classes.

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

COMPLEMENTARY INFORMATION LAYERS

Each layer contains information regarding all classes.

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



THEOREM: Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the SBM with k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of equal size.Assume the same number of labeled nodes per class. Then, (1) yields zero test error if and only if

$$m_p(\boldsymbol{\rho_\epsilon}) < 1 + \epsilon$$

where $(\boldsymbol{\rho_\epsilon})_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \epsilon$, and $t = 1, \ldots, T$.

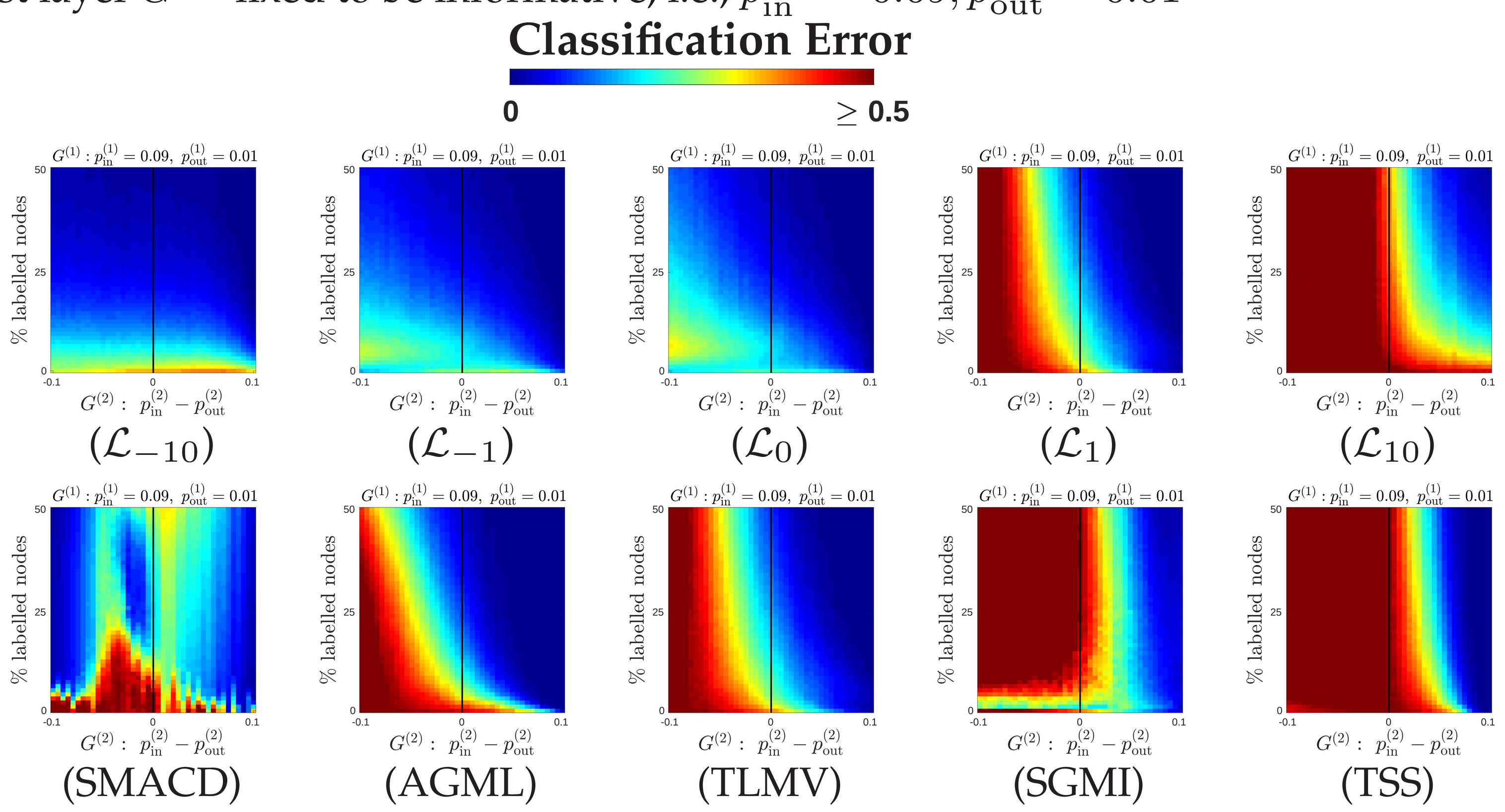
In particular, for the limit cases $p \to \pm\infty$ we have,

- For $p \to \infty$, the test error is zero $\Leftrightarrow$ $p_{\text{out}}^{(t)} < p_{\text{in}}^{(t)}$ for *all* layers.
- For $p \to -\infty$, the test error is zero $\Leftrightarrow$ there *exists* a layer s.t. $p_{\text{out}}^{(t)} < p_{\text{in}}^{(t)}$.

The limit case $p \to -\infty$ presents more general conditions than $p \to \infty$

Analysis on random graphs following the SBM. Setting:

- We show the average classification error out of 100 runs.
- Different amounts of labeled nodes: from 1% to 50% of nodes.
- First layer $G^{(1)}$ fixed to be informative, i.e., $p_{\text{in}}^{(1)} = 0.09, p_{\text{out}}^{(1)} = 0.01$



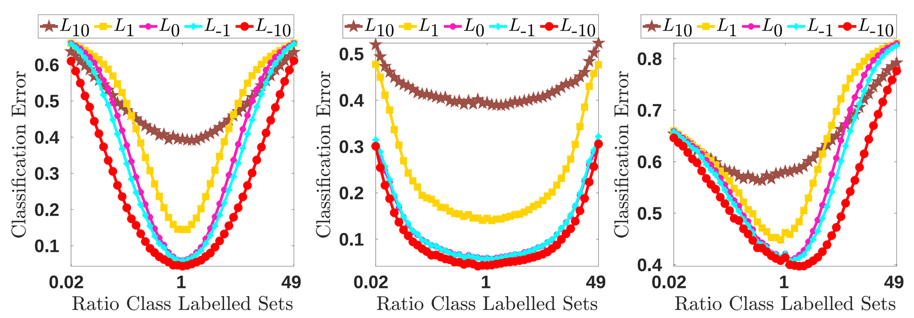
Smaller values of p induce a better performance with $\mathcal{L}_p$ under the SBM.

DIFFERENT AMOUNT OF LABELED NODES PER CLASS

THEOREM: Let $n_1, \ldots, n_k$ be the number of labeled nodes per class. Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $C_{ii} = n/n_r$ for $v_i \in \mathcal{C}_r$. Then

$$f^{(r)} = \arg\min_{f\in\mathbb{R}^n} \|f - \mathbf{C}Y^{(r)}\|^2 + \lambda f^T \mathcal{L}_\mathbf{p} f.$$

yields a zero test classification error if and only if $m_p(\boldsymbol{\rho_\epsilon}) < 1 + \epsilon$.

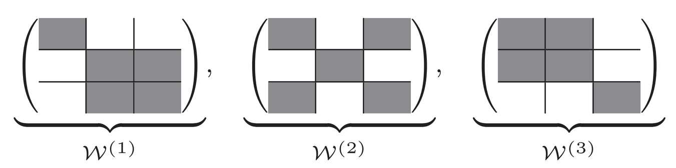


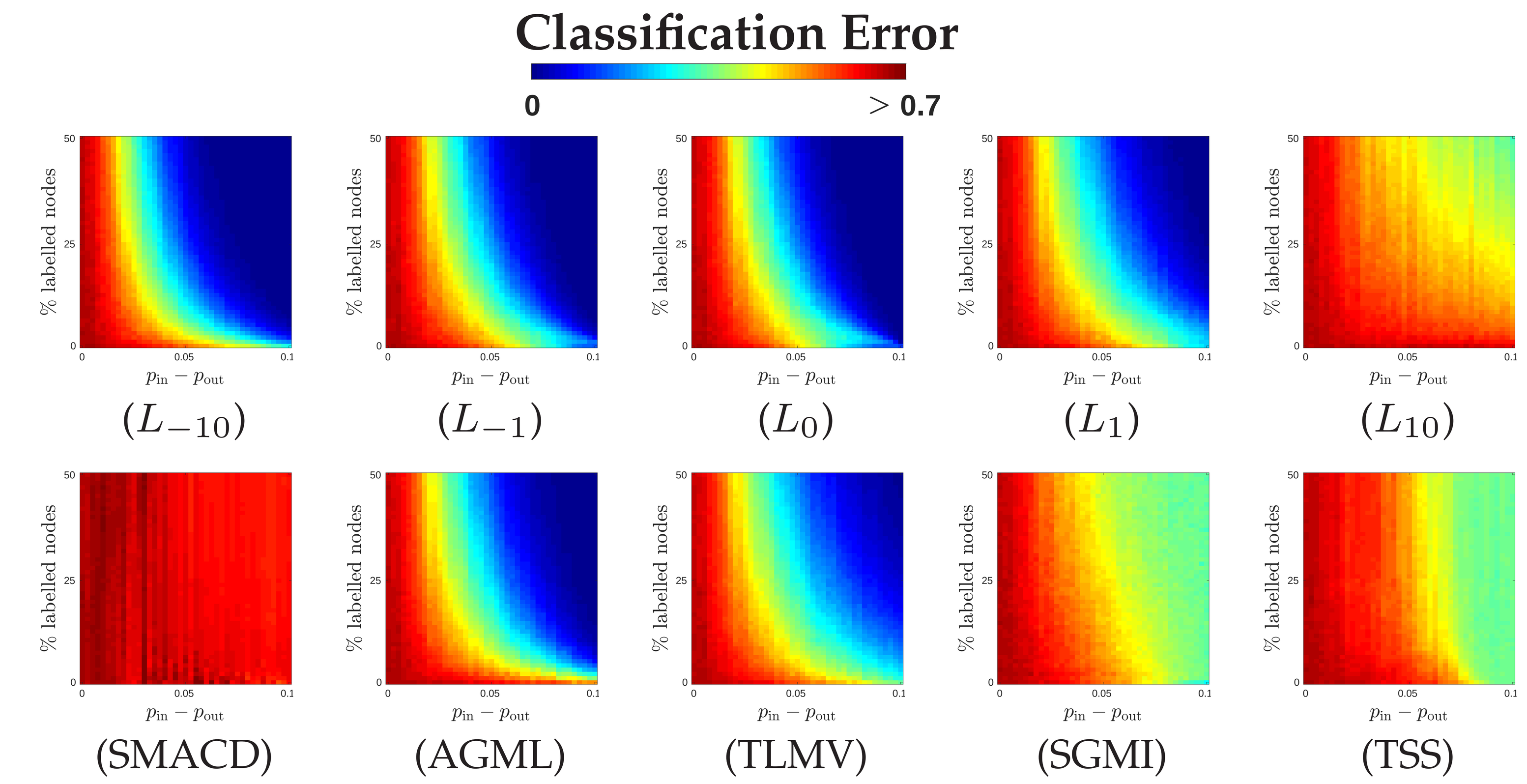
- Layer $G^{(1)}$: informative, i.e. $p_{\text{in}}^{(1)} - p_{\text{out}}^{(1)} = 0.08$
- Layer $G^{(2)}$: uninformative, i.e. $p_{\text{in}}^{(2)} - p_{\text{out}}^{(2)} = 0$.
- 2 classes:$n_1, n_2 = 1, \ldots, 49$.

Our approach can be adjusted to match class priors under the SBM.

NO LAYER CONTAINS FULL INFORMATION

Each layer contains information regarding one particular class. Single layers do not recover ground truth classes.

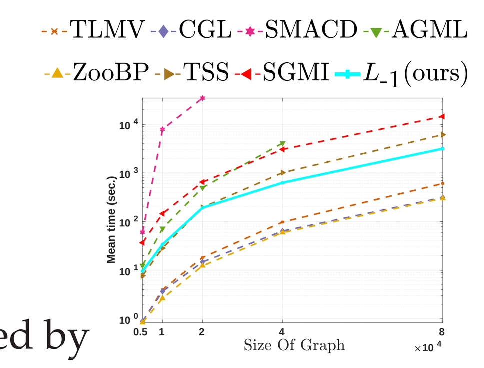




For small p our method effectively merges information where others fail.

MATRIX-FREE METHOD FOR $(I + \lambda \mathbf{L}_p)^{-1} Y$

- 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 $(I + \lambda \mathbf{L}_p)^{-1} Y$ with $p < 0$:

- We solve the system via Krylov methods (GMRES/PCG)
- At iteration h the problem is projected onto the space spanned by
$$\{\mathbf{y}, \lambda L_p\mathbf{y}, (\lambda L_p)^2\mathbf{y}, \ldots, (\lambda L_p)^h\mathbf{y}\}$$
- The product $L_p\mathbf{y}$ is approxiamted via the Cauchy integral of the function $\phi(x) = (x)^{1/p}$:
$$\phi_N(S_p)\mathbf{y} = \alpha S_p \operatorname{Im}\left\{\sum_{i=1}^N \gamma_i (z_i^2 I - S_p)^{-1}\mathbf{y}\right\}$$
- To solve the linear system $(zI - S_p)^{-1}\mathbf{y}$ we employ again a Krylov method, by projecting onto the space spanned by $\{Y, S_pY, S_p^2Y, \ldots, S_p^hY\}$.

We compute $(I + \lambda \mathbf{L}_p)^{-1} Y$ without computing the matrix $\mathbf{L}_p$ itself

EXPERIMENTS

For each layer we build $W^{(t)}$ from the k-nearest neighbour graph based on the Pearson linear correlation between nodes with $k = 10$. We show the average test error out of 10 samples of labeled nodes.

BBCS	1%	5%	10%	15%	20%	25%
TLMV[2]	25.6	12.6	10.5	7.5	6.4	5.4
CGL[3]	79.2	51.6	34.9	23.4	16.5	12.7
SMACD[4]	77.8	80.6	82.4	96.4	98.4	98.3
AGML[5]	34.6	17.4	12.1	7.0	6.0	5.4
ZooBP[6]	33.8	13.9	11.3	8.8	7.6	6.2
TSS[7]	23.9	13.2	14.1	12.3	13.1	12.2
SGMI[8]	31.9	19.6	16.6	15.5	14.8	12.1
L_1	29.9	15.0	13.5	10.6	8.7	7.2
L_{-1}	**23.8**	**11.6**	**8.7**	**6.3**	**5.8**	**5.1**
L_{-10}	48.7	22.5	14.2	9.1	7.8	6.1

- $\boldsymbol{L_{-1}}$ reaches the smallest test error.
- $\boldsymbol{L_{-1}}$ outperforms the arithmetic mean Laplacian $\boldsymbol{L_1}$.
- $\boldsymbol{L_{-1}}$ presents the 1st or 2nd best performance in 77% of cases across eight different datasets.

REFERENCES

[1] K. V. Bhagwat and R. Subramanian. Inequalities between means of positive operators. *Math. Proc. of Cambridge Phil. Soc.*, 83(3):393–401, 1978.
[2] Dengyong Zhou and Christopher JC Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, 2007.
[3] Andreas Argyriou, Mark Herbster, and Massimiliano Pontil. Combining graph Laplacians for semi–supervised learning. In *NeurIPS*, 2006.
[4] Ekta Gujral and Evangelos E Papalexakis. SMACD: Semi-supervised multi-aspect community detection. In *SDM*, 2018.
[5] F. Nie et al. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, 2016.
[6] Dhivya Eswaran et al. Zoobp: Belief propagation for heterogeneous networks. In *VLDB*, 2017.
[7] Koji Tsuda, Hyunjung Shin, and Bernhard Schölkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21(2):59–65, 2005.
[8] M. Karasuyama et al. Multiple graph label propagation by sparse integration. *IEEE Trans. Neur. Nets. & Lear. Sys.*, 24(12):1999–2012, 2013.