EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Node Classification for Signed Social Networks Using Diffuse Interface Methods

Pedro Mercado[1], Jessica Bosch[2], Martin Stoll[3]
[1]University of Tübingen [2]University of British Columbia [3]TU Chemnitz

TECHNISCHE UNIVERSITÄT CHEMNITZ

Introduction

GOAL: Extend graph-based semi-supervised learning (**SSL**) to networks that have both positive and negative interactions via diffuse interface methods.

MAIN CONTRIBUTIONS:

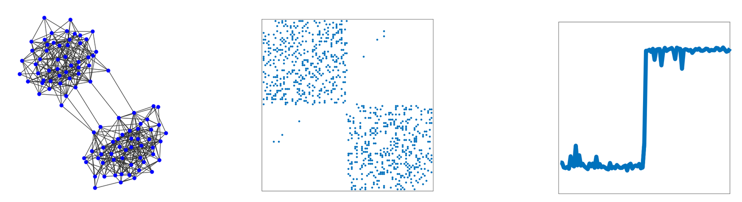
1. We present the first extension of diffuse interface methods to the task of node classification for signed graphs.
2. We present a systematic comparison considering different state of the art of signed graph Laplacians.
3. We present extensive numerical experiments showing that better classification performance is obtained by taking negative edges in consideration.

Laplacians and Spectral Clustering

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 $C_1, \ldots, C_k$.

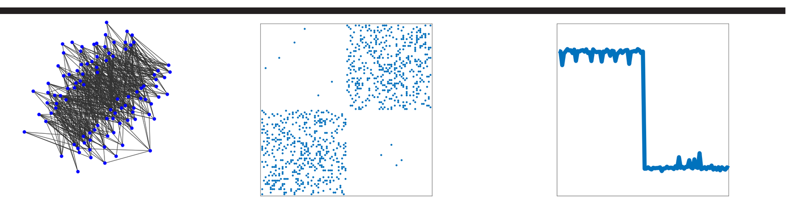
Assortative Case

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



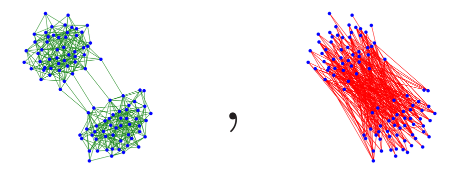
Disassortative Case

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



LAPLACIANS OF SIGNED NETWORKS:

A signed graph is the pair $G^{\pm} = (G^+, G^-)$ where G^+ and G^- encode positive and the negative relations, respectively.

$G^{\pm} = ($  $)$

Signed Laplacians and its motivating discrete problem:

- **cut**$(C, \overline{C})$: number of edges between sets $(C, \overline{C})$,
- **assoc**(C): number of edges inside set C

$$\mathbf{L}_{\mathbf{SR}} = \mathbf{D}^+ - \mathbf{W}^+ + \mathbf{D}^- + \mathbf{W}^- \text{[1]}$$
$$\min_{C \subset V} \left(2\mathbf{cut}^+(C, \overline{C}) + \mathbf{assoc}^-(C) + \mathbf{assoc}^-(\overline{C}) \right) \left(\frac{1}{|C|} + \frac{1}{|\overline{C}|} \right)$$

$$\mathbf{L}_{\mathbf{BR}} = \mathbf{D}^+ - \mathbf{W}^+ + \mathbf{W}^- \text{[2]}$$
$$\min_{C \subset V} \frac{1}{|C|} \left(\mathbf{cut}^+(C, \overline{C}) + \mathbf{assoc}^-(C) \right) + \frac{1}{|\overline{C}|} \left(\mathbf{cut}^+(C, \overline{C}) + \mathbf{assoc}^-(\overline{C}) \right)$$

$$\mathbf{L}_{\mathbf{SP}} = (\mathbf{L}^- + \mathbf{D}^+)^{-1}(\mathbf{L}^+ + \mathbf{D}^-) \text{[3]}$$
$$\min_{C \subset V} \left(\frac{\mathbf{cut}^+(C, \overline{C}) + \mathbf{vol}^-(C)}{\mathbf{cut}^-(C, \overline{C}) + \mathbf{vol}^+(C)} \right)$$

$$\mathbf{L}_{\mathbf{AM}} = (\mathbf{L}^+_{\text{sym}} + \mathbf{Q}^-_{\text{sym}}) \text{[4]}$$

Different signed Laplacians identify different clustering structures.

References

[1] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. Luca, and S. Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *ICDM*, pages 559–570, 2010.
[2] K. Chiang, J. Whang, and I. Dhillon. Scalable clustering of signed networks using balance normalized cut. CIKM, pages 615–624, 2012.
[3] M Cucuringu, P Davies, A Glielmo, and H Tyagi. SPONGE: A generalized eigenproblem for clustering signed networks. *AISTATS*, 2019.
[4] Pedro Mercado, Francesco Tudisco, and Matthias Hein. Clustering signed networks with the geometric mean of Laplacians. In *NIPS*. 2016.
[5] A. L. Bertozzi and A. Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Model. Simul.*, 10(3):1090–1118, 2012.
[6] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data, June 2014.
[7] Shuhan Yuan, Xintao Wu, and Yang Xiang. SNE: Signed network embedding. In *PAKDD*, 2017.
[8] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.
[9] Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *COLT*, 2004.
[10] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.
[11] Andrew B. Goldberg, Xiaojin Zhu, and Stephen Wright. Dissimilarity in graph-based semi-supervised classification. In *AISTATS*, 2007.
[12] Jiliang Tang, Charu Aggarwal, and Huan Liu. Node classification in signed social networks. In *SDM*, 2016.

Diffuse Interface Methods

GRAPH GINZBURG-LANDAU FUNCTIONAL FOR SSL

Graph extension of the continuous Ginzburg-Landau (**GL**) Functional to SSL [5]:

$$E_S(u) = \frac{\varepsilon}{2}\mathbf{u}^{\mathbf{T}}\mathbf{S}\mathbf{u} + \frac{1}{4\varepsilon}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{u}_\mathrm{i}^2 - 1)^2 + \sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\omega_\mathrm{i}}{2}(\mathrm{f}_\mathrm{i} - \mathrm{u}_\mathrm{i})^2,$$

- $\mathbf{u}^{\mathbf{T}}\mathbf{S}\mathbf{u}$ induces clustering information of the signed graph. Different choices of S convey information about different cluster assumptions,
- $\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{u}_\mathrm{i}^2 - 1)^2$ has minimizers with entries in $+1$ and -1. It induces a minimizer u with entries corresponding to the class assignment,
- $\sum_{\mathrm{i}=1}^{\mathrm{n}}\omega_\mathrm{i}(\mathrm{f}_\mathrm{i} - \mathrm{u}_\mathrm{i})^2$ is a fitting term to labeled nodes, where $\omega_i = 0$ for unlabeled nodes and $\omega_i = w_0$ for labeled nodes,
- $\varepsilon > 0$: interface parameter controls trade-off between clustering structure and labeled nodes: **large** values of ε give priority to graph information, whereas **small** values of ε give priority to labeled nodes

Modified Graph Allen-Cahn Equation for SSL: is the gradient flow of the functional E_S

$$\frac{\partial u}{\partial t} = -\nabla E_S(u)$$

Solution via convex-splitting scheme, $E_S(u) = E_1(u) - E_2(u)$, with implicit treatment of the convex part $E_1(u)$ and explicit treatment for the concave part $E_2(u)$:

$$\frac{u^{(t+1)} - u^{(t)}}{\tau} = -\nabla E_1(u^{(t+1)}) + \nabla E_2(u^{(t)})$$

Let (λ_l, ϕ_l) be eigenpairs of S. Let $u^{(t+1)} = \sum_{l=1}^{n} a_l\phi_l$, $u^{(t)} = \sum_{l=1}^{n} \bar{a}_l\phi_l$. Equivalently,

$$(1 + \varepsilon\tau\lambda_l + c\tau)\, a_l = -\frac{\tau}{\varepsilon}\bar{b}_l + (1 + c\tau)\bar{a}_l + \tau\bar{d}_l \quad \text{for } l = 1, \ldots, n$$

where $\bar{b} = [\phi_1, \ldots, \phi_n]^T \nabla\psi\left(\sum_{l=1}^{n}\bar{a}_l\phi_l\right)$ with $\psi(u) = \sum_{i=1}^{n}(u_i^2 - 1)^2$, and $\bar{d} = [\phi_1, \ldots, \phi_n]^T \nabla\varphi\left(\sum_{l=1}^{n}\bar{a}_l\phi_l\right)$ with $\varphi(u) = \sum_{i=1}^{n}\frac{\omega_i}{2}(f_i - u_i)^2$.

Experiments

DATASETS: Statistics of largest connected components of G^+, G^- and $G^{\pm}$.

	Wikipedia RfA [6]			Wikipedia Elections [6]			Wikipedia Editor [7]		
	G^+	G^-	$G^{\pm}$	G^+	G^-	$G^{\pm}$	G^+	G^-	$G^{\pm}$
# nodes	3024	3124	3470	1997	2040	2325	17647	14685	20198
+ nodes	55.2%	42.8%	48.1%	61.3%	47.1%	52.6%	38.5%	33.5%	36.8%
# edges	204035	189343	215013	107650	101598	111466	620174	304498	694436
+ edges	100%	0%	78.2%	100%	0%	77.6%	100%	0%	77.3%

Parameter Setting: $\omega_0 = 10^3, \varepsilon = 10^{-1}, c = \frac{3}{\varepsilon} + \omega_0$, time step-size $dt = 10^{-1}$, maximum number of iterations 2000, stopping tolerance 10^{-6}.

COMPARISON WITH STATE OF THE ART

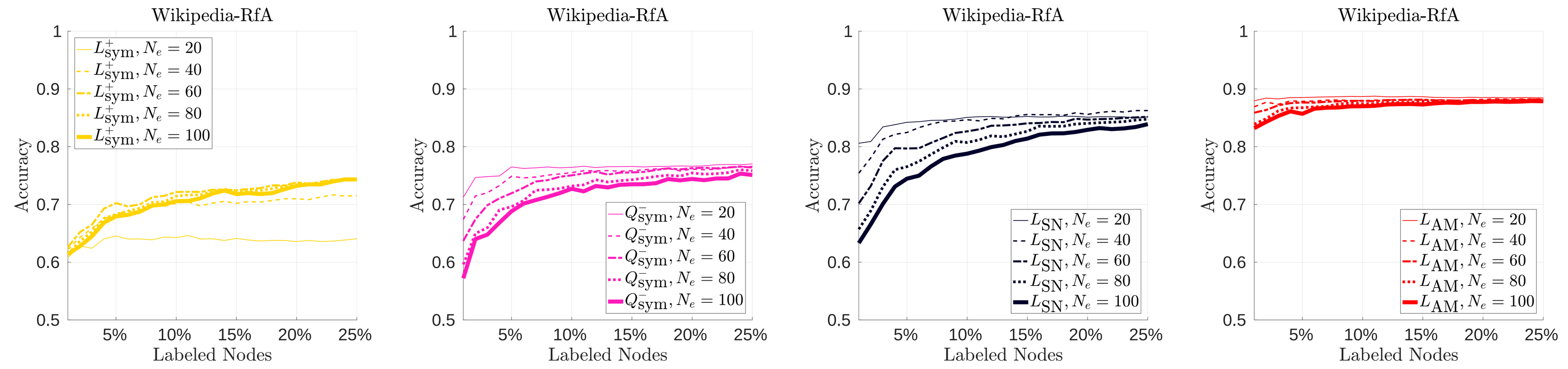
- Different amounts of labeled nodes are considered: 1%, 5%, 10%, 15%,
- Presence of negative edges improves accuracy in 2 out of 3 datasets,
- **GL**(L_{SN}) and **GL**(L_{AM}) performs best among signed graph methods.

	Wikipedia RfA				Wikipedia Elections				Wikipedia Editor			
Labeled nodes	1%	5%	10%	15%	1%	5%	10%	15%	1%	5%	10%	15%
LGC(L^+) [8]	0.554	0.553	0.553	0.553	0.614	0.614	0.613	0.613	0.786	0.839	0.851	0.857
TK(L^+) [9]	0.676	0.697	0.681	0.660	0.734	0.763	0.742	0.723	0.732	0.761	0.779	0.791
HF(L^+) [10]	0.557	0.587	0.606	0.619	0.616	0.623	0.637	0.644	0.639	**0.848**	**0.854**	**0.858**
GL(L^+_{sym})	0.577	0.564	0.570	0.584	0.608	0.622	0.626	0.614	0.819	0.759	0.696	0.667
DGB [11]	0.614	0.681	0.688	0.650	0.648	0.602	0.644	0.609	0.692	0.714	0.721	0.727
NCSSN [12]	0.763	0.756	0.745	0.734	0.697	0.726	0.735	0.776	0.491	0.533	0.559	0.570
GL(Q^-_{sym})	0.788	0.800	0.804	0.804	0.713	0.765	0.764	0.766	0.739	0.760	0.765	0.770
GL(L_{SP})	0.753	0.761	0.763	0.765	0.789	0.793	0.797	0.798	0.748	0.774	0.779	0.779
GL(L_{SN})	0.681	0.752	0.759	0.764	0.806	0.842	0.851	0.852	**0.831**	0.841	0.846	0.847
GL(L_{AM})	**0.845**	**0.847**	**0.848**	**0.849**	**0.879**	**0.885**	**0.887**	**0.887**	0.787	0.807	0.814	0.817

Our approaches GL(L_{SN}) and GL(L_{AM}) outperforms signed graph methods.

EFFECT OF THE NUMBER OF LABELED NODES

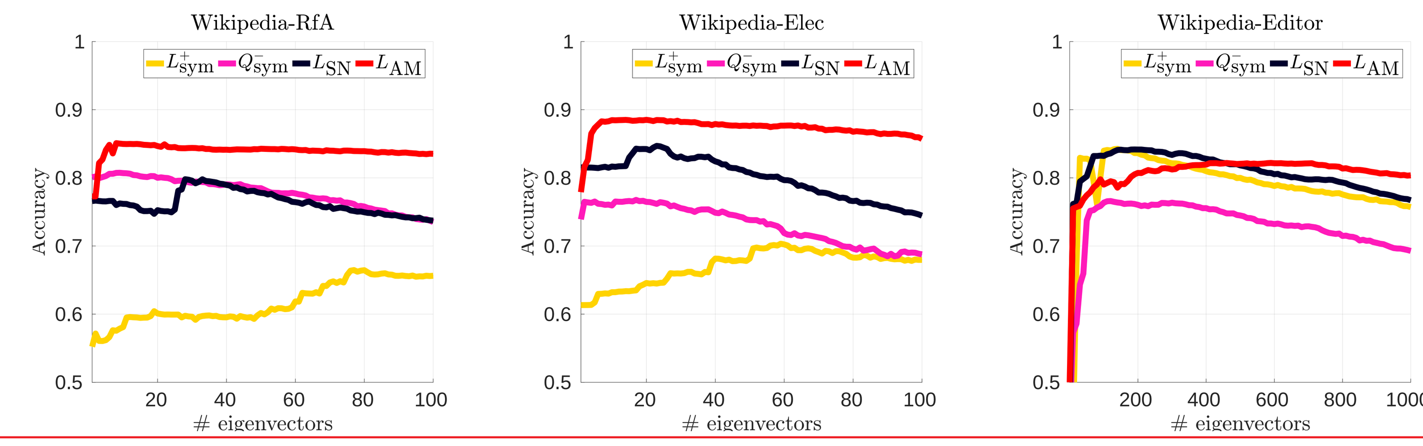
- Fix the number of eigenvectors to $N_e \in \{20, 40, 60, 80, 100\}$,
- Proportion of labeled nodes: from 1% to 25%,



Larger amount of labeled nodes improve performance.

EFFECT OF THE NUMBER OF EIGENVECTORS

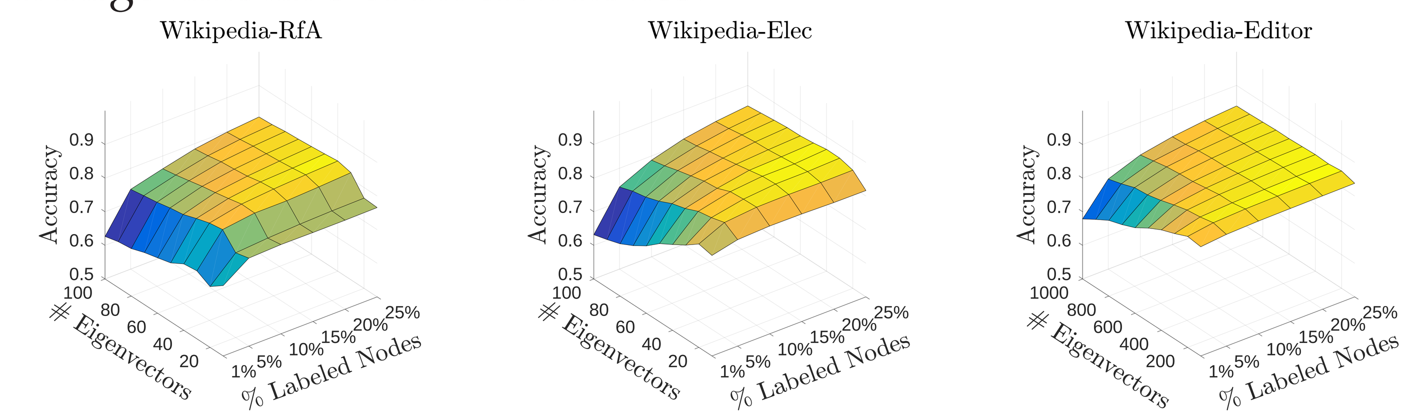
- # of eigenvectors $1, \ldots, 100$ (Wiki-RfA/Elec), and $10, \ldots, 1000$ (Wiki-Editor).
- Proportion of labeled nodes: 5%,



A small amount of eigenvectors of the corresponding Laplacian is required.

JOINT EFFECT: NUMBER OF EIGENVECTORS AND LABELED NODES

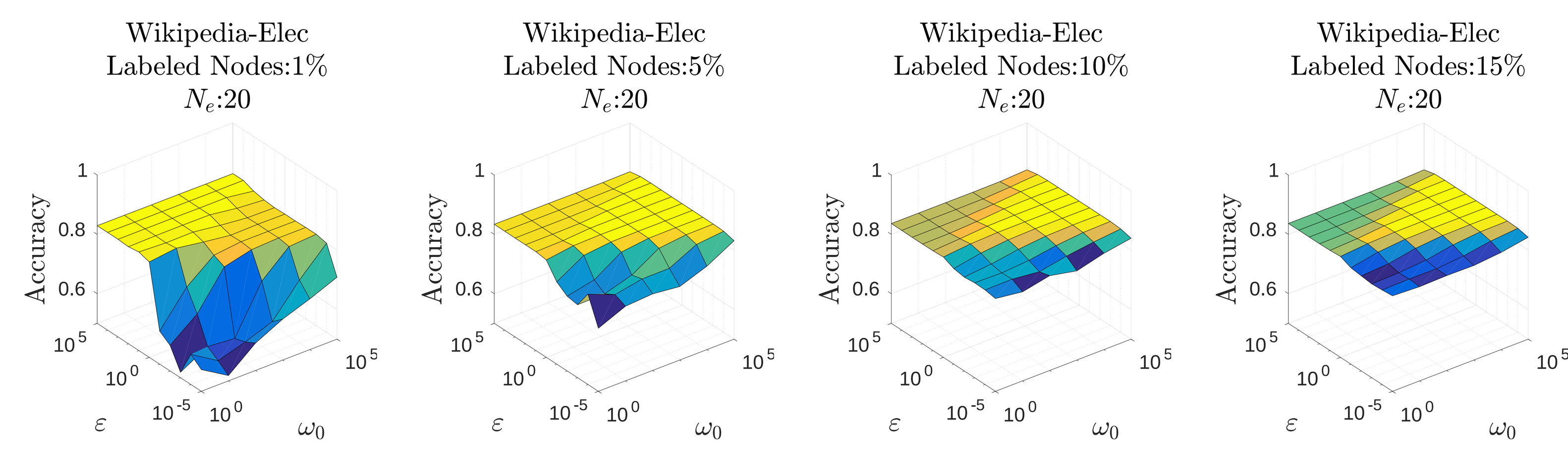
- # of eigenvectors $10, \ldots, 100$ (Wiki-RfA/Elec), and $100, \ldots, 1000$ (Wiki-Editor).
- Proportion of labeled nodes: $1, \ldots, 25\%$,
- When too many eigenvectors are taken, more labeled nodes are required.
- The best performance is achieved with few eigenvectors are taken together with large amounts of labeled nodes.



A large amount of eigenvectors has a systematic negative effect.

JOINT EFFECT: FIDELITY (ω_0) AND INTERFACE (ε) PARAMETERS

- # of eigenvectors: 20.
- Proportion of labeled nodes: 1%, 5%, 10%, 15%,
- Fidelity parameter $\omega_0 \in \{10^0, 10^1, \ldots, 10^5\}$
- Interface parameter $\varepsilon \in \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$



The smaller the amount of labeled nodes, the larger the impact of the interface parameter ε