$\mathbf{L^2}$-Based PageRank for Graph-Based Semi-Supervised Learning

E. Bautista[1,2], S. de Nigris[1], P. Abry[2], P. Gonçalves[1]
[(1)] Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, FRANCE
[(2)] Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France

This work was supported by CONACYT, Labex MILyon and ANR-14-CE27-0001 GRAPHSIP grant.

Introduction. Graph-Based Semi-Supervised Learning (G-SSL) exploits the structure of unlabeled data in addition to expert data to develop better classifiers. Both the increasing availability of large datasets and the ability of graphs to naturally code relations among such data have made the field of G-SSL of utmost importance and a rapidly evolving alternative to supervised classifiers that rely on expert data which may be hard to obtain. Succesful applications range from classification of BitTorrent content and users [1], text categorization [2] and medical diagnosis [3], among others. Despite its unquestionable success, cases such as fuzzy graph topologies or unbalanced class settings still pose challenging issues for G-SSL, that we aim to address.

Related works, Objective. G-SSL methods, such as PageRank (PR) or Standard Laplacian (SL), can endorse a random walk interpretation propagating the labeled information through the graph structure [4]. In this vein, the so-called Fractional G-SSL [5], [6] extended the framework of [4] to include the dynamics emmanating from the γ-th powers of the combinatorial Laplacian matrix, $\mathbf{L}^\gamma$. In those works, it was shown that, on the $0 < \gamma < 1$ regime, the non-local dynamics induced by $\mathbf{L}^\gamma$ lead to super diffusive random walks fuelling the force of G-SSL. As a result, Fractional G-SSL was able to overcome badly constructed graphs and provide nevertheless meaningful classification. The goal of the present work is to unveil the potential benefits of the Fractional G-SSL when considering a modification of the PR method to account for the dynamics induced by the $\gamma = 2$ case.

Semi-Supervised Learning. Consider a weighted undirected graph of N nodes where the graph structure is encoded by the adjacency matrix $\mathbf{W}$. Let $d_i = \sum_j W_{ij}$ denote the degree of node i and $\mathbf{D} = diag(d_1, \cdots, d_N)$ be the diagonal matrix of degrees. $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the combinatorial graph Laplacian. For a K-class problem the labeled information is encoded in a matrix $\mathbf{Y} \in \mathbb{R}^{N\times K}$, where $Y_{ik} = 1$ if node i belongs to class k and zero elsewhere. Then, the classification challenge is to estimate the classification matrix $\mathbf{X} \in \mathbb{R}^{N\times K}$ leading to the final prediction: node i is assigned to the class k that satisfies $\text{argmax}_k X_{ik}$. In particular, the PageRank-based G-SSL is defined as the solution to the regularization problem [7]:

$$\min_{\mathbf{x}} \left\{ \mathbf{x}^T\mathbf{D}^{-1}\mathbf{L}\mathbf{D}^{-1}\mathbf{x} + \mu\left(\mathbf{x}-\mathbf{y}\right)^T \mathbf{D}^{-1}\left(\mathbf{x}-\mathbf{y}\right)\right\}, \quad (1)$$

applied columnwise to $\mathbf{X}_{*k} = \mathbf{x}$ and to $\mathbf{Y}_{*k} = \mathbf{y}$, respectively.

$\mathbf{L}^2$-Based PageRank. Posing $\mathbf{L}^2 = \mathbf{D}_{(2)} - \mathbf{W}_{(2)}$ with $(D_{(2)})_{ii} = (L^2)_{ii}$ and $(W_{(2)})_{ij} = -(L^2)_{ij}$, $\mathbf{D}_{(2)}$ naturally identifies to a generalized degree matrix and $\mathbf{W}_{(2)}$ to a generalized adjacency matrix, verifying the Laplacian property $(D_{(2)})_{ii} = \sum_j (W_{(2)})_{ij}$, ensuring thus $\mathbf{L}^2$ to encode for a new graph. We then revamp the regularization problem (1) with $\mathbf{L}^2$ to get the $\mathbf{L}^2$-PageRank ($\mathbf{L}^2$-PR) optimization scheme:

$$\min_{\mathbf{x}} \left\{ \mathbf{x}^T\mathbf{D}_{(2)}^{-1}\mathbf{L}^2\mathbf{D}_{(2)}^{-1}\mathbf{x} + \mu\left(\mathbf{x}-\mathbf{y}\right)^T \mathbf{D}_{(2)}^{-1}\left(\mathbf{x}-\mathbf{y}\right)\right\}. \quad (2)$$

As shown in [5], problem (2) is convex with closed form solution given by

$$\mathbf{x} = \mu\left(\mathbf{L}^2\mathbf{D}_{(2)}^{-1} + \mu\mathbb{I}\right)^{-1}\mathbf{y}. \quad (3)$$

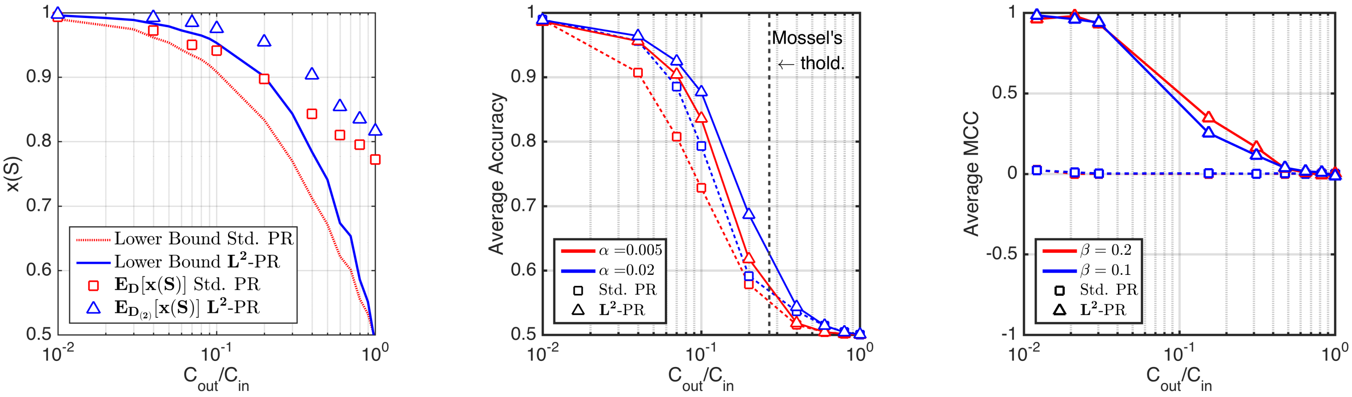
However, we note that $\mathbf{W}_{(2)}$ entails negative entries, hindering a possible interpretation of (3) as a random walk process. In consequence, to analyse the proposed $\mathbf{L}^2$-PR method, we identify the resolvent of the operator $\mathbf{L}^2\mathbf{D}_{(2)}^{-1}$ and its relation to the Green's function via the inverse Laplace transform [8]:

$$\frac{1}{\mathbf{L}^2\mathbf{D}_{(2)}^{-1} + \mu\mathbb{I}} = \int_0^\infty e^{-t\left(\mathbf{L}^2\mathbf{D}_{(2)}^{-1}+\mu\mathbb{I}\right)}dt = \mathbb{G}_\mu^{(2)}. \quad (4)$$

Compared to the eigenvalues $(\lambda_0, \ldots \lambda_i, \ldots) = diag(\Lambda)$ of $\mathbf{L}$, we conjecture[1] that the spectrum of $\mathbf{L}^2\mathbf{D}_{(2)}^{-1}$ tends to squeeze down towards zero for the λ_i's smaller than 1 and to stretch out beyond 1 for the λ_i's that are larger than 1. As a result, there is a rescaling on the times of relaxation towards equilibrium, implying a faster smoothing out of high oscillating modes and the opposite effect for the low frequency components. Therefore, the first modes which substantially capture the graph community structure, like the Fiedler vector, may hence have larger contributions to the final solution of $\mathbf{L^2}$-PR.

Relation to graph topology. In the following, we denote by S any arbitrary subset of vertices of the graph. We recall the edge boundary of a node, with respect to S, is defined as $\partial(S)_i = \{j \sim i : i \in S, i \in S^c\}$, and the volume of S, $\text{vol}(S) = \sum_{i\in S} d_i$. Similarly, posing $d_i^{(2)} = (D_{(2)})_{ii}$, a generalized volume is defined as $\text{vol}_{(2)}(S) = \sum_{i\in S} d_i^{(2)}$. The Green's function of the standard PR was employed in [9] to derive a sampling strategy for the nodes to be tagged by the expert, such that it can be bounded from below the expected amount of labeled information distributed by the PR method to a set of nodes of interest S.

[1] Strict proof is still missing but we verified it numerically.



(a) Bounds comparison

(b) Balanced classes classification

(c) Unbalanced classes classification

Fig. 1: Numerical assessment on the SBM. The inter-class and intra-class mean degrees are denoted by C_{out} and C_{in}, respectively. For the figures the average degree is set to 3. (a) Comparison of Th. 1 and Prop. 1, where we note that for a fixed topology, the latter improves the bound and $\mathbb{E}\left[x(S)\right]$; ($\mu = 1$, $|S| = 5000$). (b) Classification perofrmance in a balanced setting. α refers to the fraction of labeled nodes; ($\mu = 1 \times 10^{-3}, N = 10^4$). (c) Classification performance in an unbalanced setting. Perofrmance is measured in terms of Matthews correlation coefficient (MCC); ($\beta = |S|/|S^c|, \mu = 1 \times 10^{-3}$, $\alpha = 0.01$).

Theorem 1 (Standard PageRank [9])**.** *Let x be the PageRank solution for an unweighted graph, and pose $x(S) = \sum_{i \in S} x_i$. Then, if unit mass labeled points are randomly placed at node i with probability proportional to d_i, $\forall\, i \in S$, we have*

$$\mathbb{E}\left[x(S)\right] \geq 1 - \frac{1}{\mu \operatorname{vol}(S)} \sum_{i \in S} |\partial(S)_i|. \tag{5}$$

We extend the previous result to the L^2-PageRank.

Proposition 1 ($\mathbf{L^2}$-PageRank)**.** *Under the same notations as in theorem 1, if unit mass labeled points are randomly placed at node i with probability proportional to $d_i^{(2)}$, $\forall\, i \in S$, we have*

$$\mathbb{E}\left[x(S)\right] \geq 1 - \frac{1}{\mu \operatorname{vol}_{(2)}(S)} \left(\sum_{i \in S} |\partial(S)_i|^2 + \sum_{\ell \in S^c} |\partial(S^c)_\ell|^2 \right). \tag{6}$$

If we let S denote a class of interest, then Theorem 1 and Proposition 1 acquire importance as the rationale behind G-SSL is to diffuse most of the labeled information to nodes of the same class and prevent any label information leakage to adjacent classes. Thus, as will be shown numerically in the following section, Proposition 1 goes exactly in that direction, since it provides a strategy on the placement of labeled points that leads to a new bound in Eq. (6) that is tighter than the one implied by the strategy of Theorem 1 in Eq. (5).

Performance assessment. We contrast the performances of the standard PR and the $\mathbf{L}^2$-PR by means of numerical investigation on the stochastic block model. Firstly, Fig. 1a compares Theorem 1 and Proposition 1, with results suggesting that the diffusion of labeled information using the strategy of Proposition 1 is more significant than the one of Theorem 1. Our second experiment assesses the quality of the methods in a balanced setting. The results show significant gains in accuracy when considering $\mathbf{L}^2$-PR classification. Lastly, an unbalanced class size situation is addressed in is presented in Fig. 1c with results displaying that the dynamics offered by the $\mathbf{L}^2$ significantly help to overcome the unbalanced class constraint.

Conclusion. In this work we presented a modification of the PageRank algorithm to adopt the dynamics emmanating from $\mathbf{L}^2$. Our results point that this change in the dynamics bring potential benefits for classification. More precisely, from a theoretical standpoint we obtained improved guarantees on how effectively the labeled information is diffused. Furthermore, numerical simulations indicate that the theoritcal gains translate to accuracy gains in challenging settings. The promising results open the door for a more in depth exploration of the dynamics brought by $\mathbf{L}^2$, and, perspectively, towards tackling other γ regimes to, potentially, obtain more flexible classification schemes.

References

[1] K. Avrachenkov, P. Gonçalves, A. Legout, and M. Sokol, "Classification of content and users in bittorrent by semi-supervised learning methods," in *Int. Wireless Comm. and Mobile Comp. Conf.*, Cyprus, 2012.

[2] A. Subramanya and J. Bilmes, "Soft-supervised learning for text classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 1090–1099.

[3] M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang, "Compact graph based semi-supervised learning for medical diagnosis in alzheimer's disease," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1192–1196, Oct 2014.

[4] M. Sokol, *Graph-based Semi-supervised Learning Methods and Quick Detection of Central Nodes*, Ph.D. thesis, Université de Nice, Ecole Doctorale STIC, Inria Sophia Antipolis, Maestro, April 2014.

[5] S. de Nigris, E. Bautista, P. Abry, K. Avrachenkov, and P. Gon¸calves, "Fractional graph-based semi-supervised learning," 08 2017, pp. 356–360.

[6] E. Bautista, S. De Nigris, P. Abry, K. Avrachenkov, and P. Gonçalves, "Lévy flights for graph based semi-supervised classification," in *26th colloquium GRETSI*, 2017.

[7] K. Avrachenkov, A.Mishenin, P. Gonçalves, and M. Sokol, *Generalized Optimization Framework for Graph-based Semi-supervised Learning*, pp. 966–974.

[8] M. Schmidt, "Global properties of Dirichlet forms on discrete spaces," *ArXiv e-prints*, Jan. 2012.

[9] F. Chung, "Pagerank as a discrete green's function," *Geometry and Analysis I ALM*, vol. 17, 01 2010.