

RESEARCH

Open Access

L^γ -PageRank for semi-supervised learning



Esteban Bautista^{1,2*} , Patrice Abry² and Paulo Gonçalves¹

*Correspondence:

esteban.bautista-ruiz@ens-lyon.fr

¹Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342 Lyon, France

²Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS, Laboratoire de Physique, F-69342 Lyon, France

Abstract

PageRank for Semi-Supervised Learning has shown to leverage data structures and limited tagged examples to yield meaningful classification. Despite successes, classification performance can still be improved, particularly in cases of graphs with unclear clusters or unbalanced labeled data. To address such limitations, a novel approach based on powers of the Laplacian matrix L^γ ($\gamma > 0$), referred to as L^γ -PageRank, is proposed. Its theoretical study shows that it operates on signed graphs, where nodes belonging to one same class are more likely to share positive edges while nodes from different classes are more likely to be connected with negative edges. It is shown that by selecting an optimal γ , classification performance can be significantly enhanced. A procedure for the automated estimation of the optimal γ , from a unique observation of data, is devised and assessed. Experiments on several datasets demonstrate the effectiveness of both L^γ -PageRank classification and the optimal γ estimation.

Keywords: Semi-supervised learning, PageRank, Laplacian powers, Diffusion on graphs, Signed graphs, Optimal tuning, MNIST

Introduction

Context

Graph-based Semi-Supervised Learning (G-SSL) is a modern important tool for classification. While Unsupervised Learning fully relies on the data structure and Supervised Learning demands extensive labeled examples, G-SSL combines limited tagged examples and the data structure to provide satisfactory results. This makes the field of G-SSL of utmost importance as nowadays large and structured datasets can be readily accessed in comparison to expert data which may be hard to obtain. Examples where G-SSL provide state of the art results are vast, ranging from classification of BitTorrent contents and users (Avrachenkov et al. 2012a), text categorization (Subramanya and Bilmes 2008), medical diagnosis (Zhao et al. 2014), or zombie hunting under BGP protocol (Fontugne et al. 2019). Algorithmically, PageRank constitutes the reference tool in G-SSL. It has spurred a deluge of theory (Chung 2010; Avrachenkov et al. 2018; Litvak et al. 2009; Chung 2007), applications (Avrachenkov et al. 2008; Graham et al. 2009; Avrachenkov et al. 2012a; Fontugne et al. 2019) and implementations (Andersen et al. 2007; Andersen and Chung 2007). Despite successes, the performance of G-SSL can still be improved, particularly for graphs with unclear clusters or imbalance of labeled datasets, two situations that we aim to address in this work.

Related works

In graphs, a ground truth class is represented by a subset of graph nodes, denoted S_{gt} . Thus, in graphs, the classification challenge corresponds to finding the binary partition of the graph vertices: $\mathcal{V} = S_{gt} \cup S_{gt}^c$. If the data is structured, then S_{gt} forms a cluster, i.e., a densely and strongly connected graph region that is weakly connected to the rest of the graph. This is exploited by G-SSL methods that essentially amount to diffuse information placed on the tagged nodes of S_{gt} , through the graph, expecting a concentration of information in S_{gt} that reveals its members. Among the family of G-SSL propositions obeying this rationale (Zhou et al. 2004; Zhou and Burges 2007; Avrachenkov et al. 2012b), PageRank is considered the state of the art approach in terms of performance, algorithms and theoretical understanding. The PageRank algorithm can be interpreted as random walkers that start from the labeled points and, at each step, diffuse to an adjacent node with probability α or restart to the starting point with probability $(1-\alpha)$. In the limit (of infinite steps), each node is endowed a score proportional to the number of visits to it. Thus, vertices of S_{gt} are expected to get larger scores as walkers get trapped for a long time by the connected structure of S_{gt} . The capacity of PageRank to confine the random walks within S_{gt} depends on a topological parameter of S_{gt} known as the Cheeger ratio, or conductance, counting the ratio of external and internal connections of S_{gt} . More precisely, it is shown in Andersen et al. (2007) that the probability of a PageRank random walker leaving S_{gt} is upper bounded by the Cheeger ratio of S_{gt} . In other terms, a small Cheeger ratio designates a strongly disconnected cluster that PageRank can eventually easily detect. Based on the scores, a binary partitioning via a sweep-cut procedure allows to retrieve an estimate \hat{S}_{gt} . This procedure is granted to obtain an estimate \hat{S}_{gt} with a small Cheeger ratio if a sharp drop in magnitude appears on the sorted scores, then \hat{S}_{gt} is potentially a good estimation of the ground truth S_{gt} (Andersen and Chung 2007). In Zhou and Belkin (2011), an issue affecting G-SSL methods, coined as the ‘curse of flatness’, was highlighted. Such work proposes to extend PageRank by iterating the *random walk* Laplacian in the PageRank solution, as a mean to enforce Sobolev regularity to the vertex scores and amend the aforementioned problem. However, with this approach, guarantees that a sweep-cut still leads to a meaningful clustering remains unproven and it can be given neither diffusion nor topological interpretations. Thus, preventing insights on the properties and qualities of partitions it retrieves. This makes it hard to build upon and to address the issues listed above.

Goals, contributions and outline

In this work, we revisit Laplacian powers as a way to improve G-SSL and to address the issues listed above. As our first contribution, we propose a generalization of PageRank by using (non necessarily integers) powers of the *combinatorial* Laplacian matrix L^γ ($\gamma > 0$). We call this generalization the L^γ -PageRank method. In contradistinction to (Zhou and Belkin 2011), our approach (i) enables us to have an explicit closed form expression of the underlying optimization problem (see the L^γ -PageRank definition and solution in Definition 4 and Lemma 4, respectively); (ii) permits a diffusion and a topological interpretation (see the L^γ -PageRank properties in Lemmas 5 and 6). In our approach, we show that, for each γ , a new graph is generated. These new graphs, which we refer to as L^γ -graphs, reweight the links of the original structure and create edges, which can be positive or negative, between initially far-distant nodes. This topological change has

the potential to improve classification as the signed edges introduce what can be seen as agreements (positive edges) or disagreements (negative edges) between nodes, allowing to revamp clusters as groups of nodes agreeing between them and disagreeing with the rest of the graph. This paper investigates the potential of these L^γ -graphs to better delineate a targeted S_{gt} , compared to PageRank. The theoretical analysis of our proposition permits to extend the Cheeger ratio to L^γ -graphs and to prove that if there is a L^γ -graph in which S_{gt} has a smaller Cheeger ratio, then we can more accurately identify it with our generalized L^γ -PageRank procedure using the sweep-cut technique. Then, by means of numerical investigations, we point the existence of an optimal γ value that maximizes performance. As a second contribution, we propose an algorithm that allows to estimate the optimal γ directly from the initial graph and the labeled points. Lastly, we demonstrate the classification improvements permitted by L^γ -PageRank on several real world datasets commonly used in classification, as well as the relevance of the estimation procedure for the optimal tuning.

The paper is organized as follows: “State of the art” section sets definitions and recalls classical results on G-SSL. “ L^γ -PageRank for semi-Supervised learning” section presents the main contributions of the paper: “The L^γ -graphs” section introduces L^γ -graphs; “The L^γ -PageRank method” section defines L^γ -PageRank and its theoretical analysis; “The selection of γ ” section discusses the existence of an optimal γ and its estimation. “ L^γ -PageRank in practice” section evaluates L^γ -PageRank and the algorithm for the optimal γ estimation in practice.

State of the art

Preliminaries

Let $\mathcal{G}(\mathcal{V}, \mathcal{E}, w)$ denote a weighted undirected graph with no self-loops in which: \mathcal{V} refers to the set of vertices of cardinality $|\mathcal{V}| = N$; \mathcal{E} denotes the set of edges, where a connected pair $u, v \in \mathcal{V}$, denoted $u \sim v$, implies $(u, v), (v, u) \in \mathcal{E}$; and $w : \mathcal{E} \rightarrow \mathbb{R}^+$ is a weight function. The graph adjacency matrix is denoted by W in which $W_{uv} = w(u, v)$ if $u \sim v$ and $W_{uv} = 0$ otherwise. For a vertex $u \in \mathcal{V}$ we let $d_u = \sum_v W_{uv}$ denote the degree of u and $D = \text{diag}(d_1, \dots, d_N)$ be the diagonal matrix of degrees. Let Δ_{uv} denote the geodesic distance between u and v . Given a set of nodes $S \subseteq \mathcal{V}$, we denote by $\mathbb{1}_S$ the indicator function of such set, meaning that $(\mathbb{1}_S)_u = 1$ if $u \in S$ and $(\mathbb{1}_S)_u = 0$ otherwise. The volume of S is defined to be $\text{vol}(S) = \sum_{u \in S} d_u$. We refer to the volume of the entire graph by $\text{vol}(G)$. Let $f : \mathcal{V} \rightarrow \mathbb{R}$ be a signal lying on the graph vertices. Graph signals are represented as column vectors, where f_u refers to the signal value at node u . The sum of signal values in the set S is denoted by $f(S) = \sum_{u \in S} f_u$. We denote by $L = D - W$ the combinatorial graph Laplacian which, by construction, is a real symmetric matrix with eigendecomposition of the form $L = Q\Lambda Q^T$. The positivity of the Dirichlet form $f^T L f = \sum_{u \sim v} W_{uv} (f_u - f_v)^2 \geq 0$ implies that L has real non-negative eigenvalues.

A random walk on a graph is a Markov chain where the nodes form the state space. Thus, when a walker is located at a node u at a specific time t , at time step $t + 1$ the walker moves to a neighbor v with probability P_{uv} , where $P = D^{-1}W$. If the graph signal χ represents the distribution for the random walk starting point, then the signal $x^T = \chi^T P^t$ denotes the distribution of the walker position at time t . Independently of the starting distribution, if the graph is connected and not bipartite, the random walk converges to a stationary distribution $\pi^T = \pi^T P$, where $\pi_u = d_u / \text{vol}(G)$.

Clustering is the search of groups of nodes that are strongly connected between them and weakly connected to the rest of the graph. The Cheeger ratio is a metric that counts the ratio of external and internal connections of a group of nodes, thus assessing its pertinence as a cluster, while penalizing uninteresting solutions that may fit the cluster criteria, like isolated nodes linked by a few edges. It is defined as follows.

Definition 1 For a set of nodes $S \subseteq \mathcal{V}$, the Cheeger ratio, or conductance, of S is defined as:

$$h_S := \frac{\sum_{u \in S} \sum_{v \in S^c} W_{uv}}{\min\{\text{vol}(S), \text{vol}(S^c)\}}. \quad (1)$$

Thus, we define clustering as finding the binary partition of the graph vertices: $\mathcal{V} = S \cup S^c$ such that S has low h_S .

PageRank-based semi-Supervised learning

Let $\mathcal{V}_{S_{gt}} \subseteq S_{gt}$ denote the set of nodes tagged to belong to the ground truth S_{gt} and y be indicator function of $\mathcal{V}_{S_{gt}}$, i.e. $y_u = 1$ if node $u \in \mathcal{V}_{S_{gt}}$ and $y_u = 0$ otherwise. The PageRank G-SSL is defined as the solution to the optimization problem (Avrachenkov et al. 2012b):

$$\arg \min_f \left\{ f^T D^{-1} L D^{-1} f + \mu (f - y)^T D^{-1} (f - y) \right\}. \quad (2)$$

Optimization problem (2) can be seen as the search of a smooth graph signal in the sense that strongly connected nodes should have similar values (left term), while the labeled data is respected (right term), and a regularization parameter μ tunes the trade off between both terms. Notably, problem (2) is convex with closed form solution given by Avrachenkov et al. (2012b):

$$f = \mu (L D^{-1} + \mu \mathbb{I})^{-1} y. \quad (3)$$

We present the PageRank solution in this form as it will simplify derivations in the remainder of the paper, but it is not hard to rewrite (3) to its more popular version: $f^T = (1 - \alpha) \sum_{k=0}^{\infty} \alpha^k y^T P^k$ where $\alpha = 1/(1 + \mu)$. This latter helps to expose the connection between PageRank and diffusion processes. Namely, it corresponds to the equilibrium state of a random walk that decides either to continue with probability α , or to restart to the starting distribution y with probability $(1 - \alpha)$. As $y = \sum_{u \in \mathcal{V}_{S_{gt}}} \delta_u$ is the combination of different random walk initial distributions, it is clear that the PageRank score assigned to a particular node is (up to re-normalization) the probability of finding a walker that started from the labels, at equilibrium, at this node. PageRank diffusion satisfies the following properties (Tsiatas 2012): (i) mass preservation: $\sum_{u \in \mathcal{V}} f_u = \sum_{u \in \mathcal{V}} y_u$; (ii) stationarity: $f = \pi$ if $y = \pi$; and (iii) limit behavior: $f \rightarrow \pi$ as $\mu \rightarrow 0$ and $f \rightarrow y$ as $\mu \rightarrow \infty$.

In Andersen et al. (2007), it is shown that the behavior of this type of random walks is tightly related to the cluster structure of graphs. This connection between PageRank and clustering is quantified in the following result.

Lemma 1 Andersen et al. (2007) Let $S \subset \mathcal{V}$ be an arbitrary set with $\text{vol}(S) \leq \text{vol}(G)/2$. For a labeled point placed at a node $u \in S$ selected with probability proportional to its

degree in S , i.e. $d_u/vol(S)$, the PageRank satisfies:

$$\mathbb{E}[f(S^c)] \leq \frac{h_S}{\mu}. \tag{4}$$

This lemma implies that if we apply PageRank diffusion to the labels of S_{gt} and it has a small $h_{S_{gt}}$, then the probability of finding a walker outside S_{gt} is small and the nodes with largest PageRank value should index S_{gt} . This is formalized in Andersen et al. (2007) and Andersen and Chung (2007). The former shows that a proxy \hat{S}_{gt} that has small $h_{\hat{S}_{gt}}$ can be found by looking for regions of high concentration of PageRank mass. The latter improves that result, showing that \hat{S}_{gt} can be found more easily by looking for a sharp drop in the PageRank scores. To state their result, we first introduce the sweep-cut technique.

Definition 2 A sweep-cut is a procedure to retrieve a partition $\mathcal{V} = \hat{S}_{gt} \cup \hat{S}_{gt}^c$ from the PageRank vector. The procedure is as follows:

- Let v_1, \dots, v_N be a rearrangement of the vertices in descending order, so that the permutation vector q satisfies $q_{v_i} = f_{v_i}/d_{v_i} \geq q_{v_{i+1}} = f_{v_{i+1}}/d_{v_{i+1}}$;
- Let $S_j = \{v_1, \dots, v_j\}$ be the set of vertices indexed by the first j elements of q ;
- Let $\tau(f) = \min_j h_{S_j}$;
- Retrieve $\hat{S}_{gt} = S_j$ for the set S_j achieving $\tau(f)$.

Now, we state the result of Andersen and Chung (2007), showing that if there is a sharp drop in rank at S_j , then the set S_j has small Cheeger ratio.

Lemma 2 Andersen and Chung (2007) Let $h \in (0, 1)$, j be any index in $[1, N]$ and $\alpha \in (0, 1]$ denote the PageRank restarting probability. Let $C(S_j, S_j^c) = \sum_{u \in S_j} \sum_{v \in S_j^c} W_{uv}$ be the numerator of the Cheeger ratio. Then, S_j satisfies one of the following: (a) $C(S_j, S_j^c) < 2h vol(S_j)$; or (b) there is some index $k > j$ such that $vol(S_k) \geq vol(S_j)(1 + h)$ and $q_k \geq q_j - \alpha/h vol(S_j)$.

In other words, this lemma implies that either S_j has a small Cheeger ratio, or there is no sharp drop at q_j .

Generalization to multiple classes

PageRank G-SSL can be readily generalized to a multi-class setting in which labeled points of K classes are used to find a partition $\mathcal{V} = S_1 \cup S_2 \cup \dots \cup S_K$. Let \mathcal{V}_{S_k} denote the labeled points of class k and the indicator function of \mathcal{V}_{S_k} be placed as the k -th column of a matrix Y . Then, the multi-class PageRank is computed in matrix form as (Sokol 2014): $\min_F \{F^T D^{-1} L D^{-1} F + \mu (F - Y)^T D^{-1} (F - Y)\}$, with classification matrix given in closed form by $F = \mu (L D^{-1} + \mu \mathbb{I})^{-1} Y$. This leads a node u to have K associated scores and it is assigned to the cluster k satisfying $\arg \max_k F_{uk}$. In Sokol (2014), the following rule explaining the classification is provided: let pr_{uv} denote the probability that a random walk reaches node v before restarting to node u , then v is assigned to the class k that satisfies the inequality:

$$\sum_{u \in \mathcal{V}_k} pr_{uv} \geq \sum_{w \in \mathcal{V}_{k'}} pr_{vw}, \quad \forall k' \neq k. \tag{5}$$

This inequality highlights an important issue of the multi-class approach as the sums depend on the cardinality of the sets of labeled points. Thus, cases of unbalanced number of labeled points can potentially bias the classification.

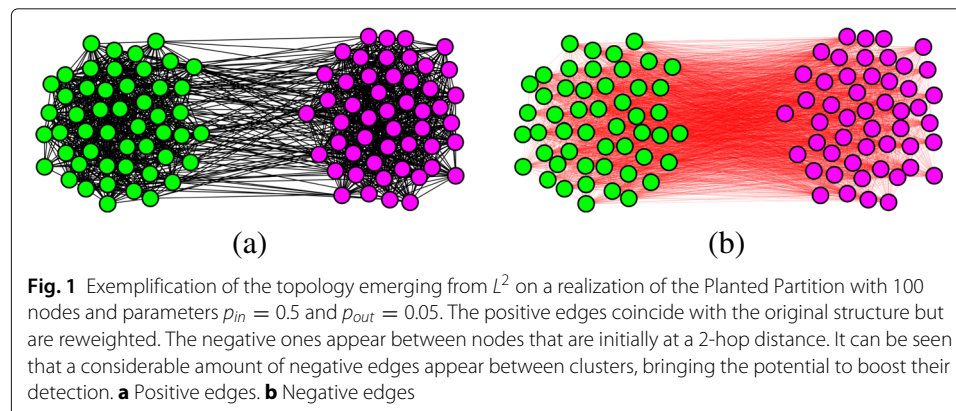
L^γ -PageRank for semi-Supervised learning

The L^γ -graphs

In this work, we propose to change the graph topology in which the problem is solved as a means to improve classification. We evoke such change by considering powers of the Laplacian matrix, noting that the L^γ operator, for $\gamma > 0$, generates a new graph for every fixed γ value. More precisely, the Laplacian definition indicates that $L^\gamma = V\Lambda^\gamma V^T = D_\gamma - W_\gamma$ codes for a new graph, where $[D_\gamma]_{uu} = [L^\gamma]_{uu}$ refers to a generalized degree matrix and $[W_\gamma]_{uv} = -[L^\gamma]_{uv}$, with $u \neq v$, to a generalized adjacency matrix that satisfies the Laplacian property $[D_\gamma]_{uu} = \sum_v [W_\gamma]_{uv}$ since $L^\gamma \mathbf{1} = 0$. We refer to such graphs as L^γ -graphs.

The L^γ -graphs reweight the edges of the original structure and creates links between originally far-distant nodes. Indeed, for $\gamma \in \mathbb{Z}$ the new edges can be related to paths of different lengths. To have a grasp on this, let us take the topology from $\gamma = 2$ as an example: for $L^2 = (D - W)^2 = D^2 + W^2 - (DW + WD)$, the elements of the emanating graph are given as $[D_2]_{uu} = [L^2]_{uu} = D_{uu}^2 + \sum_v W_{uv}^2$ and $[W_2]_{uv} = -[L^2]_{uv} = (D_{uu} + D_{vv})W_{uv} - \sum_{l \neq u,v} W_{ul}W_{lv}$, showing that, in W_2 , nodes originally connected get their link reweighted (still, remaining positive) while those at a 2-hop distance become linked by a negatively weighted edge.

This change in the topology has the potential to impact clustering, as the emergence of positive and negative edges opens the door for an interpretation in terms of an agreement (positive edge) or a disagreement (negative edge) between datapoints. Hence, clustering can be revamped to assume that nodes agreeing should belong to the same cluster and nodes disagreeing should belong to different ones. From this perspective, a revisit to the case of $\gamma = 2$ shows that this is indeed a potentially good topology since, for several graphs, it is more likely that vertices having a 2-hop distance lie in different clusters than in the same one, thus creating a considerable amount of disagreements between clusters, that may enhance their separability. This idea is illustrated in Fig. 1, where for a realization of the planted partition model we show that with $\gamma = 2$ a big amount of negative edges appear between clusters.



Thus, in the reminder of the paper we investigate if for a target set of nodes S_{gt} , the detection of S_{gt} can be enhanced by solving the clustering problem in some of these new graphs.

Remark 1 The graphs emerging in the regime $0 < \gamma < 1$ have already been studied in Pérez Riascos and Mateos (2014); de Nigris et al. (2017); Bautista et al. (2017), where it is shown that such graphs remain within the class of graphs with only positive edges, hence preserving the random walk framework. In these works, the graphs were shown to embed the so-called Lévy flight, permitting random walkers to perform long-distant jumps in a single step.

The L^γ -PageRank method

The signed graphs emerging from L^γ preclude the employment of the random walk-based approaches to find clusters, as ‘negative transitions’ appear. Thus, the L^γ -graphs call for a technique to find clusters in such graphs. In this subsection, we introduce L^γ -PageRank, a generalization of PageRank to find clusters on the L^γ -graphs. Further, we analyze the L^γ -PageRank theoretical properties and clustering capabilities.

For our analysis, it is useful to first extend some of the graph topological definitions to the L^γ -graphs. Let $vol_\gamma(S) = \sum_{u \in S} [D_\gamma]_{uu}$ denote the generalized volume of S . Let π_γ denote a generalized stationary distribution with entries given by $(\pi_\gamma)_u = [D_\gamma]_{uu} / vol_\gamma(G)$. It is important to stress that $[D_\gamma]_{uu} = \sum_k \lambda_k^\gamma Q_{uk}^2 \geq 0$. Thus, for all $\gamma > 0$, the generalized volume and the generalized stationary distribution are non-negative quantities.

The Cheeger ratio metric, lacking the ability to account for the sign of edges, cannot be employed to assess the presence of clusters in the L^γ -graphs. Thus, we generalize the Cheeger ratio definition to the new graphs as follows.

Definition 3 For a set of nodes $S \subseteq V$, the generalized Cheeger ratio, or generalized conductance, of S is defined as:

$$h_S^{(\gamma)} = \frac{\sum_{u \in S} \sum_{v \in S^c} [W_\gamma]_{uv}}{\min(vol_\gamma(S), vol_\gamma(S^c))}. \tag{6}$$

This generalization of the Cheeger ratio is mathematically sound. First, it is a non-negative quantity since $\sum_{u \in S} \sum_{v \in S^c} [W_\gamma]_{uv} = \mathbb{1}_S^T L^\gamma \mathbb{1}_{S^c} \geq 0$. Second, the set S attaining the minimum value coincides with a sensible clustering. To show the latter, let the edges in W_γ be split according to their sign as $W_\gamma = W_\gamma^+ + W_\gamma^-$. Let $\mathcal{A}_{in}^{(\gamma)}(S) = \sum_{u \in S} \sum_{w \in S} |[W_\gamma^+]_{uw}|$ be the sum of agreements within S , $\mathcal{A}_{out}^{(\gamma)}(S) = \sum_{u \in S} \sum_{v \in S^c} |[W_\gamma^+]_{uv}|$ the agreements between S and S^c , $\mathcal{D}_{in}^{(\gamma)}(S) = \sum_{u \in S} \sum_{w \in S} |[W_\gamma^-]_{uw}|$ the disagreements within S , and $\mathcal{D}_{out}^{(\gamma)}(S) = \sum_{u \in S} \sum_{v \in S^c} |[W_\gamma^-]_{uv}|$ the disagreements between S and S^c . Then we state the following lemma.

Lemma 3 Let $S^* = \arg \min_S h_S^{(\gamma)} \forall S$ s.t. $vol_\gamma(S) \leq vol_\gamma(G)/2$. Then, S^* is the set that attains the best balance of:

- Maximal $\mathcal{D}_{out}^{(\gamma)}(S^*)$ and $\mathcal{A}_{in}^{(\gamma)}(S^*)$;
- Minimal $\mathcal{D}_{in}^{(\gamma)}(S^*)$ and $\mathcal{A}_{out}^{(\gamma)}(S^*)$.

The proof is provided in Appendix [Proof of Lemma 3](#).

Lemma 3 shows that, for a given L^γ -graph, sets of small generalized Cheeger ratio are bound to have strong between-cluster disagreements and strong within-cluster agreements as well as small between-cluster agreements and small within-cluster disagreements, thus coinciding with our definition of clusters in signed graphs.

Now, we introduce the L^γ -PageRank formulation. Departing from the optimization problem in (2), we revamp PageRank to operate on the L^γ -topology as follows.

Definition 4 *The L^γ -PageRank G-SSL is defined as the solution to the optimization problem:*

$$\arg \min_f \left\{ f^T D_\gamma^{-1} L^\gamma D_\gamma^{-1} f + \mu (f - y)^T D_\gamma^{-1} (f - y) \right\}. \tag{7}$$

The two following Lemmas show that, for any $\gamma > 0$, the L^γ -PageRank solution exists in closed form and such solution preserves the PageRank properties.

Lemma 4 *Let $\gamma > 0$. Then, problem (7) is convex with closed form solution given as:*

$$f = \mu \left(L^\gamma D_\gamma^{-1} + \mu \mathbb{I} \right)^{-1} y. \tag{8}$$

The proof is provided in Appendix [Proof of lemma 4](#).

Remark 2 Eq. (8) emphasizes the difference between our approach and the one in Zhou and Belkin (2011): they propose to iterate the operator in the G-SSL solution as $f = \mu \left([LD^{-1}]^m + \mu \mathbb{I} \right)^{-1} y$, for $m \in \mathbb{Z}_{>0}$, for which the formulation of the optimization problem having this expression as solution remains unknown.

Remark 3 The solution of L^γ -PageRank in Eq. (8) can be easily cast as a low-pass graph filter, allowing a fast and distributed approximation via Chebyshev polynomials (Shuman et al. 2018).

Lemma 5 *Let $\gamma > 0$. The L^γ -PageRank solution in (8) satisfies the following properties: (i) mass preservation: $\sum_{u \in \mathcal{V}} f_u = \sum_{u \in \mathcal{V}} y_u$; (ii) stationarity: $f = \pi_\gamma$ if $y = \pi_\gamma$; and (iii) limit behavior: $f \rightarrow \pi_\gamma$ as $\mu \rightarrow 0$ and $f \rightarrow y$ as $\mu \rightarrow \infty$.*

The proof is provided in Appendix [Proof of lemma 5](#).

The previous Lemmas are important because they show that our generalization, for any $\gamma > 0$, is a well-posed problem. Indeed, the properties of Lemma 5 imply that, while not necessarily modeled by random walkers, L^γ -PageRank remains a diffusion process having π_γ as stationary state and diffusion rate controlled by the μ parameter.

Our next results shows that it is hard for such diffusion process to escape clusters in the L^γ -graphs.

Lemma 6 *Let $\gamma > 0$ and let $S \subset \mathcal{V}$ be an arbitrary set with $\text{vol}_\gamma(S) \leq \text{vol}_\gamma(G)/2$. For a labeled point placed at node $u \in S$ with probability proportional to its generalized degree*

in S , i.e. $\frac{[D_\gamma]_{uu}}{vol_\gamma(S)}$, L^γ -PageRank satisfies:

$$\mathbb{E}[f(S^c)] \leq \frac{h_S^{(\gamma)}}{\mu}. \tag{9}$$

The proof is provided in Appendix [Proof of lemma 6](#).

Lemma 6 admits a similar interpretation as Lemma 1. Namely, if L^γ -PageRank is applied to the labeled points of some set S with small $h_S^{(\gamma)}$, then diffusion is confined to S and the score values outside of S are expected to be small. Thus, by looking at the nodes with largest score values we should be able to retrieve a good estimation of S . If such score concentration phenomenon takes place, then a sharp drop must appear after sorting the L^γ -PageRank scores in descending order. We will use the following lemma to show that if a sharp drop is present, then the sweep cut procedure applied on the L^γ -PageRank vector retrieves a partition \hat{S} that has small $h_{\hat{S}}^{(\gamma)}$.

Lemma 7 *Let q denote the permutation vector and S_j denote the set associated to q_j obtained by applying the sweep-cut procedure on the L^γ -PageRank vector. Then, the partition $\mathcal{V} = S_j \cup S_j^c$ satisfies the inequality:*

$$\begin{aligned} \mathcal{A}_{out}^{(\gamma)}(S_j) \left(2 - \frac{(q_j - q_{j+1})}{(q_1 - q_N)} \right) - \mathcal{D}_{out}^{(\gamma)}(S_j) \left(2 \frac{(q_j - q_{j+1})}{(q_1 - q_N)} - 1 \right) &\geq \frac{\mu (y(S_j) - f(S_j))}{(q_1 - q_n)} \\ &\geq \mathcal{A}_{out}^{(\gamma)}(S_j) \left(2 \frac{(q_j - q_{j+1})}{(q_1 - q_N)} - 1 \right) - \mathcal{D}_{out}^{(\gamma)}(S_j) \left(2 - \frac{(q_j - q_{j+1})}{(q_1 - q_N)} \right). \end{aligned} \tag{10}$$

The proof is provided in Appendix [Proof of lemma 7](#).

We have that $\sum_{u \in S_j} \sum_{v \in S_j^c} [W_\gamma]_{uv} = \mathcal{A}_{out}^{(\gamma)}(S_j) - \mathcal{D}_{out}^{(\gamma)}(S_j) \geq 0$. Thus, the generalized Cheeger ratio of S_j is small if $\mathcal{A}_{out}^{(\gamma)}(S_j)$ is not much larger than $\mathcal{D}_{out}^{(\gamma)}(S_j)$. In the inequality above, we have two cases in which $(q_j - q_{j+1})/(q_1 - q_N) \approx 1$: (a) q is approximately constant; and (b) q has a drop that satisfies $q_j \approx q_1$ and $q_{j+1} \approx q_N$. The former can only occur if $f \rightarrow \pi_\gamma$ and clearly no cluster can be retrieved from that vector, as confirmed by the inequality growing unbounded. The latter case is what we coin as having a sharp drop between q_j and q_{j+1} . In such case, the inequality is controlled by the difference $y(S_j) - f(S_j)$ which, due to the mass preserving property and the assumption that $q_{j+1} \approx q_N$, should be small. Thus, granting that $\mathcal{A}_{out}^{(\gamma)}(S_j)$ is not much larger than $\mathcal{D}_{out}^{(\gamma)}(S_j)$ and S_j has a small $h_{S_j}^{(\gamma)}$.

Discussion. The previous results show that L^γ -PageRank is a sensible tool to find clusters in the L^γ -graphs, i.e. groups of nodes with small generalized Cheeger ratio. Thus, revisiting the classification case in which we target group of nodes S_{gt} , we have that the smaller the value of $h_{S_{gt}}^{(\gamma)}$, the better the L^γ -PageRank method can recover it. This observation, in addition to noting that standard PageRank emerges as the particular case of $\gamma = 1$, indicate that we should be able to enhance the performance of G-SSL in the detection of S_{gt} by finding the graph, i.e. the γ value, in which $h_{S_{gt}}^{(\gamma)} < h_{S_{gt}}^{(1)}$.

The selection of γ

Case of $\gamma = 2$: analytic study

In “[The \$L^\gamma\$ -graphs](#)” section, it was argued that the topology emerging from L^2 places a negatively weighted link between nodes at a 2-hop distance, thus carrying the potential to

place a big amount of disagreements between clusters that may enhance their separability. Our next result formalizes this claim, demonstrating that on graphs from the Planted Partition model it is expected that the L^2 -graph improves the generalized Cheeger ratio.

Theorem 1 *Consider a Planted Partition model of parameters (p_{in}, p_{out}) and cluster sizes $|S_{gt}| = |S_{gt}^c| = n$. Then, as $n \rightarrow \infty$ we have that*

$$\mathbb{E} \left[h_{S_{gt}}^{(2)} \right] = 2\mathbb{E} \left[h_{S_{gt}}^{(1)} \right]^2, \tag{11}$$

where $\mathbb{E} \left[h_{S_{gt}}^{(1)} \right] = p_{out} / (p_{in} + p_{out})$.

The proof is provided in Appendix [Proof of theorem 1](#).

Corollary 1 *If $p_{in} \geq p_{out}$, then $\mathbb{E} \left[h_{S_{gt}}^{(2)} \right] \leq \mathbb{E} \left[h_{S_{gt}}^{(1)} \right]$, with equality occurring in the case $p_{in} = p_{out}$.*

The proof is provided in Appendix [Proof of corollary 1](#).

Theorem 1 and Corollary 1 open the door to investigate, on arbitrary graphs, in which cases the L^2 -graph improves the generalized Cheeger ratio of a set. In the next Proposition, we provide a sufficient condition in which the L^2 -graph improves the generalized Cheeger ratio a set.

Proposition 1 *Let $\langle D_{S_{gt}} \rangle$ denote the mean degree of S_{gt} . A sufficient condition on S_{gt} so that $h_{S_{gt}}^{(2)} \leq h_{S_{gt}}^{(1)}$ is that*

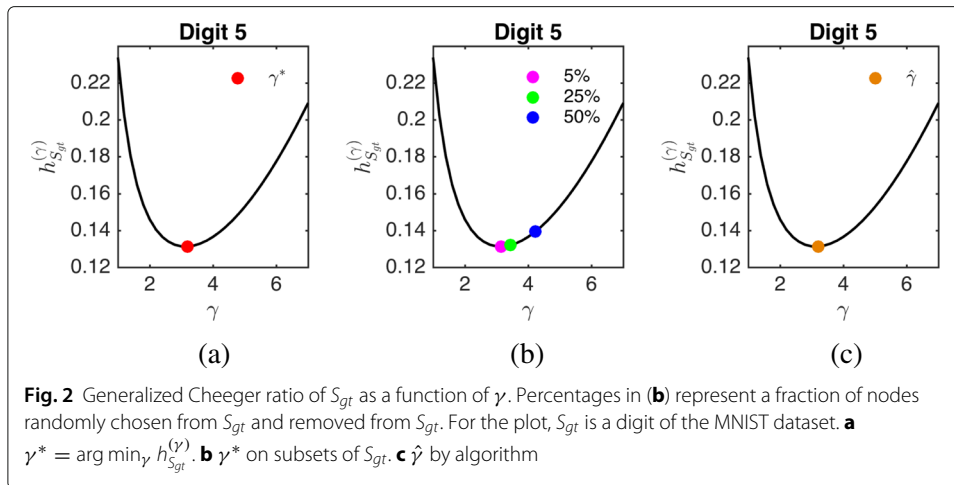
$$\langle D_{S_{gt}} \rangle \geq \max_{u \in S_{gt}} \sum_{v \in S_{gt}^c} W_{uv} + \max_{w \in S_{gt}^c} \sum_{\ell \in S_{gt}} W_{w\ell}. \tag{12}$$

The proof is provided in Appendix [Proof of proposition 1](#).

This proposition points in the same direction as Theorem 1, saying that graphs having a cluster structure are bound to benefit from L^2 . Concretely, the first term on the right hand side of the inequality searches, among all the nodes of S_{gt} , the one that has the maximum number of connections towards S_{gt}^c . The second term does the reverse for the nodes of S_{gt}^c . Hence, asking for the nodes of S_{gt} to have, on average, more connections than the maximum possible boundary implies that S_{gt} should have a cluster structure.

An algorithm for the estimation of the optimal γ

Numerical experiments show that increasing γ can further decrease the generalized Cheeger ratio up to a point where it starts increasing. We show an example of this phenomenon in Fig. 2a, displaying the evolution of $h_{S_{gt}}^{(\gamma)}$ as a function of γ when S_{gt} corresponds to a digit of the MNIST dataset. From the figure, it is evident that an optimal value appears, denoted $\gamma^* = \arg \min_{\gamma} h_{S_{gt}}^{(\gamma)}$, raising the question of how to find such value. Since the behavior of $h_{S_{gt}}^{(\gamma)}$ depends on S_{gt} , in practice, the derivative or a greedy search to find γ^* cannot be employed since S_{gt} is unknown. A second question that arises is whether the optimal value changes drastically or smoothly with changes in S_{gt} . We perform the following test: for a given S_{gt} (same MNIST digit), we remove some percentage of the nodes in S_{gt} and record the optimal value on subsets of S_{gt} . More precisely, recall that $h_{S_{gt}}^{(\gamma)} = \mathbb{1}_{S_{gt}}^T L^\gamma \mathbb{1}_{S_{gt}} / \mathbb{1}_{S_{gt}}^T D_\gamma \mathbb{1}_{S_{gt}}$, hence we randomly select some percentage of the entries



indexing S_{gt} in $\mathbb{1}_{S_{gt}}$, set them to zero and obtain a new indicator function indexing a subset of S_{gt} . Mean results are evaluated in the original curve and displayed in Fig. 2b. The figure suggest that it is not necessary to know S_{gt} to find a proxy $\hat{\gamma}$ of γ^* , it suffices to know a subset of S_{gt} . Based on the last observation, we propose Algorithm 1 for the estimation of γ^* . The rationale of the algorithm is to exploit the labeled points and the graph to find a proxy \hat{S} of S_{gt} on which we can compute the estimate. The procedure consists in letting walkers started from the label points, run for a number of steps that is determined by the maximum geodesic distance between the labels. This allows walkers to explore S_{gt} without escaping too far from it. After running the walk, we list the nodes in descending order according to the probability of finding a walk at a node. We take the first element on the list (the one where it is more likely to find a walker), add it to \hat{S} and remove it from the list, so that the former second element becomes the first in the listing. We repeat the procedure until the probability of finding a walker in the nodes conforming \hat{S} is 0.7.

In Table 1, we evaluate the performance of Algorithm 1 on the estimation of γ^* for all the digits of the MNIST. The first row displays, as γ^* , the value of γ (from the input range) attaining the minimum generalized Cheeger ratio. The second row displays the performance of the algorithm when estimating such value. The last three rows show the value of the generalized Cheeger ratio evaluated at γ^* , $\hat{\gamma}$ and $\gamma = 1$, respectively. The estimator finds values of $\hat{\gamma}$ whose Cheeger ratios are: (a) significantly smaller than those of $\gamma = 1$; (b) close to the optimal.

Table 1 Evaluation of Algorithm 1 on the MNIST Dataset

Digit	1	2	3	4	5	6	7	8	9
γ^*	7.0	3.0	7.0	3.2	3.2	7.0	7.0	3.2	4.2
$\hat{\gamma}$	5.45 (0.15)	3.10 (0.14)	6.41 (0.11)	4.92 (0.16)	3.20 (0.14)	6.04 (0.15)	4.98 (0.17)	4.40 (0.18)	5.08 (0.15)
$h_{S_{gt}}^{(\gamma^*)}$	0.065	0.166	0.035	0.141	0.131	0.011	0.052	0.116	0.135
$h_{S_{gt}}^{(\hat{\gamma})}$	0.073 (9e-4)	0.174 (8e-4)	0.041 (1e-3)	0.185 (4e-3)	0.148 (2e-3)	0.017 (1e-3)	0.074 (2e-3)	0.142 (2e-3)	0.149 (9e-4)
$h_{S_{gt}}^{(1)}$	0.175	0.248	0.216	0.258	0.233	0.107	0.203	0.215	0.285

Mean values (95% confidence interval) are shown. The graph construction guidelines are provided in "Real world datasets" section. For the experiment, 500 realizations of labeled points and a grid of γ ranging from 1 to 7 with a resolution of 0.2 were used

Algorithm 1 Estimation of γ^* **Input:** \mathcal{G} , $\mathcal{V}_{S_{gt}}$ and a grid of γ values.**Output:** $\hat{\gamma}$ Compute $\Delta_{uv} \forall u, v \in \mathcal{V}_{S_{gt}}$.Set $k = \max_{u,v} \{\Delta_{uv}\}$ Set $\chi = y/\|y\|_1$ Run a k -step walk with seed χ : $x^T = \chi^T P^k$ Reorder the vertices as v_1, \dots, v_N , so that $x_{v_i} \geq x_{v_{i+1}}$ **for** $i = 1 : N$ **do** **if** $\sum_{j=1}^i x_{v_j} < 0.7$ **then** Set $(\mathbb{1}_{\hat{S}})_{v_i} = 1$ **else** Set $(\mathbb{1}_{\hat{S}})_{v_i} = 0$ **end if****end for**Compute $h_{\hat{S}}^{(\gamma)} = \frac{\mathbb{1}_{\hat{S}}^T L^{\gamma} \mathbb{1}_{\hat{S}}}{\mathbb{1}_{\hat{S}}^T D_{\gamma} \mathbb{1}_{\hat{S}}} \forall \gamma$ Return $\hat{\gamma} = \arg \min_{\gamma} h_{\hat{S}}^{(\gamma)}$. **L^{γ} -PageRank in practice****Planted partition**

Experimental setup and goals. In the following experiment, we show that L^{γ} -PageRank can increase the performance of G-SSL as the graph approaches the Planted Partition detectability transition. More precisely, it is shown in Mossel et al. (2015) that the Planted Partition possesses a detectability threshold above which unsupervised methods are unable to retrieve a meaningful clustering. Indeed, if the clusters sizes are denoted as $|S_{gt}| = |S_{gt}^c| = n$, the mean degree of a node is given as $C_{avg} = C_{in} + C_{out}$, where $C_{out} = (p_{out})(n)$ and $C_{in} = (p_{in})(n - 1)$. It is then possible to recover a cluster that is positively correlated with the true partition, in an unsupervised manner, if $(C_{in} - C_{out})^2 > 2(C_{in} + C_{out})$, and impossible otherwise. As for G-SSL, the work in Zhang et al. (2014) showed that such threshold can be overcome when a fraction of labeled points is introduced to the task. Nonetheless, the performance of G-SSL drastically degrades when approaching the detectability transition.

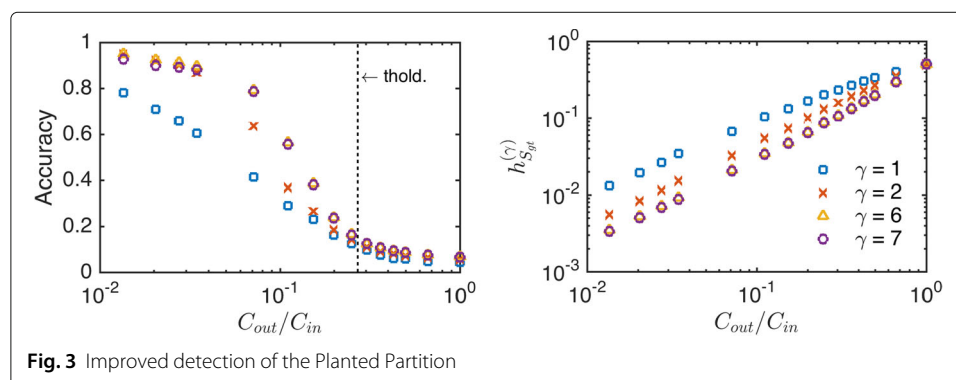
The experimental setup is the following: for a given C_{out}/C_{in} , a realization of the Planted Partition is drawn with $n = 500$ and $C_{avg} = 3$. Then, 1% of labeled points are sampled at random and the L^{γ} -PageRank method is applied for different values of μ lying on a discrete grid. The clusters are determined via a sweep-cut procedure, and the best performance is retained. The whole procedure is repeated for 10 different realizations of the labeled points. Finally, all the preceding steps are repeated for 100 graph realizations. Performance is assessed in terms of the Matthews Correlation Coefficient (MCC) (Matthews 1975), so that a value of 1 implies perfect agreement with the true partition and 0 a random decision.

Results and discussion. Figure 3 displays the performance of L^γ -PageRank at recovering the Planted Partition as a function of the ratio C_{out}/C_{in} . Standard PageRank ($\gamma = 1$) performs poorly as the configuration approaches the phase transition (referred by the vertical line) since $h_S^{(1)}$ becomes large. Clearly, the introduction of γ allows to decrease $h_{S_{gt}}^{(\gamma)}$, which, accordingly, enhances the clustering performance. Furthermore, the figure verifies that the smaller the value of $h_{S_{gt}}^{(\gamma)}$ (right plot), the better the L^γ -PageRank recovers the true partition (left plot). It is important to remark that, for this experiment, while $\gamma = 2$ shows good improvements, larger values of γ keep improving $h_S^{(\gamma)}$, until it reaches a saturation plateau, designating a region of optimal γ values ($\gamma \geq 6$).

Real world datasets

Experimental setup and goals. In our following experiment, we assesses the performance of L^γ -PageRank and Algorithm 1 on real world datasets.

The experimental setup is as follows: graphs are build connecting the K-Nearest Neighbors (KNN) with distances computed via the Gaussian kernel, so that the weight between points \mathbf{x}_u and \mathbf{x}_v is given by $W_{uv} = \exp\{-\|\mathbf{x}_u - \mathbf{x}_v\|_2^2/\sigma^2\}$. For each class, 2% of labeled points are randomly selected, L^γ -PageRank is applied for a grid of μ values, partitions are retrieved via the sweep-cut, and the best performance, assessed in terms of MCC, is retained. Such procedure is repeated for 100 realization of labeled points, except for the MNIST on which 30 realizations only are employed. In all cases, classes are balanced in size and the graph construction parameters are selected to provide a good distribution of weights as follows: (a) MNIST (Lecun et al. 1998): Images of handwritten digits (1 to 9). From the entire dataset, 200 images of each digit are selected and used to build the graph with KNN = 10 and $\sigma = 10^4$; (b) Gender Images (Hond and Spacek 1997): Images of male and female subjects for gender recognition. From the entire dataset, 200 images of each gender are selected and used to build the graph with KNN = 60 and $\sigma = 10^4$. The large value of KNN is to avoid disconnected components; (c) BBC articles (Greene and Cunningham 2006): Word frequency attributes from news media articles. From the entire dataset, 200 business and 200 entertainment articles are used to build the graph with KNN = 5 and $\sigma = 50$; and (d) Phoneme (The phoneme database): Five attributes to discern nasal sounds from oral sounds. From the entire dataset, 200 oral and 200 nasal sounds are used to build the graph with KNN = 10 and $\sigma = 2$.



Results and discussion. Table 2 shows the performance of L^γ -PageRank on the classification of these real world datasets. Clearly, the introduction of γ can significantly improve performance and, in general, the estimation $\hat{\gamma}$ performs close to the optimal value γ^* . It can be seen that some datasets are more sensitive to γ than others. For instance, in the BBC articles we observe that a small change in γ , going from $\gamma = 1$ to $\gamma^* = 1.1$, increases performance, and going further to $\hat{\gamma} = 1.3$ and $\gamma = 2$ significantly worsens the classification. On the other hand, the MNIST dataset is less sensitive to γ , obtaining similar performances with larger variations in γ .

It is important to stress that, thus far, we have assumed possession of the proper tuning of the diffusion rate (μ) that attains the best results. However, when working with real data, clusters may have intricate local structures, e.g. sub-clusters, that play an important role in the way information diffuses, and that can make more difficult the finding of the optimal diffusion rate μ . As a result, two clusters may have equal Cheeger ratios but one of them being harder to find if its local structure is complex. Digit 8 poses an example of this phenomenon, where the mean performance for $\hat{\gamma}$ is slightly better than that of γ^* . This anomaly can be explained as an aftereffect of using a finite grid on μ : for some realization of labeled points, the best performance for γ^* falls in a region not covered by the grid.

Unbalanced labeled data

Experimental setup and goals In our last experiment, we show that L^γ -PageRank, adapted to the multi-class setting described in “[Generalization to multiple classes](#)” section, can improve the performance of G-SSL in the presence of unbalanced labeled data.

The experimental setup is as follows: graphs with two balanced classes (in size) are built using the datasets from the preceding experiments. The parameters of the graphs’ construction follow the guidelines provided in “[Real world datasets](#)” section. For the Planted Partition, the configuration is $n = 200$, $C_{avg} = 3$, $C_{out} = 0.1$. Then, unbalanced labeled points are drawn at random: 2% from one class and 6% from the other. Lastly, L^γ -PageRank, in the multi-class setting, is applied for a grid of μ values and the best

Table 2 Performance on real world datasets: each cell reports MCC, 95% confidence interval (parenthesis) and the value of γ [squared brackets]

	S_{gt}	$\gamma = 1$	$\gamma = 2$	$\gamma = \hat{\gamma}$	$\gamma = \gamma^*$
MNIST	Digit 1	0.67 (0.075)	0.78 (0.032)	0.78 (0.034) [5.4]	0.80 (0.027) [7.0]
	Digit 2	0.38 (0.042)	0.60 (0.064)	0.64 (0.059) [3.3]	0.64 (0.059) [3.0]
	Digit 3	0.47 (0.040)	0.61 (0.032)	0.61 (0.028) [6.0]	0.61 (0.028) [7.0]
	Digit 4	0.39 (0.022)	0.48 (0.036)	0.53 (0.044) [4.7]	0.53 (0.037) [3.2]
	Digit 5	0.44 (0.036)	0.56 (0.046)	0.61 (0.036) [3.3]	0.64 (0.035) [3.2]
	Digit 6	0.90 (0.039)	0.94 (0.003)	0.94 (0.002) [6.0]	0.94 (0.002) [7.0]
	Digit 7	0.43 (0.027)	0.66 (0.043)	0.71 (0.042) [4.8]	0.75 (0.032) [7.0]
	Digit 8	0.47 (0.062)	0.65 (0.057)	0.74 (0.038) [4.8]	0.72 (0.050) [3.2]
	Digit 9	0.43 (0.020)	0.52 (0.026)	0.53 (0.023) [4.9]	0.56 (0.026) [4.2]
Gender images	Female	0.51 (0.039)	0.57 (0.028)	0.57 (0.020) [3.0]	0.57 (0.028) [2.0]
	Male	0.55 (0.028)	0.61 (0.021)	0.60 (0.022) [3.3]	0.61 (0.021) [2.4]
BBC articles	Business	0.80 (0.020)	0.53 (0.038)	0.72 (0.040) [1.3]	0.81 (0.021) [1.1]
	Entmt.	0.84 (0.027)	0.57 (0.040)	0.76 (0.047) [1.5]	0.86 (0.025) [1.3]
Phoneme	Nasal	0.37 (0.030)	0.41 (0.028)	0.43 (0.025) [2.9]	0.43 (0.025) [3.0]
	Oral	0.41 (0.025)	0.44 (0.022)	0.46 (0.019) [2.8]	0.46 (0.019) [3.0]

Table 3 Performance on unbalanced labeled data: each cell reports MCC, 95% confidence interval (parenthesis) and the value of γ [squared brackets]

	Planted Partition	MNIST 4vs9	MNIST 3vs8	BBC articles	Gender images	Phoneme
$\gamma = 1$	0.81 (1.1e-2)	0.51 (1.5e-2)	0.70 (1.4e-2)	0.66 (1.8e-2)	0.63 (2.1e-2)	0.44 (2.3e-2)
$\gamma = 2$	0.87 (8.7e-3)	0.56 (1.5e-2)	0.76 (1.2e-2)	0.92 (5.0e-3)	0.73 (1.6e-2)	0.48 (1.4e-2)
$\gamma = \text{Best}$	0.90 (7.0e-3) [6]	0.57 (1.5e-2) [3]	0.78 (1.2e-2) [4]	0.93 (1.5e-3) [3]	0.75 (1.7e-2) [3]	0.48 (1.4e-2) [1.9]

performance, assessed by MCC, is recorded. For the planted partition, the procedure is repeated over 15 realizations of the labeled points and for 100 graph realizations. For the other datasets, 100 realizations of labeled points are employed.

Results and discussion Table 3 displays the performance L^γ -PageRank in the presence of unbalanced labeled data. It is important to stress that, in this framework, a unique value of γ is used to retrieve all the clusters at the same time, precluding the notion of an optimal γ as defined in “The selection of γ ” section. However, one value of γ seems to perform better, we denote it as $\gamma = \text{Best}$. The results confirm that the introduction of γ helps to improve the classification in the presence of the unbalanced labeled data.

Conclusion

This work proposed L^γ -PageRank, an extension of PageRank based on (non necessary integer) powers of the (combinatorial) Laplacian matrix. Our analysis shows that the added degree of freedom offers more versatility than standard PageRank, providing the potential to address some of the limitations of G-SSL. Precisely, we showed that when clusters are obtained via the sweep-cut procedure, L^γ -PageRank can significantly outperform standard PageRank. Further, we showed that the multi-class approach also benefits from our proposition, as performance was enhanced in the presence of unbalanced labeled data. These improvements were possible due to the L^γ ($\gamma > 0$) operator coding for graphs whose topology can reinforce the separability of clusters. The richness of such graphs comes from the sign of edges, allowing to code for similarities but also to emphasize dissemblance between individuals. Thus, while 2 nodes can only be disconnected on the initial graph, they can ‘repulse’ themselves in these topologies. Notably, we have shown that there is an optimal graph (related to an optimal γ) on which the classification will lead to a maximal performance. We proposed a simple yet efficient algorithm to estimate the optimal γ and hence determine the best topology for analyzing a given dataset. The procedures proposed in this work open the door for more in-depth study of the L^γ -graphs and what determines their optimal topology. They also pave the way towards the extension of other standard clustering tools, such as Unsupervised Learning via Spectral Clustering, to exploit these richer topologies.

Appendix: Proofs

Proof of Lemma 3

Proof For an arbitrary and fixed $\gamma > 0$, let S denote an arbitrary set of $\text{vol}_\gamma(S) \leq \text{vol}_\gamma(G)/2$. Let $r = (\mathcal{A}_{out}^{(\gamma)}(S) - \mathcal{D}_{out}^{(\gamma)}(S))/(\mathcal{A}_{in}^{(\gamma)}(S) - \mathcal{D}_{in}^{(\gamma)}(S))$. It is easy to show that

$h_S^{(\gamma)} = r/(r + 1)$, which is monotonically increasing with r . Thus, the set S that minimizes r also minimizes $h_S^{(\gamma)}$. \square

Proof of lemma 4

Proof It suffices to show the positive semi-definiteness of the functional and to apply the first order optimality condition. Let $\tilde{f} = Q^T D_\gamma^{-1} f$. Then, the left term satisfies $\sum_j \lambda_j^\gamma \tilde{f}_j^2 \geq 0$. It can be shown that $[D_\gamma]_{uu} = \sum_j Q_{uj}^2 \lambda_j^\gamma \geq 0$ granting the right term satisfies $\sum_u (f_u - y_u)^2 / [D_\gamma]_{uu} \geq 0$. Now, computing the derivative of the functional with respect to f and equating to 0 leads to: $L^\gamma D_\gamma^{-1} f + \mu(f - y) = 0$. The lemma is proved after isolating f . \square

Proof of lemma 5

Proof From the demonstration of Lemma 4 we have that $L^\gamma D_\gamma^{-1} f + \mu(f - y) = 0$. Then, $\mathbb{1}^T L^\gamma D_\gamma^{-1} + \mu \mathbb{1}^T f = \mu \mathbb{1}^T y$. Since $\mathbb{1}^T L^\gamma = 0$ we have that $\mathbb{1}^T f = \mathbb{1}^T y$, proving (i). We prove property (iii) using the same expression. We only develop the case $\mu \rightarrow 0$ since the case $\mu \rightarrow \infty$ follows the same steps: taking $\lim_{\mu \rightarrow 0} \{L^\gamma D_\gamma^{-1} f + \mu(f - y) = 0\}$ leads to $L^\gamma D_\gamma^{-1} f = 0$, whose solution is proportional to $\pi_\gamma = D_\gamma \mathbb{1} / \text{vol}_\gamma(G)$. Lastly, we prove (ii) by noting that the operator $L^\gamma D_\gamma^{-1}$ has a positive real spectrum as it is similar to $D_\gamma^{-1/2} L^\gamma D_\gamma^{-1/2}$ which is positive semi-definite. Thus, we can use the inverse Laplace transform of the resolvent $(L^\gamma D_\gamma^{-1} + \mu \mathbb{1})^{-1} = \int_0^\infty e^{-t} e^{-t L^\gamma D_\gamma^{-1} / \mu} dt$, which, after using its Taylor expansion, allows to rewrite the PageRank solution as $f = \sum_{k=0}^\infty \frac{(-1)^k}{\mu^k} (L^\gamma D_\gamma^{-1})^k y$. If $y = \pi_\gamma$, the previous equation is only non-zero for $k = 0$, proving (ii). \square

Proof of lemma 6

Proof Let $y = D_\gamma \mathbb{1}_S / \text{vol}_\gamma(S)$. Using (8) we can see that

$$\mathbb{1}_{S^c}^T f = \sum_{u \in S} \frac{[D_\gamma]_{uu}}{\text{vol}_\gamma(S)} \mathbb{1}_{S^c}^T \left[\mu (L^\gamma D_\gamma^{-1} + \mu \mathbb{1})^{-1} \delta_u \right], \tag{13}$$

showing that $\mathbb{1}_{S^c}^T f$ can be interpreted as $\mathbb{E}[f(S^c)]$ when labels are selected with probability proportional to their generalized degree in S . Using the fact that

$$(L^\gamma D_\gamma^{-1} + \mu \mathbb{1})^{-1} (L^\gamma D_\gamma^{-1} + \mu \mathbb{1}) = \mathbb{1}, \tag{14}$$

we express

$$f = \left(\mathbb{1} - \frac{1}{\mu} L^\gamma D_\gamma^{-1} + \frac{1}{\mu} L^\gamma D_\gamma^{-1} (L^\gamma D_\gamma^{-1} + \mu \mathbb{1})^{-1} L^\gamma D_\gamma^{-1} \right) y. \tag{15}$$

The upper bound is thus obtained by substituting y and summing over S .

$$\begin{aligned} \mathbb{1}_S^T f &= \frac{\mathbb{1}_S^T D_\gamma \mathbb{1}_S}{\text{vol}_\gamma(S)} - \frac{\mathbb{1}_S^T L^\gamma \mathbb{1}_S}{\mu \text{vol}_\gamma(S)} + \frac{\mathbb{1}_S^T L^\gamma (L^\gamma + \mu D_\gamma)^{-1} L^\gamma \mathbb{1}_S}{\mu \text{vol}_\gamma(S)} \\ &\geq \frac{\mathbb{1}_S^T D_\gamma \mathbb{1}_S}{\text{vol}_\gamma(S)} - \frac{\mathbb{1}_S^T L^\gamma \mathbb{1}_S}{\mu \text{vol}_\gamma(S)} \\ &= 1 - \frac{h_S^{(\gamma)}}{\mu}. \end{aligned} \tag{16}$$

Employing property (i) from Lemma 5 finishes the proof. \square

Proof of lemma 7

Proof We only show the proof of the lower bound as the upper bound follows a similar derivation. We recast (4) as $L^\gamma D_\gamma^{-1} f = \mu (y - f)$. Thus, the set S_j satisfies:

$$\begin{aligned}
 \mu (y(S_j) - f(S_j)) &= \mathbb{1}_{S_j}^T L^\gamma D_\gamma^{-1} f \\
 &= \mathbb{1}_{S_j}^T L^\gamma q \\
 &= \sum_{u \in S_j, v \in S_j^c} [W_\gamma]_{uv} (q_u - q_v) \\
 &= \sum_{u \in S_j, v \in S_j^c} |[W_\gamma^+]_{uv}| (q_u - q_v) - \sum_{u \in S_j, v \in S_j^c} |[W_\gamma^-]_{uv}| (q_u - q_v) \\
 &\quad + \sum_{u \in S_j, v \in S_j^c} |[W_\gamma^+]_{uv}| (q_j - q_{j+1}) - \sum_{u \in S_j, v \in S_j^c} |[W_\gamma^-]_{uv}| (q_j - q_{j+1}) \\
 &\quad + \sum_{u \in S_j, v \in S_j^c} |[W_\gamma^-]_{uv}| (q_j - q_{j+1}) - \sum_{u \in S_j, v \in S_j^c} |[W_\gamma^+]_{uv}| (q_j - q_{j+1}) \\
 &\geq (q_j - q_{j+1}) \left(2\mathcal{A}_{out}^{(\gamma)}(S_j) + \mathcal{D}_{out}^{(\gamma)}(S_j) \right) \\
 &\quad - (q_1 - q_N) \left(2\mathcal{D}_{out}^{(\gamma)}(S_j) + \mathcal{A}_{out}^{(\gamma)}(S_j) \right). \tag{17}
 \end{aligned}$$

Re-ordering terms finishes the proof. □

Proof of theorem 1

Proof Let $n = |S|$. For $u, v \in S$ and $w \in S^c$ the Planted Partition satisfies $\sum_v W_{uv} \sim B(n - 1, p_{in})$ and $\sum_w W_{uw} \sim B(n, p_{out})$. The key step in the proof is to show that, in the limit $n \rightarrow \infty$, $\mathbb{E} [h_S^{(1)}] = \mathbb{E} \left[\frac{\mathbb{1}_S^T L \mathbb{1}_S}{vol(S)} \right] = \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S]}{\mathbb{E} [vol(S)]}$, and the same for $h_S^{(2)}$. By application of the Chebyshev inequality we have that

$$Pr (d_u - \mathbb{E} [d_u] \geq \mathbb{E} [d_u]) \leq \frac{var(d_u)}{var(d_u) + \mathbb{E} [d_u]^2} = \mathcal{O}(n^{-1}). \tag{18}$$

Thus, in the limit of $n \rightarrow \infty$ we can establish the inequality $d_u < 2\mathbb{E} [d_u]$ and further that $vol(S) < 2\mathbb{E} [vol(S)]$. This latter allows to express $\mathbb{E} [h_S^{(1)}]$ as follows (Rice 2008):

$$\begin{aligned}
 \mathbb{E} [h_S^{(1)}] &= \mathbb{E} \left[\frac{\mathbb{1}_S^T L \mathbb{1}_S}{vol(S)} \right] \\
 &= \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S]}{\mathbb{E} [vol(S)]} + \sum_{i=1}^{\infty} (-1)^i \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S] \langle \langle i vol(S) \rangle \rangle + \langle \langle \mathbb{1}_S^T L \mathbb{1}_S, i vol(S) \rangle \rangle}{\mathbb{E} [vol(S)]^{i+1}} \\
 &= \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S]}{\mathbb{E} [vol(S)]} + \sum_{i=1}^{\infty} (-1)^i \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S (vol(S) - \mathbb{E} [vol(S)])^i]}{\mathbb{E} [vol(S)]^{i+1}} \\
 &= \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S]}{\mathbb{E} [vol(S)]} + \sum_{i=1}^{\infty} (-1)^i \mathbb{E} \left[\frac{\mathbb{1}_S^T L \mathbb{1}_S}{\mathbb{E} [vol(S)]} \left(\frac{vol(S)}{\mathbb{E} [vol(S)]} - 1 \right)^i \right] \\
 &= \frac{\mathbb{E} [\mathbb{1}_S^T L \mathbb{1}_S]}{\mathbb{E} [vol(S)]} + \sum_{i=1}^{\infty} (-1)^i c_i, \tag{19}
 \end{aligned}$$

where $\langle\langle a, b \rangle\rangle = \mathbb{E}[(a - \mathbb{E}[a])(b - \mathbb{E}[b])^i]$. The fact that $vol(S) < 2\mathbb{E}[vol(S)]$ and the monotonicity of the expected value imply that the sequence $\sum_i |c_i|$ decreases monotonically. Also, it can be shown that its dominant term: $c_1 = \mathcal{O}(n^{-2})$. Replacing the expectations and evaluating the limit leads to:

$$\lim_{n \rightarrow \infty} \mathbb{E}[h_S^{(1)}] = \frac{p_{out}}{p_{in} + p_{out}}. \tag{20}$$

The case of $\mathbb{E}[h_S^{(2)}]$ follows a similar derivation. Since $[D_2]_{uu} = d_u^2 + d_u$, the Jensen inequality implies that $[D_2]_{uu} < 2\mathbb{E}[[D_2]_{uu}]$ and consequently that $vol_2(S) < 2\mathbb{E}[vol_2(S)]$. Thus, we cast:

$$\begin{aligned} \mathbb{E}[h_S^{(2)}] &= \mathbb{E}\left[\frac{\mathbb{1}_S^T L^2 \mathbb{1}_S}{vol_2(S)}\right] \\ &= \frac{\mathbb{E}[\mathbb{1}_S^T L^2 \mathbb{1}_S]}{\mathbb{E}[vol_2(S)]} + \sum_{i=1}^{\infty} (-1)^i \mathbb{E}\left[\frac{\mathbb{1}_S^T L^2 \mathbb{1}_S}{\mathbb{E}[vol_2(S)]} \left(\frac{vol_2(S)}{\mathbb{E}[vol_2(S)]} - 1\right)^i\right] \\ &= \frac{\mathbb{E}[\mathbb{1}_S^T L^2 \mathbb{1}_S]}{\mathbb{E}[vol_2(S)]} + \sum_{i=1}^{\infty} (-1)^i c_i^{(2)}. \end{aligned} \tag{21}$$

Let the random variable $O_u = \sum_{w \in S^c} W_{uw}$. Then we have that $\mathbb{1}_S^T L^2 \mathbb{1}_S = 2 \sum_{u \in S} (O_u)^2$. This fact, in addition to $vol_2(S) = \sum_{u \in S} d_u^2 + d_u$, allow to show that the sequence $\sum_i |c_i^{(2)}|$ is monotonically decreasing with $c_1^{(2)} = \mathcal{O}(n^{-1})$. Replacing the expectations and evaluating the limit leads to:

$$\lim_{n \rightarrow \infty} \mathbb{E}[h_S^{(2)}] = 2 \left(\frac{p_{out}}{p_{in} + p_{out}}\right)^2. \tag{22}$$

□

Proof of corollary 1

Proof Let $p_{in} = p_{out} + \epsilon$ and assume that $h_S^{(1)} \geq h_S^{(2)}$. Thus $p_{out}/(p_{in} + p_{out}) \geq 2(p_{out}/(p_{in} + p_{out}))^2$, which can be further simplified to $1 \geq 2p_{out}/(2p_{out} + \epsilon)$. We observe that such expression holds for $\epsilon \geq 0$ and equality occurs when $\epsilon = 0$. □

Proof of proposition 1

Proof We search a condition on S that permits $\frac{\mathbb{1}_S^T L \mathbb{1}_S}{\mathbb{1}_S^T D \mathbb{1}_S} \geq \frac{\mathbb{1}_S^T L^2 \mathbb{1}_S}{\mathbb{1}_S^T D_2 \mathbb{1}_S}$, or equivalently, that satisfies the inequality $\frac{\mathbb{1}_S^T D_2 \mathbb{1}_S}{\mathbb{1}_S^T D \mathbb{1}_S} - \frac{\mathbb{1}_S^T L^2 \mathbb{1}_S}{\mathbb{1}_S^T L \mathbb{1}_S} \geq 0$. We have

$$\begin{aligned} \frac{\mathbb{1}_S^T D_2 \mathbb{1}_S}{\mathbb{1}_S^T D \mathbb{1}_S} - \frac{\mathbb{1}_S^T L^2 \mathbb{1}_S}{\mathbb{1}_S^T L \mathbb{1}_S} &\geq \frac{\mathbb{1}_S^T D^2 \mathbb{1}_S}{\mathbb{1}_S^T D \mathbb{1}_S} - \frac{\mathbb{1}_S^T L^2 \mathbb{1}_S}{\mathbb{1}_S^T L \mathbb{1}_S} \\ &\geq \frac{\mathbb{1}_S^T D^2 \mathbb{1}_S}{\mathbb{1}_S^T D \mathbb{1}_S} - \left(\max_{u \in S} \sum_{w \in S^c} W_{uw} + \max_{\ell \in S^c} \sum_{v \in S} W_{\ell v}\right) \\ &\geq \frac{\mathbb{1}_S^T D \mathbb{1}_S}{\mathbb{1}_S^T \mathbb{1}_S} - \left(\max_{u \in S} \sum_{w \in S^c} W_{uw} + \max_{\ell \in S^c} \sum_{v \in S} W_{\ell v}\right) \\ &= \frac{vol(S)}{|S|} - \left(\max_{u \in S} \sum_{w \in S^c} W_{uw} + \max_{\ell \in S^c} \sum_{v \in S} W_{\ell v}\right), \end{aligned} \tag{23}$$

where we have used Lehmer's and Holder's inequalities and that $\mathbb{1}_S^T L^2 \mathbb{1}_S = \sum_{u \in S} (\sum_{w \in S^c} W_{uw})^2 + \sum_{\ell \in S^c} (\sum_{v \in S} W_{\ell v})^2$. Thus, it is sufficient that S satisfies:

$$\frac{\text{vol}(S)}{|S|} - \left(\max_{u \in S} \sum_{w \in S^c} W_{uw} + \max_{\ell \in S^c} \sum_{v \in S} W_{\ell v} \right) \geq 0. \quad (24)$$

□

Acknowledgements

Not applicable

Author's contributions

EB, PA and PG participated equally in designing and developing the project, and in writing the paper. All authors read and approved the final manuscript.

Funding

This work was supported by CONAcYT and the Labex MILyon

Availability of data and materials

All data generated or analysed during this study are included the following articles (Lecun et al. 1998; Hond and Spacek 1997; Greene and Cunningham 2006; The phoneme database). The code to replicate the results is available in the GitHub repository, https://github.com/estbautista/Lgamma-PageRank_Paper.

Competing interests

The authors declare that they have no competing interests.

Received: 1 March 2019 Accepted: 22 July 2019

Published online: 19 August 2019

References

- Andersen R, Chung FRK, Lang KJ (2007) Using pagerank to locally partition a graph. *Int Math* 4:35–64. <https://doi.org/10.1080/15427951.2007.10129139>
- Andersen R, Chung F (2007) Detecting sharp drops in pagerank and a simplified local partitioning algorithm. In: Cai J-Y, Cooper SB, Zhu H (eds). *Theory and Applications of Models of Computation*. Springer, Berlin. pp 1–12
- Avrachenkov K, Dobrynin V, Nemirovsky D, Pham SK, Smirnova E (2008) Pagerank based clustering of hypertext document collections. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Singapore. pp 873–874
- Avrachenkov K, Gonçalves P, Legout A, Sokol M (2012a) Classification of Content and Users in BitTorrent by Semi-supervised Learning Methods. In: *International Wireless Communications and Mobile Computing Conference (3rd International Workshop on Traffic Analysis and Classification)*. IEEE, Cyprus. Best paper award
- Avrachenkov K, Gonçalves P, Mishenin A, Sokol M (2012b) Generalized Optimization Framework for Graph-based Semi-supervised Learning. In: *Proceedings of the 2012 SIAM International Conference on Data Mining*. pp 966–974. <https://epubs.siam.org/doi/abs/10.1137/1.9781611972825.83>
- Avrachenkov K, Kadavankandy A, Litvak N (2018) Mean Field Analysis of Personalized PageRank with Implications for Local Graph Clustering. *J Stat Phys* 173(3-4):895–916. <https://doi.org/10.1007/s10955-018-2099-5>
- Bautista E, De Nigris S, Abery P, Avrachenkov K, Gonçalves P (2017) Lévy Flights for Graph Based Semi-Supervised Classification. In: *26th Colloquium GRETSI*. GRETSI, 2017 - Proceeding of the 26th colloquium. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, Juan-Les-Pins
- Chung F (2007) Four cheeger-type inequalities for graph partitioning algorithms. In: *Proceedings of ICCM*. Sciotech Publisher, Hiroshima
- Chung F (2010) Pagerank as a discrete green's function. *Geom Anal I ALM* 17:285–302
- de Nigris S, Bautista E, Abery P, Avrachenkov K, Gonçalves P (2017) Fractional graph-based semi-supervised learning. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. pp 356–360. <https://doi.org/10.23919/EUSIPCO.2017.8081228>
- Fontugne R, Bautista E, Petrie C, Nomura Y, Abery P, Gonçalves P, Fukuda K, Aben E (2019) BGP Zombies: an Analysis of Beacons Stuck Routes. In: *PAM 2019 - 20th Passive and Active Measurements Conference*. Springer, Puerto Varas. pp 1–13. <https://hal.inria.fr/hal-01970596>
- Graham FC, Horn P, Tsiatas A (2009) Distributing antidote using pagerank vectors. *Int Math* 6:237–254
- Greene D, Cunningham P (2006) Practical solutions to the problem of diagonal dominance in kernel document clustering. In: *Proceedings of the 23rd International Conference on Machine Learning. ICML '06*. ACM, New York. pp 377–384. <https://doi.org/10.1145/1143844.1143892>
- Hond D, Spacek L (1997) Distinctive descriptions for face processing. In: Clark AF (ed). *BMVC*. British Machine Vision Association, Essex
- Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324. <https://doi.org/10.1109/5.726791>
- Litvak N, Scheinhardt W, Volkovich Y, Zwart B (2009) Characterization of tail dependence for in-degree and pagerank. In: Avrachenkov K, Donato D, Litvak N (eds). *Algorithms and Models for the Web-Graph*. Springer, Berlin. pp 90–103
- Matthews BW (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta Protein Struct* 405(2):442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)

- Mossel E, Neeman J, Sly A (2015) Reconstruction and estimation in the planted partition model. *Probab Theory Relat Fields* 162(3):431–461. <https://doi.org/10.1007/s00440-014-0576-6>
- Pérez Riascos A, Mateos J (2014) Fractional dynamics on networks: Emergence of anomalous diffusion and lévy flights. *Phys Rev E* 90:032809. <https://doi.org/10.1103/PhysRevE.90.032809>
- Rice SH (2008) A stochastic version of the Price equation reveals the interplay of deterministic and stochastic processes in evolution. *BMC evolutionary biology* 8(1):262
- Shuman DI, Vandergheynst P, Kressner D, Frossard P (2018) Distributed signal processing via chebyshev polynomial approximation. *IEEE Trans Signal Inf Proc Over Netw* 4(4):736–751. <https://doi.org/10.1109/TSIPN.2018.2824239>
- Sokol M (2014) Graph-based semi-supervised learning methods and quick detection of central nodes. PhD thesis. Université de Nice, Ecole Doctorale STIC, Inria Sophia Antipolis, Maestro
- Subramanya A, Bilmes J (2008) Soft-supervised learning for text classification. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08*. Association for Computational Linguistics, Stroudsburg, pp 1090–1099
- Tsiatas A (2012) Diffusion and clustering on large graphs. PhD thesis. University of California at San Diego, La Jolla
- The phoneme database. <https://www.openml.org/d/1489>. Accessed 1 Feb 2019
- Zhang P, Moore C, Zdeborova L (2014) Phase transitions in semisupervised clustering of sparse networks. *Phys Rev E Stat Nonlinear Soft Matter Phys* 90. <https://doi.org/10.1103/PhysRevE.90.052802>
- Zhao M, Chan RHM, Chow TWS, Tang P (2014) Compact graph based semi-supervised learning for medical diagnosis in alzheimer's disease. *IEEE Signal Proc Lett* 21(10):1192–1196. <https://doi.org/10.1109/LSP.2014.2329056>
- Zhou X, Belkin M (2011) Semi-supervised learning by higher order regularization. In: Gordon G, Dunson D, Dudík M (eds). *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*. PMLR, Fort Lauderdale Vol. 15. pp 892–900. <http://proceedings.mlr.press/v15/zhou11b/zhou11b.pdf>
- Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2004) Learning with local and global consistency. In: Thrun S, Saul LK, Schölkopf B (eds). *Advances in Neural Information Processing Systems* 16. pp 321–328. <http://papers.nips.cc/paper/2506-learning-with-local-and-global-consistency.pdf>
- Zhou D, Burges CJC (2007) Spectral clustering and transductive learning with multiple views. In: *Proceedings of the 24th International Conference on Machine Learning. ICML '07*. ACM, New York. pp 1159–1166. <https://doi.org/10.1145/1273496.1273642>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
