Lévy Flights for Graph Based Semi-Supervised Classification

Esteban BAUTISTA[1], Sarah DE NIGRIS[1], Patrice ABRY[1], Konstantin AVRACHENKOV[2], Paulo GONÇALVES[1]

[1]Univ Lyon, Ens de Lyon, Inria, CNRS, UCB Lyon 1, F-69342, Lyon, FRANCE

[2]Inria Sophia Antipolis, 2004 Route des Lucioles, Sophia-Antipolis, France

{esteban.bautista-ruiz,sarah.de-nigris,patrice.abry,paulo.goncalves}@ens-lyon.fr,
konstantin.avratchenkov@inria.fr

Résumé – Les algorithmes de classification semi-supervisée sur graphes peuvent être vus comme des processus de diffusion avec réinitialisation aux points labélisés. En partant de cette interprétation, nous proposons un nouvel algorithme inspiré d'un processus de diffusion non local, basé sur la puissance γ de la matrice Laplacienne standard, où $0 < \gamma < 1$. En permettant à la marche de relier en un seul saut, des nœuds distants du graphe, cette approche induit des transitions à longue portée, aussi appelées vols de Lévy, qui accélèrent l'exploration du graphe. Dans cet article, nous montrons que ces processus peuvent améliorer les performances des classifieurs semi-supervisés dans certains cas de figure pathologiques tels que celui des classes déséquilibrées et proposons une règle théorique de classification.

Abstract – Classification through Graph-based semi-supervised learning algorithms can be viewed as a diffusion process with restart on the labels. In this work, we exploit this equivalence to introduce a novel algorithm which relies on the formulation of a non-local diffusion process, fueled by the γ-th power of the standard Laplacian matrix $\mathbf{L}^\gamma$, with $0 < \gamma < 1$. This approach allows to jump in one step between far apart nodes and such long-range transitions, called Lévy Flights, entail a wider exploration of the graph. In the present contribution, we embed such mechanism in graph based semi-supervised algorithms to improve the classification outcome, even in settings traditionally poorly performing such as unbalanced classes, and we derive a theoretical rule for classification decision.

1 Introduction

Graph-based semi-supervised learning (G-SSL) allows for data classification blending two ingredients : the graph structure and the labeled data. By levering the graph, one is allowed to classify data in situations where only a handful amount of labeled data is available, presenting an advantage with respect to the popular paradigm of supervised learning where only the label data prime. The usefulness of this approach is evident in contexts where the data structure is easily accessible, while label data might not due to expensive expertise. As a consequence, G-SSL has been successfully used in tasks like classification of BitTorrent content and users [1], text categorization [2], medical diagnosis [3], among others. In this, the widely used methods of *Standard Laplacian* (SL) and *PageRank* (PR), on which we will focus, admit a closed solution that can be viewed under the light of random walks (RW) theory : from this perspective, the class attribution of a node depends on how many times, on average, it is visited from a labeled node. Albeit its successes, G-SSL presents some drawbacks as, for instance, when hubs skew the classes. [1]

Contributions and Outline The main contribution of the article is to profit from the RW interpretation of G-SSL methods to derive new classification rules reminiscent of more efficient diffusion processes, such as Lévy Flights, with the aim of overcoming some of the limitations of standard G-SSL methods. In Sec. 2.1 we state the G-SSL problem and outline the generalized optimization formulation and its particular cases of SL and PR. Further, we present in Sec. 2.2 the interpretation of G-SSL in the terminology of RW theory and motivate the use of long-range transition RWs. The developments in Sec. 3 present novel definitions of G-SSL build on Lévy Flight-based operators and are the main contribution of the article. Numerical experiments conducted on a synthetic graph are presented in Sec. 4 where we illustrate the potential of our formulation.

1. We thank CONACYT and Labex MILyon for their financial support.

2 State of the art and related work

2.1 Graph-based Semi-Supervised Learning

Given an undirected graph $G(V, E, W)$, a label set $\mathcal{Y} = \{1, \ldots, K\}$, and a labeled subset of the vertices $S \subset V$, we want to classify the points in the complement of S. For the sake of simplicity, all theoretical derivations below consider $w_{i,j} = 1$ if nodes i and j are connected, and zero otherwise, although the extension to weighted graphs is direct (cfr. pg. 12 of [4]). Let $D = diag(d_1 \ldots d_N)$ be a matrix whose entries are the nodes' degrees $d_i = \sum_j w_{i,j}$. Therefore, the Standard (or Combinatorial) Laplacian operator defined as $L = D - W$ is diagonalizable according to $L = Q^T \Lambda Q$. Also, let V_k denote the set of labeled points that belong to class $k \in \mathcal{Y}$ with $|S| = |V_1| + \cdots + |V_K|$. Consider $Y \in \mathbb{R}^{N \times K}$ to be the ground

truth matrix encoding the labeled nodes by setting $[Y]_{ik} = 1$ if node i belongs to class $k \in \mathcal{Y}$ and zero otherwise. Lastly, the $F \in \mathbb{R}^{N\times K}$ matrix denotes the classification functions we look for and finally, the decision rule affects node i to the class k that satisfies $\text{argmax}_k F_{ik}$.

We build upon a series of works [1, 5, 6] that present a generalized expression for G-SSL that embraces the different standard G-SSL methods, namely SL, PR, and Normalized Laplacian (NL). The expression proposed in [5] reads

$$\min_F \left\{ 2F^T D^{\sigma-1} L D^{\sigma-1} F + \mu \left(F-Y\right)^T D^{2\sigma-1} (F-Y) \right\} . \tag{1}$$

We recall that throughout this work, with F and Y we intend the column vectors F_{*k} and Y_{*k}. In [5] is also shown that the minimization problem (1) has a closed form solution that takes the form

$$F^T = (1-\alpha) Y^T D^{\sigma} \left(I - \alpha D^{-1} W\right)^{-1} D^{-\sigma} , \tag{2}$$

where $\alpha = \frac{2}{2+\mu}$. Properly tuned, the σ parameter allows to get back the SL and PR methods as described next.

- SL : Replacing $\sigma = \mathbf{1}$ in (1) leads to the SL classification problem cast as

$$\min_F \left\{ 2F^T L F + \mu \left(F-Y\right)^T D \left(F-Y\right) \right\} , \tag{3}$$

with solution

$$F^T = (1-\alpha) Y^T D \left(I - \alpha D^{-1} W\right)^{-1} D^{-1}. \tag{4}$$

- PR : Replacing $\sigma = \mathbf{0}$ in (1) leads to the PR classification problem cast as

$$\min_F \left\{ 2F^T D^{-1} L D^{-1} F + \mu \left(F-Y\right)^T D^{-1} \left(F-Y\right) \right\} , \tag{5}$$

with solution

$$F^T = (1-\alpha) Y^T \left(I - \alpha D^{-1} W\right)^{-1} . \tag{6}$$

We conclude this subsection by noting that very recently a similar framework, implementing the Laplacian matrix normalized by $D^{-\alpha}$, was successfully used in the context of community detection [7].

2.2 Random Walks interpretation

We recognize from (4) and (6) the $D^{-1}W$ matrix that is indeed the *transition matrix* of a RW on a graph, and also the operator $I - \alpha D^{-1} W$ gives the stationary solution of a *RW with restart*, with restarts occurring with probability $p_r = 1 - \alpha$. Concretely, a classification rule in RW terminology is derived in [8] stating that unlabeled node i is attributed to class k if

$$\sum_{p\in V_k} d_p^{\sigma} q_{pi} > \sum_{s\in V_{k'}} d_s^{\sigma} q_{si}, \ \forall k' \neq k, \tag{7}$$

where q_{pi} is the probability that RWs starting from any labeled point p in class k reach the node i, before reinitialization due to the absorption state with probability $(1-\alpha)$. We refer the reader to [8] for the derivation and we would like to emphasise that ours, presented in Sec. 3, resembles theirs closely.

Lévy Flights Random Walks The random walk interpretation of G-SSL offers an entry point to embed other types of diffusion dynamics, like Lévy Flights into the classification context. Lévy Flights spurred a remarkable stream of research due to its efficient exploration properties [9]. In metric spaces, the basic step is natural to define : the walkers are able to perform jumps, the "flights", whose length ℓ is drawn from a probability distribution $P(\ell)$, enhancing the overall space exploration through these *long-range transitions*. However a degeneracy in definition arises on networks since the "length" of a jump is not univocally defined on networks. To overcome this drawback, various generalizations of the diffusion operator L have been proposed :

Random Walk-like operators We recall that the generalization of L must be associated to a stochastic adjacency matrix, i.e. with non-negative entries and the sum over the rows (or the columns) has to be zero to entail probability conservation. The two following operators satisfy this prescription, thus being RW operators *strictu sensu :*

$\diamond$ **$\mathbf{L^\gamma}$ with $\mathbf{0 < \gamma \leq 1}$ - Fractional Laplacian**

In [10], it has been analytically demonstrated, on rings, that the fractional powers of L, defined as $L^\gamma = Q^T \Lambda^\gamma Q$, lead, in the $0 < \gamma \leq 1$ regime, to long-range transitions with probability of transitioning from node i to node j given by $(\tau_\gamma)_{i\to j} \sim d_{i,j}^{-(1+2\gamma)}$, with $d_{i,j}$ defined as the shortest-path distance between nodes i and j. Furthermore, the long-range nature of the process was shown, by numerical investigation, also on more general random and small-world networks.

More generally, the Laplacian belongs to the class of Stieljes matrices and taking its fractional version L^γ, i.e. elevating to the power γ the spectrum, preserves the property of belonging to this class when $0 < \gamma \leq 1$ [11]. Indeed, we then can derive an approximation of the fractional Laplacian matrix

$$\begin{aligned} L^\gamma &= (D-W)^\gamma \\ &= \left(D^{1/2}D^{1/2} - D^{1/2}D^{-1/2}WD^{-1/2}D^{1/2}\right)^\gamma \\ &= D^{\gamma/2}\left(I - D^{-1/2}WD^{-1/2}\right)^\gamma D^{\gamma/2} \\ &= D^{\gamma/2}\Big[I - \gamma D^{-1/2}WD^{-1/2} \\ &\quad + \frac{\gamma(\gamma-1)}{2}(-D^{-1/2}WD^{-1/2})^2 \\ &\quad - \frac{\gamma(\gamma-1)(\gamma-2)}{6}(D^{-1/2}WD^{-1/2})^3 + \cdots \Big] D^{\gamma/2}, \end{aligned} \tag{8}$$

which implies that, in the range $0 < \gamma \leq 1$, the infinite sum of negative terms entail non-positive off-diagonal elements in L^γ that can be associated to a stochastic adjacency matrix.

$\diamond$ **$\tilde{\mathbf{L}} \equiv \tilde{\mathbf{D}} - \tilde{\mathbf{W}}$ - Laplacian from Biased RW**

The principle behind *biased random walks*, is to tailor the transition probabilities according to some node property, like the degree. This approach effectively introduces a bias in the way the walk is performed [12] and, this way, it is possible to "force" long-range transitions : this approach was explored

in [13] and it relies on the construction of a transition matrix informed by the *geodesic distances* between nodes [13, 14].

3 Fractional G-SSL

We now exploit the random walk interpretation of the G-SSL by amending the optimization formulation to attain more efficient label propagation processes. We proceed by replacing the L operator with its generalization $L^\gamma = Q^T \Lambda^\gamma Q$ in the functional (1), where $\Lambda^\gamma = diag(\lambda_0^\gamma, \ldots, \lambda_{N-1}^\gamma)$. We also will have, in the following, its decomposition in diagonal and off-diagonal components, which respectively corresponds to the generalized D_γ matrix and a new weighted adjacency matrix $-(W_\gamma)_{ij} = (L^\gamma)_{ij}$ with $i \neq j$. We observe that, in order to recast (1) in a proper RW, we need to have some consistent diagonal matrix D_γ in the fitting term. We recall that when we revert to the class of L^γ operators, the matrix D_γ, which originally was the degree matrix, now has to be generalized to $(D_\gamma)_{ii} = (L^\gamma)_{ii}$. Therefore, G-SSL with the generalized Laplacian L^γ is the minimization of the new functional $S(F)$

$$S(F) = 2F^T D_\gamma^{\sigma-1} L^\gamma D_\gamma^{\sigma-1} F + \mu \left(F - Y\right)^T D_\gamma^{2\sigma-1} \left(F - Y\right). \tag{9}$$

We refer the reader to Sec. III-A of [15] for a proof that, for any $\gamma > 0$, the Fractional G-SSL formulation remains a convex optimization problem. Since $S(F)$ is convex, we can proceed further and apply the first optimality condition $\partial_F S(F) = 0$ in order to obtain the F functions.

Proof : The first optimality condition reads

$$2F^T D_\gamma^{\sigma-1} \left(L^\gamma + (L^\gamma)^T\right) D_\gamma^{\sigma-1} + 2\mu \left(F - Y\right)^T D_\gamma^{2\sigma-1} = 0$$

Multiplying on the right hand side the above equation by $D_\gamma^{1-2\sigma}$

$$2F^T D_\gamma^{\sigma-1} \left(L^\gamma + (L^\gamma)^T\right) D_\gamma^{-\sigma} + 2\mu \left(F - Y\right)^T = 0\,. \tag{10}$$

Thus, substituting $L^\gamma = D_\gamma - W_\gamma$ into the previous equation

$$F^T D_\gamma^{\sigma} \left(2I - D_\gamma^{-1}\left(W_\gamma + (W_\gamma)^T\right) + \mu I\right) D_\gamma^{-\sigma} - \mu Y^T = 0.$$

Since W^γ is symmetric we finally arrive to

$$F^T D_\gamma^{\sigma} \left(2I - 2D_\gamma^{-1} W_\gamma + \mu I\right) D_\gamma^{-\sigma} - \mu Y^T = 0.$$

Therefore, we can conclude that the classification function with the generalized standard Laplacian L^γ takes the form

$$F^T = (1-\alpha) Y^T D_\gamma^{\sigma} \left(I - \alpha D_\gamma^{-1} W_\gamma\right)^{-1} D_\gamma^{-\sigma}, \tag{11}$$

corresponding to the generalization of the solution for F in (2). As before we can thus consider the two cases of SL and PR, obtaining for F :

$\sigma = 1$ - Fractional Standard Laplacian (FSL) :

$$F^T = (1-\alpha) Y^T D_\gamma \left(I - \alpha D_\gamma^{-1} W_\gamma\right)^{-1} D_\gamma^{-1}. \tag{12}$$

$\sigma = 0$ - Fractional PageRank (FPR) :

$$F^T = (1-\alpha) Y^T \left(I - \alpha D_\gamma^{-1} W_\gamma\right)^{-1}. \tag{13}$$

It clearly appears from the fractional solutions (12) and (13) the formal symmetry with (4) and (6) : indeed the labels' attribution is, as before, the result of diffusion on the labels driven by a new generalized "fractional transition matrix" $D_\gamma^{-1} W_\gamma$ and restarted with probability $p_r = 1 - \alpha$.

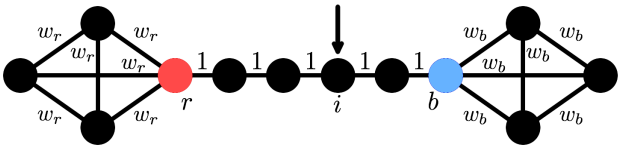

FIGURE 1 – Lollipop graph. Two class problem with the labeled points colored in red and blue. The node under study is pointed by the arrow.

We further extend this complete symmetry between the results in Sec. 2 by deriving the classification rule for the fractional G-SSL that reads as follows : node i is classified to class k by the fractional G-SSL (9) if

$$\sum_{p \in V_k} (d_\gamma)_p^\sigma q_{pi} > \sum_{s \in V_k'} (d_\gamma)_s^\sigma q_{si}, \ \forall k' \neq k. \tag{14}$$

Proof : We recall that F and Y intend the column vectors F_{*k} and Y_{*k}. Let $\text{ppr}(j) = (1-\alpha) e_j^T (I - \alpha D_\gamma^{-1} W_\gamma)^{-1}$ denote the personalized PageRank vector of the fractional RW and also observe that $Y_{*k}^T = \sum_{p \in V_k} e_p^T$ and $F_{ik} = F_{*k}^T e_i$. We replace these in (11) to obtain

$$F_{ik} = \frac{1}{(d_\gamma)_i^\sigma} \sum_{p \in V_k} (d_\gamma)_p^\sigma \text{ppr}_i(p). \tag{15}$$

In [16] it was proved that $\text{ppr}_i(p) = q_{pi} \text{ppr}_i(i)$, thus after replacing it in (15) we conclude that data point i is classified to class k if

$$F_{ik} - F_{ik'} \propto \left(\sum_{p \in V_k} (d_\gamma)_p^\sigma q_{pi} - \sum_{s \in V_k'} (d_\gamma)_s^\sigma q_{si}\right) > 0, \ \forall k \neq k'.$$

4 Numerical Experiments

We perform numerical simulations on the lollipop graph presented in Fig. 1. This simple toy example is meant to illustrate the behavior of (14) and expose how the fractional operator is able to compensate for cases in which the graph presents biases.

Fractional PageRank ($\sigma = 0$) - In this case, we would like to focus only on the dynamical part of (14) , i.e. the probabilities q_{bi} and q_{ri} : we thus revert to the FRP method. For our test we set $w_b > w_r$ and our aim is to have i classified blue since b is in a highly connected region and i is closer to b than to r, thus priming the spatial proximity. From (14), for node i to be classified as blue by PR, we need $q_{bi} > q_{ri}$. However, in our setting, the i node is more likely to become red since the stronger w_b weight attracts walkers starting from b, trapping them in the blue cluster while, on the other hand, walkers from r are freer due to the smaller w_r, so practically we are in a $q_{bi} < q_{ri}$ situation. With FPR, as displayed in Fig. 2a, we can alter this inequality : indeed, for a fixed α, small values of the fractional parameter γ increase q_{bi}, allowing walkers to escape the blue cluster and reach node i more frequently.

Consequently, the classification of i changes as shows the evolution of $\text{sign}(q_{bi} - q_{ri})$ in Fig. 2c : for a proper tuning of the

γ parameter ($\gamma \to 0$), we can bring the walkers from the blue class to avoid the 'sink' and classify node i as blue, regardless of the confidence we give to the labels. We remark that standard PR is also capable of recovering node i as blue, nonetheless this effect appears only in a framework of frequent restarts. Another setup of RW biased by the network structure is the case of unbalanced density of intra-class links. In this setting, the long-range transition properties of FPR seem like a straightforward tool to compensate for this type of unbalancedness.

Fractional Standard Laplacian ($\sigma = 1$) **-** Recalling (14), FSL incorporates the generalized degree of labeled points to the classification decision. In this case, we choose $w_r \gg w_b$ in order to highlight the cohesion of the cluster near to r in contrast to the cluster next to b, where the small w_b is a proxy of a small similarity between the nodes. Therefore, we would like node i to be red because, heuristically, we trust more the cluster near to r. For $\sigma = 1$, which is the present FSL case, we can heavily reduce the influence of the degree with γ, as displayed in Fig.2b and, thus, tweak the classification of i. In this setting of $w_r > w_b$ (Fig.2d), interestingly, standard SL is never able to classify the point as red since a normal RW gets stuck in the cluster adjacent to r, and this prevents the red walkers from attaining node i often enough. On the other hand, the FSL is able to revert the classification outcome, and for a reasonable range of alpha, we can classify node i as red, as we aimed to. We have this way privileged the red class because we interpreted the higher w_r as a measure of confidence and, interestingly, this interpretation is equivalent to having multiple ground truth labels, which might be difficult or expensive to obtain in practice.

5 Conclusion

In this work we presented a generalization of the G-SSL framework originally attached to classical random walk process, to richer diffusion dynamics with *long-range transitions*.

From a theoretical standpoint, those jumps can be ignited on networks by various choices for the Laplacian operator L and here we focused on the Fractional Laplacian L^γ with $0 < \gamma \leq 1$. Thus, we were able to extend the framework for G-SSL, developed in [1, 5, 6], to the fractional case and, in particular, an extension of the criterion for labels attribution was derived.

In the numerical section, we showed that the parameter γ introduces a new degree of freedom that helps circumventing the biases of standard methods such as hubs, which usually act as "sinks" for the labels. Our toy example, which complements the consistency of our theoretical derivation, allows to grasp heuristically the power of our method : this type of rich dynamics gives a tool to balance between the strength of the ground truth labels Y and the necessity of widely diffusing through the graph, avoiding to some extent the attraction of misleading nodes or the effect of unbalanced classes. This flexibility with respect to the structural constraints of the graph can therefore be a key to an improved classification in difficult settings such as when the graph is badly constructed or it is flawed by spurious or missing edges.

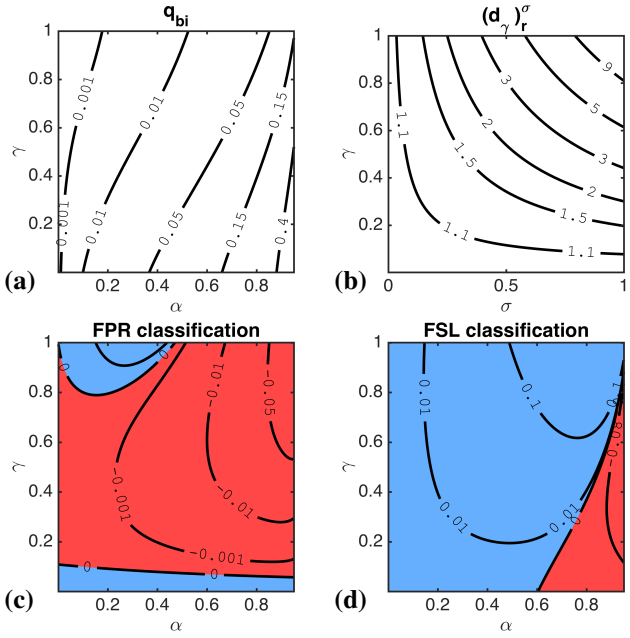

FIGURE 2 – (a) Dependence on the γ and α parameters of the probability q_{bi}. (b) Dependence on the γ and σ parameters of the generalized degree of node r $(d_\gamma)^\sigma_b$. Classification of node i with respect to γ and α for the two fractional methods : FPR (c) and FSL (d). We set the weights as follows : (a) and (c) : $w_r = 1$ and $w_b = 5$ - (b) and (d) : $w_r = 5$ and $w_b = 1$.

Références

[1] K. Avrachenkov, P. Gonçalves, A. Legout, M. Sokol, *Int. Wireless Comm. and Mobile Comp. Conf.* (Cyprus, 2012).

[2] A. Subramanya, J. Bilmes, *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, P. Tang, *IEEE Signal Processing Letters* **21**, 1192 (2014).

[4] F. R. Chung, *Spectral graph theory*, vol. 92 (American Mathematical Soc., 1997).

[5] K. Avrachenkov, P. Gonçalves, A. Mishenin, M. Sokol, *SIAM Data Mining* (2012).

[6] K. Avrachenkov, P. Gonçalves, M. Sokol, *10th WS on Algorithms and Models for the Web Graph* (Harvard U., USA, 2013).

[7] H. T. Ali, R. Couillet, *arXiv preprint :1611.01096* (2016).

[8] 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 (2014).

[9] R. Klages, G. Radons, I. M. Sokolov, *Anomalous transport : foundations and applications* (John Wiley & Sons, 2008).

[10] A. Riascos, J. L. Mateos, *Phys. Rev. E* **90**, 032809 (2014).

[11] C. A. Micchelli, R. Willoughby, *Linear Algebra and its Applications* **23**, 141 (1979).

[12] R. Lambiotte, *et al.*, *Phys. Rev. E* **84**, 017102 (2011).

[13] A. Riascos, J. L. Mateos, *Phys. Rev. E* **86**, 056110 (2012).

[14] E. Estrada, *et al.*, *arXiv preprint :1612.08631* (2016).

[15] S. de Nigris, E. Bautista, P. Abry, K. Avrachenkov, P. Gonçalves, *2017 25th European Signal Processing Conference (EUSIPCO) (EUSIPCO 2017)* (Kos, Greece, 2017).

[16] K. Avrachenkov, N. Litvak, *Stochastic Models* **22**, 319 (2006).