Efficient community detection in sparse networks with non-backtracking random walkers

Candidate Number : 1029205

This manuscript was compiled on March 27, 2019

The task of detecting communities on undirected networks has attracted much attention in the past decades thanks to the progress of computational power. Spectral methods, techniques based on the study of eigenvalues and eigenvectors of particular matrices, represent one possible avenue to solve this problem. However, "classic" processes based on the adjacency matrix and related matrices such as the normalized laplacian or the modularity matrix, fail to perform this task on "sparse" networks, which depict a common type of network. In this paper we illustrate this issue and review a spectral method based on the "non-backtracking" matrix, that is more suited for this type of network, as well as other alternative matrices. We quantify the performance of those methods on artificial generated networks and real data sets.

Community Detection | Spectral Methods | Sparse Networks

Detecting communities is a key challenge in numerous situations, from biology (1) to social science (2), or machine learning optimization (3). Many approaches have been explored in that respect, using statistical inference, graph-cut or modularity optimization (4) for example. Spectral methods focus on the spectrum of particular matrices to uncover the structure of the corresponding graph. Introduced in the 70's, they combine effectiveness and expertise about the network beyond the simple modular structure. They originally involve the Adjacency matrix, the Laplacian matrix or the Modularity matrix for instance. However, these techniques are unable to perform community detection in the particular case, though frequently encountered, of "sparse" networks, whose communities are theoretically identifiable, and identified with other statistically-based methods. Fortunately, this curse is not inherent to the entire class of spectral methods, as a different operator, the non-backtracking operator, slashes this issue.

We first describe the context of the use of spectral methods in community detection, and their limitation in a certain regime of network "sparsity" that we quantify and illustrate. This constraint justifies the introduction of the non-backtracking operator, or Hashimoto matrix, that solves the previously identified issue. Next, we bring in variants of this operator that behave more appropriately in certain given situations. Finally, we compare the performance of the formerly described methods in the case of artificial and real world data sets.

Spectral Methods and their limitation

In the case of graph balanced bi-partitioning, if we consider a graph G, we can assign each node i a label $s_i \in \{-1, 1\}$. The problem can then be formulated as the minimization the graph cut

$$R = \frac{1}{2}\sum_{i,j} A_{ij} - \frac{1}{4}\sum_{i,j}(s_i s_j + 1)A_{ij} = \frac{1}{4}s^T Ls \qquad [1]$$

with L the Laplacian matrix, subject to $\sum_i s_i = 0$. To solve this, we can use a *Relaxation method* in which we minimize $R_x = \frac{1}{4}x^T Lx$ instead, where x can take real values, and is subject to $\mathbf{1}^T x = 0$ and $x^T x = n$. With the Lagrangian multipliers, we can find that

$$Lx = \lambda x \qquad [2]$$

In this case,

$$R_x = \frac{\lambda n}{4} \qquad [3]$$

Since the lowest eigenvector is forbidden by our constraints (it is proportional to $\mathbf{1}$), the optimal solution is proportional to the eigenvector of the second-lowest eigenvalue, the so-called Fiedler vector. Since the components of the vector x are not necessarily integers, taking the labels as the sign of the components gives a good approximation of the original problem.

A generalization of this algorithm, though not formally well justified (5), has been given for a partition of $k \geq 3$ clusters. In this method, one takes the $2^{nd}-$ to $k^{th}-$ lowest eigenvalues of the Laplacian L, and build a $n \times (k-1)$ matrix whose columns are precisely the selected vectors. This matrix stores n vectors in $\mathbb{R}^{k-1}$, on which we perform a clustering algorithm such as k-means. We then assign each node of the graph to the cluster in which the corresponding vector has been assigned.

In the case of "dense" networks (theoretically speaking, in the asymptotic network in which the average degree of a node is proportional to n), these spectral methods are reliable. However, they suffer from some drawbacks in the "sparse" case, where the average degree of nodes is constant over the size of the graph. Unfortunately, this type of network is often encountered in real life problems, see Table II in (6). We will use the stochastic block model to illustrate those deficiencies.

The stochastic block model (SBM) is a generalization of the classic Erdős-Rényi (ER) random graph model. The ER model presents two parameters n and p that defines respectively the size of the network and the probability of an edge between

Significance Statement

Network clustering is primordial in a wide range of applications. We review spectral methods, a particular approach that exploits the spectrum of certain matrices of the graph to solve this task, and introduce a "non-backtracking" operator that performs efficiently on a common class of "sparse" graphs where classic operators based on the adjacency matrix fail to detect communities. Variants of this operator, the Flow and Reluctant matrices, balance the rationale behind the non-backtracking matrix and certain of its drawbacks. Numerical performance analyses give further insight and credit to these methods.

two nodes. The SBM allows a block structure and provides an easy way to generate networks with specified inter and intra-community probabilities. Formally, the SBM is defined by n, the number of nodes, r, the number of communities, the partition of the nodes in r communities, and the matrix $[P_{ij}]_{ij} = p_{ij}$ corresponding to the probability of an edge between a node in community i and a node in community j.
The "sparsity" of the network can be translated to $p_{ii} = \frac{c_{ij}}{n}$ in which the degrees of a node does not increase with n. In the simplified case where there are only two balanced communities with equal average degrees (in the case of different average degrees, a study of the degree distribution of the nodes clearly separates two modes. These modes define a strategy of separation of the two communities), we have

$$P = \begin{bmatrix} \frac{c_{in}}{n} & \frac{c_{out}}{n} \\ \frac{c_{out}}{n} & \frac{c_{in}}{n} \end{bmatrix} \qquad [4]$$

We can show (7) that the adjacency matrix, which is used in spectral method for clustering, has an asymptotic spectrum defined by a continuous part, where the density of the eigenvalues is given by $\rho(\lambda) = \frac{\sqrt{4c-\lambda^2}}{2\pi c}$, the Wigner semi-circle law (8), and a discrete part, where the largest engenvalue is bounded by the maximum degree, and the second eigenvalue is given by

$$\lambda_c = \frac{c_{in} - c_{out}}{2} + \frac{c_{in} + c_{out}}{c_{in} - c_{out}} \qquad [5]$$

The eigenvector linked to this eigenvalue is informative about the community (7). The density previously formulated is the density of the eigenvalues of the matrix representing the "noise" in the stochastic block model, or the spectrum of the matrix defined by the difference between the adjacency matrix and the expected adjacency matrix (7). This "bulk" of noisy eigenvalues is included in $[-2\sqrt{c}, 2\sqrt{c}]$. Hence, by setting $\lambda_c = 2\sqrt{c}$, which is equivalent to

$$c_{in} - c_{out} = \sqrt{2(c_{in} + c_{out})} \qquad [6]$$

we have an eigenvalue indistinguishable from the part of the spectrum induced by randomness. Thus, a spectral method cannot identify the eigenvector encapsulating the information about the communities, even if the communities exist ($c_{in} - c_{out} > 0$). This threshold becomes $\sqrt{q(c_{in} + (q-1)c_{out})}$, in the case of multiple communities. It has been proven that this impossibility to detect communities under this threshold applies, in fact, to any detection algorithm (9). However, above this threshold, there is a regime in which the spectral method fails, but the other methods work. This is due to the localization of high degree nodes, which generate eigenvalues that can shadow the eigenvalues linked to community information. The eigenvectors corresponding to those eigenvalues are uninformative about the community structure, but localize high degree nodes. We illustrated this phenomenon in Fig 1.

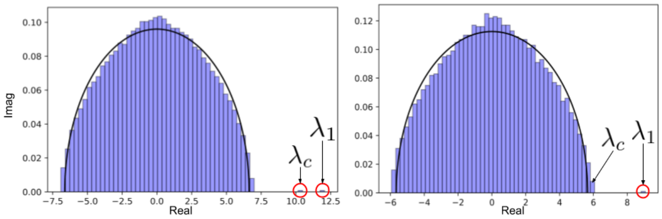

Fig. 1. Density plot of the spectrum of the adjacency matrix of a graph with $n = 4000$ nodes, drawn from a Stochastic Block Model with two balanced communities. (**Left**) $c_{in} = 20$ and $c_{out} = 2$. (**Right**) $c_{in} = 11$ and $c_{out} = 5$. In the first case, λ_c is clearly separated from the "bulk", whereas, in the second case, λ_c is out of the asymptotic distribution of the noisy eigenvalues, but is shadowed by eigenvalues linked to high degree nodes (compared to their expected value). In both case the bound
$c_{in} - c_{out} > \sqrt{2(c_{in} + c_{out})}$ is verified.

Spectral Redemption

Hence there is a regime in which a community detection is possible, but the "classic" spectral methods, that make use of the adjacency matrix or a transformation of it, fail to fulfill this task. A large number of schemes have been introduced to cope with this issue, such as removing high degree nodes (10), or add constants to the matrix, but they cause a loss of information about the communities. A new operator called the "Non-Backtracking" operator (11) can overcome this difficulty. This operator can be though as the study of a random walker that is forbidden to return to its direct previous state. This method works because the limitations of the previous designs concerned the influence of high degree nodes in sparse networks that biased the spectrum of the adjacency matrix and made impossible the identification of the eigenvectors that encapsulate the community structure. In the case of a non-backtracking random walker, the agent cannot return back to its previous state, which weaken the influence of those outliers. This random walker is no longer a Markov chain over the space of nodes, since the probability to reach a particular node depends on the two previous visited nodes, thus we cannot formulate a simple transition matrix over this space. However, it becomes back a Markov chain in the space of directed edges, in which we can formulate transition matrices and apply the same resolution pipeline. Hence, in the case of non-backtracking random walks, we introduce the non-backtracking matrix, or the Hashimoto matrix (11), on the $2m \times 2m$ space of induced directed edges on the undirected graph as:

$$B_{(u \to v),(w \to x)} = \begin{cases} 1 & \text{if } v = w \text{ and } u \neq x \\ 0 & \text{otherwise} \end{cases} \qquad [7]$$

This matrix has interesting properties on its spectrum, that make it convenient for the development of spectral methods. We will develop the main results linked to its spectrum, and particularly in the case of the Erdős-Rényi and Stochastic block model (12).

The first important theorem attributable to Hashimoto (13), is that the spectrum of the non-backtracking matrix is embedded in the Ihara zeta function of a graph, function defined in the corresponding paper.

Theorem 1 (Spectrum of the non-backtracking matrix) *Let B denotes the non-backtracking matrix of a graph G, and $\zeta_G(z)$ its Ihara zeta function. Then the following identity holds:*

$$det(I - zB) = \frac{1}{\zeta_G(z)}$$

Hence, studying the spectrum of B is equivalent to localizing the poles of the Ihara zeta function of random graphs. From this, one can show the following theorems in the case of the Erdős-Rényi and Stochastic block model (12).

Theorem 2 (Spectrum in Erdős-Rényi model) *Let G be an Erdős-Rényi graph with parameters $(n, \frac{\alpha}{n})$ for some fixed parameter $\alpha > 1$. Then, with probability tending to 1 as $n \to \infty$, the eigenvalues $\lambda_i(B)$ of its non-backtracking matrix B satisfy*

$$\lambda_1(B) = \alpha + o(1) \text{ and } |\lambda_2(B)| \leq \sqrt{\alpha} + o(1)$$

Hence, the spectrum of a "sparse" graph coming from the Erdős-Rényi model tends to be confined in a ball of radius $r = \sqrt{\alpha}$ where α is the average degree of the nodes of the graph, except for one real eigenvalue, that account for the degree of the nodes, which tends to α.

In the case of the stochastic block model with r communities, we need to introduce the expected adjacency matrix $\tilde{A} = \mathbb{E}(A)$ where A is drawn from the SBM. This matrix has exactly the same eigenvalues as the matrix of probability defining the probability of edges between classes. We denote them :

$$|\mu_r| \leq ... \leq |\mu_2| \leq \mu_1$$

μ_1 is the eigenvalue given by the Perron-Frobenius theorem, since all the elements in $\tilde{A}$ are positive. We then define r_0 as following :

$$\mu_k^2 > \mu_1 \text{ for all } k \in [r_0] \text{ and } \mu_{r_0+1}^2 \leq \mu_1$$

Theorem 3 (Spectrum in the Stochastic Block Model) *Under the assumption that each vertex type has the same asymptotic average degree α, and two other assumptions described in (12), if we let G be a graph drawn from the SBM, then, with probability tending to 1 as $n \to \infty$*

$$\begin{cases} \lambda_k(B) = \mu_k + o(1) & \text{for } k \in [r_0] \\ |\lambda_k(B)| \leq \sqrt{\alpha} + o(1) & \text{for } k > r_0 \end{cases}$$

Hence, from this theorem, we can deduce that there is a threshold r_0 separating "visible" eigenvalues from "invisible" ones in the expected adjacency matrix, according to their comparison to the square root of the Perron-Frobenius eigenvalue. This induces asymptotic properties on the non-backtracking matrix, namely that the r_0 first eigenvalues are equals to the eigenvalues of the expected adjacency matrix, and that the others are confined in a "bulk" defined by a circle of radius $\sqrt{c}$ where c is the average degree of a node.

In the case of two communities of equal size in the stochastic block model studied previously, with the block structure elucidate in Eq 4 and $c_{in} > c_{out}$, we have two eigenvalues, $\mu_1 = \frac{c_{in}+c_{out}}{2}$ which is the average degree, and $\mu_2 = \frac{c_{in}-c_{out}}{2}$. By Theorem 3, $\mu_2^2 > \mu_1 \iff (c_{in} - c_{out}) > \sqrt{2(c_{in} + c_{out})}$. In this situation, the second eigenvalue of B is asymptotically equal to μ_2, which is out of the "bulk", and the other eigenvalues of B are asymptotically inside. Hence, there is a phenomena of "Spectral redemption" (14), in the sense that the regime above the threshold where the classical spectral methods failed to detect communities no longer exists with community detection based on the non-backtracking matrix, at least asymptotically.

Now, we justify the use of a spectral method for the non-backtracking matrix, in the case of bi-partitioning. The approach used in (14) is to introduce a vector that is both correlated to the communities, and near the eigenvector of μ_c.

We introduce

$$g^{(r)}_{u\to v} = \mu_c^{-r} \sum_{(w,x):d(u\to v,w\to x)=r} \sigma_x \qquad [8]$$

where r is an integer, $\sigma_x \in \{-1, +1\}$ is the label of x, and $d(.,.)$ represents the distance between edges in terms of number of steps required to go to one edge to the other.
We can easily interpret this vector. In the asymptotic case of a tree-like structure, we can manually compute, with elementary induction and probabilities, that

$$\mathbb{E}[\sigma_x]_{d(u,x)=r} = \sigma_u \left[\frac{c_{in-c_{out}}}{c_{in+c_{out}}}\right]^r$$

Hence,

$$g^{(r)}_{u\to v} \simeq \sigma_u \left(\frac{2 \times c}{c_{in} + c_{out}}\right)^r = \sigma_u$$

More rigorously, in the case where the communities are distinguishable, it is established that $\left\langle g^r_{u\to v}, \sigma_u \right\rangle$ is bounded away from zero as $n \to \infty$ (14).
Now that we have a vector asymptotically correlated to the communities, we have to show that this vector is, in fact, close to the identifiable eigenvector that we can extract from the spectrum of B. We have:

$$(Bg^{(r)})_{u\to v} = \sum_{x\in V(v)/\{u\}} g^{(r)}_{v\to x} \qquad [9]$$

In the case where the graph has a tree-like structure around u, in a radius of $r + 1$, which is true asymptotically, then the σ_x arising from the previous sum are exactly coming from the vertices at the end of the edges $r + 1$ steps from u, hence :

$$(Bg^{(r)})_{u\to v} = \mu_c^{-r} \sum_{(w,x):d(u\to v,w\to x)=r+1} \sigma_x = \mu_c g^{(r+1)}_{u\to v} \qquad [10]$$

Hence, $g^{(r)}$ is "almost" the eigenvector associated to μ_c, in the sense that $g^{(r)} \neq g^{(r+1)}$, but, they are close with high probability. Indeed :

$$g^{(r)}_{u\to v} - g^{(r+1)}_{u\to v} = \mu_c^{-r} \sum_{(w,x):d(u\to v,w\to x)=r} \left[\sigma_x - \mu_c^{-1} \sum_{y\in N(x)/\{w\}} \sigma_y\right] \qquad [11]$$

There is c^r terms on average in this sum, and we have directly that

$$\mathbb{E}_{\sigma_x}[\sigma_y] = \frac{c_{in}}{c_{in} + c_{out}} \sigma_x - \frac{c_{out}}{c_{in} + c_{out}} \sigma_x \qquad [12]$$

Hence the elements of the sum have mean zero and finite variance, which implies that

$$\mathbb{E}[(g^{(r)}_{u\to v} - g^{(r+1)}_{u\to v})^2] = \mathcal{O}(c^r \mu_c^{-2r}) \qquad [13]$$

And, summing over the edges,

$$\mathbb{E}[(g^{(r)} - g^{(r+1)})^2] = \mathcal{O}(c^r \mu_c^{-2r} |E|) \qquad [14]$$

With the detectability condition, $c < \mu_c^2$, the error term tends to zero as r grows. 10 becomes, since $|g^{(r)}|$ is bounded,

$$|Bg^{(r)} - \mu_c g^r| = o(1)|g^{(r)}| \qquad [15]$$

we can see, from the definition of $g^{(r)}$, that in the tree-like approximation the elements of $g^{(r)}_{u\to v}$ are constant over u. Hence if we sum them, we obtain the sum of the labels at r distance from v, multiplied by the number of incoming edges. This sum has, on average, the same sign as the label of v. From this study, a good approximation of the label of a vertex would be the sign of the sum over the incoming edges of the elements of the eigenvectors associated with μ_c elicited with a spectral decomposition.

For computation interest, one can also show (11) that the $2n \times 2n$ reduced matrix :

$$B' = \begin{bmatrix} 0 & D - \mathbb{1} \\ -\mathbb{1} & A \end{bmatrix}$$

has $2n$ eigenvalues shared with the non-backtracking matrix, and that the $2(m-n)$ remaining eigenvalues are either 1 or -1. The eigenvectors corresponding to the eigenvalues are defined by :

$$Bg = \mu g \implies B' \begin{pmatrix} g^{in} \\ g^{out} \end{pmatrix} = \mu \begin{pmatrix} g^{in} \\ g^{out} \end{pmatrix} \qquad [16]$$

with $g_u^{out} = \sum_{v \in N(u)} g_{u \to v}$ and $g_u^{in} = \sum_{v \in N(u)} g_{v \to u}$. The litterature (11) finds that g^{in} and g^{out} are inverted in the previous equation, but a manual computation shows the reverse:

$$(Bg)_{u \to v} = \sum_{x \in \mathcal{N}(v)/\{u\}} g_{v \to x}$$

$$\begin{aligned}(Bg)_u^{out} &= \sum_{v \in \mathcal{N}(u)} (Bg)_{u \to v} = \sum_{v \in \mathcal{N}(u)} \sum_{x \in \mathcal{N}(v)/\{u\}} g_{v \to x} \\ &= \sum_{v \in \mathcal{N}(u)} g_v^{out} - g_{v \to u} = (\sum_{v \in \mathcal{N}(u)} g_v^{out}) - g_u^{in}\end{aligned}$$

$$\begin{aligned}(Bg)_u^{in} &= \sum_{v \in \mathcal{N}(u)} (Bg)_{v \to u} = \sum_{v \in \mathcal{N}(u)} \sum_{x \in \mathcal{N}(u)/\{v\}} g_{u \to x} \\ &= (d_u - 1) \sum_{x \in \mathcal{N}(u)} g_{u \to x} = (d_u - 1) g_u^{out}\end{aligned}$$

Hence, in the community detection method, one can use B' instead of B, with a great economy in complexity. The inferred label are then given by the n first components of the eigenvectors. In Fig 2 we plotted the spectrum of a graph drawn from the same SBM as in Fig 1.

In the case of $q > 2$ communities, a similar generalization as in the "classic" spectral methods can be expressed. There is $q-1$ independent vectors after the leading eigenvector with eigenvalue c. And we can, again, perform a clustering algorithm on the formed vectors in $\mathbb{R}^{q-1}$. However, in this case, there is an additional condition to ensure that all the eigenvectors linked to communities are identifiable in the spectrum:

$$|c_{in} - c_{out}| > q\sqrt{c} \qquad [17]$$

defining an "easily detectable" threshold (11).

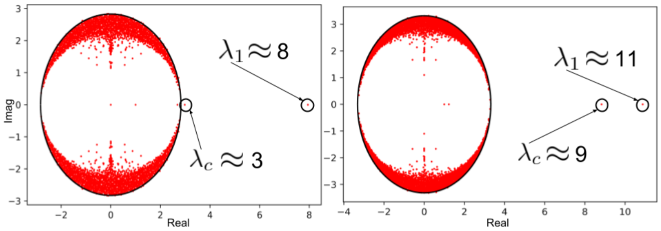

Fig. 2. Spectrum of the non-backtracking matrix of a graph with $n = 4000$ nodes, drawn from a Stochastic Block Model with two balanced communities. (**Right**) $c_{in} = 20$ and $c_{out} = 2$. (**Left**) $c_{in} = 11$ and $c_{out} = 5$. With the non-backtracking matrix, λ_c on the left figure is not "blurred" by interference eigenvalues.

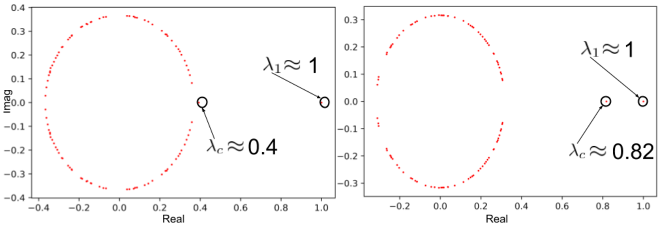

Fig. 3. 500 largest eigenvalues in the spectrum of the Flow matrix of a graph with $n = 4000$ nodes, drawn from a Stochastic Block Model with two balanced communities. (**Right**) $c_{in} = 20$ and $c_{out} = 2$. (**Left**) $c_{in} = 11$ and $c_{out} = 5$. With the non-backtracking matrix, λ_c on the left figure is not "blurred" by interference eigenvalues.

Alternative operator

Flow matrix. We have seen the limitation of classic spectral methods using the adjacency matrix in order to infer communities from its spectrum. We have also seen that the non-backtracking matrix filled the gap. An alternative matrix has been introduced by (15), the flow matrix, similar to the non-backtracking matrix but behave more appropriately to sparse networks that displays high degree nodes. The Flow matrix F is defined by :

$$F_{(u \to v),(w \to x)} = \frac{B_{(u \to v),(w \to x)}}{d_v - 1} \qquad [18]$$

This matrix is similar to the non-backtracking matrix, but weigh inversely proportionally the pair of edges by the degree of the nodes they traverse. To justify the formulation of a spectral algorithm on this matrix, we can easily show, as for the bi-partition problem for the spectral method based on the adjacency matrix, that :

$$\mathbf{u^T}(\mathbf{F} - \mathbf{1}\mathbf{1^T})\mathbf{v} = Q \qquad [19]$$

where $u_{(i \to j)} = v_{(i \to j)} = s_j$ and $\mathbf{1} = (1,..,1)/\sqrt{2m}$ the normalized unit vector. And

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{i,j} - \frac{d_i d_j}{2m} \right] \delta_{g_i g_j} \qquad [20]$$

is the modularity of a given partition of the networks.

In the bi-partition problem, we find that the leading eigenvector of $(\mathbf{F} - \mathbf{1}\mathbf{1^T})$ maximizes the modularity and is non-trivial, and we can apply the same algorithm as previously elucidated with the non-backtracking matrix. One can notice (15) that studying the leading eigenvector of $(\mathbf{F} - \mathbf{1}\mathbf{1^T})$ is equivalent to studying the second largest eigenvector of $\mathbf{F}$, which is more convenient here for comparing the spectrum of the different matrices. In the same example as before, we plotted the spectrum of the Flow matrix in the case of the stochastic block model. As shown in (15), the eigenvalues of the "bulk" lie within $\sqrt{\frac{c/(c-1)}{c}}$.The results are given in Fig 3.

Reluctant-backtracking. A limitation of the fundamental constituents of the non-backtracking operators, is that they are based on a non-backtracking random walk that cannot return back to its previous state. Hence, a non-backtracking random walker is typically stuck in hanging trees and leaves. As a result those structures are ignored in the spectrum of the non-backtracking matrix whereas they can be candidate for communities. To cope with this issue, we can introduce a "reluctant backtracking" operator (16), defining a random walker that can return back to its previous state with a small probability. This matrix allows a good trade between the noise of high degree nodes in classic random walker and the total deletion of leaves and hanging trees in the case of the non-backtracking matrix in community detection. Following the previous definition of the non-backtracking matrix and its normalized version with the Flow matrix, the reluctant backtracking operators are defined as such :

$$R_{(u\to v),(w\to x)} = B_{(u\to v),(w\to x)} + \delta_{vw}\delta ux\frac{1}{d_u} \qquad [21]$$

$$P_{(u\to v),(w\to x)} = \frac{R_{(u\to v),(w\to x)}}{d_v - 1 + \frac{1}{d_u}} \qquad [22]$$

$\mathbf{R}$ is equivalent to the non-backtracking matrix, with a small term in the case of two identical edges with opposite direction, where the corresponding term in the matrix is equal to the inverse of the degree of the source. This value discourages a random walker to return back to a high degree node. $\mathbf{P}$ is a normalized version of this matrix, similar to the Flow matrix with the non-backtracking matrix.

In the case of a bi-partitioning:

$$Q = \frac{1}{2m}\mathbf{u}^{\mathbf{T}}(\mathbf{P} - \mathbf{1}\mathbf{1}^{\mathbf{T}})\mathbf{v} \qquad [23]$$

where $u_{(i\to j)} = v_{(i\to j)} = s_j$, with $s_j \in \{-1, +1\}$ the labels of the nodes in the network.

As a result we can use either the leading eigenvector of $(\mathbf{P} - \mathbf{1}\mathbf{1}^{\mathbf{T}})$ or the second leading eigenvectors of $\mathbf{P}$ in order to apply the same algorithm elucidated earlier. Again, we plotted the spectrum of the normalized reluctant matrices in the same configuration as for the previous spectra in Fig 4.

One important difference with the non-backtracking matrix is that, while performing community detection with the reluctant matrix, we approximate the solution of modularity maximization by setting:

$$s_i = sgn(\sum_j v_{i\to j}) \qquad [24]$$

The label of a node is approximated with the sum of the elements of the eigenvector corresponding to the outgoing edges.

Results

First, we justify the introduction of the Reluctant matrix in the context of community detection where the presence of hanging trees intervenes in the community detection. To do this, we formed binary trees of a given depth, constructed by induction, and we computed the overlap between the partition constituted of the two trees, and the partition given by the spectral method applied to the non-backtracking matrix and the Reluctant matrix. The results are given in Fig 5. In this particular case, the Reluctant matrix find perfectly the two communities. In the case of the non-backtracking matrix, there is no second real eigenvalues and the partition is random.

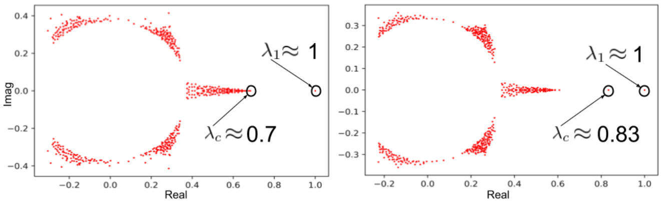

Fig. 4. 500 largest eigenvalues of the spectrum of the normalized reluctant matrix of a graph with $n = 4000$ nodes, drawn from a Stochastic Block Model with two balanced communities. (**Right**) $c_{in} = 20$ and $c_{out} = 2$. (**Left**) $c_{in} = 11$ and $c_{out} = 5$. With the non-backtracking matrix, λ_c on the left figure is not "blurred" by interference eigenvalues.

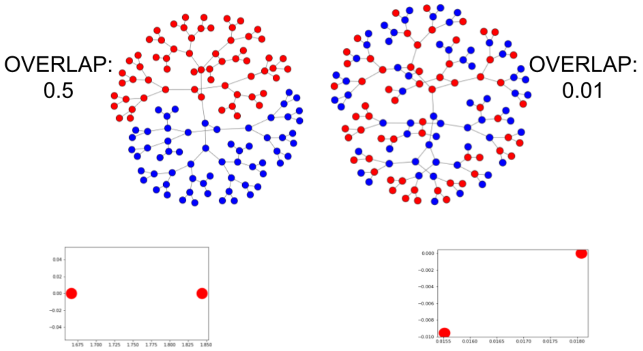

Fig. 5. Comparison of the bi-partition found in a graph composed of two concatenated binary trees, with the non-backtracking matrix and the Reluctant matrix, together with the eigenvalues of the respective matrices.

We also compared all the operators previously introduced by their performance at recovering the partitions of the SBM with two communities, the plot is displayed in Fig. 7. We can clearly identify the "sparse" regime in which the adjacency matrix performs poorly despite the identifiability of the two communities, better recovered by the other matrices. We can also notice that the Flow matrix performs poorly as well compared to the non-backtracking matrix, at least in the regime where the two communities begin to be identifiable. The Reluctant and Normalized Reluctant matrices perform similarly and are efficient.

Finally, we applied the operators to real-world data sets, first the Zachary's karate club data set (17), and then the political book network of Krebs (unpublished). We plotted the resulting graphs, with nodes colored according to their inferred community, as well as the spectrum of the matrices of the operators.

Discussion

We have reviewed spectral methods in the context of graph partitioning and showed their limitation in a special regime of sparsity in graphs. In this regime, the interesting parts of the spectrum of classic matrices such as the adjacency matrix are shadowed by noisy eigenvalues. We illustrated this with the Stochastic Block Model. Fortunately the non-backtracking matrix has a spectrum that behave more conveniently and overcomes this difficulty. We presented the Flow matrix, that better separates clusters in the case of degree distributed

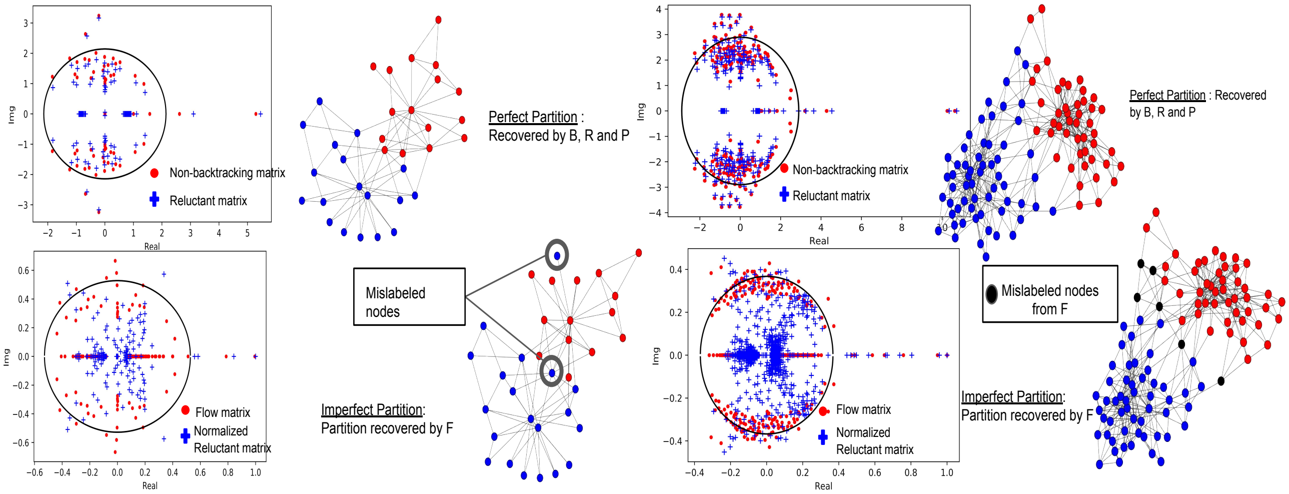

Fig. 6. Application of spectral methods on real-world data sets. **Left** : Karate Club network, the operators B, R, P perfectly recover the communities elucidated in the original paper, F mislabels two nodes. **Right** : Political Book network, the operators B, R, P recover the same communities, F mislabels six nodes compared to the communities inferred by the other operators

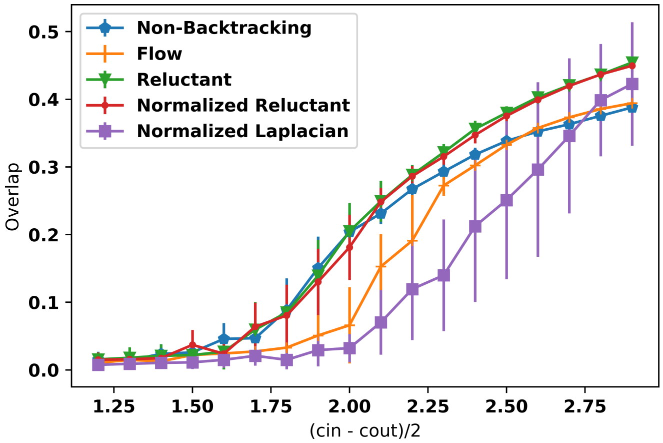

Fig. 7. Overlap of the recovered communities in the SBM with two communities in function of $(c_{in} - c_{out})/2$, and for different matrices. Here we have fixed $c = 3$. Each point is computed with 20 networks of size $n = 4000$.

with "fat tail" distribution, where there are high degree nodes. Finally we presented the Reluctant matrix, that aims at preserving information contained in hanging trees. We applied those operators to artificial and real word networks to illustrate their benefits.

Multiple difficulties remain to be solved. First, we have not extended our analysis to the case of multi-clustering. Moreover such task is not explicitly handled in the case of the Reluctant matrix. Second, we have seen that in the context of clustering, we can replace the non-backtracking matrix with a reduced version of it with a dimension near the number of nodes. While this brings huge computational saving, no such matrix has been explicited for the Flow matrix and the Reluctant matrix. Finally, we notice that the Flow matrix perform much less than the non-backtracking matrix, a deeper understanding of this difference could be advantageous.

Materials and Methods

Overlap. The overlap is a measure initially introduced in physics to measure the relevance of inferred communities, in the case of the SBM with two communities we have: Overlap $= \frac{1}{n}\sum_{u=1}^{n}\mathbf{1}_{\pi(\hat{\sigma}(u))=\sigma(u)} - Max_{\sigma\in[r]}a_\sigma$ where a_σ is the proportion of nodes in category σ.

Implementation. We used Python, and specifically the classic libraries such as Numpy and Scipy. For the manipulation of graphs, we used networkX, and iGraph for the plots. For computational efficiency, we encoded matrices in scipy sparse matrix representations. This enables us to use the Implicitly Restarted Arnoldi Method to find a limited number of eigenvalues and eigenvectors. In particular, we just compute the one corresponding to the two largest real values. This way of doing allows us to infer communities on large network of 4000 nodes in a few seconds, whereas it takes about 10 minutes when taking the dense matrix and using classic numpy methods for computing the spectrum of matrices.
We used data sets from the website http://konect.uni-koblenz.de/.

1. Girvan M, Newman ME (2002) Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99(12):7821–7826.
2. Pizzuti C (2008) Ga-net: A genetic algorithm for community detection in social networks in *International conference on parallel problem solving from nature.* (Springer), pp. 1081–1090.
3. Wedel M (1990) Ph.D. thesis (Wedel).
4. Newman ME (2013) Spectral methods for community detection and graph partitioning. *Physical Review E* 88(4):042822.
5. Riolo MA, Newman M (2014) First-principles multiway spectral partitioning of graphs. *Journal of Complex Networks* 2(2):121–140.
6. Newman ME (2003) The structure and function of complex networks. *SIAM review* 45(2):167–256.
7. Nadakuditi RR, Newman ME (2012) Graph spectra and the detectability of community structure in networks. *Physical review letters* 108(18):188701.
8. Wigner EP (1958) On the distribution of the roots of certain symmetric matrices. *Ann. Math* 67(2):325–327.
9. Mossel E, Neeman J, Sly A (2015) Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields* 162(3-4):431–461.
10. Coja-Oghlan A (2010) Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing* 19(2):227–284.
11. Krzakala F, et al. (2013) Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences* 110(52):20935–20940.
12. Bordenave C, Lelarge M, Massoulié L (2015) Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science.* (IEEE), pp. 1347–1357.
13. Hashimoto Ki (1989) Zeta functions of finite graphs and representations of p-adic groups in *Automorphic forms and geometry of arithmetic varieties.* (Elsevier), pp. 211–280.
14. Nadakuditi RR, Newman ME (2013) Spectra of random graphs with arbitrary expected degrees. *Physical Review E* 87(1):012803.
15. Newman M (2013) Spectral community detection in sparse networks. *arXiv preprint arXiv:1308.6494.*
16. Singh A, Humphries MD (2015) Finding communities in sparse networks. *Scientific reports* 5:8828.
17. Zachary WW (1977) An information flow model for conflict and fission in small groups. *Journal of anthropological research* 33(4):452–473.