Dario Fasino; Francesco Tudisco

Localization of dominant eigenpairs and planted communities by means of Frobenius inner products

Czechoslovak Mathematical Journal, Vol. 66 (2016), No. 3, 881–893

Persistent URL: `http://dml.cz/dmlcz/145877`

Czechoslovak Mathematical Journal, 66 (141) (2016), 881–893

LOCALIZATION OF DOMINANT EIGENPAIRS AND PLANTED COMMUNITIES BY MEANS OF FROBENIUS INNER PRODUCTS

Dario Fasino, Udine, Francesco Tudisco, Saarbrücken

(Received December 30, 2015)

Dedicated to the memory of Professor Miroslav Fiedler

Abstract. We propose a new localization result for the leading eigenvalue and eigenvector of a symmetric matrix A. The result exploits the Frobenius inner product between A and a given rank-one landmark matrix X. Different choices for X may be used, depending on the problem under investigation. In particular, we show that the choice where X is the all-ones matrix allows to estimate the signature of the leading eigenvector of A, generalizing previous results on Perron-Frobenius properties of matrices with some negative entries. As another application we consider the problem of community detection in graphs and networks. The problem is solved by means of modularity-based spectral techniques, following the ideas pioneered by Miroslav Fiedler in mid-'70s.

We show that a suitable choice of X can be used to provide new quality guarantees of those techniques, when the network follows a stochastic block model.

Keywords: dominant eigenpair; cone of matrices; spectral method; community detection

MSC 2010: 15A18, 15B48

1. Introduction

Consider the following result, found in [10]:

Theorem 1.1. *If a symmetric $n \times n$ matrix A satisfies*

$$\mathbf{1}^{\mathrm{T}} A \mathbf{1} \geqslant \sqrt{(n-1)^2+1}\,\|A\|_{\mathrm{F}}, \tag{1.1}$$

then its spectral radius is a simple eigenvalue and the corresponding eigenvector is nonnegative.

The first author has been supported by Istituto Nazionale di Alta Matematica (Italy) and the second author by the ERC grant NOLEPRO.

The symbol $\mathbf{1}$ above stands for a vector whose entries are 1's. This theorem, whose original proof relies upon studying specific convex cones of nonnegative matrices and the solution of certain linear programming problems, shows various intriguing aspects. For example, it reveals that a result which is typical of the Perron-Frobenius theory of nonnegative matrices can be valid for matrices having some negative entries. Moreover, it can establish a localization property of the dominant eigenvector with respect to the central ray of the positive orthant, without involving the spectral separation of the associated eigenvalue. On the other hand, the authors of [10] missed other interesting consequences. For example, in the hypotheses of the aforementioned theorem, the rightmost eigenvalue of A is larger than but quite close to r/n where r denotes the right-hand side of (1.1), as we shall prove in the following.

The purpose of this paper is to shed light on some imaginable generalizations of Theorem 1.1, along with a few more consequences and a possible application in network analysis. It is well known that on many occasions the analysis of graphs and networks interacts profitably with matrix theory, see e.g., [8]. Also in these circumstances we discovered an unexpected contact between these subjects.

The paper is organized as follows. Before concluding this introduction, we collect below some notation and preliminary results. In Section 2, we state, prove and discuss our new localization result for the leading eigenvalue and an associated eigenvector of a symmetric matrix A. The result is given in terms of the angle between A and a given nonzero rank-one matrix $X = xx^{\mathrm{T}}$, measured in terms of the Frobenius inner product between A and X. The generality of the landmark matrix X allows to apply the theorem in various contexts. In Section 3, we observe that the choice $X = \mathbf{1}\mathbf{1}^{\mathrm{T}}$ can be used to obtain an improved version of Theorem 1.1, with an alternative proof and further details on the size of the leading eigenvalue of A, and on the signature of the associated leading eigenvector. In the last section we apply our main results to the analysis of a simple spectral method in community detection. We consider the planted partition model, which is a widespread benchmark for community detection methods, and estimate the fraction of correctly classified nodes that is almost certainly attained in large graphs.

1.1. The Frobenius inner product and matrix norm. The space of $n \times n$ real matrices is naturally endowed with the Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}(AB^{\mathrm{T}})$ and the associate matrix norm $\|A\|_{\mathrm{F}} = \langle A, A\rangle^{1/2}$. Correspondingly, the angle between A and B is defined by the respective cosine as follows:

$$\cos(A, B) = \frac{\mathrm{Tr}(AB^{\mathrm{T}})}{\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}}. \tag{1.2}$$

For any fixed matrix B let $\mathcal{P}_B(A)$ denote the orthogonal projection of a generic matrix A onto the linear space spanned by B. We have the explicit expression $\mathcal{P}_B(A) = \tau(A)B$, where

$$\tau(A) = \frac{\mathrm{Tr}(AB^{\mathrm{T}})}{\|B\|_{\mathrm{F}}^2}.$$

Indeed, consider the decomposition $A = \tau(A)B + Z$. The two terms of that decomposition are orthogonal with respect to the Frobenius inner product. In fact,

$$\langle A - \mathcal{P}_B(A), B\rangle = \langle A, B\rangle - \langle \mathcal{P}_B(A), B\rangle = \langle A, B\rangle - \tau(A)\|B\|_{\mathrm{F}}^2 = 0,$$

and Z is the residual of the projection.

We conclude this introduction by recalling a few useful results. If A is symmetric, we consider its eigenvalues in nonincreasing order, $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \ldots$, and we have the formula $\|A\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \lambda_i(A)^2$. In what follows we will use a special case of the Hoffman-Wielandt theorem, see e.g., [2], Theorem 9.21:

Theorem 1.2. *Let A and B be symmetric $n \times n$ matrices, then*

$$\sum_{i=1}^{n} (\lambda_i(A) - \lambda_i(B))^2 \leqslant \|A - B\|_{\mathrm{F}}^2.$$

Finally, we borrow from [11], Lemma 2, the following result.

Lemma 1.1. *When $B = vv^{\mathrm{T}}$ is a symmetric, positive semidefinite rank-one matrix, equation* (1.2) *becomes*

$$\cos(A, vv^{\mathrm{T}}) = \frac{v^{\mathrm{T}}Av}{v^{\mathrm{T}}v\|A\|_{\mathrm{F}}}. \tag{1.3}$$

In this case we also have $\tau(A) = v^{\mathrm{T}}Av/(v^{\mathrm{T}}v)^2$ and the only nontrivial eigenvalue of $\mathcal{P}_B(A)$ is $v^{\mathrm{T}}Av/(v^{\mathrm{T}}v)$.

2. Localization of a dominant eigenpair

Hereafter we prove that if the angle between a symmetric matrix A and a symmetric rank-one matrix is sufficiently small, then A has one simple, dominant eigenvalue, and the associated spectral projector is close to a multiple of the rank-one matrix. If u and v are vectors, the notation $\cos(u, v)$ has the obvious meaning given by the Euclidean inner product.

Theorem 2.1. *Let A be symmetric and let $X = xx^{\mathrm{T}}$ be a nonzero rank-one matrix. Moreover, let $c = \cos(A, X) > 0$ and $s = \sqrt{1-c^2}$. Then:*

(1) *$\lambda_1(A) \geqslant \mu$, where $\mu = c\|A\|_{\mathrm{F}}$. In particular, if $c > 1/\sqrt{2}$, then $\lambda_1(A)$ is simple and dominant;*

(2) *$|\lambda_1(A) - \mu|/|\lambda_1(A)| \leqslant s$;*

(3) *let v_1 be an eigenvector of $\lambda_1(A)$. If $c^2 \geqslant 1/2$, then $\cos(v_1, x)^2 \geqslant \xi$, where $\xi \in [1/2, 1]$ is the largest root of $c^2 = 2\xi^2 - 2\xi + 1$.*

Proof. First, note the equivalent formulas

$$\mu = \cos(A, X)\|A\|_{\mathrm{F}} = \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x}.$$

Hence, the inequality $\lambda_1(A) \geqslant \mu$ is due to the variational characterization of the largest eigenvalue of a symmetric matrix, see e.g., [2], Theorem 2.30.

Let $Z = A - \mathcal{P}_X(A)$. By simple trigonometry, $\|Z\|_{\mathrm{F}} = s\|A\|_{\mathrm{F}}$. Owing to Lemma 1.1, the matrix $\mathcal{P}_X(A)$ has only one (positive) eigenvalue, which is equal to μ. From Theorem 1.2 we have

$$s^2 \sum_{i=1}^{n} \lambda_i(A)^2 = s^2\|A\|_{\mathrm{F}}^2 = \|Z\|_{\mathrm{F}}^2 = \|A - \mathcal{P}_X(A)\|_{\mathrm{F}}^2 \geqslant (\lambda_1(A) - \mu)^2 + \sum_{i=2}^{n} \lambda_i(A)^2.$$

Consequently, $s^2\lambda_1(A)^2 \geqslant (1-s^2)\sum\limits_{i=2}^{n} \lambda_i(A)^2$ and if $c > 1/\sqrt{2}$, then $1 - s^2 > 1/2$, and we obtain $\lambda_1(A)^2 > \sum\limits_{i\neq 1} \lambda_i(A)^2$, completing the proof of the first claim. We observe in passing that if $c > 1/\sqrt{2}$, then the numbers $\lambda_i(A)^2$ fulfill a reversed polygonal inequality. In fact, Fiedler used to name the relation $2\max\limits_i \alpha_i \leqslant \sum\limits_i \alpha_i$ among nonnegative numbers α_i a polygonal inequality, see e.g. [4].

Rearranging terms we also get

$$(\lambda_1(A) - \mu)^2 \leqslant s^2\lambda_1(A)^2 - (1-s^2)\sum_{i=2}^{n} \lambda_i(A)^2 \leqslant s^2\lambda_1(A)^2,$$

which proves the second claim. Finally, let

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^{\mathrm{T}}$$

be the spectral decomposition of A. For notational simplicity, let $c_i = v_i^{\mathrm{T}} x / \|x\|$ and $\mu_i = \lambda_i / \|A\|_{\mathrm{F}}$. Note that c_i is the cosine of the angle between the vectors v_i and x. With this auxiliary notation we obtain

$$\cos(A, X) = \frac{\sum_{i=1}^{n} \lambda_i (v_i^{\mathrm{T}} x)^2}{x^{\mathrm{T}} x \, \|A\|_{\mathrm{F}}} = \sum_{i=1}^{n} \mu_i c_i^2.$$

Owing to the equations

$$\sum_{i=1}^{n} \mu_i^2 = \sum_{i=1}^{n} c_i^2 = 1$$

and the inequalities $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_n$, we have

$$c = \cos(A, X) \leqslant \mu_1 c_1^2 + \mu_2 (1 - c_1^2) \leqslant \mu_1 c_1^2 + \sqrt{1 - \mu_1^2}\,(1 - c_1^2). \tag{2.1}$$

Arguing by contradiction, suppose $c_1^2 < \xi$. By hypothesis and the first claim, we have $\mu_1^2 \geqslant c^2 \geqslant 1/2$. Consequently $\sqrt{1 - \mu_1^2} \leqslant \mu_1$ and

$$\mu_1 c_1^2 + \sqrt{1 - \mu_1^2}\,(1 - c_1^2) < \mu_1 \xi + \sqrt{1 - \mu_1^2}\,(1 - \xi).$$

Consider the function $f(\mu) = \mu\xi + \sqrt{1 - \mu^2}\,(1 - \xi)$. Simple computations prove that

$$\max_{0 \leqslant \mu \leqslant 1} f(\mu) = \sqrt{2\xi^2 - 2\xi + 1}.$$

Finally,

$$\mu_1 c_1^2 + \sqrt{1 - \mu_1^2}\,(1 - c_1^2) < f(\mu_1) \leqslant \sqrt{2\xi^2 - 2\xi + 1} = c,$$

which contradicts (2.1). Hence we must have $c_1^2 \geqslant \xi$. □

The foregoing theorem can be used as a localization tool for the extreme eigenvalues of A and the corresponding eigenvectors. In fact, the quantity $\cos(A, xx^{\mathrm{T}})$ is maximized when x is an eigenvector associated to the rightmost eigenvalue of A. On the other hand, if $\cos(A, xx^{\mathrm{T}}) < 0$ we obtain immediately a corresponding result for the leftmost eigenvalue, by replacing A with $-A$. It is worth noting that, unlike in Perron-Frobenius theory, Theorem 2.1 makes no assumption about the positivity of A. Furthermore, differently from classical results in spectral perturbation theory, the estimate on $\cos(v_1, x)$ does not depend on the spectral separation of $\lambda_1(A)$.

The following examples show that the hypotheses of Theorem 2.1 cannot be weakened in general. Indeed, let n be an even integer, let $x = (y, y)^{\mathrm{T}}$, where $y \neq 0$ is a vector with $n/2$ entries, and

$$A = \begin{pmatrix} yy^{\mathrm{T}} & O \\ O & yy^{\mathrm{T}} \end{pmatrix},$$

where all blocks have size $n/2$. This matrix has rank two and its two nonzero eigenvalues are equal to $\|y\|_2^2$. Moreover, A fulfills the equality $\cos(A, xx^{\mathrm{T}}) = 1/\sqrt{2}$, thus showing that the strict inequality $c > 1/\sqrt{2}$ in point (1) of Theorem 2.1 can be necessary to have a simple eigenvalue. Furthermore, the vector $v_1 = (y, 0)^{\mathrm{T}}$ is an eigenvector associated to the nontrivial eigenvalue, and we have $\cos(v_1, x)^2 = 1/2$. That value meets the lower bound in point (3) of the theorem. Another counterexample in the same vein is given by

$$A = \begin{pmatrix} O & yy^{\mathrm{T}} \\ yy^{\mathrm{T}} & O \end{pmatrix}.$$

Here, the nonzero eigenvalues are simple but have opposite sign and the same modulus.

3. The signature of a leading eigenvector

In this section we derive Theorem 1.1 as a special case of a much more general result, using arguments very different from those in [10]. Our proof, which is based on Theorem 1.2, extends immediately to eigenvectors having a prescribed minimal number of nonnegative entries.

In what follows we denote by $\mathbf{1}$ the all-ones vector of appropriate size, and we let $J = \mathbf{1}\mathbf{1}^{\mathrm{T}}$ be an all-ones matrix. Moreover, we use the notation $\mathbb{P}_k^n$ to indicate the set of real n-vectors having no more than k (strictly) positive entries, for $0 \leqslant k \leqslant n$. In particular, $\mathbb{P}_n^n = \mathbb{R}^n$; $\mathbb{P}_{n-1}^n$ is the complement of the positive orthant in $\mathbb{R}^n$; and $\mathbb{P}_{k_1}^n \subset \mathbb{P}_{k_2}^n$ for $k_1 < k_2$. In the following lemma we exploit a trick found in [11], Lemma 1.

Lemma 3.1. *For $0 < k \leqslant n$ it holds that*

$$\pi_{k,n} := \max_{x \in \mathbb{P}_k^n} \cos(x, \mathbf{1}) = \sqrt{k/n}.$$

Moreover, we have $\cos(x, \mathbf{1}) = \pi_{k,n}$ if and only if x is any vector with exactly k entries having the same positive value and the remaining $n - k$ entries being zeros.

P r o o f. Let $x \in \mathbb{P}_k^n$ and let m be the number of its positive entries. Let x^+ be the positive m-dimensional vector made by the positive entries of x. By hypothesis $m \leqslant k$, therefore

$$\cos(x, \mathbf{1}) = \frac{x^{\mathrm{T}}\mathbf{1}}{\sqrt{n}\|x\|_2} \leqslant \frac{\|x^+\|_1}{\sqrt{n}\|x^+\|_2} \leqslant \sqrt{k/n}. \tag{3.1}$$

The rightmost inequality follows from the fact that for any vector $v \in \mathbb{R}^m$ one has $\|v\|_1 \leqslant \sqrt{m}\|v\|_2$. The last part of the claim is verified by requiring that both inequalities in (3.1) hold as equalities. □

Remark 3.1. By the preceding lemma, if a vector $x \in \mathbb{R}^n$ fulfills $\cos(x, \mathbf{1}) \geqslant \pi_{k,n}$, then either $\cos(x, \mathbf{1}) = \pi_{k,n}$, whence x is a special nonnegative vector having k positive entries, or $x \notin \mathbb{P}_k^n$, so it has at least $k+1$ positive entries. In both cases x has at least $k+1$ nonnegative entries.

Theorem 3.1. *Let A be a symmetric $n \times n$ matrix such that*

$$\mathbf{1}^{\mathrm{T}}A\mathbf{1} \geqslant \sqrt{(n-k)^2 + k^2}\,\|A\|_{\mathrm{F}} \tag{3.2}$$

for some $1 \leqslant k < n/2$. Then, the spectral radius of A is a simple eigenvalue, and the corresponding eigenvector can be oriented so that it has at least $n-k+1$ nonnegative entries, of which at least $n-k$ are positive.

P r o o f. Let $c = \cos(A, J)$. By hypothesis and (1.3),

$$c = \frac{\mathbf{1}^{\mathrm{T}}A\mathbf{1}}{n\|A\|_{\mathrm{F}}} \geqslant \sqrt{\frac{(n-k)^2 + k^2}{n^2}}.$$

In particular, $c > 1/\sqrt{2}$. From Theorem 2.1 we derive that the rightmost eigenvalue of A is positive, simple and dominant, therefore it coincides with $\varrho(A)$. Moreover, the corresponding eigenvector v_1 has $\cos(v_1, \mathbf{1})^2 \geqslant \xi$, where

$$\xi = \frac{1}{2} + \frac{1}{2}\sqrt{2c^2 - 1} \geqslant \frac{n-k}{n} = \pi_{n-k,n}^2.$$

By Remark 3.1, v_1 must have at least $n-k+1$ nonnegative entries, of which at least $n-k$ are positive, and the proof is complete. □

The hypothesis of the last theorem cannot be weakened, as shown by the following construction. Consider the matrix

$$A = \begin{pmatrix} J & O \\ O & J \end{pmatrix},$$

whose diagonal blocks have order $k \times k$ and $(n-k) \times (n-k)$, respectively. With this matrix the inequality (3.2) is fulfilled as equality. Moreover, an eigenvector corresponding to the largest eigenvalue of A is $(0, \ldots, 0, 1, \ldots, 1)^{\mathrm{T}}$, with exactly $n-k$ positive entries, thus showing optimality of the claim. Finally, we see immediately that Theorem 1.1 is exactly the case $k = 1$ of the foregoing theorem.

Any map $A \mapsto A + \alpha I$ leaves unchanged the eigenvectors of A and translates its eigenvalues without affecting their relative ordering. This fact suggests the next consequence.

Corollary 3.1. *Let A be a symmetric matrix such that for some $\alpha \in \mathbb{R}$ and $1 \leqslant k < n/2$ we have*

$$\mathbf{1}^{\mathrm{T}} A \mathbf{1} \geqslant \sqrt{(n-k)^2 + k^2}\, \|A + \alpha I\|_{\mathrm{F}} - n\alpha.$$

Then the rightmost eigenvalue of A is simple, and the corresponding eigenvector can be oriented so that it has at least $n-k+1$ nonnegative entries.

P r o o f. It is sufficient to apply Theorem 3.1 to the matrix $B = A + \alpha I$. □

A convenient, almost optimal value for α to be used in the last corollary is $\alpha = -\operatorname{Tr}(A)/n$. In fact, with that value the inequality simplifies considerably, as shown below.

Corollary 3.2. *Let μ and σ^2 be the mean and variance of the eigenvalues of A,*

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(A), \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (\lambda_i(A) - \mu)^2.$$

If for some $1 \leqslant k < n/2$ we have

$$\frac{1}{n}\sum_{i \neq j} A_{ij} \geqslant \sigma \sqrt{\frac{(n-k)^2 + k^2}{n}},$$

then the rightmost eigenvalue of A is simple, and the corresponding eigenvector can be oriented so that it has at least $n-k+1$ nonnegative entries.

P r o o f. Recall that $\operatorname{Tr}(A) = \sum_i \lambda_i(A)$ and $\|A\|_{\mathrm{F}}^2 = \sum_i \lambda_i(A)^2$. Now,

$$\begin{aligned}\|A - \mu I\|_{\mathrm{F}}^2 &= \|A\|_{\mathrm{F}}^2 - 2\mu \sum_i A_{ii} + n\mu^2 = \sum_i \lambda_i(A)^2 - 2\mu \sum_i \lambda_i(A) + n\mu^2 \\ &= \sum_i (\lambda_i(A) - \mu)^2 = n\sigma^2.\end{aligned}$$

The claim follows by rearranging terms and letting $\alpha = -\mu$ in the preceding corollary.
□

The stochastic block model, also referred to as the planted partition model, is a popular generative model for random graphs having a prescribed clustering structure, see e.g., [5], [9]. It is often considered as a benchmark in the context of graph clustering and community detection problems in networks.

Given a set of nodes V, the model in its most general form assumes that a partition $\{C_1, \dots, C_k\}$ of V is given, together with a set of connection probabilities p_{ij}, for $1 \leqslant i, j \leqslant k$. Each p_{ij} defines the probability that any two nodes $u \in C_i$ and $v \in C_j$ are connected. For our purposes the interesting case is that of a bipartition $\{C, \overline{C}\}$, where $|C| = |\overline{C}| = n/2$, $\overline{C} = V \setminus C$, and a pair p_{in}, p_{out} of connection probabilities: p_{in} is the probability that there exists an edge between any two nodes both belonging to the same subset, whereas p_{out} is the probability that there exists an edge between two nodes belonging to different banks of the bipartition.

According to this model, A is an $n \times n$ symmetric $\{0, 1\}$-matrix whose entry a_{ij} takes the value 1 with probability p_{in} if both i and j belong to the same cluster, or p_{out} if i and j belong to different clusters. When p_{in} is bigger than p_{out}, the edges tend to accumulate inside the clusters C and $\overline{C}$. This phenomenon tends to set up C and $\overline{C}$ as communities inside the network, and the adjacency matrix of the graph tends to show a block diagonal predominance.

Various popular and effective techniques for revealing the community structure of a given graph or network with vertex set $V = \{1, \dots, n\}$ are based on the spectral analysis of M, the modularity matrix of the graph. Several formulations of this matrix have been proposed in recent literature, see [1] for an overview. We consider here the one based on the Erdős-Rényi random graph model, defined as follows:

$$M = A - \mathcal{P}_J(A) = A - \frac{\operatorname{vol} V}{n^2} J$$

being A the adjacency matrix of the given graph and $\operatorname{vol} V = \mathbf{1}^{\mathrm{T}} A \mathbf{1}$ its volume. The definition of this matrix is based on the following remark: Let $\mathbf{1}_S$ be the characteristic vector of the set $S \subseteq V$. Then,

$$\mathbf{1}_S^{\mathrm{T}} M \mathbf{1}_S = \mathbf{1}_S^{\mathrm{T}} A \mathbf{1}_S - \frac{\operatorname{vol} V}{n^2} (\mathbf{1}^{\mathrm{T}} \mathbf{1}_S)^2 = \operatorname{vol} S - \operatorname{vol} V \frac{|S|^2}{n^2},$$

where $\operatorname{vol} S = \mathbf{1}_S^{\mathrm{T}} A \mathbf{1}_S$ is the volume of S, that is, the overall weight of internal edges, and the term $\operatorname{vol} V |S|^2/n^2$ quantifies the expected number of edges lying in S, if edges were placed uniformly at random in the graph. Hence, the maximization of the Rayleigh quotient $q(S) = \mathbf{1}_S^{\mathrm{T}} M \mathbf{1}_S / \mathbf{1}_S^{\mathrm{T}} \mathbf{1}_S$ with respect to S arises naturally as a theoretically based method to detect the presence of a cluster in the graph:

If $q(S)$ is "large" then the edge density in the subgraph induced by S is higher than expected. However, the combinatorial nature of the maximization procedure makes it an NP-hard problem. Effective algorithms can be based on a continuous relaxation, leading to the so-called spectral methods.

Loosely speaking, spectral methods for community detection locate clusters in a given graph according to the sign of the entries of the leading eigenvector of M. The idea, first introduced in physics literature (see [7], [6]), mirrors the analogous approach widely used for solving graph partitioning problems by exploiting spectral properties of Laplacian matrices, as pioneered by Fiedler in [4], [3].

Given the bipartition $\{C, \overline{C}\}$ of the vertex set $V = \{1, \ldots, n\}$, define the vector $z = (z_1, \ldots, z_n)^{\mathrm{T}}$, where $z_i = 1$ if $i \in C$ and $z_i = -1$ if not, and let $Z = zz^{\mathrm{T}}$. By using Theorem 2.1 we can show that for certain values of p_{in} and p_{out}, the value of $\cos(M, Z)$ is large, and the leading eigenvector of M is almost parallel to z. Consequently, partitioning vertices on the basis of the sign of the corresponding entries of the leading eigenvector yields a reliable approximation of the true bipartition.

Since $z^{\mathrm{T}}\mathbf{1} = 0$, we have

$$\mathrm{Tr}(MZ) = z^{\mathrm{T}}Mz = z^{\mathrm{T}}Az - \frac{\mathbf{1}^{\mathrm{T}}A\mathbf{1}}{n^2} z^{\mathrm{T}}Jz = z^{\mathrm{T}}Az.$$

Moreover, $a_{ij} \in \{0, 1\}$ implies $a_{ij}^2 = a_{ij}$, therefore $\|A\|_{\mathrm{F}}^2 = \mathbf{1}^{\mathrm{T}}A\mathbf{1}$ and

$$\|M\|_{\mathrm{F}}^2 = \|A\|_{\mathrm{F}}^2 - \frac{(\mathbf{1}^{\mathrm{T}}A\mathbf{1})^2}{n^2} = (\mathbf{1}^{\mathrm{T}}A\mathbf{1})\left(1 - \frac{\mathbf{1}^{\mathrm{T}}A\mathbf{1}}{n^2}\right).$$

As the entries of A are independent Bernoulli random variables, the quantities $\mathbf{1}^{\mathrm{T}}A\mathbf{1}$ and $z^{\mathrm{T}}Az$ are independent random variables as well. In fact, let $\nu_C = \mathbf{1}_C^{\mathrm{T}}A\mathbf{1}_C + \mathbf{1}_{\overline{C}}^{\mathrm{T}}A\mathbf{1}_{\overline{C}}$ and $\partial_C = 2(\mathbf{1}_C^{\mathrm{T}}A\mathbf{1}_{\overline{C}})$. Then we have $\mathrm{Tr}(MZ) = \nu_C - \partial_C$ and $\mathbf{1}^{\mathrm{T}}A\mathbf{1} = \nu_C + \partial_C$. Both ν_C and ∂_C are the sum of identically and independently distributed Bernoulli trials, thus they follow a binomial distribution:

$$\nu_C \sim B\left(\frac{n^2}{2}, p_{\mathrm{in}}\right), \quad \partial_C \sim B\left(\frac{n^2}{2}, p_{\mathrm{out}}\right).$$

Consequently, we have the following statistics:

$$\begin{aligned}
\mathbb{E}(\mathbf{1}^{\mathrm{T}}A\mathbf{1}) &= \mathbb{E}(\nu_C) + \mathbb{E}(\partial_C) = n^2\,\frac{p_{\mathrm{in}} + p_{\mathrm{out}}}{2}, \\
\mathbb{E}(z^{\mathrm{T}}Az) &= \mathbb{E}(\nu_C) - \mathbb{E}(\partial_C) = n^2\,\frac{p_{\mathrm{in}} - p_{\mathrm{out}}}{2}, \\
\mathrm{Var}(\mathbf{1}^{\mathrm{T}}A\mathbf{1}) &= \mathrm{Var}(\nu_C) + \mathrm{Var}(\partial_C) = \frac{n^2}{2}(p_{\mathrm{in}}(1 - p_{\mathrm{in}}) + p_{\mathrm{out}}(1 - p_{\mathrm{out}})).
\end{aligned}$$

Moreover, we also have $\mathrm{Var}(z^{\mathrm{T}}Az) = \mathrm{Var}(\mathbf{1}^{\mathrm{T}}A\mathbf{1})$. With the help of the foregoing formulas, the subsequent lemma estimates the average value of $\cos(M, Z)$ when M is the modularity matrix of a graph belonging to the stochastic block model introduced before.

Lemma 4.1. *Let $M = A - (\mathrm{vol}\, V/n^2)J$ be the Erdős-Rényi modularity matrix of a graph belonging to the stochastic block model with two equally sized clusters C and $\overline{C}$ and edge probabilities p_{in} and p_{out}. Let $z = (z_1, \ldots, z_n)^{\mathrm{T}}$, where $z_i = 1$ if $i \in C$ and $z_i = -1$ if not. Moreover, let*

$$\gamma = \frac{p_{\mathrm{in}} - p_{\mathrm{out}}}{\sqrt{(p_{\mathrm{in}} + p_{\mathrm{out}})(2 - p_{\mathrm{in}} - p_{\mathrm{out}})}}.$$

For any fixed $\varepsilon > 0$, with probability converging to 1 as $n \to \infty$ we have

$$\cos(M, Z)^2 \geqslant \gamma^2 - \varepsilon.$$

P r o o f. Since $z^{\mathrm{T}}z = n$, we have the equivalent formulas

$$\cos(M, Z) = \frac{z^{\mathrm{T}}Mz}{n\|M\|_{\mathrm{F}}} = \frac{z^{\mathrm{T}}Az}{n^2}\,\frac{n}{\|M\|_{\mathrm{F}}}.$$

Hence, the inequality $\cos(M, zz^{\mathrm{T}})^2 \geqslant \gamma^2 - \varepsilon$ is equivalent to the condition

$$\frac{(z^{\mathrm{T}}Az)^2}{n^4} \geqslant (\gamma^2 - \varepsilon)\frac{\|M\|_{\mathrm{F}}^2}{n^2}. \tag{4.1}$$

From the equation $\mathbb{E}((\mathbf{1}^{\mathrm{T}}A\mathbf{1})^2) = \mathrm{Var}(\mathbf{1}^{\mathrm{T}}A\mathbf{1}) + \mathbb{E}(\mathbf{1}^{\mathrm{T}}A\mathbf{1})^2$ and the preceding expressions we can obtain the expectation of $\|M\|_{\mathrm{F}}^2/n^2$ in the considered stochastic block model:

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\big(\|M\|_{\mathrm{F}}^2\big) &= \frac{1}{n^2}\mathbb{E}(\mathbf{1}^{\mathrm{T}}A\mathbf{1}) - \frac{1}{n^4}(\mathrm{Var}(\mathbf{1}^{\mathrm{T}}A\mathbf{1}) + \mathbb{E}((\mathbf{1}^{\mathrm{T}}A\mathbf{1}))^2)\\
&= \frac{p_{\mathrm{in}} + p_{\mathrm{out}}}{2} - \frac{1}{n^4}\Big(\frac{n^2(p_{\mathrm{in}} - p_{\mathrm{in}}^2 + p_{\mathrm{out}} - p_{\mathrm{out}}^2)}{2} + \frac{n^4(p_{\mathrm{in}} + p_{\mathrm{out}})^2}{4}\Big)\\
&= \frac{(p_{\mathrm{in}} + p_{\mathrm{out}})(2 - p_{\mathrm{in}} - p_{\mathrm{out}})}{4} + \mathcal{O}(n^{-2}).
\end{aligned}$$

Since $\mathrm{Var}(z^{\mathrm{T}}Az) = \mathrm{Var}(\mathbf{1}^{\mathrm{T}}A\mathbf{1})$, we also have

$$\begin{aligned}
\frac{1}{n^4}\mathbb{E}((z^{\mathrm{T}}Az)^2) &= \frac{1}{n^4}\big(\mathrm{Var}(z^{\mathrm{T}}Az) + \mathbb{E}(z^{\mathrm{T}}Az)^2\big)\\
&= \frac{(p_{\mathrm{in}} - p_{\mathrm{out}})^2}{4} + \mathcal{O}(n^{-2}).
\end{aligned}$$

Owing to the independence of $\|M\|_{\mathrm{F}}^2$ and $(z^{\mathrm{T}}Az)^2$, with the value of γ given in the claim the inequality (4.1) is certainly fulfilled in the limit of large n. $\square$

Finally, we exploit Theorem 2.1 and the preceding lemma to estimate certain spectral properties of the modularity matrix M that are relevant for the community detection problem. In particular, we focus on $\lambda_1(M)$, which is an important indicator of the detectability of the planted partition $\{C, \overline{C}\}$ (see [5], [6]), and the overlap between that partition and the nodal sets of a leading eigenvector of M.

Theorem 4.1. *With the hypotheses and notations of* Lemma 4.1, *with probability approaching* 1 *in the limit for large* n *we have:*

(1) $\lambda_1(M) \geqslant \mu$ *and* $|\lambda_1(M) - \mu|/\lambda_1(M) \leqslant \sqrt{1-\gamma^2}$, *where* $\mu = (p_{\rm in} - p_{\rm out})n/2$.

(2) *Let* v_1 *be an eigenvector of* $\lambda_1(M)$. *If* $\gamma^2 \geqslant 1/2$, *then* $\cos(v_1, z)^2 \geqslant \bar{\xi}$ *where*

$$\bar{\xi} = \frac{1}{2} + \frac{1}{2}\sqrt{2\gamma^2 - 1}.$$

(3) *The fraction of vertices classified correctly by the partition* $\{i\colon (v_1)_i \geqslant 0\}$ *and* $\{i\colon (v_1)_i < 0\}$ *is at least* $\bar{\xi}$.

P r o o f. The first two claims are easy consequences of Lemma 4.1 and Theorem 2.1. In fact, the value of μ comes from the average value of $z^{\rm T}Mz/(z^{\rm T}z)$, which is equal to $z^{\rm T}Az/n$, while $\bar{\xi}$ is the largest root of $\gamma^2 = 2\xi^2 - 2\xi + 1$.

Let v_1 be an eigenvector of $\lambda_1(M)$ oriented so that $\cos(v_1, z) > 0$. Define $w \in \mathbb{R}^n$ as $w_i = 1$ if $v_i \geqslant 0$ and -1 if not. Clearly, $0 < \cos(v_1, z) \leqslant \cos(w, z) = \cos(w \circ z, \mathbf{1})$, where $w \circ z$ is the Hadamard (componentwise) product of w and z. Note that the number of positive entries in $w \circ z$ gives exactly the number of correctly classified vertices. Let k be the integer part of $\bar{\xi}n$, so that we have $\sqrt{\bar{\xi}} \geqslant \sqrt{k/n} = \pi_{k,n}$. Hence, $\cos(w \circ z, \mathbf{1}) \geqslant \cos(v_1, z) \geqslant \pi_{k,n}$. The last claim follows from Remark 3.1. $\square$

References

[1] *D. Fasino, F. Tudisco*: Generalized modularity matrices. Linear Algebra Appl. *502* (2016), 327–345.

[2] *M. Fiedler*: Special Matrices and Their Applications in Numerical Mathematics. Martinus Nijhoff Publishers, Dordrecht, 1986, SNTL, Praha, 1986.

[3] *M. Fiedler*: A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czech. Math. J. *25* (1975), 619–633.

[4] *M. Fiedler*: Algebraic connectivity of graphs. Czech. Math. J. *23* (1973), 298–305.

[5] *R. R. Nadakuditi, M. E. J. Newman*: Graph spectra and the detectability of community structure in networks. Phys. Rev. Lett. *108* (2012), 188701, 5 pages.

[6] *M. E. J. Newman*: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E *74* (2006), 036104, 19 pages.

[7] *M. E. J. Newman, M. Girvan*: Finding and evaluating community structure in networks. Phys. Rev. E *69* (2004), 026113, 15 pages.

[8] *V. Nikiforov*: The influence of Miroslav Fiedler on spectral graph theory. Linear Algebra Appl. *439* (2013), 818–821.

[9] *K. Rohe, S. Chatterjee, B. Yu*: Spectral clustering and the high-dimensional stochastic blockmodel. Ann. Stat. *39* (2011), 1878–1915.

[10] *P. Tarazaga, M. Raydan, A. Hurman*: Perron-Frobenius theorem for matrices with some negative entries. Linear Algebra Appl. *328* (2001), 57–68.

[11] *P. Tarazaga*: Eigenvalue estimates for symmetric matrices. Linear Algebra Appl. *135* (1990), 171–179.

Authors' addresses: Dario Fasino, Department of Mathematics, Computer Science and Physics, University of Udine, Via delle Scienze 206, 33100 Udine, Italy, e-mail: dario.fasino@uniud.it; Francesco Tudisco, Department of Mathematics and Computer Science, Saarland University, Campus E24, 66123 Saarbrücken, Germany, e-mail: tudisco@cs.uni-saarland.de.