Applications to network analysis: Graph partitioning and community detection Lecture notes

Dario Fasino, University of Udine (Italy)

Lecture notes for the third part of the course "Nonnegative and spectral matrix theory with applications to network analysis", held within the Rome-Moscow school on Matrix Methods and Applied Linear Algebra, August–September 2014. The author is partially supported by Istituto Nazionale di Alta Matematica. This document is intended for internal circulation only.

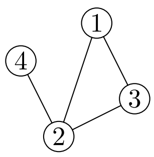
1 Nodal domains

In what follows, the graph $G = (V, E)$ is assumed to be undirected (so that A_G is symmetric). Hereafter, the following notations will be used in correspondence with an arbitrary set $S \subseteq V$:

- Denote by $|S|$ its cardinality (that is, the number of its elements), by $\bar{S}$ its complement (that is, $\bar{S} = V \setminus S$) and by $\mathbf{1}_S$ its characteristic vector, that is $(\mathbf{1}_S)_i = 1$ if $i \in S$ and 0 otherwise.
- Let $\operatorname{vol} S = \sum_{i \in S} d_i$ be the *volume* of S (recall that d_i is the degree of node i). Note: $\operatorname{vol} S = d^T \mathbf{1}_S$.
- Let $e_{\text{in}}(S) = \mathbf{1}_S^T A \mathbf{1}_S$ and $e_{\text{out}}(S) = \mathbf{1}_S^T A(\mathbf{1} - \mathbf{1}_S) = \operatorname{vol} S - e_{\text{in}}(S)$. Note: $e_{\text{out}}(S)$ is the number of edges joining S with $\bar{S}$ while $e_{\text{in}}(S)$ is twice the number of edges whose endpoints are both in S.
- The *subgraph induced by S* is the graph $G(S)$ whose adjacency matrix is $[A]_{i,j \in S}$.

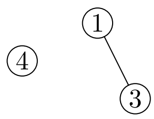
Let $0 \neq v \in \mathbb{R}^n$ and consider the set $S = \{i : v_i \geq 0\}$. The subgraph $G(S)$ may result in a collection of subgraphs which are disconnected one from the other. These components are called *nodal domains* of v. For example, for the following graph G and vector v,

G :



$$v = \begin{pmatrix} 1 \\ -2 \\ 0 \\ 0.1 \end{pmatrix} \quad \rightsquigarrow$$

$G(S)$:



the resulting nodal domains are the subgraphs $G(\{1,3\})$ and $G(\{4\})$.

Let $A = A_G$. A Perron vector v has positive entries, so that v has only nodal domain which is G itself. Obviously, we cannot say the same about other eigenvectors (why?). The goal of this section is to show something interesting about the nodal domains of eigenvectors associated to non-dominant eigenvalues of A [3]. Before going further, a basic fact in matrix theory must be recalled:

Lemma 1.1. [1] *Let $M \in \mathbb{R}^{p \times p}$ be a symmetric matrix, and let $N \in \mathbb{R}^{q \times q}$ be one of its principal submatrices. Let $\lambda_1(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_p(M)$ and $\lambda_1(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_q(N)$ denote the eigenvalues of M and N counted with their multiplicity, respectively. Then, $\lambda_i(M) \geq \lambda_i(N)$ for $i = 1, \ldots, q$.*

Theorem 1.2. *Let $A \geq O$ be irreducible and symmetric. Let $\rho(A) = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$ be its eigenvalues, let v be an eigenvector associated to λ_2, and let $S = \{i \in V : v \geq 0\}$. Then $G(S)$ is connected.*

[1] See e.g., [6, §5.7].

PROOF. Proceed by contradiction. Assume that $S = S_1 \cup S_2$ with $S_1 \cap S_2 = \varnothing$, both $G(S_1)$ and $G(S_2)$ are connected but there is no edge joining V_1 with V_2.

By a suitable permutation of rows and columns, we can assume that $v = (v_1, v_2, v_3)^T$ where $v_1 \geq 0$ and $v_2 \geq 0$ are the entries with indices in S_1 and S_2, respectively, and $v_3 < 0$ are the entries with indices in $\bar{S}$. Accordingly, the structure of A is

$$A = \begin{pmatrix} A_{11} & O & A_{13} \\ O & A_{22} & A_{23} \\ * & * & * \end{pmatrix}$$

where A_{11} and A_{22} are irreducible, and both A_{13} and A_{23} are nonzero (because A is irreducible). Then, equation $Av = \lambda_2 v$ leads to

$$\begin{aligned} A_{11}v_1 + A_{13}v_3 &= \lambda_2 v_1 \\ A_{22}v_2 + A_{23}v_3 &= \lambda_2 v_2. \end{aligned}$$

Let y_1 and y_2 be left Perron eigenvectors of A_{11} and A_{22}, respectively: $y_i^T A_{ii} = \rho(A_{ii}) y_i^T$. Then,

$$\underbrace{y_i^T A_{ii} v_i}_{=\rho(A_{ii}) y_i^T v_i} + \underbrace{y_i^T A_{i3} v_3}_{<0} = \lambda_2 y_i^T v_i, \qquad i = 1, 2.$$

Since $y_i^T v_i > 0$ we get $\rho(A_{ii}) > \lambda_2$ for $i = 1, 2$. Hence, the submatrix $\begin{pmatrix} A_{11} & O \\ O & A_{22} \end{pmatrix}$ hat at least 2 eigenvalues that are $> \lambda_2$. By Lemma 1.1 we deduce that also A has at least two eigenvalues $> \lambda_2$, thus contradicting the fact that $\rho(A)$ is simple. ■

Remarks:

- By applying Theorem 1.2 to $-v$ in place of v, you can deduce easily that also the set $\{i : v_i \leq 0\}$ induces a connected subgraph.
- The argument of the proof of Theorem 1.2 can be extended naturally to eigenvalues λ_i with $i \geq 2$. The result is that, if $Av = \lambda_i v$ and $S = \{i : v_i \geq 0\}$ then $G(S)$ is composed by no more than $i - 1$ connected components, see e.g., [3].

The subsequent sections outline two applicative contexts where nodal domains play an important role; see [5] for a reference.

2 Graph partitioning problems

A graph partitioning problem requires to partition the nodes of a given graph $G = (V, E)$ into pairwise disjoint sets (clusters) so that the number of edges running across different sets is minimized, in some sense.

Hereafter, I consider the special graph partitioning problem where we want to split V into two subsets S and $\bar{S}$, with $S \cup \bar{S} = V$ and $S \cap \bar{S} = \varnothing$. The pair $\{S, \bar{S}\}$ is a *cut* in G.

For any $S \subseteq V$ consider the number

$$H(S) = e_{\text{out}}(S)/|S|,$$

which is sometimes called the *conductance of* S. A set with high conductance has a relatively large amount of edges connecting it to its complement, with respect to the number of nodes. Conversely, a set having low conductance is a set that can be easily separated from the rest of the graph, by removing a quite small number of edges.

In the framework of graph partitoning preblems, a useful merit function of the graph cut $\{S, \bar{S}\}$ (which is easily generalized to more than two sets) is the following:

$$h(S, \bar{S}) = H(S) + H(\bar{S}) = \ldots = \frac{n}{|S||\bar{S}|} e_{\text{out}}(S).$$

As an exercise, you may fill in the blanks in the previous equality.[2]

[2] Note: $e_{\text{out}}(S) = e_{\text{out}}(\bar{S})$.

One of the main graph partitioning problems consists in computing

$$h_G = \min_{S\subseteq V} h(S,\bar{S}) \tag{1}$$

which is an important graph invariant. Indeed, a set attaining that minimum splits the graphs into two parts that are comparable in size and are connected by relatively few edges. The task of finding the set S which minimizes $h(S)$ is very hard. To help its solution, there exists an heuristic technique based on nodal domains that often goes very close to the true solution.

2.1 The Laplacian matrix

Let $D = \mathrm{Diag}(d_1,\dots,d_n)$. The matrix $L = D - A$ is called *Laplacian matrix of G.* This is one of the most useful matrices associated to a graph. The study of its spectral properties and applications has been pioneered by M. Fiedler, see e.g., [2].

For every $v \in \mathbb{R}^n$ we have

$$v^T L v = \sum_{ij\in E} (v_i - v_j)^2, \tag{2}$$

where the sum runs over the set of edges, every edge being counted only once. Thus, L is positive semidefinite; the vector $\mathbf{1}$ is in the kernel of L, that is $L\mathbf{1} = 0$; and the dimension of $\mathrm{ker}L$ is 1 if and only if G is connected.

Exercise 2.1. Prove (2). Deduce from it that the dimension of $\mathrm{ker}L$ is equal to the number of connected components of G.
Hint: let S be the nodes in a connected component of G and consider $v = \mathbf{1}_S$ in (2).

For any given $S \subseteq V$ we have

$$\mathbf{1}_S^T L\mathbf{1}_S = \mathbf{1}_S^T D\mathbf{1}_S - \mathbf{1}_S^T A\mathbf{1}_S = \mathrm{vol}\, S - e_{\mathrm{in}}(S) = e_{\mathrm{out}}(S).$$

Define $v \in \mathbb{R}^n$ as $v = \mathbf{1}_S - (|S|/n)\mathbf{1}$, that is

$$v_i = \begin{cases} |\bar{S}|/n & i \in S \\ -|S|/n & i \notin S. \end{cases}$$

You can easily verify the following equalities:

$$\mathbf{1}^T v = 0, \qquad v^T v = \frac{|S||\bar{S}|}{n}, \qquad v^T L v = e_{\mathrm{out}}(S), \qquad h(S,\bar{S}) = \frac{v^T L v}{v^T v}. \tag{3}$$

We obtain a nontrivial lower bound for the number h_G defined in (1):

Theorem 2.2. *Let G be connected, and let $0 = \lambda_1 < \lambda_2 \le \dots \lambda_n$ be the eigenvalues of L. Then, $h_G \le \lambda_2$.*

Proof. Owing to the variational characterization of the eigenvalues of a symmetric matrix,[3] we have exactly

$$\lambda_2 = \min_{v:\mathbf{1}^T v = 0} \frac{v^T L v}{v^T v}.$$

Moreover, by (3), all possible values of $h(S,\bar{S})$ are contained in the right hand side of the previous equality. ■

Hence, the eigenvalue λ_2, which is named the *algebraic connectivity of G* after [2], tells us how easy is to split the graph into two (roughly balanced) pieces. Indeed, if $\lambda_2 \approx 0$ then G can be easily disconnected (in particular, if $\lambda_2 = 0$ then G is already divided into at least two parts) while, if h_G is large then also λ_2 must be large.

[3] See e.g., [6, §5.6].

2.2 The spectral cut

The nodal domains of an eigenvector associated to λ_2 often provide good approximations to the cut $\{S, \bar{S}\}$ which optimizes $h(S, \bar{S})$. Their connectedness is considered in the following result:

Theorem 2.3. *Let G be a connected, undirected graph. Suppose that the Laplacian matrix L has eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$. Let f an eigenvector associated to λ_2 and let $S = \{i : f_i \geq 0\}$. Then $G(S)$ is connected.*

PROOF. By choosing a sufficiently large positive constant α, the matrix $M = \alpha I - L = \alpha I - D + A$ is nonnegative and irreducible. Moreover, any eigenvector of M is also an eigenvector of L, and conversely. Indeed, $Mv = \mu v \iff Lv = (\alpha - \mu)v$. In particular we see that the eigenvalues of M are the numbers $\alpha > \alpha - \lambda_2 \geq \ldots \geq \lambda_n$. The claim follows immediately from Theorem 1.2. ■

3 Community detection

The goal of a community detection problem is to reveal the presence of "communities", that is, groups of nodes that are tightly connected. Community detection is different from graph partitioning, under many respects. Indeed, a good division of a network into communities is not merely one in which there are few edges between communities; it is one in which edges between communities are fewer than expected. Indeed, according to the original idea of Newman and Girvan [4], a set $S \subset V$ can be recognized as a community only if the number $e_{\text{out}}(S)$ is smaller than the average value of that number, if edges are placed at random.

Here comes an important question: How we quantify the expected number of edges between two arbitrary subsets of a random graph? One of the most convenient and widespread solutions to this problem is based on the following argument, which supposes that we know the degrees $d_1, \ldots, d_n$ of all nodes of G but not the way they are connected.

The total number of (undirected) edges in the graph is $\frac{1}{2}\text{vol}\, V$ (why?). Let S and T be two arbitrary disjoint subsets of V. Pick any of the edges in E, say ij. If edges are placed at random, then

- the probability that $i \in S$ is $\text{vol}\, S/\text{vol}\, V$
- the probability that $j \in T$ is $\text{vol}\, T/\text{vol}\, V$
- the probability that ij connects S and T is $2\,\text{vol}\, S\,\text{vol}\, T/(\text{vol}\, V)^2$.

But there are exactly $\frac{1}{2}\text{vol}\, V$ edges in G. Hence the average number of edges running between S and T can be estimated as $\text{vol}\, S\,\text{vol}\, T/\text{vol}\, V$. That estimate is not rigorous (because the argument allows the presence of multiple edges between two nodes) but is a good approximation of the exact value, in particular when G is sparse, that is $\text{vol}\, V \ll n^2$, as is often the case with complex networks found in real world.

3.1 The modularity matrix

Define the *modularity* of $S \subseteq V$ as

$$Q(S) = \frac{\text{vol}\, S\,\text{vol}\, \bar{S}}{\text{vol}\, V} - e_{\text{out}}(S).$$

This is the difference between the number of edges connecting S with its exterior and the expectation of that number if edges were placed at random. Hence, the inequality $Q(S) > 0$ may indicate that S is a "community" inside G, and we say that S is a *module*. On the other hand, if $Q(S) \leq 0$ then S is well connected with its exterior, and it is unlikely to be a "community".

Note that $Q(S) = Q(\bar{S})$. Moreover, we have the alternative formula (prove it!)

$$Q(S) = e_{\text{in}}(S) - \frac{(\text{vol}\, S)^2}{\text{vol}\, V}.$$

Introduce the *modularity matrix* $M = A - dd^T/\text{vol}\, V$. Then,

$$\mathbf{1}_S^T M \mathbf{1}_S = \mathbf{1}_S^T A \mathbf{1}_S - \frac{(d^T \mathbf{1}_S)^2}{\text{vol}\, V} = e_{\text{in}}(S) - \frac{(\text{vol}\, S)^2}{\text{vol}\, V} = Q(S).$$

Note: $M\mathbf{1} = 0$, whence $Q(V) = 0$. The following result tells us that if $\rho(M)$ is small then G "looks like a random graph."

Theorem 3.1. *Let S and T be any two disjoint subsets of V, and let $e(S,T)$ denote the number of edges joining S and T. Then,*

$$\left| e(S,T) - \frac{\operatorname{vol} S \operatorname{vol} V}{\operatorname{vol} V} \right| \leq \frac{\sqrt{|S||\bar{S}||T||\bar{T}|}}{n} \rho(M). \tag{4}$$

PROOF. Noting that $e(S,T) = \mathbf{1}_S^T A \mathbf{1}_T$, straightforward computations prove that the left hand side of (4) is exactly $|\mathbf{1}_S^T M \mathbf{1}_T|$. Introduce the vectors $v = \mathbf{1}_S - (|\bar{S}|/n)\mathbf{1}$ and $w = \mathbf{1}_T - (|\bar{T}|/n)\mathbf{1}$. We have

$$\mathbf{1}^T v = 0, \qquad \|v\|_2^2 = v^T v = \frac{|S||\bar{S}|}{n},$$

and analogous formulas for w. Finally, using $M\mathbf{1} = 0$ we have

$$|\mathbf{1}_S^T M \mathbf{1}_T| = |v^T M w| \leq \|v\|_2 \|w\|_2 \|M\|_2 \leq \frac{\sqrt{|S||\bar{S}||T||\bar{T}|}}{n} \|M\|_2.$$

To complete the proof it suffices to observe that $\|M\|_2 = \rho(M)$ since M is symmetric. ∎

Note that the right hand side of (4) is not larger than $\frac{n}{4}\rho(M)$, independently on S and T; but becomes a small multiple of $\rho(M)$ when both S and T are tiny.

3.2 The cut-modularity problem

In what follows, I will consider the simplest version of the community detection problem, where we look for a cut $\{S, \bar{S}\}$ which maximizes the merit function

$$q(S, \bar{S}) = \frac{Q(S)}{|S|} + \frac{Q(\bar{S})}{|\bar{S}|} = \ldots = Q(S) \frac{n}{|S||\bar{S}|}.$$

By arguing exactly as in Theorem 2.2 we can obtain the following result:

Theorem 3.2. *Let M be the modularity matrix of a connected graph, and let*

$$m_G = \max_{v : \mathbf{1}^T v = 0} \frac{v^T M v}{v^T v}. \tag{5}$$

Then, $\max_{S \subseteq V} q(S, \bar{S}) \leq m_G$.

Remark 3.3. *The equation $M\mathbf{1} = 0$ tells us that M has 0 as an eigenvalue; but that eigenvalue may not be simple. On the basis of the variational characterization of the eigenvalues of a symmetric matrix,[4] it is not difficult to conclude that the number m_G defined in (5) is the largest eigenvalue of M that remains after deflation of one zero eigenvalue from the spectrum of M.*

Analogously to the graph partitioning problem, the most popular and successful heuristic method to approximate the solution of the cut-modularity problem $\max_{S \subseteq V} q(S, \bar{S})$ is to compute an eigenvector v such that $Mv = m_G v$, $\mathbf{1}^T v = 0$ and then set $S = \{i : v_i \geq 0\}$ [5]. Actually, one can prove that[5]

- at least one subgraph among $G(S)$ and $G(\bar{S})$ is connected;
- there exist graphs such that only one among $G(S)$ and $G(\bar{S})$ is connected, while the other subgraph splits into any arbitrary number of nodal domains;
- there is a relationship between the number of positive eigenvalues of M and the number of distinct modules in G.

[4] See e.g., [6, §5.6].
[5] See the lecture by F. Tudisco "Spectral inequalities for the modularity of a graph". For a reference, see [1].

4 Exercises and problems

Exercises marked with a star ($\star$) are requested for the final evaluation.

1. ($\star$) Let G be a star graph with n nodes.
   - Compute the spectral decomposition of its modularity matrix M.
   - Compute the number m_G from (5).
   - Use the preceding results to prove that G has no modules.
2. Repeat the preceding exercise for a clique, that is the graph whose adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}.$$

3. Let i and j two distinct nodes in a loop-free graph G that are joined by an (undirected) edge, $i \sim j$. Let d_i and d_j be their respective degrees. Prove that if $d_i + d_j < \sqrt{\operatorname{vol} V}$ then the set $S = \{i, j\}$ is a module.
4. Let i and j two distinct nodes in a undirected graph G. Let $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$ be the eigenvalues of the Laplacian matrix. Prove that $\lambda_2 \leq \frac{1}{2}(d_i + d_j) + \delta \leq \lambda_n$ where $\delta = 1$ if $i \sim j$ and $\delta = 0$ otherwise.

References

[1] D. Fasino, F. Tudisco. An algebraic analysis of the graph modularity. *SIAM J. Matrix Anal. Appl.* 35 (2014), 997–1018.

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

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

[4] M. Newman, M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69 (2004), 026113.

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

[6] E. E. Tyrtyshnikov. *A Brief Introduction to Numerical Analysis.* Birkhäuser, 1997.

`Last update: September 17, 2014`