Matrix methods in the analysis of complex networks

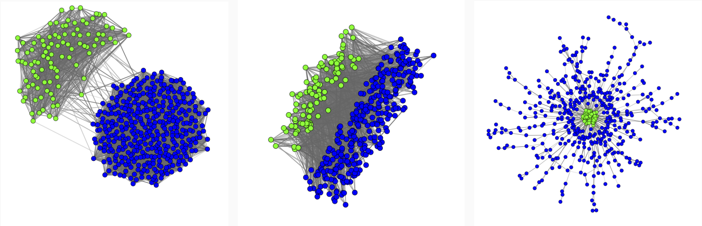
Spectral methods in community detection

Dario Fasino

Rome, Univ. “Tor Vergata”, November 22–24, 2022

Meso-scale structures

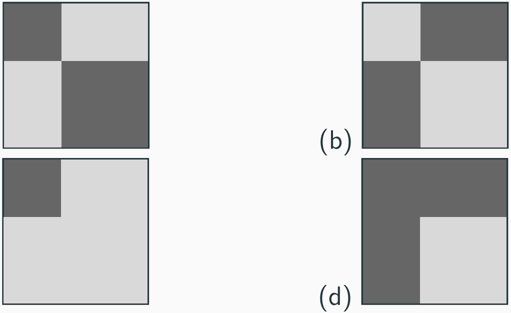
Interesting sub-structures in complex networks: communities, (almost-)bipartite subgraphs, and core-periphery



Random walks and spectral methods are powerful tools to discover them.

Meso-scale structures - block models

(a) (b) (c) (d)



Examples of block models for meso-scale structures in undirected graphs. Shaded areas represent densities of non-zero entries in idealized adjacency matrices.

(a) A block model with two communities. (b) A block model with two anti-communities. (c,d) The two protypical core-periphery block models.

Variational properties of eigenvalues

Let $A = A^{\mathrm{T}} \geq O$, eigenvalues $\rho(A) = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, associate eigenvectors: $v_1, v_2, \ldots, v_n$. Rayleigh quotient:

$$\mathcal{R}(v) = \frac{v^{\mathrm{T}} A v}{v^{\mathrm{T}} v} = \frac{\sum_{i \sim j} v_i v_j}{\sum_i v_i^2}.$$

- $\lambda_1 = \sup \mathcal{R}(v)$. Note: we can choose $v_1 \geq 0$.
- $\lambda_2 = \sup_{v_1^{\mathrm{T}} v = 0} \mathcal{R}(v)$. Note: v_2 cannot have constant sign.
- $\lambda_n = \inf \mathcal{R}(v)$. Note: v_n cannot have constant sign.

Try this procedure!

Compute v_i for $i \in \{1, 2, n\}$.
Permute nodes so that $(v_i)_1 \geq (v_i)_1 \geq \ldots \geq (v_i)_n$.
Do `spy(A)`.

Meso-scale structures - basic spectral methods

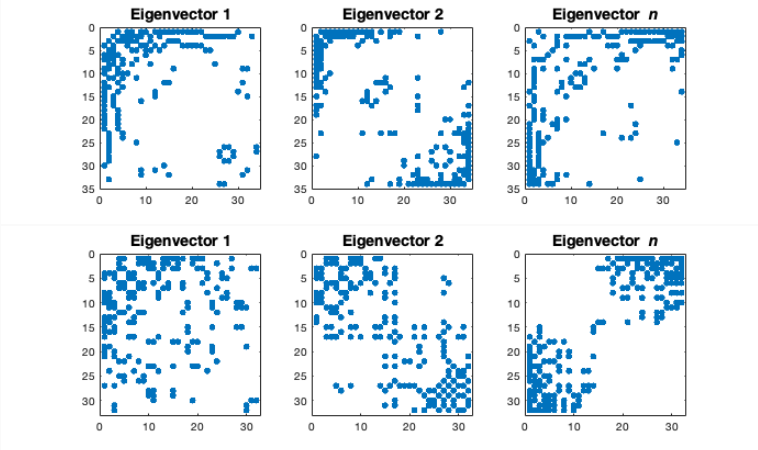
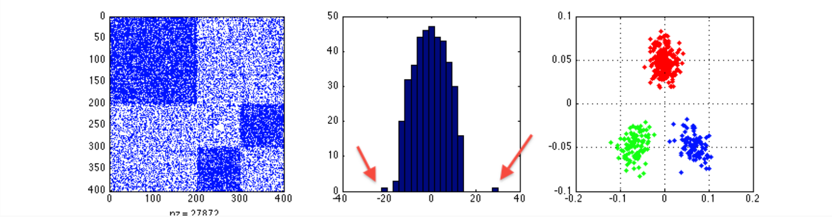


Figure 1: Node reordering of networks `karate` (top) and `Davis` (bottom).

Simultaneous community/anti-community detection



- Left: adjacency matrix a graph with one community and two anti-communities.
- Center: the eigenvalue hystogram reveals two extreme, well separated eigenvlaues.
- Right: the entries of the eigenvectors corresponding to the extreme eigenvalues cluster the nodes belonging to each group.

Graph clustering – Community detection

A relevant problem in graph theory and network science:

Locate one or more groups of nodes which are tightly connected internally but sparsely connected to each other

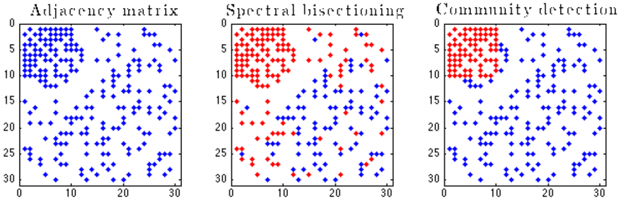
Applications

- Identify people with similar interests/behaviours
- graph compression
- automatic document classification, topic extraction
- identification of functional modules

Modules as communities

How to identify "communities" inside a graph?

- Many answers available; trade-off betwen intercluster edges (many) and intracluster edges (few)
- A different problem from graph partitioning:
  "communities" are densely linked subgraphs
- number and size of clusters are not apriori specified.



Community detection - first steps

Idea [Newman, Girvan '04]

"A good division of a network into communities (...) is one in which there are fewer than expected edges between communities."

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

Idea (rephrased)

Let $\mathcal{G} = (V, E)$ be an undirected graph. A subset $S \subseteq V$ is a "community" if it contains more edges than expected if edges were placed at random in $\mathcal{G}$.

Not only communities!

" (...) exist **nearly complete bipartite subgraphs** *within the protein-protein interaction networks, i.e. two groups of proteins with little or no intra-group connections but strong inter-group connections."*

J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert. Bioinformatics, 2 (2006), 2012–2019.

"In an anti-community, vertices have most of their connections outside their group and have no or fewer connections with the members within the same group."

L. Chen, Q. Yu, B. Chen. Information Sciences 275 (2014), 293–313.

Notation

Let A be the adjacency matrix of $\mathcal{G} = (V, E)$,

$d = (d_1, \ldots, d_n)^T = Ae$ the degree vector.

For $S \subseteq V$ let χ_S be its characteristic vector,

$$(\chi_S)_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S. \end{cases}$$

$\operatorname{vol} S = \sum_{i \in S} d_i$ is the volume of S.

$E(S) = \sum_{i,j \in S} A_{ij} = \chi_S^{\mathrm{T}} A \chi_S$ is the number of edges internal to S.

The modularity of $S \subseteq V$: $Q(S) = E(S) - \mathbb{E}(\text{edges inside } S)$

The modularity of a subgraph

The modularity of $S \subseteq V$: $Q(S) = E(S) - \mathbb{E}(\text{edges inside } S)$

The rightmost term depends on the meaning of the phrase "placing edges at random".

Erdös–Rényi random graph model

The probability that $(i,j) \in E$ is $\alpha = \sum_k d_k / n^2$.

$$\mathbb{E}(\dots) = p|S|^2 \quad \rightsquigarrow \quad Q(S) = E(S) - \alpha|S|^2.$$

Chung–Lu random graph model

Fixed $d_1, \dots, d_n$, the probability that $(i,j) \in E$ is $d_i d_j / \sum_k d_k$.

$$\mathbb{E}(\dots) = \sum_{i,j \in S} \frac{d_i d_j}{\sum_k d_k} \quad \rightsquigarrow \quad Q(S) = E(S) - \frac{(\mathrm{vol}\, S)^2}{\mathrm{vol}\, V}$$

The modularity matrix

In both cases there exists a matrix M such that $Q(S) = \chi_S^T M \chi_S$.

- E–R model:

$$
\begin{aligned}
Q(S) &= \chi_S^T A \chi_S - \alpha |S|^2 \\
&= \chi_S^T A \chi_S - \alpha (e^{\mathrm{T}} \chi_S)^2 = \chi_S^T [A - \alpha e e^{\mathrm{T}}] \chi_S
\end{aligned}
$$

  where $\alpha = \sum_i d_i / n^2$.

- C–L model: Let $d = Ae$ be the degree vector. Then,

$$
\begin{aligned}
Q(S) &= \chi_S^T A \chi_S - \frac{(\mathrm{vol}\, S)^2}{\mathrm{vol}\, V} \\
&= \chi_S^{\mathrm{T}} A \chi_S - \frac{(d^{\mathrm{T}} \chi_S)^2}{\mathrm{vol}\, V} = \chi_S^{\mathrm{T}} [A - \sigma d d^{\mathrm{T}}] \chi_S
\end{aligned}
$$

  where $\sigma = 1/\mathrm{vol}\, V$.

Generalized modularity matrices

Definition

A *generalized modularity matrix* M is a matrix of the form

$$M = A + D - \sigma x x^T$$

where:

- A is symmetric and (entrywise) nonnegative
- D is a diagonal matrix
- σ is a positive scalar
- $x \neq 0$ is a nonnegative vector

All modularity functions having the form

$$Q(S) = E(S) + \sum_{i \in S} D_{ii} - \sigma \Big(\sum_{i \in S} x_i \Big)^2$$

can be restated as quadratic forms: $Q(S) = \chi_S^T M \chi_S$.

Simultaneous community/anti-community detection

Let $Q(S) = \chi_S^{\mathrm{T}} M \chi_S$ be a modularity measure induced by a modularity matrix. In practice, it is best to consider relative modularity measures $q(S) = Q(S)/\mu(S)$ where $\mu(S)$ is an additive measure of S:

$$\mu(S) = |S|, \qquad \text{or} \qquad \mu(S) = \mathrm{vol}\,(S).$$

A successful approach

- $q(S) \gg 0 \rightsquigarrow S$ is a "community"
- $q(S) \ll 0 \rightsquigarrow S$ is an "anti-community"

Simultaneous community/anti-community detection

Let $Q(S) = \chi_S^{\mathrm{T}} M \chi_S$ be a modularity measure induced by a modularity matrix. In practice, it is best to consider relative modularity measures $q(S) = Q(S)/\mu(S)$ where $\mu(S)$ is an additive measure of S:

$$\mu(S) = |S|, \qquad \text{or} \qquad \mu(S) = \mathrm{vol}\,(S).$$

Define the corresponding measure vector

$$m_S = \chi_S \qquad \text{or} \qquad m_S = \mathrm{Diag}(d)^{-1/2} \chi_S,$$

respectively. Then

$$q(S) = \frac{m_S^{\mathrm{T}} \widehat{M} m_S}{m_S^{\mathrm{T}} m_S} = \mathcal{R}(m_S),$$

where $\widehat{M}$ is a suitable diagonal scaling of M. Thus locating "good" modules reduces to computing extremal eigenvalues of $\widehat{M}$.

Main results - in a nutshell

We say that $C \subset V$ is a module if $|q(S)|$ is "large".

Pseudo-theorem (see references for rigorous statements!)

Let $C_1, \ldots, C_k$ be pairwise disjoint modules,
$|q(C_1)| \geq \ldots \geq |q(C_k)|$.
Sort eigenvalues of $\widehat{M}$ as $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$
with corresponding eigenvectors $v_1, v_2, \ldots, v_n$.
Then $m_{C_1}, \ldots, m_{C_k}$ are are "close" to $\langle v_1, \ldots, v_k \rangle$.
Moreover, the relative error between $q(C_i)$ and λ_i is "small."

Thus well separated, extreme eigenvalues of $\widehat{M}$ indicate the presence of good modules.

Nodal sets

How to reconstruct the modules from eigenvectors of a (generalized) modularity matrix?

Nodal sets

Let $\mathcal{G} = (V, E)$ and $v \in \mathbb{R}^{|V|}$ be given. The sets

$$\{i \in V : v_i \geq 0\}, \qquad \{i \in V : v_i < 0\},$$

are the nodal sets induced by v.

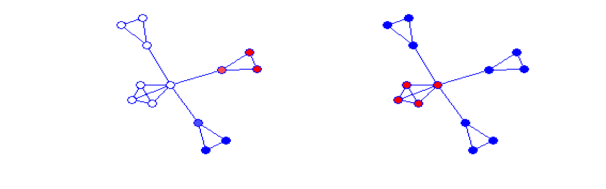
Idea: use nodal sets (or intersections thereof) induced by extreme eigenvectors.

Modularity nodal theorems

Theorem

- $M = A + D - \sigma xx^T$ generalized modularity matrix
- $M\, v_{max} = \lambda_{max}(M) v_{max}$ oriented so that $x^T v_{max} \geq 0$.

Then the subgraph induced by $\{i : v_{\max,i} \geq 0\}$ is connected.



Nodal domains in a small graph. Left: Fiedler vector. Right: Leading modularity eigenvector.

Modularity nodal theorems

Theorem

- $M = A + D - \sigma x x^T$ generalized modularity matrix
- $M\, v_{max} = \lambda_{max}(M) v_{max}$ oriented so that $x^T v_{max} \geq 0$.

Then the subgraph induced by $\{i : v_{\mathrm{max},i} \geq 0\}$ is connected.

More generally:

Let $M = A + D - \sigma x x^T$ be any generalized modularity matrix, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Let v_k be an eigenvector of λ_k oriented so that $v_k^T x \geq 0$. Then the subgraph induced by $\{i : v_{k,i} \geq 0\}$ has at most k connected components.

Modularity nodal theorems

Theorem

- $M = A + D - \sigma xx^T$ generalized modularity matrix
- $M\, v_{max} = \lambda_{max}(M) v_{max}$ oriented so that $x^T v_{max} \geq 0$.

Then the subgraph induced by $\{i : v_{\max,i} \geq 0\}$ is connected.

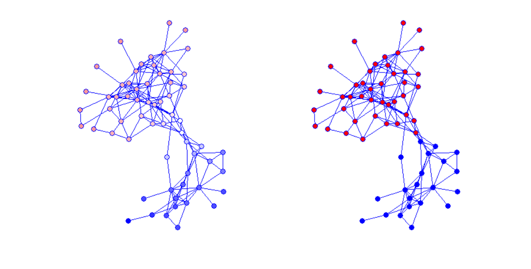


Figure 2: The principal eigenvector of the Newman-Girvan modularity matrix and its nodal domains for the `dolphins` network.

If $\lambda_{max}(M)$ is large enough...

So we can obtain connected subgraphs by thresholding $v_{\max}$.

Typically $\{i : v_{\max,i} \geq 0\}$ is a good indicator of the leading module and it has positive modularity (experimentally).

Actually, if $\lambda_{max}(M)$ is large enough and $v_{\max}$ is not localized, then the subset $\{i : v_{\max,i} \geq 0\}$ is a good module.

Theorem

Let $Mv = \lambda v$ with $\lambda > 0$.
For any $S \subset V$ let $\alpha = \measuredangle(v, \mathrm{Span}\{\chi_S, e\})$. Then,

$$Q(S) \geq \frac{|S||\bar{S}|}{n}\left[\lambda \cos^2\alpha + \underbrace{\lambda_{\min}(M)}_{<0} \sin^2\alpha\right].$$

If $Q(C_+)$ is large enough...

Let C_+ be the node set with largest modularity. If $Q(C_+)$ is large enough, then C_+ is the set obtained by thresholding v_{max}.

Let $Q(C_1, C_2) = e_{C_1}^T M e_{C_2}$ be the *joint* modularity of the subsets C_1, C_2.

Theorem

If the subset C is such that

$$Q(C) + Q(\bar{C}) - 2Q(C, \bar{C}) \geq \sqrt{(n-1)^2 + 1}\, \|M + \alpha I\|_F - n\alpha$$

for some $\alpha \in \mathbb{R}$, then

$$C = \{i : v_{\max,i} \geq 0\} = C_+$$

being $M\, v_{max} = \lambda_{max}(M) v_{max}$.

How many communities?

Positive eigenvalues of M are related to the number of distinct communities in G.

Theorem

Let $\{S_1, \ldots, S_k\}$ be an *optimal*(*) partitioning of V into modules,

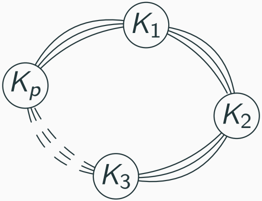
$$Q(S_i) > 0, \qquad Q(S_i \cup S_j) \leq Q(S_i) + Q(S_j).$$

Then $k - 1$ does not exceed the number of positive eig.s of M.

(*) with respect to the overall modularity $\sum_i Q(S_i)$.

Example

The inequality $\#\text{communities} \le \#\text{positive eig.s} + 1$ is sharp:

G : 

$$A = \begin{pmatrix} K_1 & \frac{1}{2}I & & \frac{1}{2}I \\ \frac{1}{2}I & K_2 & \ddots & \\ & \ddots & \ddots & \frac{1}{2}I \\ \frac{1}{2}I & & \frac{1}{2}I & K_p \end{pmatrix}$$

There are $n = pq$ nodes

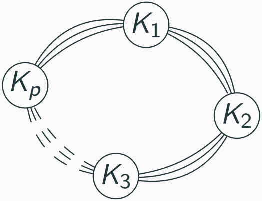
For $i = 1, \ldots, p$ each K_i is a clique with q nodes

consecutive clusters joined by q edges with weight $< 1/2$.

Example

The inequality $\#\text{communities} \leq \#\text{positive eig.s} + 1$ is sharp:

G :



$$A = \begin{pmatrix} K_1 & \frac{1}{2}I & & \frac{1}{2}I \\ \frac{1}{2}I & K_2 & \ddots & \\ & \ddots & \ddots & \frac{1}{2}I \\ \frac{1}{2}I & & \frac{1}{2}I & K_p \end{pmatrix}$$

There are $n = pq$ nodes

For $i = 1, \ldots, p$ each K_i is a clique with q nodes

consecutive clusters joined by q edges with weight $< 1/2$.

$\rightsquigarrow$ A has p positive eig.s, $M = A - \frac{q}{n}ee^T$ has one less

$\rightsquigarrow$ Nodal domains of M's leading eigenvectors separate $K_1, \ldots, K_p$.

References

> *The best research in the future will follow the same patterns: find a problem on a social network, determine a realistic model, and then decide on a computable method to solve the model. Hopefully, I've convinced you of the usefulness of stating problems as matrix problems.*
>
> David Gleich, *ACM Crossroads* **19** (2013) 32–36.

- D. F., F. Tudisco. An algebraic analysis of the graph modularity. *SIAM J. Matrix Anal. Appl.*, 35 (2014), 997–1018.
- D. F., F. Tudisco. Generalized modularity matrices. *Lin. Algebra Appl.*, 502 (2016), 327–345.
- D. F., F. Tudisco. A modularity based spectral method for simultaneous community and anti-community detection. *Lin. Algebra Appl.*, 542 (2018), 605–623.