A modularity-based spectral graph analysis

Dario Fasino (Udine), Francesco Tudisco (Roma TV)

Cagliari, VDM60

A *complex network* is a (di-)graph found in real world.

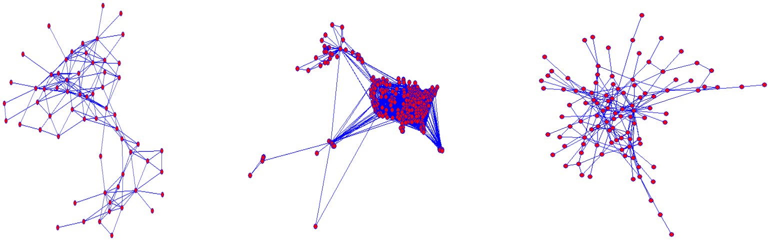


Figure: Small complex networks: `dolphins`, `USAir97`, `Householder93`.

A *complex network* is a (di-)graph found in real world.

Outline:

1. Elements of algebraic graph theory
2. Two problems on complex networks:
   1. graph partitioning — Laplacian matrices
   2. community detection — modularity matrices
3. Spectral analysis of modularity matrices
4. Complements, comments, conclusion

D. F., F. Tudisco.
An algebraic analysis of the graph modularity.
Preprint (2013).

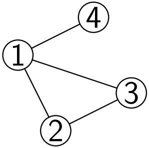
A *complex network* is a (di-)graph found in real world.

Notations:

- $G = (V, E)$: (unoriented) graph, vertices $V = \{1, \dots, n\}$, edges $E \subseteq V \times V$
- A subset $S \subseteq V$ induces a subgraph, having edge set $E(S)$ and edge boundary ∂S
- if $S \subseteq V$ then $\bar{S}$ denotes complement, $|S|$ denotes cardinality
- the degree of vertex i is $d_i = \deg(i)$. The volume of $S \subseteq V$ is $\operatorname{vol} S = \sum_{i \in S} d_i$;

$$\operatorname{vol} S = 2|E(S)| + |\partial S|.$$

A few special matrices are usually associated to a graph G: the adjacency matrix A and the graph Laplacian $L = \mathrm{Diag}(d_1, \ldots, d_n) - A$:

$G =$ 

$$d = \begin{pmatrix} 3 \\ 2 \\ 2 \\ 1 \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \qquad L = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

Note: $L\mathbf{1} = 0$.

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

Graph partitioning

Graph partitioning problem

Find a partitioning of the vertices into clusters, which minimizes the total weight (e.g., number) of intercluster edges.

- Number and size of subsets are (roughly, at least) fixed;
- most familiar quality measure of a cut $\{S, \bar{S}\}$:

$$h(S) = \frac{|\partial S|}{\min\{|S|, |\bar{S}|\}}, \quad \text{conductance of } S$$

- Minimize $h(S) \rightsquigarrow$ NP-hard $\rightsquigarrow$ spectral techniques

Let $\mathbf{1}_S$ denote the characteristic vector of S.
Then $|\partial S| = \mathbf{1}_S^T L \mathbf{1}_S$, $|S| = \mathbf{1}_S^T \mathbf{1}_S$.

Graph partitioning problem

Find a partitioning of the vertices into clusters, which minimizes the total weight (e.g., number) of intercluster edges.

Spectral partitioning technique

Instead of $\min_S h(S)$ solve

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

Then set $S = \{i : v_i \geq \sigma\}$.

The solution is the Fiedler vector: $Lf = a(G)f$
$a(G)$ = smallest positive e.value of L = algebraic connectivity of G.

Level sets of Fiedler vectors

Theorem

Let G be a connected graph with $a(G)$ simple eigenvalue, $Lf = a(G)f$. For $\sigma \leq 0$, let $S = \{i : f_i \geq \sigma\}$. Then S induces a connected subgraph.

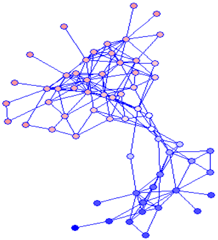


Figure: Spectral bisection of the `dolphins` network. Left: Fiedler vector. Right: level sets, $\sigma = 0$.

Level sets of Fiedler vectors

Theorem

Let G be a connected graph with $a(G)$ simple eigenvalue, $Lf = a(G)f$. For $\sigma \le 0$, let $S = \{i : f_i \ge \sigma\}$. Then S induces a connected subgraph.

More generally, if $\lambda_i(L)$ is simple and $\sigma = 0$ then the connected components of S and $\bar{S}$ are no more than $i + 1$.

Analogous results hold also for Schrödinger operators on weighted graphs, i.e., $\mathrm{Diag}(v) - A$.

Davies, Gladwell, Leydold, Stadler.
Discrete nodal domain theorems.
Lin. Alg. Appl., 336 (2001), 51–60.

Community detection

How to partition a graph into "communities"?

- Many answers available; trade-off betwen intercluster edges (many) and intracluster edges (few)
- number and size of clusters are not a priori specified.

Idea [Newman, Girvan 06]

"*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 (2006), 026113.

Community detection — modularity

We need a null model to define the expected number of edges in a subgraph; e.g., the Erdös-Renyi random graph model.
A better choice:

Chung-Lu random graph model

Fixed integers $d_1, \ldots, d_n$, the probability that the edge (i,j) exists is $d_i d_j / \sum_k d_k$.

Accordingly, the expected number of edges supported in $S \subseteq V$ is

$$\sum_{i,j \in S} \frac{d_i d_j}{\sum_k d_k} = \frac{(\mathrm{vol}\, S)^2}{\mathrm{vol}\, G}.$$

The difference between that number and $|E(S)|$ is a quality measure for S as a "community".

Community detection — modularity

Modularity of $S \subseteq V$:

$$Q(S) = 2|E(S)| - \frac{(\mathrm{vol}\, S)^2}{\mathrm{vol}\, G}$$
$$= \frac{\mathrm{vol}\, S\, \mathrm{vol}\, \bar{S}}{\mathrm{vol}\, G} - |\partial S| = Q(\bar{S}).$$

What is a "community"?

A *community* is a subset $S \subset V$ having positive modularity.

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

$$Q(S) = \mathbf{1}_S^T M \mathbf{1}_S.$$

Indeed, $\mathbf{1}_S^T A \mathbf{1}_S = 2|E(S)|$ and $\mathbf{1}_S^T d = \mathrm{vol}\, S$. Note: $M\mathbf{1} = 0$.

Algebraic modularity

Community detection problem (simplified: just one cluster)

Find $S \subset V$ which maximizes the modularity $Q(S)$.

Instead of $\max_{S \subset V} Q(S)$ (NP-hard) solve

$$m(G) := \max_{v^T \mathbf{1} = 0} \frac{v^T M v}{v^T v}$$

Then set $S = \{i : v_i \geq \sigma\}$. By far, the most popular and successful heuristic for community detection [Newman'06, Fortunato'10, VanDooren+'12...]
The solution is $Mv = m(G)v$
$m(G)$ = algebraic modularity of G.
Very informally, v = Newman vector. $v^T \mathbf{1} = 0$.

Spectral properties of M

$Q(S) = \mathbf{1}_S^T M \mathbf{1}_S = \mathrm{trace}(M(\mathbf{1}_S^T \mathbf{1}_S))$. Owing to $Q(S) = Q(\bar{S})$,

$$Q(S) = \alpha Q(S) + (1-\alpha) Q(\bar{S}) = \mathrm{trace}(MB)$$

for all $0 \le \alpha \le 1$, where $B = \alpha \mathbf{1}_S \mathbf{1}_S^T + (1-\alpha)\mathbf{1}_{\bar{S}} \mathbf{1}_{\bar{S}}^T$.
Let $\alpha = |\bar{S}|/n$. From Wieland-Hoffman theorem,

$$\begin{aligned} Q(S) &\le \lambda_1(M)\lambda_1(B) + \lambda_2(M)\lambda_2(B) \\ &= (\lambda_1(M) + \lambda_2(M))\frac{|S||\bar{S}|}{n} \\ &\le \lambda_1(M)\frac{n}{4}, \end{aligned}$$

independently of S.
Owing to $M\mathbf{1} = 0$ we can replace $\lambda_1(M)$ by $m(G)$.

Spectral properties of M

Let $G_0 = (V, V \times V, \omega_0)$ the *null model* weighted graph with $\omega_0(i,j) = d_i d_j / \mathrm{vol}\, G$, and let L_0 be its Laplacian:

$$(L_0)_{ij} = \begin{cases} -\omega_0(i,j) & i \neq j \\ \sum_{k \neq i} \omega_0(i,k) & i = j. \end{cases}$$

Then, $L_0 = D - dd^T / \mathrm{vol}\, G$. Moreover,

$$M = A - D + D - dd^T / \mathrm{vol}\, G = L_0 - L.$$

We also obtain:

$$d_{\min} - a(G) \leq a(G_0) - a(G) \leq m(G) \leq d_{\max} - a(G).$$

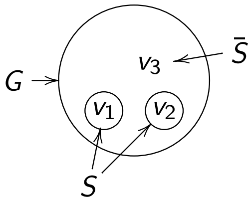
In particular, $m(G) \geq -d_{\min}/(n-1)$, optimal bound.

Level sets of Newman vectors

Theorem

Let $Mv = m(G)v$ with $m(G)$ simple eigenvalue and $d^T v \geq 0$.
For all $\sigma \leq 0$, $S = \{i : v_i \geq \sigma\}$ induces a connected subgraph.

PROOF (sketch, $\sigma = 0$).
$m(G)v = Mv = Av - (d^T v/\mathrm{vol}\, G)d \leq Av$.
By contradiction, assume that S consists of 2 disjoint subgraphs:
Reorder entries of v according to partitioning:



Level sets of Newman vectors

Theorem

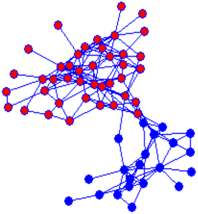
Let $Mv = m(G)v$ with $m(G)$ simple eigenvalue and $d^T v \geq 0$.
For all $\sigma \leq 0$, $S = \{i : v_i \geq \sigma\}$ induces a connected subgraph.

PROOF (sketch, $\sigma = 0$).
$m(G)v = Mv = Av - (d^T v/\mathrm{vol}\, G)d \leq Av$.
By contradiction, assume that S consists of 2 disjoint subgraphs:
Reorder and partition consistently A, M, v. Then,

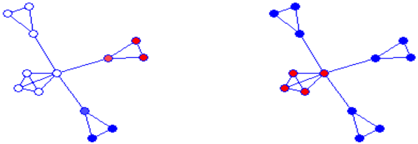
$$\begin{pmatrix} m(G)v_1 \\ m(G)v_2 \\ m(G)v_3 \end{pmatrix} \leq \begin{pmatrix} A_{11} & & * \\ & A_{22} & * \\ * & * & * \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \leq \begin{pmatrix} A_{11}v_1 \\ A_{22}v_2 \\ * \end{pmatrix}.$$

By nonnegativity and eigenvalue interlacing,
A has at least 2 eigenvalues $> m(G)$, absurd. □

Nodal domains: Examples



The `dolphins` network. Left: Fiedler vector. Right: Newman vector.



A small graph. Left: Fiedler vector. Right: Newman vector.

The Householder93 collaboration graph

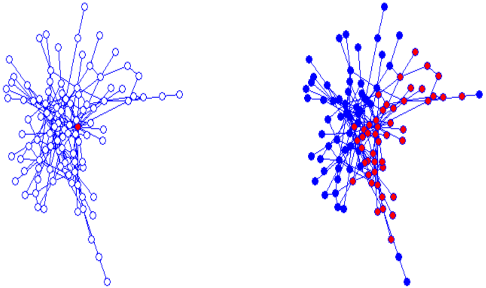


Figure: Community detection in `Householder93`.

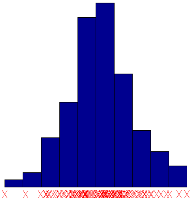


Figure: Spectral distribution of M

The Householder93 collaboration graph

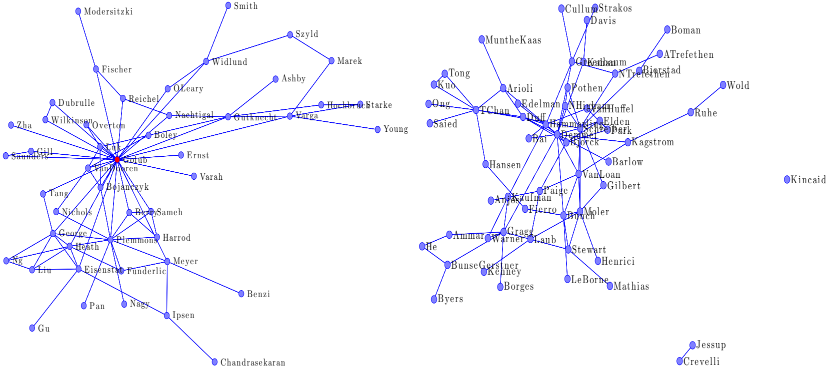


Figure: Community detection in the `Householder93` network.
Left: positive cluster. Right: negative cluster.

Cheeger-type inequalities — conductance

Definition

The isoperimetric constant (aka *Cheeger number*) of G is

$$h_G = \min_{S \subset V} \frac{|\partial S|}{\min\{|S|, |\bar{S}|\}}.$$

Theorem (Dodziuk'84, Alon-Milman'85, Mohar'89...)

If G is k-regular and $a(G)$ its algebraic connectivity then

$$\frac{a(G)}{2} \leq h_G \leq \sqrt{a(G)(2k - a(G))}.$$

Definition (Newman, Girvan 2004)

The modularity of a graph G is

$$Q_G = \frac{2}{\mathrm{vol}\, G} \max_{S \subset V} Q(S), \qquad Q(S) = \mathbf{1}_S^T M \mathbf{1}_S.$$

Theorem

If G is k-regular and $m(G)$ its algebraic modularity then

$$\frac{1}{2n} - \sqrt{\frac{k - m(G)}{2k}} \leq Q_G \leq \frac{m(G)}{2k}.$$

Conclusions

Spectral properties of modularity matrices:

- difference of two Laplacians $\rightsquigarrow$ bounds for the algebraic modularity $m(G)$, relations with $a(G)$
- level sets of (leading) eigenvectors $\rightsquigarrow$ Fiedler-type results, theoretical support to spectral community detection algorithms
- Cheeger-type inequalities.

Best wishes, Cor!

Thank you.