Community detection with spectral methods

Laurent Massoulié

MSR-Inria Joint Centre

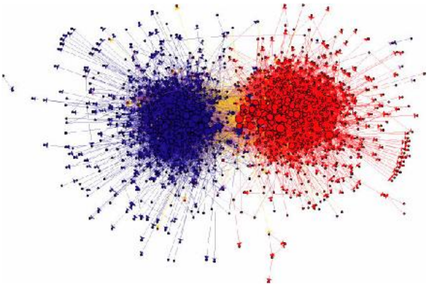
October 27, 2016

Community detection

Definition: Identification of groups of similar items, based on observed interactions between items

Observation: a graph (e.g. represented by its adjacency matrix); edges could be directed or not, labeled or not.

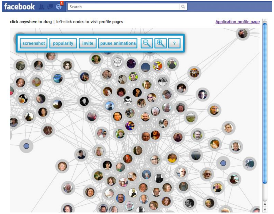
→ Same as clustering, specialized to "graphical" observations



Example: From observations of citation graph between blog posts during US presidential campaign, partitioning into democrats / republicans communities

Example 1

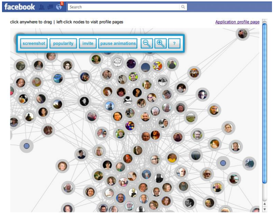
From "friendship" graph of facebook infer communities of "similar" users to guide recommendations of potential new contacts



An example of *assortative* communities: stronger connectivity within than across communities

Example 1

From "friendship" graph of facebook infer communities of "similar" users to guide recommendations of potential new contacts



An example of *assortative* communities: stronger connectivity within than across communities

Variation: NSA's "co-traveler programme": spot group of suspect persons meeting regularly in unusual places

Example 2

From matrix of user ratings of items, infer communities of "similar" items to guide item recommendations ("users who liked this also liked that")

User / Movie	f_1	f_2	…	f_m
u_1	?	$**$		$***$
u_2	$***$	?		?
…				
u_n	$*****$	$**$		$**$

$\rightarrow$ e.g. Netflix "Cinematch" movie recommendation engine (see Netflix prize)

Example 3

Lists of proteins involved in chemical reactions of cell biology

$\rightarrow$ Graph: co-involvement in some reaction

$\rightarrow$ Infer groups of proteins with same functional role (see "Functional cartography of complex metabolic networks", R. Guimera, L. Nunes Amaral, Nature 2005 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2175124/)

Potentially *disassortative* communities (connections within community may be rarer than across)

Knowledge graphs

A generic representation of data:

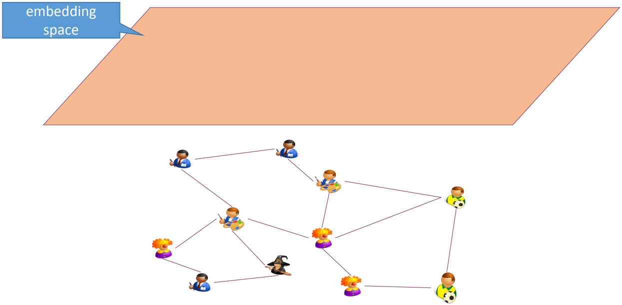
A1 has with **B1** interaction of type **C1**
A2 has with **B2** interaction of type **C2**
A3 has with **B3** interaction of type **C3**
...

$\rightarrow$ Popularized by companies like Google as a universal format

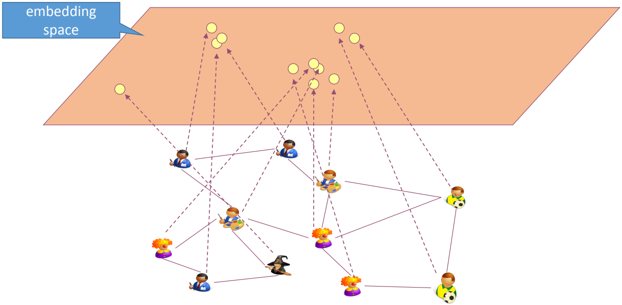
Outline

- Spectral methods
- The Stochastic Block Model (SBM)
- Consistent community detection in a “high signal” regime
- Tools: linear algebra and random matrices

Embedding



Embedding



Spectral embedding

- Extract top two (more generally top k) eigenvalues λ_1, λ_2 of graph's adjacency matrix $A \in \mathbb{R}^{n \times n}$ (ordered by absolute value: $|\lambda_1| \geq \lambda_2 \geq \cdots$)

- Let $x_1, x_2 \in \mathbb{R}^n$: corresponding normalized eigenvectors

- Embed vertex $k \in [n]$ into $\mathbb{R}^2$ by letting $z_k := \sqrt{n}(x_1(k), x_2(k))$

Spectral embedding

- Extract top two (more generally top k) eigenvalues λ_1, λ_2 of graph's adjacency matrix $A \in \mathbb{R}^{n\times n}$ (ordered by absolute value: $|\lambda_1| \geq \lambda_2 \geq \cdots$)

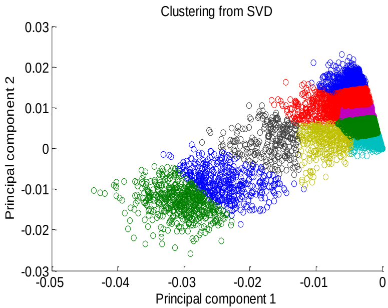
- Let $x_1, x_2 \in \mathbb{R}^n$: corresponding normalized eigenvectors

- Embed vertex $k \in [n]$ into $\mathbb{R}^2$ by letting $z_k := \sqrt{n}(x_1(k), x_2(k))$

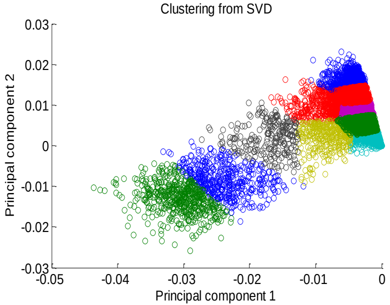
$\rightarrow$ Also known as Principal Component Analysis (PCA)

Due to Karl Pearson "On Lines and Planes of Closest Fit to Systems of Points in Space", 1901

Example: 2D-spectral embedding of Netflix prize data



Example: 2D-spectral embedding of Netflix prize data



From embedding to clustering: K-means algorithm

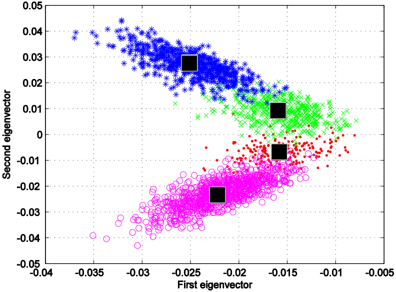
1. Choose K cluster centers $y_1, \ldots, y_K \in \mathbb{R}^2$
2. Form clusters
   $C_k := \left\{ i \in [n] : ||y_k - z_i|| = \min_{\ell \in [K]} \{||y_\ell - z_i||\} \right\}$
3. Reset y_k to cluster average $\frac{1}{|C_k|} \sum_{i \in C_k} z_i$
4. Go back to step 2

The stochastic block model

A multi-type version of the Erdős-Rényi random graph

- n vertices partitioned into K communities
- Type (community) of node $i : \sigma(i) \in [K]$
- For each $k \in [K]$, number $\sum_i \mathbb{I}_{\sigma(i)=k}$ of type-k-nodes: $\sim \alpha_k n$ for fixed $\alpha_k > 0$
- For each pair $i, j \in [n]$: edge (i, j) present with probability $B_{\sigma(i),\sigma(j)} \frac{d}{n}$, where $B \in \mathbb{R}_+^{K \times K}$: fixed matrix, and d: may increase as $n \to \infty$

Example: spectral embedding for SBM



A case with $K = 4$ communities

Spectral embedding seems to reflect community structure

$\rightarrow$ Why / when do spectral methods work?

Theorem

Assume communities are **distinguishable**, *i.e. for each* $k \neq \ell \in [K]$, *there exists* $m \in [K]$ *such that* $B_{km} \neq B_{\ell m}$.
Assume $d \sim n^{\delta}$ *for some* $\delta \in]0,1[$. *Let* R: *rank of matrix* B. *Then with high probability:*
(i) the spectrum of A *consists of* R *eigenvalues of order* $\Theta(d)$ *and* $n - R$ *eigenvalues of order* $o(d)$.
(ii) R*-dimensional spectral embedding reveals underlying communities: except for vanishing fraction of nodes* $i \in [n]$,

$$||z_i - z_j|| = \begin{cases} o(1) & \text{if } \sigma(i) = \sigma(j), \\ \Omega(1) & \text{if } \sigma(i) \neq \sigma(j) \end{cases}$$

Theorem

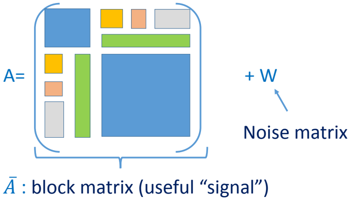
Assume communities are **distinguishable**, *i.e. for each $k \neq \ell \in [K]$, there exists $m \in [K]$ such that $B_{km} \neq B_{\ell m}$.*
Assume $d \sim n^{\delta}$ for some $\delta \in]0,1[$. Let R: rank of matrix B. Then with high probability:
(i) the spectrum of A consists of R eigenvalues of order $\Theta(d)$ and $n - R$ eigenvalues of order $o(d)$.
(ii) R-dimensional spectral embedding reveals underlying communities: except for vanishing fraction of nodes $i \in [n]$,

$$||z_i - z_j|| = \begin{cases} o(1) & \text{if } \sigma(i) = \sigma(j), \\ \Omega(1) & \text{if } \sigma(i) \neq \sigma(j) \end{cases}$$

Corollary

Under these conditions any sensible clustering scheme (eg K-means properly initialized) correctly classifies all but vanishing fraction of nodes.

Proof strategy



- Write adjacency matrix as $A = \overline{A} + W$ with $\overline{A}_{ij} = \frac{d}{n} B_{\sigma(i),\sigma(j)}$

- R leading eigen-elements of $\overline{A}$ capture community structure

- Control perturbation of eigen-elements of a symmetric matrix $\overline{A}$ by addition of symmetric matrix W in terms of **spectral radius** $\rho(W)$ of noise matrix

- Prove bound on $\rho(W)$ for random noise matrix W

Eigenstructure of $\overline{A}$

Block structure of $\overline{A} \Rightarrow \overline{A}x$ constant on each block $\Rightarrow$ eigenvectors associated to non-zero eigenvalue are block-constant.

For $t \in \mathbb{R}^K$ define $x := \phi(t) = (t_{\sigma(i)})_{i \in [n]} \in \mathbb{R}^n$.

Then $\overline{A}\phi(t) = d\phi(Mt)$, where $M_{uv} := B_{uv}\alpha_v$.

Lemma

Spectrum of $\overline{A}$:
R eigen-pairs $(\lambda_u = d\mu_u, \bar{x}_u = \phi(t_u))$ where (μ_u, t_u): eigen-pairs of M with $\mu_u \neq 0$;
0: eigenvalue with multiplicity $n - R$

Eigenstructure of $\overline{A}$ (continued)

Lemma

Under distinguishability hypothesis there exists $\epsilon > 0$ *function of* B, α *such that for any choice of normalized leading eigenvectors* $\overline{x}_1, \ldots, \overline{x}_R$, $\overline{z}_i = \sqrt{n}(\overline{x}_1(i), \ldots, \overline{x}_R(i))^T$ *verify*

$$\sigma(i) \neq \sigma(j) \Rightarrow ||\overline{z}_i - \overline{z}_j|| \geq \epsilon > 0$$

Proof: Let $t_u \in \mathbb{R}^K$ be such that $\sqrt{n}\overline{x}_u = \phi(t_u)$, and $\sqrt{\alpha} = \text{Diag}(\sqrt{\alpha_u})$.
Then: $\{\sqrt{\alpha} t_u\}_{u \in [R]}$: orthonormal family by orthonormality of the $\overline{x}_u$.
t_u eigenvectors of matrix $M = B\alpha$, hence $\sqrt{\alpha} t_u$: orthonormal family of eigenvectors of matrix $\sqrt{\alpha} B \sqrt{\alpha}$.
Thus $\sqrt{\alpha} B \sqrt{\alpha} = \sum_{u \in [R]} \mu_u (\sqrt{\alpha} t_u)(\sqrt{\alpha} t_u)^T$.
Equivalently: $B = \sum_{u \in [R]} \mu_u t_u t_u^T$.
Hence minimum of $||\overline{z}_i - \overline{z}_j||$ over $\sigma(i) \neq \sigma(j)$ strictly positive, for otherwise B has two identical rows, i.e. distinguishability fails.

Controlling perturbation of eigenvalues

Lemma

(Weyl's inequality) Order eigenvalues of A (resp. $\overline{A} = A + W$ for symmetric A, W as $\lambda_1 \geq \lambda_2 \geq \cdots$ (respectively, $\overline{\lambda}_1 \geq \overline{\lambda}_2 \geq \cdots$). Then for all $i \in [n]$, $|\lambda_i - \overline{\lambda}_i| \leq \rho(W)$

Proof: by Courant-Fisher theorem,

$$\lambda_i = \sup_{\dim(E)=i} \inf_{x \in E, ||x||=1} x^T A x$$

Apply to $E = \text{Vect}\{\overline{x}_1, \ldots, \overline{x}_i\}$ to obtain

$$\begin{aligned}
\lambda_i &\geq \inf_{x \in E, ||x||=1} x^T A x \\
&\geq \inf_{x \in E, ||x||=1} x^T \overline{A} x + \inf_{x \in E, ||x||=1} x^T W x \\
&\geq \overline{\lambda}_i - \rho(W).
\end{aligned}$$

By symmetry, $\overline{\lambda}_i \geq \lambda_i - \rho(W)$ hence the result

Controlling perturbation of eigenvectors

Lemma

Let $\Delta := \inf_{\overline{\lambda}_i \neq \overline{\lambda}_j} |\overline{\lambda}_i - \overline{\lambda}_j|$*. Assume* $\rho(W) < \Delta/2$*. Then for any normed eigenvector* x_i *of* A *associated with* λ_i *there exists* $\overline{x}_i$ *normed eigenvector of* $\overline{A}$ *associated with* $\overline{\lambda}_i$ *such that*

$$\langle x_i, \overline{x}_i \rangle \geq \sqrt{1 - \left(\frac{\rho(W)}{\Delta - \rho(W)} \right)^2}$$

Proof: Decomposition $x_i = \sum_j \theta_j \overline{x}_j$ yields
$Ax_i = \lambda_i x_i = \sum_j \theta_j \overline{\lambda}_j \overline{x}_j + W x_i$ hence $W x_i = \sum_j (\lambda_i - \overline{\lambda}_j) \theta_j \overline{x}_j$
By Weyl's inequality, $|\lambda_i - \overline{\lambda}_j| \geq \Delta - \rho(W)$ if $\overline{\lambda}_i \neq \overline{\lambda}_j$. Thus

$$\rho(W) \geq (\Delta - \rho(W)) \sqrt{1 - \sum_{k : \overline{\lambda}_k = \overline{\lambda}_i} |\theta_k|^2}$$

$$\Rightarrow \sum_{k : \overline{\lambda}_k = \overline{\lambda}_i} |\theta_k|^2 \geq 1 - \left(\frac{\rho(W)}{\Delta - \rho(W)} \right)^2$$

Summary of argument

- Matrix $\overline{A}$ of rank R, spectral gaps $|\overline{\lambda}_i - \overline{\lambda}_j| = \Omega(d)$, R-dimensional spectral embedding with $\overline{x}_1, \ldots, \overline{x}_R$ reveals clusters
- Assuming $\rho = \rho(A - \overline{A}) << d$, Weyl's inequality: R eigenvalues λ_i close to $\overline{\lambda}_i = \Omega(d)$, others of order $\rho << d$
- Associated eigenvectors x_i such that $\langle x_i, \overline{x}_i \rangle = 1 - O((\rho/d)^2)$

Then $\sum_{i\in[n]} ||z_i - \overline{z}_i||^2 = n \sum_{u \in [R]} ||x_u - \overline{x}_u||^2 = n\theta$ with $\theta = O((\rho/d)^2) = o(1)$
Hence (Tchebitchev inequality):
$|\{i : ||z_i - \overline{z}_i|| \geq \theta^{1/3}\}| \leq n\theta^{1/3} = o(n)$
Yields desired conclusion: except for vanishing fraction $\theta^{1/3}$ of nodes, spectral representatives z_i $\theta^{1/3}$-close of corresponding $\overline{z}_i$, themselves clustered according to community structure

Bounding $\rho = \rho(W)$

Lemma

Let $W \in \mathbb{R}^{n\times n}$: symmetric matrix with entries independent up to symmetry, bounded by 1, and such that $\mathbb{E}(W_{ij}^2) \le O(d/n)$. Then for any fixed ϵ, with high probability, $\rho(W) \le O(\sqrt{d} n^{\epsilon})$.

Corollary

Under assumptions of main result, $d = n^{\delta}$ for some $\delta \in]0,1[$, then with high probability $\rho(W) = o(d)$.

Proof: Take $\epsilon < \delta/2$ to obtain $\rho(W) = O(n^{\delta/2+\epsilon}) = o(n^{\delta})$.

For fixed $k \in \mathbb{N}$, write $\rho^{2k} \leq \sum_{i\in[n]} \lambda_i(W)^{2k} = \mathrm{Trace}(W^{2k})$

Thus $\mathbb{P}(\rho \geq x) \leq x^{-2k}\mathbb{E}(\rho^{2k}) \leq x^{-2k}\mathbb{E}\mathrm{Trace}(W^{2k})$
Combinatorial view of trace:

$$\mathrm{Trace}(W^{2k}) = \sum_{i_0^{2k}\in[n]^{2k+1}: i_0=i_{2k}} \prod_{j=1}^{2k} W_{i_{j-1}i_j}$$

Recall W_{ij}: centered and independent
$\rightarrow$ Only paths contributing non-zero expectation: traverse each edge at least twice

$$\Rightarrow \mathbb{E}\mathrm{Trace}(W^{2k}) \leq \sum_{e=1}^{k}\sum_{v=2}^{e+1} C(e,v)n^v O((d/n)^e)$$

Yields $\mathbb{E}\mathrm{Trace}(W^{2k}) = O(nd^k)$.
For $x = \sqrt{d}n^{\epsilon}$, yields $\mathbb{P}(\rho \geq x) \leq O(n^{1-2k\epsilon})$
Result follows by taking $k > 1/(2\epsilon)$

Stronger bounds on $\rho = \rho(W)$

Theorem (Feige and Ofek, 2005)

Let $A \in \mathbb{R}^{n \times n}$: symmetric matrix with entries independent up to symmetry, $A_{ij} \in [0,1]$, and such that $\mathbb{E}(A_{ij}) \leq d/n$, where $d \leq n^{1/5}$.
Then for some (universal) constant $\kappa > 0$, with high probability $\rho(A - \mathbb{E}(A)) \leq \kappa\sqrt{\max(d, \log(n))}$.

Hence result on spectral methods for Stochastic Block Model still valid as long as $d >> \sqrt{\max(d, \log(n))}$, i.e. $d >> \sqrt{\log(n)}$.

Takeaway messages

- Community detection a generic inference problem with many applications
- Basic spectral methods successful in scenarios well described by stochastic block model with strong enough signal and fixed number of large communities
- Variants with improved efficiency under weaker signal and in more difficult scenarios (many communities, small communities, overlapping communities,...) subject of ongoing research