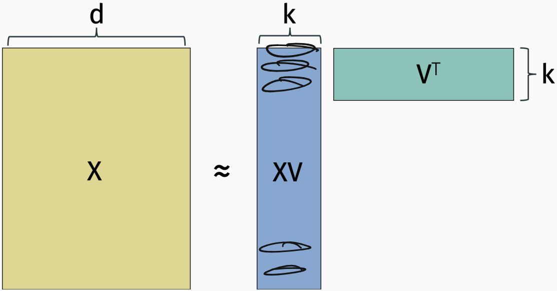
CS-GY 9223 I: Lecture 10
Spectral clustering, spectral graph theory.

NYU Tandon School of Engineering, Prof. Christopher Musco

- Project proposal feedback.
- Problem set.

LOW-RANK APPROXIMATION

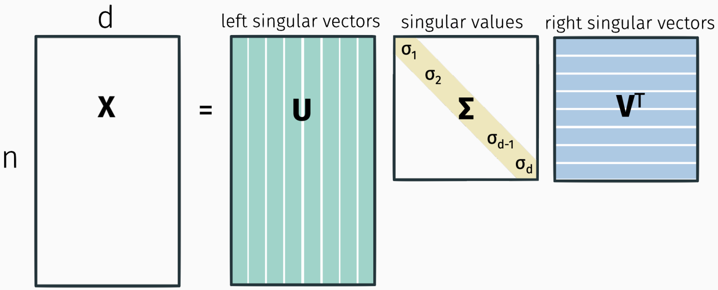
Write **X** as a rank k factorization by projecting onto the subspace spanned by an orthonormal matrix $\mathbf{V} \in \mathbb{R}^{d \times k}$



SINGULAR VALUE DECOMPOSITION

One-stop shop for computing optimal low-rank approximations.

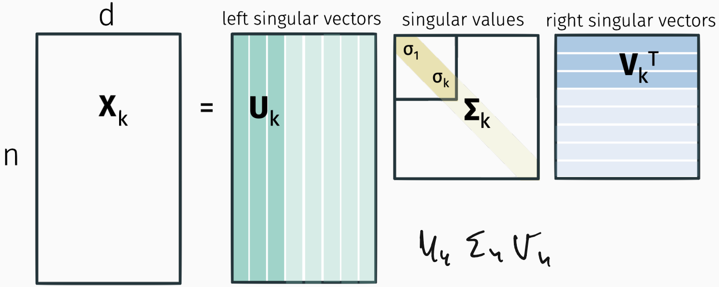
Any matrix $\mathbf{X}$ can be written:



Where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, and $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d \geq 0$.

SINGULAR VALUE DECOMPOSITION

Can read off optimal low-rank approximations from the SVD:



$U_k \Sigma_k V_k$

$$X_k = U_k U_k^T X = \boxed{X V_k V_k^T.}$$

$$V_k = \underset{\text{orthonormal } V \in \mathbb{R}^{d \times k}}{\arg\min} \|X - XVV^T\|_F^2 = \underset{\text{orthonormal } V \in \mathbb{R}^{d \times k}}{\arg\max} \|XVV^T\|_F^2$$

POWER METHOD CONVERGENCE

Theorem (Power Method Convergence)

Let $\gamma = \frac{\sigma_1 - \sigma_2}{\sigma_1}$ be parameter capturing the "gap" between the first and second largest singular values. If Power Method is initialized with a random Gaussian vector then, with high probability, after $T = O\left(\frac{\boxed{\log d/\epsilon}}{\boxed{\gamma}}\right)$ steps, we have:

$$\|\mathbf{v}_1 - \underline{\underline{\mathbf{z}^{(T)}}}\|_2 \leq \epsilon.$$

Total runtime: $O(T \cdot \mathrm{nnz}(\mathbf{X})) \leq \underline{\underline{O(T \cdot nd)}}$

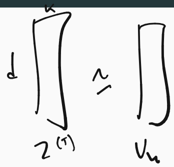
$\downarrow$

number of non-zeros

BLOCK POWER METHOD FOR LARGER k

Block power method:

$Z^T Z = I$



- Choose $\mathbf{G} \in \mathbb{R}^{d \times k}$ be a random Gaussian matrix.
- $\mathbf{Z}_0 = \text{orth}(\mathbf{G})$.
- For $i = 1, \ldots, T$
  - $\mathbf{Z}^{(i)} = \mathbf{X}^T \cdot (\mathbf{X}\mathbf{Z}^{(i-1)})$
  - $\mathbf{Z}^{(i)} = \underline{\text{orth}(\mathbf{Z}^{(i)})}$

  Return $\mathbf{Z}^{(T)}$

$qr(Z)$

$\underline{\underline{Z^{(T)}}} \approx V_k$

$\approx O\left(\frac{1}{\epsilon}\right)$

Convergence Guarantee: $T = O\left(\frac{\log d/\epsilon}{\epsilon}\right)$ iterations to obtain a nearly optimal low-rank approximation:

$$\underline{\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T}\|_F^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}\mathbf{V_k}\mathbf{V_k}^T\|_F^2.$$

Runtime: $O(\text{nnz}(\mathbf{X}) \cdot k \cdot T) \leq O(nd\underline{\underline{k}} \cdot T)$.

number of nonzeros in X

KRYLOV METHODS

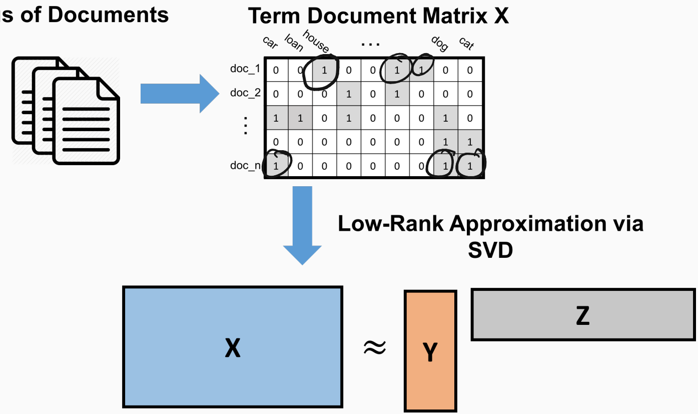
Possible to "accelerate" these methods. $1/\sqrt{\varepsilon}$

Convergence Guarantee: $T = O\left(\frac{\log d/\epsilon}{\sqrt{\epsilon}}\right)$ iterations to obtain a nearly optimal low-rank approximation:

$$\|\mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^T\|_F^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}\mathbf{V_k}\mathbf{V_k}^T\|_F^2.$$

Runtime: $O(\text{nnz}(\mathbf{X}) \cdot k \cdot T) \leq O(ndk \cdot T)$.

APPLICATION OF PARTIAL SVD: ENTITY EMBEDDINGS



Corpus of Documents

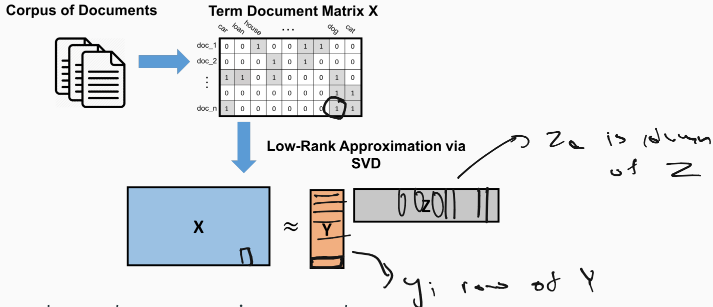
Term Document Matrix X

	car	loan	house	…				dog	cat
doc_1	0	0	1	0	0	1	1	0	0
doc_2	0	0	0	1	0	1	0	0	0
⋮	1	1	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	1	1
doc_n	1	0	0	0	0	0	0	1	1

Low-Rank Approximation via SVD

$$X \approx Y \; Z$$

EXAMPLE: LATENT SEMANTIC ANALYSIS



Corpus of Documents

Term Document Matrix X

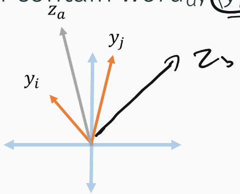
	car	loan	house	…				dog	cat
doc_1	0	0	1	0	0	1	1	0	0
doc_2	0	0	0	1	0	1	0	0	0
⋮	1	1	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	1	1
doc_n	1	0	0	0	0	0	0	1	1

Low-Rank Approximation via SVD

z_a is column of Z

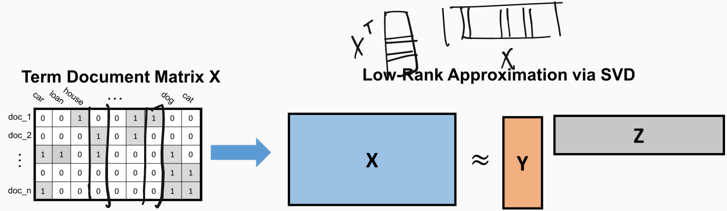
y_i row of Y

- $\langle \vec{y}_i, \vec{z}_a \rangle \approx 1$ when doc_i contains $word_a$.
- If doc_i and doc_j both contain $word_a$, $\langle \vec{y}_i, \vec{z}_a \rangle \approx \langle \vec{y}_j, \vec{z}_a \rangle = 1$.



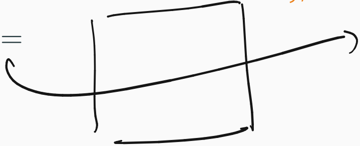
EXAMPLE: LATENT SEMANTIC ANALYSIS

XX^T



	car	loan	house		…			dog	cat
doc_1	0	0	1	0	0	1	1	0	0
doc_2	0	0	0	1	0	1	0	0	0
⋮	1	1	0	1	0	0	0	1	0
	0	0	0	0	0	0	0	1	1
doc_n	1	0	0	0	0	0	0	1	1

- The columns $\vec{z}_1, \vec{z}_2, \ldots$ give representations of words, with $\vec{z}_i$ and $\vec{z}_j$ tending to have high dot product if *word*$_i$ and *word*$_j$ appear in many of the same documents.
- **Z** corresponds to the top k right singular vectors: the eigenvectors of $\mathbf{X}\mathbf{X}^T$. Intuitively, what is $\mathbf{X}\mathbf{X}^T$?
- $(\mathbf{X}\mathbf{X}^T)_{i,j} =$



$= \langle x_i, x_j \rangle$

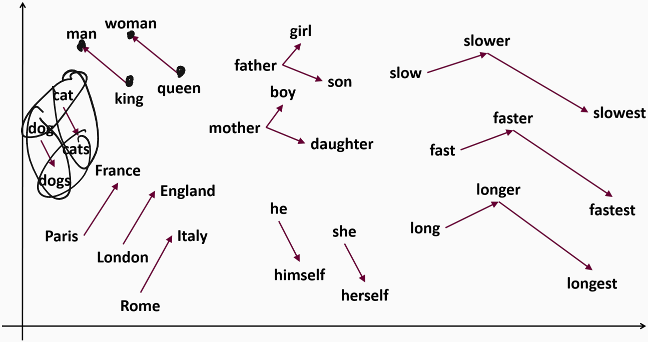
for i, j column of X

EXAMPLE: WORD EMBEDDINGS

Not obvious how to convert a word into a feature vector that captures the meaning of that word. Approach suggested by LSA: build a $d \times d$ symmetric "similarity matrix" $\mathbf{M}$ between words, and factorize: $\mathbf{M} \approx \mathbf{F}\mathbf{F}^T$ for rank k $\mathbf{F}$.

- **Similarity measures:** How often do $word_i, word_j$ appear in the same sentence, in the same window of w words, in similar positions of documents in different languages?
- Replacing $\mathbf{X}\mathbf{X}^T$ with these different metrics (sometimes appropriately transformed) leads to popular word embedding algorithms: `word2vec`, `GloVe`, etc.

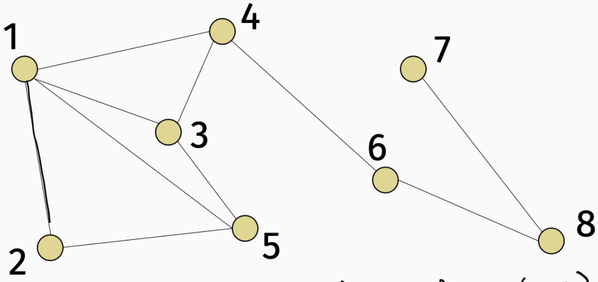
EXAMPLE: WORD EMBEDDINGS



`word2vec` was originally described as a neural-network method, but Levy and Goldberg show that it is simply low-rank approximation of a specific similarity matrix. (*Neural word embedding as implicit matrix factorization.*)

SPECTRAL GRAPH THEORY

Main idea: Understand graph data by constructing natural matrix representations, and studying that matrix's spectrum (eigenvalues/eigenvectors).



1, ... 8

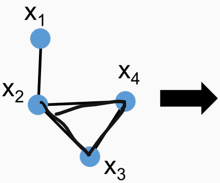
(1, 2), (2, 5) ... (7, 8)

For now assume $G = (V, E)$ is an undirected, unweighted graph with n nodes.

MATRIX REPRESENTATIONS OF GRAPHS

Two most common representations: $n \times n$ adjacency matrix $\mathbf{A}$ and graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is the diagonal degree matrix.

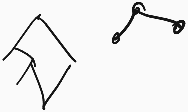
$$\begin{bmatrix} 1 & & & \\ & 1/\sqrt{3} & & \\ & & 1/\sqrt{2} & \\ & & & 1/\sqrt{2} \end{bmatrix}$$



$$\overset{\boxed{\mathbf{D}}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}} - \overset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}} = \overset{\mathbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}}$$

Also common to look at normalized versions of both of these: $\bar{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ and $\bar{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.

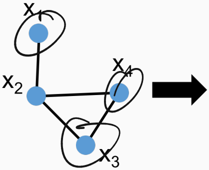
SPECTRAL GRAPH THEORY TIDBITS



- If L have k eigenvalues equal to 0, then G has k connected components.
- Sum of cubes of A's eigenvalues is equal to number of triangles in the graph.
- Sum of eigenvalues to the power q is the number of q cycles.

THE LAPLACIAN VIEW

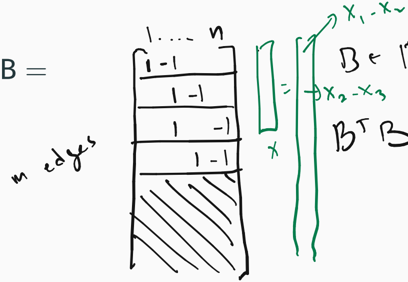
$$[B^T B]_{ii} = [B^{(i)T} B^{(i)}] = \|B^{(i)}\|_2^2$$



$$\underset{\textbf{D}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}} - \underset{\textbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}} = \underset{\textbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$\mathbf{L} = \mathbf{B}^T\mathbf{B}$ where B is the signed "edge-vertex incidence" matrix.

$\mathbf{B} =$

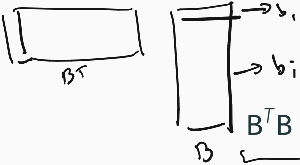


$B \in \mathbb{R}^{m \times n}$

$B^T B \quad (n \times m) \times (m \times n)$

$= (n \times n)$

THE LAPLACIAN VIEW



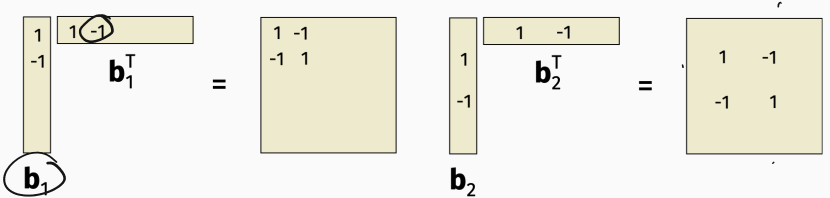
PSD

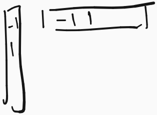
$$x^T B^T B x = x^T L x$$

$$= \|Bx\|_2^2$$

$$\underbrace{B^TB = b_1b_1^T + b_2b_2^T + \ldots + b_m b_m^T,}$$

where b_i is the i^{th} row of B (each row corresponds to a single edge).





$$= \cancel{-1 \cdot -1} \cdot b_1 b_1^T$$

$$= b_1 b_1^T$$

Conclusions from $\mathbf{L} = \mathbf{B}^T\mathbf{B}$

- $\mathbf{L}$ is positive semidefinite: $\mathbf{x}^T\mathbf{L}\mathbf{x} \geq 0$ for all $\mathbf{x}$.
- $\mathbf{L} = \underline{\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T}$ where $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ is $\mathbf{B}$'s SVD. Columns of $\mathbf{V}$ are eigenvectors of $\mathbf{L}$.

$V \Sigma U^T U \Sigma V^T$

- For any vector $\mathbf{x} \in \mathbb{R}^n$,

"smoothness of x"

$$\mathbf{x}^T L \mathbf{x} = \sum_{(i,j)\in E} (\mathbf{x}(i) - \mathbf{x}(j))^2.$$

$x \in \mathbb{B}^n$

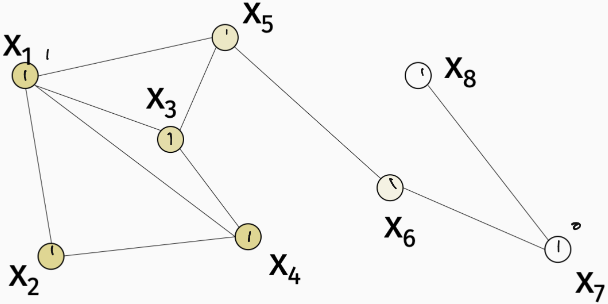
$= \|Bx\|_2^2$

$x^T L x$

THE LAPLACIAN VIEW

$\mathbf{x}^T L \mathbf{x}$ is small if $\mathbf{x}$ is a "smooth" function with respect to the graph.

$$x^T L x = 0 \quad \text{if} \quad x = \underline{\vec{1}}$$



Eigenvectors of the Laplacian with <u>small eigenvalues</u> correspond to <u>smooth functions</u> over the graph.

SMALLEST LAPLACIAN EIGENVECTOR

Courant–Fischer min-max principle

Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ be the eigenvectors of $\mathbf{L}$.

$$\boxed{\mathbf{v}_n = \mathop{\arg\min}_{\|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{L}\mathbf{v}}$$

$$\mathbf{v}_{n-1} = \mathop{\arg\min}_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_n} \underline{\mathbf{v}^T\mathbf{L}\mathbf{v}}$$

$$\mathbf{v}_{n-2} = \mathop{\arg\min}_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_n, \mathbf{v}_{n-1}} \mathbf{v}^T\mathbf{L}\mathbf{v}$$

$$\vdots$$

$$\mathbf{v}_1 = \mathop{\arg\min}_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_n, \ldots, \mathbf{v}_2} \mathbf{v}^T\mathbf{L}\mathbf{v}$$

LARGEST LAPLACIAN EIGENVECTOR

Courant–Fischer min-max principle

Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ be the eigenvectors of $\mathbf{L}$.

$$\boxed{\underline{\mathbf{v}_1} = \arg\max_{\|\mathbf{v}\|=1} \underline{\mathbf{v}^T \mathbf{L} \mathbf{v}}}$$

$$\mathbf{v}_2 = \arg\max_{\|\mathbf{v}\|=1, \underline{\mathbf{v}} \perp \enclose{circle}{\mathbf{v}_1}} \enclose{circle}{\mathbf{v}^T \mathbf{L} \mathbf{v}}$$

$$\mathbf{v}_3 = \arg\max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \mathbf{v}_2} \underline{\mathbf{v}^T \mathbf{L} \mathbf{v}}$$

$$\vdots$$

$$\mathbf{v}_n = \arg\max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

EXAMPLE APPLICATION OF SPECTRAL GRAPH THEORY

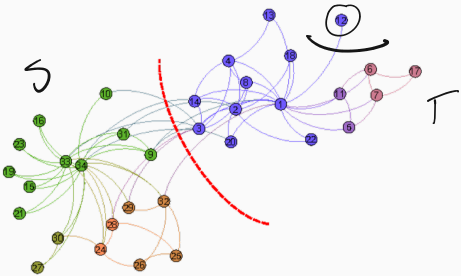
- Study graph partitioning problem important in 1) understanding social networks 2) nonlinear clustering in unsupervised machine learning (spectral clustering).
- See how this problem can be solved approximately using Laplacian eigenvectors.
- Give a full analysis of the method for a common random graph model.
- Use two tools: matrix concentration and eigenvector perturbation bounds.

BALANCED CUT

Common goal: Given a graph $G = (V, E)$, partition nodes along a cut that:

$\nearrow V \setminus S$

- Has few crossing edges: $|\{\underline{(u, v) \in E} : u \in \boxed{S}, v \in T\}|$ is small.
- Separates large partitions: $|S|, |T|$ are not too small.



(a) Zachary Karate Club Graph

Important in understanding community structure in social networks.

SOCIAL NETWORKS IN THE 1970S

Wayne W. Zachary (1977). An Information Flow Model for Conflict and Fission in Small Groups.

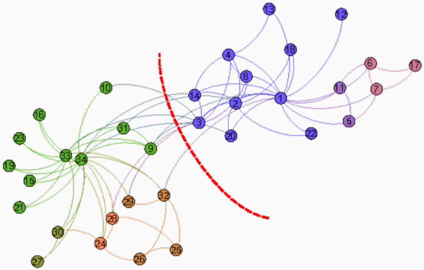
“The network captures 34 members of a karate club, documenting links between pairs of members who interacted outside the club. During the study a conflict arose between the administrator ”John A” and instructor ”Mr. Hi” (pseudonyms), which led to the split of the club into two. Half of the members formed a new club around Mr. Hi; members from the other part found a new instructor or gave up karate. Based on collected data Zachary correctly assigned all but one member of the club to the groups they actually joined after the split.” – Wikipedia

Beautiful paper – definitely worth checking out!

BALANCED CUT

Common goal: Given a graph $G = (V, E)$, partition nodes along a cut that:

- Has few crossing edges: $|\{(u, v) \in E : u \in S, v \in T\}|$ is small.
- Separates large partitions: $|S|, |T|$ are not too small.

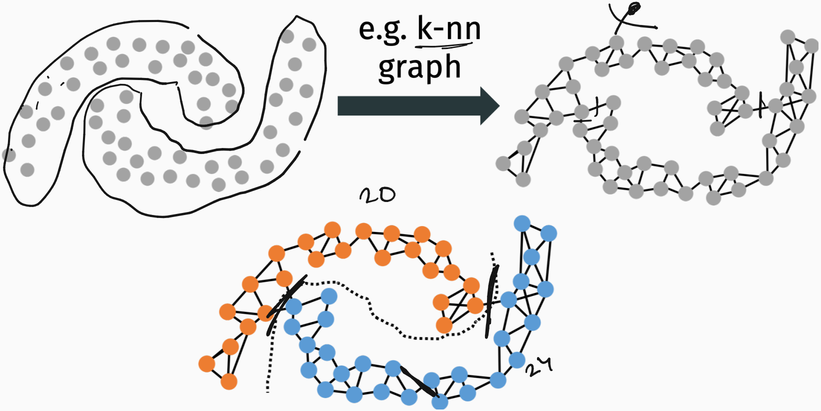


(a) Zachary Karate Club Graph

Important in understanding community structure in social networks.

SPECTRAL CLUSTERING

Idea: Construct synthetic graph for data that is hard to cluster.



Spectral Clustering, Laplacian Eigenmaps, Locally linear embedding, Isomap, etc.

SPECTRAL GRAPH THEORY

There are many way's to formalize Zachary's problem:

Sparsest Cut:

$$\min_S \frac{\overset{\nearrow 3}{[\operatorname{cut}(S, V \setminus S)]}}{\min\left(|S|, |V \setminus S|\right)} \to 20$$

$|S|$

β-Balanced Cut:

$\beta \leq 1/2$

$$\min_S \operatorname{cut}(S, V \setminus S) \qquad \text{such that} \qquad \min\left(|S|, |V \setminus S|\right) \geq \beta n$$

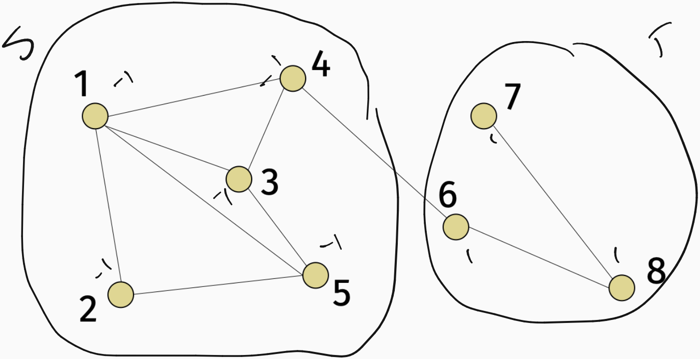
Most formalizations lead to computationally hard problems. Lots of interest in designing polynomial time approximation algorithms, but tend to be slow. In practice, much simpler methods based on the graph spectrum are used.

THE LAPLACIAN VIEW

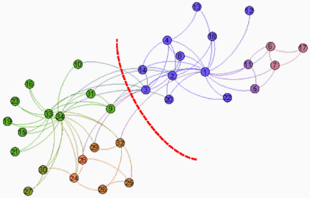
Another conclusion from $\mathbf{L} = \mathbf{B}^T\mathbf{B}$:

For a <u>cut indicator vector</u> $\mathbf{c} \in \{-1, 1\}^n$ with $\mathbf{c}(i) = -1$ for $\boxed{i \in S}$ and $\mathbf{c}(i) = 1$ for $i \in \underline{T} = V \setminus S$:

$$\underline{\mathbf{c}^T L \mathbf{c} = \sum_{(i,j)\in E} (\mathbf{c}(i) - \mathbf{c}(j))^2} = \boxed{4 \cdot \mathrm{cut}(S, T).} \tag{1}$$



THE LAPLACIAN VIEW



(a) Zachary Karate Club Graph

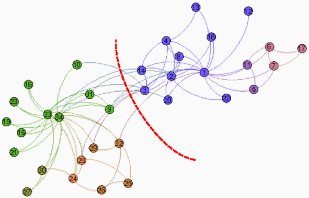
For a <u>cut indicator vector</u> $\mathbf{c} \in \{-1, 1\}^n$ with $\mathbf{c}(i) = -1$ for $i \in S$ and $\mathbf{c}(i) = 1$ for $i \in T$:

- $\mathbf{c}^T L \mathbf{c} = 4 \cdot cut(S, T)$.
- $\left|\mathbf{c}^T \vec{\mathbf{1}}\right| = \left||T| - |S|.\right|$

$\mathbf{c}^T \mathbf{1} = \sum_{i=1}^{n} c_i$

Want to minimize both $\mathbf{c}^T L \mathbf{c}$ (cut size) and $\mathbf{c}^T \mathbf{1}$ (imbalance).

THE LAPLACIAN VIEW



(a) Zachary Karate Club Graph

Equivalent formulation if we divide everything by $\sqrt{n}$ so that $\mathbf{c}$ has norm 1. Then $\mathbf{c} \in \{-\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}\}^n$ and:

- $\mathbf{c}^T L \mathbf{c} = \frac{4}{n} \cdot cut(S, T)$.
- $\mathbf{c}^T \mathbf{1} = \frac{1}{\sqrt{n}}(|T| - |S|)$.

$\|c\|_2 = 1$

Want to minimize both $\mathbf{c}^T L \mathbf{c}$ (cut size) and $\mathbf{c}^T \mathbf{1}$ (imbalance).

SMALLEST LAPLACIAN EIGENVECTOR

The smallest eigenvector/singular vector $\mathbf{v}_n$ satisfies:

$$\underline{\underline{\mathbf{v}_n}} = \underline{\frac{1}{\sqrt{n}} \cdot \mathbf{1} =} \mathop{\arg\min}_{v \in \mathbb{R}^n \text{ with } \|\mathbf{v}\|=1} \mathbf{v}^T L \mathbf{v}$$

with $\mathbf{v}_n^T L \mathbf{v}_n = 0$.

SECOND SMALLEST LAPLACIAN EIGENVECTOR

By Courant-Fischer, $\mathbf{v}_{n-1}$ is given by:

$$\underline{\mathbf{v}_{n-1}} = \underset{\|\mathbf{v}\|=1,\ \underline{\mathbf{v}_n^T}\mathbf{v}=0}{\arg\min} \mathbf{v}^T L \mathbf{v}$$

If $\mathbf{v}_{n-1}$ were binary $\{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}^n$ it would have:

- $\mathbf{v}_{n-1}^T L \mathbf{v}_{n-1} = \frac{4}{n} \text{cut}(S, T)$ as small as possible given that $\mathbf{v}_{n-1}^T \mathbf{1} = |T| - |S| = 0.$
- $\mathbf{v}_{n-1}$ would indicate the smallest perfectly balanced cut.

$\mathbf{v}_{n-1} \in \mathbb{R}^n$ is not generally binary, but a natural approach is to 'round' the vector to obtain a cut.

CUTTING WITH THE SECOND LAPLACIAN EIGENVECTOR

Find a good partition of the graph by computing

$$\mathbf{v}_{n-1} = \underset{\mathbf{v}\in\mathbb{R}^n \text{ with } \|\mathbf{v}\|=1,\ \mathbf{v}^T\mathbf{1}=0}{\arg\min} \mathbf{v}^T \underline{L}\mathbf{v}$$

$\mathbf{1} \to v_n$

$V \Lambda V^T$

$P \Lambda P^{-1}$

$v_{n-1}^T v_n = 0$

Set S to be all nodes with $\mathbf{v}_{n-1}(i) < 0$, and T to be all with $\mathbf{v}_{n-1}(i) \geq 0$.

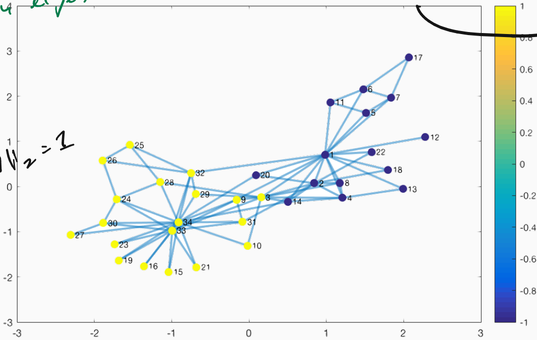
$\boxed{v^T L v = 0}$

Assume v is an eigenvector:

$$Lv = \lambda v$$

$$v^T \lambda v = \lambda v^T v = \boxed{\lambda}$$

$\|v\|_2 = 1$



$$\to v^T B^T B v = \|Bv\|_2^2 \geq 0$$

L is PSD

$x = 1$

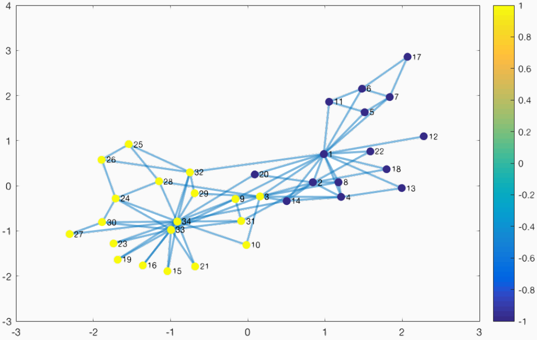
$$x^T L x = x^T B^T B x = \sum_{(i,j)\in E} (x(i) - x(j))^2$$

CUTTING WITH THE SECOND LAPLACIAN EIGENVECTOR

Find a good partition of the graph by computing

$$\mathbf{v}_{n-1} = \underset{v \in \mathbb{R}^n \text{ with } \|\mathbf{v}\|=1,\ \mathbf{v}^T\mathbf{1}=0}{\arg\min} \mathbf{v}^T L \mathbf{v}$$

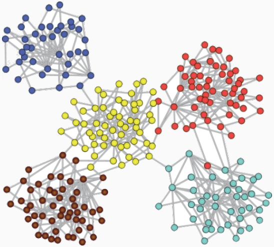
Set S to be all nodes with $\mathbf{v}_{n-1}(i) < 0$, and T to be all with $\mathbf{v}_{n-1}(i) \geq 0$.



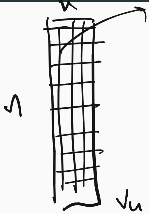
SPECTRAL PARTITIONING IN PRACTICE

The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\bar{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.

Important consideration: What to do when we want to split the graph into more than two parts?



SPECTRAL PARTITIONING IN PRACTICE



$\underline{v_{n-1}} \quad \underline{v_{n-2}} \quad \ldots v_{n-k}$

$V_n = \mathbb{1}$

Spectral Clustering:

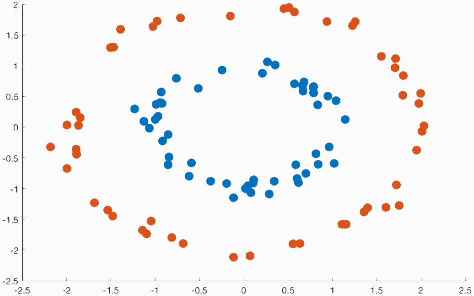
- Compute smallest k eigenvectors $\mathbf{v}_{n-1}, \ldots, \mathbf{v}_{n-k}$ of $\mathbf{L}$.
- Represent each node by its corresponding row in $\mathbf{V} \in \mathbb{R}^{n \times k}$ whose rows are $\mathbf{v}_{n-1}, \ldots \mathbf{v}_{n-k}$.
- Cluster these rows using k-means clustering (or really any clustering method).

$\min_{v} \; v^T L v \qquad \vec{1}^T L \vec{1} = \sum_{(i,j)\in E} (\overset{\to 0}{1_i - 1_j})^2 = 0$

$v^T L v = 0$

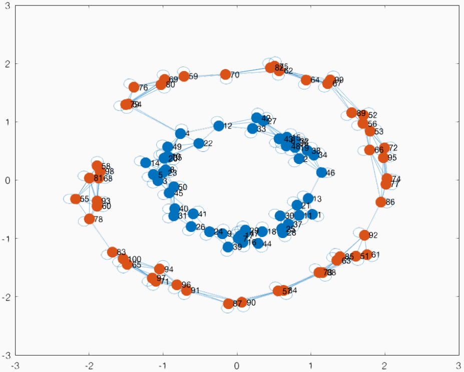
LAPLACIAN EMBEDDING

Original Data: (not linearly separable)



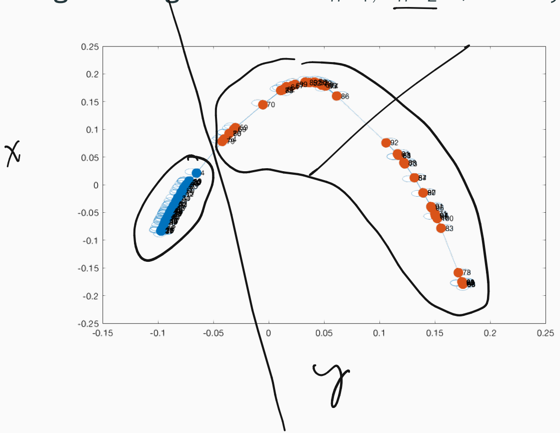
LAPLACIAN EMBEDDING

k-Nearest Neighbors Graph:



V_{n-1}

Embedding with eigenvectors $\mathbf{v}_{n-1}, \underline{\mathbf{v}_{n-2}}$: (linearly separable)

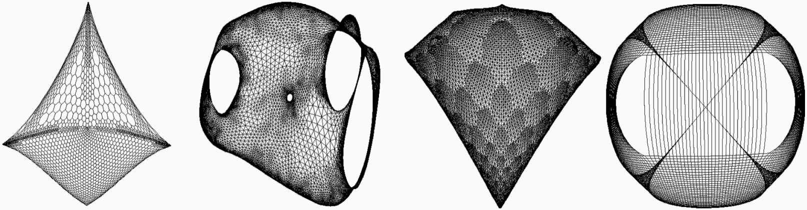


WHY DOES THIS WORK?

Intuitively, since $\mathbf{v} \in \mathbf{v}_1, \ldots \mathbf{v}_k$ are smooth over the graph,

$$\sum_{i,j \in E} (\mathbf{v}[i] - \mathbf{v}[j])^2$$

is small for each coordinate. I.e. this embedding explicitly encourages nodes connected by an edge to be placed in nearby locations in the embedding.



Also useful e.g., in graph drawing.

GENERATIVE MODELS

So far: Showed that spectral clustering partitions a graph along a small cut between large pieces.

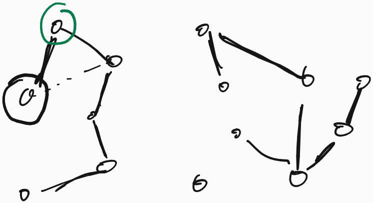
- No formal guarantee on the 'quality' of the partitioning.
- Would be difficult to analyze for general input graphs.

Common approach: Design a natural generative model that produces random but realistic inputs and analyze how the algorithm performs on inputs drawn from this model.

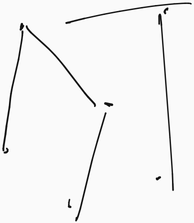
- Very common in algorithm design and analysis. Great way to start approaching a problem.
- This is also the whole idea behind Bayesian Machine Learning (can be used to justify ℓ_2 linear regression, k-means clustering, PCA, etc.)

STOCHASTIC BLOCK MODEL

Ideas for a generative model for **social network graphs** that would allow us to understand partitioning?



Erdős - Rényi



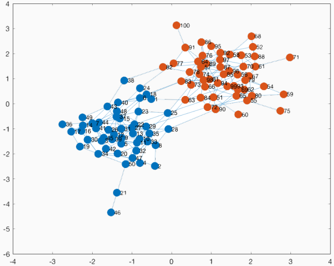
$p \in [0,1]$

STOCHASTIC BLOCK MODEL

Stochastic Block Model (Planted Partition Model):

Let $\underline{G_n(p, \overset{\in [0, 1]}{q})}$ be a distribution over graphs on n nodes, split equally into two groups $\underline{B}$ and $\underline{C}$, each with $n/2$ nodes.

- Any two nodes in the same group are connected with probability p (including self-loops).
- Any two nodes in different groups are connected with prob $\boxed{q < p}$.

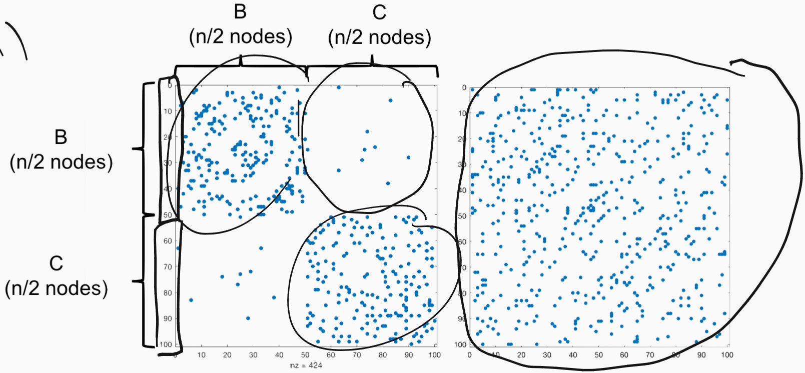


LINEAR ALGEBRAIC VIEW

Let G be a stochastic block model graph drawn from $G_n(p, q)$.

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G. What is $\mathbb{E}[\mathbf{A}]$?

spy(A)



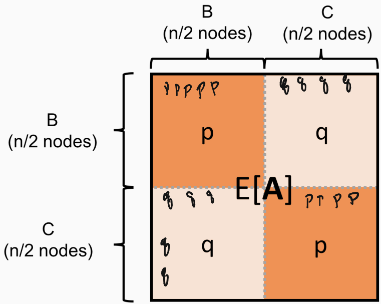
Note that we are arbitrarily ordering the nodes in A by group. In reality A would look "scrambled" as on the right.

EXPECTED ADJACENCY SPECTRUM

$L = D - A \quad A = D - L \quad \underline{\mathbb{E}[A]} = \mathbb{E}[D] - \underline{\mathbb{E}[L]}$

Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i, j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.

$$\begin{bmatrix} \left(\frac{p+q}{2}\right)n & & \\ & \ddots & \\ & & \ddots \end{bmatrix}$$



What are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

EXPECTED LAPLACIAN

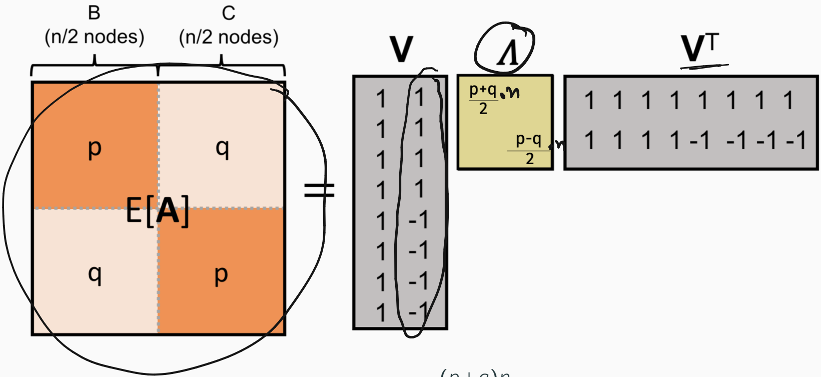
What is the expected Laplacian $G_n(p, q)$?

A and **L** have the same eigenvectors and eigenvalue are equal up to a shift.

EXPECTED ADJACENCY SPECTRUM

Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

EXPECTED ADJACENCY SPECTRUM



- $\mathbf{v}_1 \sim \mathbf{1}$ with eigenvalue $\lambda_1 = \frac{(p+q)n}{2}$.
- $\mathbf{v}_2 \sim \boldsymbol{\chi}_{B,C}$ with eigenvalue $\lambda_2 = \frac{(p-q)n}{2}$.
- $\chi_{B,C}(i) = 1$ if $i \in B$ and $\chi_{B,C}(i) = -1$ for $i \in C$.

If we compute $\mathbf{v}_2$ then we recover the communities B and C!

EXPECTED LAPLACIAN SPECTRUM

Upshot: The second smallest eigenvector of $\mathbb{E}[\mathbf{L}]$, equivalently the second largest of $\mathbb{E}[\mathbf{A}]$, is $\chi_{B,C}$ – the indicator vector for the cut between the communities.

- If the random graph G (equivilantly $\mathbf{A}$ and $\mathbf{L}$) were exactly equal to its expectation, partitioning using this eigenvector would exactly recover communities B and C.

How do we show that a matrix (e.g., $\mathbf{A}$) is close to its expectation? **Matrix concentration inequalities.**

- Analogous to scalar concentration inequalities like Markovs, Chebyshevs, Bernsteins.

MATRIX CONCENTRATION

Matrix Concentration Inequality: If $\underline{p \geq} O\left(\frac{\log^4 n}{n}\right)$, then with high probability

$$\underline{\|\boxed{\mathbf{A} - \mathbb{E}[\mathbf{A}]}\|_2} \leq \underline{O(\sqrt{pn})}.$$

where $\|\cdot\|_2$ is the matrix spectral norm (operator norm).

For $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\|\mathbf{X}\|_2 = \max_{z \in \mathbb{R}^d : \|z\|_2 = 1} \|\mathbf{X}z\|_2$.

Exercise: Show that $\|\mathbf{X}\|_2$ is equal to the largest singular value of $\mathbf{X}$. For symmetric $\mathbf{X}$ (like $\mathbf{A} - \mathbb{E}[\mathbf{A}]$) show that it is equal to the magnitude of the largest magnitude eigenvalue.

For the stochastic block model application, we want to show that the second <u>eigenvectors</u> of $\mathbf{A}$ and $\mathbb{E}[\mathbf{A}]$ are close. How does this relate to their difference in spectral norm?

EIGENVECTOR PERTURBATION

Davis-Kahan Eigenvector Perturbation Theorem: Suppose $\mathbf{A}, \bar{\mathbf{A}} \in \mathbb{R}^{d \times d}$ are symmetric with $\underline{\|\mathbf{A} - \bar{\mathbf{A}}\|_2} \leq \epsilon$ and eigenvectors $v_1, v_2, \ldots, v_d$ and $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_d$. Letting $\theta(v_i, \bar{v}_i)$ denote the angle between v_i and $\bar{v}_i$, for all i:

$$\underline{\sin[\theta(v_i, \bar{v}_i)]} \leq \frac{\epsilon}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

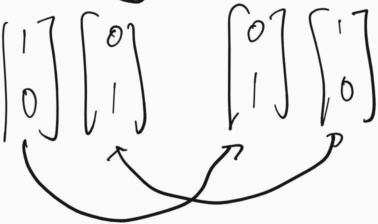
eigenvectors of A.

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\bar{\mathbf{A}}$.

The error gets larger if there are eigenvalues with similar magnitudes.

EIGENVECTOR PERTURBATION

$$\overset{\mathbf{A}}{\begin{bmatrix}1+\varepsilon & 0\\ 0 & 1\end{bmatrix}} - \overset{\bar{\mathbf{A}}}{\begin{bmatrix}1 & 0\\ 0 & 1+\varepsilon\end{bmatrix}} = \overset{\|\mathbf{A}-\bar{\mathbf{A}}\|_2 = \varepsilon}{\begin{bmatrix}\varepsilon & 0\\ 0 & \varepsilon\end{bmatrix}}$$



$$\begin{bmatrix}1\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\end{bmatrix} \qquad \begin{bmatrix}0\\1\end{bmatrix} \quad \begin{bmatrix}1\\0\end{bmatrix}$$

APPLICATION TO STOCHASTIC BLOCK MODEL

Claim 1 (Matrix Concentration): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\|_2 \leq O(\sqrt{pn}).$$

Claim 2 (Davis-Kahan): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p-q)n/2} = O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

Recall: $\mathbb{E}[\mathbf{A}]$, has eigenvalues $\lambda_1 = \frac{(p+q)n}{2}$, $\lambda_2 = \frac{(p-q)n}{2}$, $\lambda_i = 0$ for $i \geq 3$.

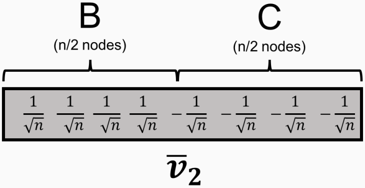
$$\min_{j \neq i} |\lambda_i - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$

Assume $\left|\frac{(p-q)n}{2} - 0\right|$ will be the minimum of the two gaps. I.e. smaller than $\left|\frac{(p+q)n}{2} - \frac{(p-q)n}{2}\right| = qn$.

APPLICATION TO STOCHASTIC BLOCK MODEL

So Far: $\sin\theta(v_2, \bar{v}_2) \leq O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$. What does this give us?

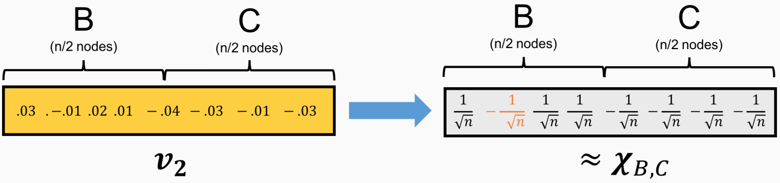
- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{v}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.


B
(n/2 nodes)
C
(n/2 nodes)
$\frac{1}{\sqrt{n}}$ $\frac{1}{\sqrt{n}}$ $\frac{1}{\sqrt{n}}$ $\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$
$\bar{v}_2$

- Every i where $v_2(i), \bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.
- So they differ in sign in at most $O\left(\frac{p}{(p-q)^2}\right)$ positions.

APPLICATION TO STOCHASTIC BLOCK MODEL

Upshot: If G is a stochastic block model graph with adjacency matrix $\mathbf{A}$, if we compute its second large eigenvector v_2 and assign nodes to communities according to the sign pattern of this vector, we will correctly assign all but $O\left(\frac{p}{(p-q)^2}\right)$ nodes.


B
(n/2 nodes)
C
(n/2 nodes)
.03 . −.01 .02 .01 −.04 −.03 −.01 −.03
v_2
B
(n/2 nodes)
C
(n/2 nodes)
$\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $\frac{1}{\sqrt{n}}$ $\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$ $-\frac{1}{\sqrt{n}}$
$\approx \chi_{B,C}$

- Why does the error increase as q gets close to p?
- Even when $p - q = O(1/\sqrt{n})$, assign all but an $O(n)$ fraction of nodes correctly. E.g., assign 99% of nodes correctly.