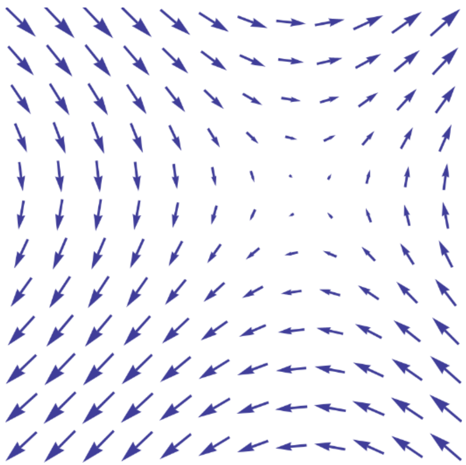
Spectral Clustering

Part 1: The Graph Laplacian

Ng Yen Kaow

Laplacian of a function

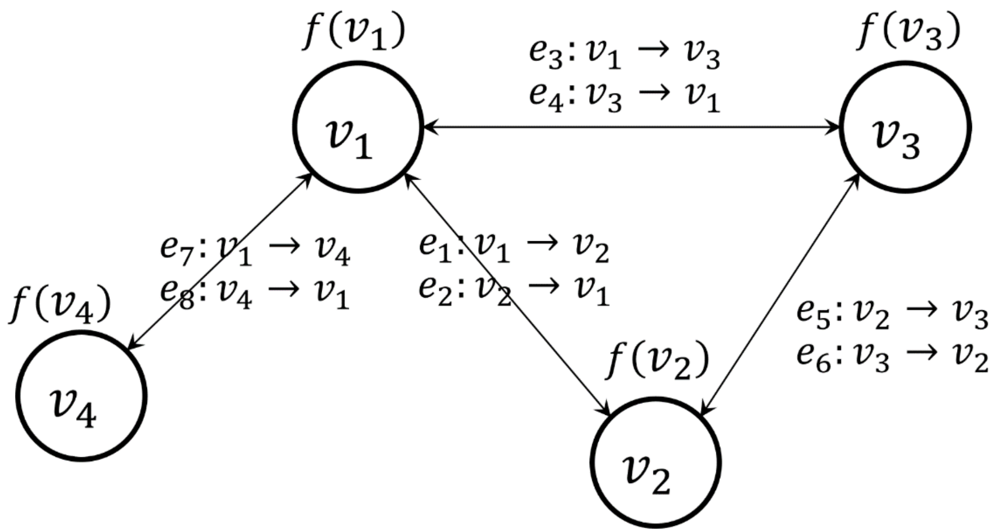
- Given a multivariate function $f: \mathbb{R}^n \to \mathbb{R}$
- $\nabla f(\boldsymbol{x})$, the gradient at $f(\boldsymbol{x})$, is a vector pointing at the steepest ascent of $f(\boldsymbol{x})$



Vector field ∇f

- Δf, the Laplacian of f, is the divergence of ∇f, that is, $\Delta f(\boldsymbol{x}) = \nabla \cdot \nabla f(\boldsymbol{x})$
  - A scalar measurement of the smoothness in $\nabla f(\boldsymbol{x})$ about point $\boldsymbol{x}$

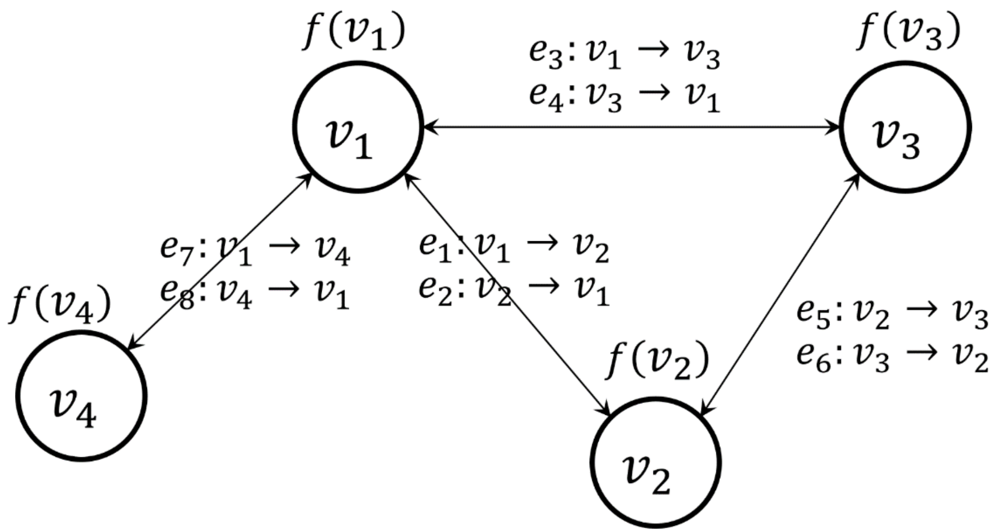
Incidence matrix



Let vector $f = \begin{bmatrix} f(v_1) \\ f(v_2) \\ f(v_3) \\ f(v_4) \end{bmatrix}$

- Consider each vertex as a point on the grid
  - The domain of f are now the vertices
  - $f(v)$ operates on each vertex v
  - The gradient from vertex v to v' is given by the edge $e: v \to v'$, more specifically, $f(v') - f(v)$
    - Denote the gradient of edge e as $w(e)$
- Define a matrix which captures all the gradients

Incidence matrix



Let vector $f = \begin{bmatrix} f(v_1) \\ f(v_2) \\ f(v_3) \\ f(v_4) \end{bmatrix}$

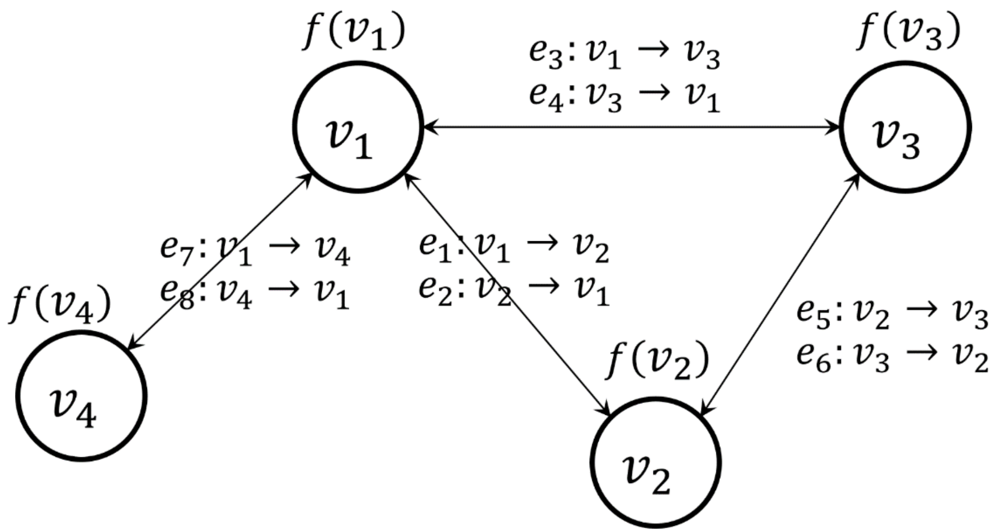
- Incidence matrix

$$M = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{array} \begin{array}{c} \begin{array}{cccccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 \end{array} \\ \begin{bmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 1 & -1 \\ -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array}$$

- Every column represents an edge in the graph

$$\underbrace{(M^\top)_1}_{\text{row 1 of } M} f = [1 \quad -1 \quad 0 \quad 0] \begin{bmatrix} f(v_1) \\ f(v_2) \\ f(v_3) \\ f(v_4) \end{bmatrix} = f(v_1) - f(v_2) = w(e_1)$$

Incidence matrix



Let vector $f = \begin{bmatrix} f(v_1) \\ f(v_2) \\ f(v_3) \\ f(v_4) \end{bmatrix}$

- Incidence matrix

$$M = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{array} \begin{array}{c} \begin{array}{cccccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 & e_8 \end{array} \\ \begin{bmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 1 & -1 \\ -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array}$$

- Every column represents an edge in the graph
- $M^\top f$ is a $|E| \times 1$ vector where each entry gives the gradient of an edge
  - $M^\top f$ contains all the gradients of the graph

The graph Laplacian L

- The graph Laplacian L is obtained by
$$\Delta f = \nabla \cdot \nabla f = MM^\top f$$
e.g.
$$(MM^\top f)_1 = \begin{bmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} w(e_1) \\ w(e_2) \\ w(e_3) \\ w(e_4) \\ w(e_5) \\ w(e_6) \\ w(e_7) \\ w(e_8) \end{bmatrix} = \underbrace{w(e_1) - w(e_2) + w(e_3) - w(e_4) + w(e_7) - w(e_8)}_{\text{divergence of vertex } v_1}$$
- $MM^\top f$ is a $|V| \times 1$ vector where each entry gives the divergence of a vertex
- $MM^\top$ is a $|V| \times |V|$ matrix where
$$MM^\top \begin{bmatrix} f(v_1) \\ f(v_2) \\ \vdots \end{bmatrix} = \begin{bmatrix} \Delta f(v_1) \\ \Delta f(v_2) \\ \vdots \end{bmatrix}$$

Properties of L

- The graph Laplacian L is obtained as $L = MM^\top$
  - Since L is of the form $MM^\top$, L **is symmetric** and **positive-semidefinite**
    - This allows us to obtain an orthogonal eigenbasis, which has special meanings (next slide)
  - $L = D - A$, where D is the degree matrix and A the adjacency matrix

Eigenvectors of L

- A eigenvector x of L fulfills $Lx = \lambda x$
- Compared with $Lx = \begin{bmatrix} \Delta f(v_1) \\ \vdots \end{bmatrix}$, we have $\lambda x = \begin{bmatrix} \Delta f(v_1) \\ \vdots \end{bmatrix}$
- The eigenvector x corresponds to the values $f(v)$ where $\lambda f(v) \approx \Delta f(v)$
  - **A small λ indicates that $f(v)$ does not vary much from $f(v')$ of its neighbors v'**
- The smallest λ (for a connected graph) is 0, indicating that $\forall v \Delta f(v) = 0$
  - In which case $f(v) = \text{const}$ (stationary state)
  - Graphs that are not fully connected will be discussed later

Eigenvectors of L

- A eigenvector $x = [f(v_1) \quad f(v_2) \quad \ldots]$ of L furthermore minimizes $\frac{x^\top L x}{x^\top x}$ (Rayleigh quotient)
- Since $Lx = \begin{bmatrix} \Delta f(v_1) \\ \vdots \end{bmatrix}$, we have

$$x^\top L x = [f(v_1) \quad \ldots] \begin{bmatrix} \Delta f(v_1) \\ \vdots \end{bmatrix} = \sum_v f(v) \Delta f(v)$$

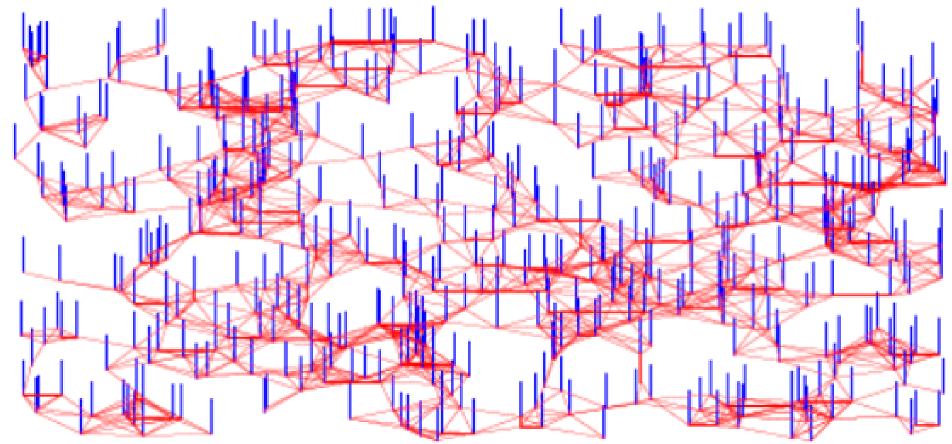
$\Rightarrow x^\top L x$ = projection of Δf on eigenvector x

$\Rightarrow \frac{x^\top L x}{x^\top x}$ = projection of Δf on unit eigenvector x

  - Furthermore the projection $\frac{x^\top L x}{x^\top x} = \lambda$ (eigenvalue of x)
- A eigenvector is **a distribution f that minimizes the total differences between neighboring $f(v)$ values**

Eigenvectors of L

- A eigenvector = a distribution f that minimizes the total differences between neighboring $f(v)$ values
- $f(v)$ values from eigenvector of $\lambda = 0$
  - $f(v) = \text{const}$
    $\Rightarrow$ zero differences

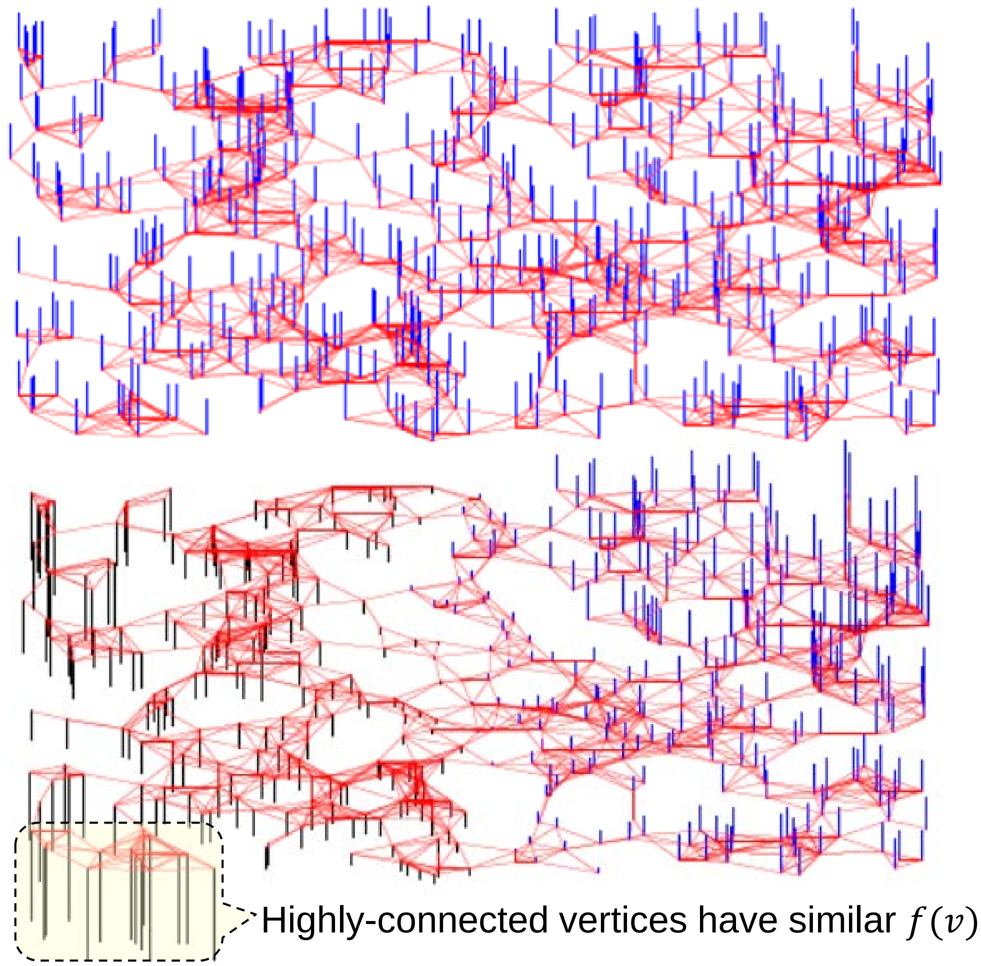


From Shuman *et al.* “The emerging field of signal processing on graphs”, 2013

- If the graph consists of two disconnected components, the $f(v)$ values of the individual components can have different constant values

Eigenvectors of L

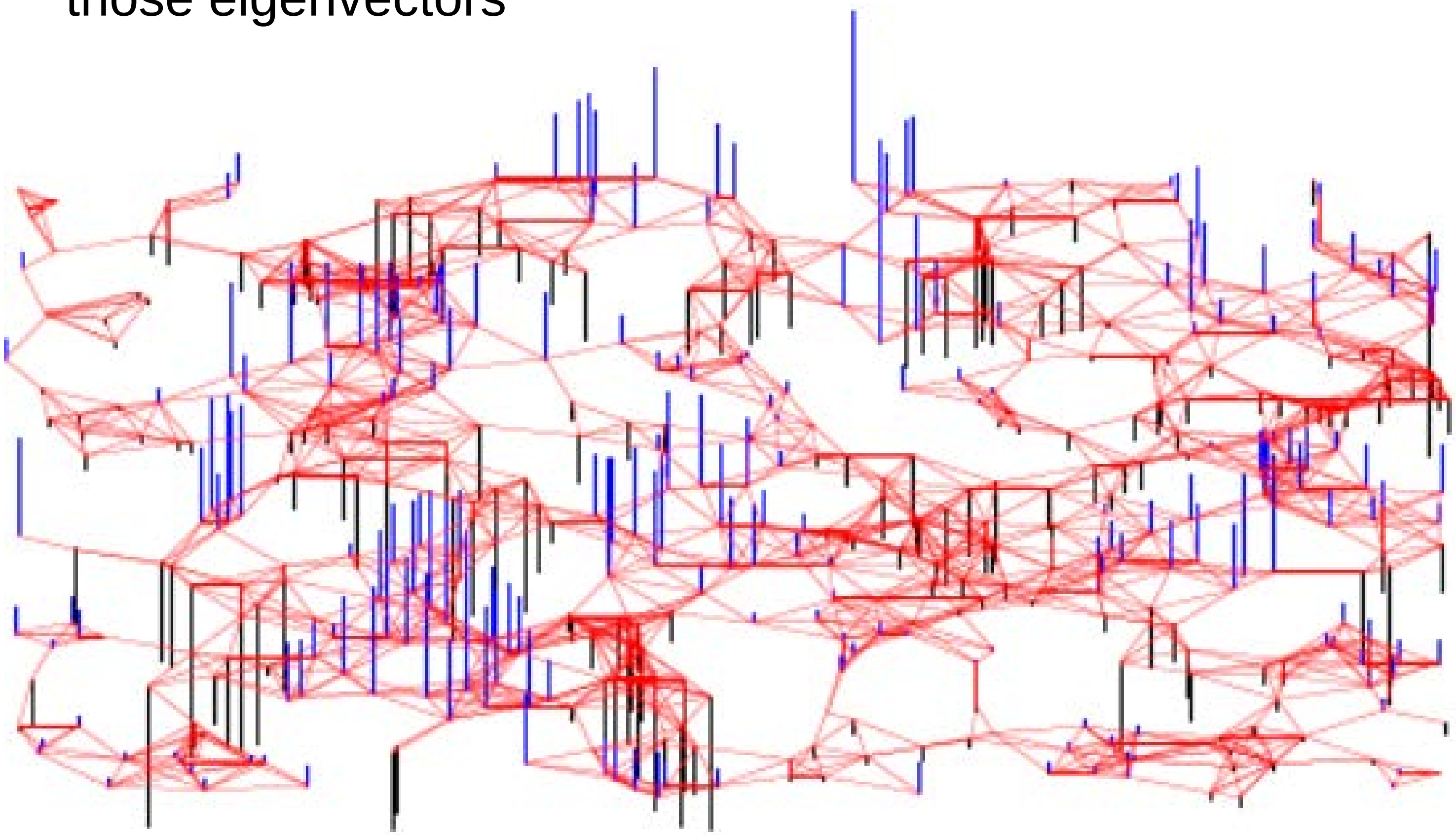
- A eigenvector = a distribution f that minimizes the total differences between neighboring $f(v)$ values
- $f(v)$ values from eigenvector of $\lambda = 0$
  - $f(v) = \text{const}$
    $\Rightarrow$ zero differences
- $f(v)$ values for eigenvector of 2nd smallest λ
  - Orthogonality with eigenvector of $\lambda = 0$ forces large variations in $f(v)$



From Shuman *et al.* "The emerging field of signal processing on graphs", 2013

Eigenvectors of L

- $f(v)$ values from eigenvector of 50th smallest λ
  - Orthogonality of this eigenvector with the 1st~49th smallest eigenvectors forces distinctly different variations in $f(v)$ from those eigenvectors



From Shuman *et al.* “The emerging field of signal processing on graphs”, 2013

Mathematical property of L

- A precise mathematical property of L relates it to "sparsest cut" problems
- Let the adjacency matrix $A = (a_{ij})$, then

$$x^\top L x = \frac{1}{2}\sum_{i,j=1}^{m} a_{ij}\left(x_i - x_j\right)^2$$

$$\boxed{\begin{aligned} x^\top L x &= x^\top D x - x^\top A x = \sum_{i=1}^{m} d_i x_i^2 - \sum_{i,j=1}^{m} a_{ij} x_i x_j \\ &= \frac{1}{2}\left(\sum_{i=1}^{m} d_i x_i^2 - 2\sum_{i,j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m} d_i x_i^2\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{m} a_{ij}\left(x_i - x_j\right)^2 \end{aligned}}$$

Mathematical property of L

- A precise mathematical property of L relates it to "sparsest cut" problems
- Let the adjacency matrix $A = (a_{ij})$, then

$$x^{\top} L x = \frac{1}{2} \sum_{i,j=1}^{m} a_{ij} (x_i - x_j)^2$$

  - Suppose x is a vector of only the values +1 and -1, indicating the membership of the vertices in a set S

$$x_i = \begin{cases} 1 & \text{if } v_i \in S \\ -1 & \text{if } v_i \in \bar{S} \end{cases}$$

  - That is, we want to use x to indicate the result of a 2-partition, S and $\bar{S}$

Mathematical property of L

- A precise mathematical property of L relates it to “sparsest cut” problems
- Let the adjacency matrix $A = (a_{ij})$, then

$$x^\top L x = \frac{1}{2}\sum_{i,j=1}^{m} a_{ij}\left(x_i - x_j\right)^2$$

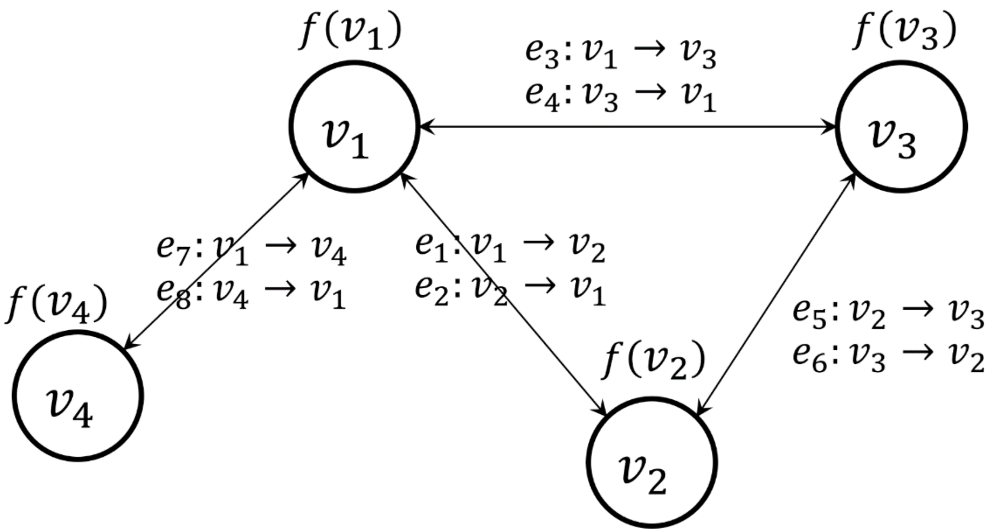
  - Suppose x is a vector of only $\{1, -1\}$, then $x^\top L x$ has special significance

$$\frac{1}{2}\sum_{i,j=1}^{m} a_{ij}\left(x_i - x_j\right)^2 = \sum_{i,j=1,i<j}^{m} a_{ij}\left(x_i - x_j\right)^2$$
$$= 4\sum_{1\le i<j\le m, x_i \ne x_j} a_{ij}$$

  - That is, $x^\top L x$ is 4 times the number of edges between adjacent vertices of each from S and $\bar{S}$

Finding x that minimizes $x^\top Lx/x^\top x$

- Compute $x^\top Lx$ for all x
  - e.g. for our earlier graph, when $x = [1 \ -1 \ -1 \ -1]$, $x^\top Lx = 12$



$$\Rightarrow \quad L = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$

$$x^\top Lx = [1 \quad -1 \quad -1 \quad -1] \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} = 12$$

Finding x that minimizes $x^\top Lx / x^\top x$

- Compute $x^\top Lx$ for all x
- $x^\top Lx = 0$ when $x = \mathbf{1} = [1\ 1\ 1\ 1]$
  (or $x = -\mathbf{1} = [-1\ -1\ -1\ -1]$)
  - However, we do not want this trivial solution

Group 1	Group 2	$x^\top Lx$
v_1	$v_2\ v_3\ v_4$	12
v_2	$v_1\ v_3\ v_4$	8
v_3	$v_1\ v_2\ v_4$	8
v_4	$v_1\ v_2\ v_3$	4
$v_1\ v_2$	$v_3\ v_4$	12
$v_1\ v_3$	$v_2\ v_4$	12
$v_1\ v_4$	$v_2\ v_3$	8
$v_1\ v_2\ v_3\ v_4$	$\emptyset$	0

- Next we compute the $\frac{x^\top Lx}{x^\top x}$ values from these

Finding x that minimizes $x^\top Lx / x^\top x$

- Complete $\frac{x^\top Lx}{x^\top x}$ values
  (x is of only +1 and -1 $\Rightarrow$ $x^\top x = |x| = 4$)

Group 1	**Group 2**	$\boldsymbol{x^\top Lx}$	$\boldsymbol{\frac{x^\top Lx}{x^\top x}}$
v_1	$v_2\ v_3\ v_4$	12	3
v_2	$v_1\ v_3\ v_4$	8	2
v_3	$v_1\ v_2\ v_4$	8	2
v_4	$v_1\ v_2\ v_3$	**4**	**1**
$v_1\ v_2$	$v_3\ v_4$	12	3
$v_1\ v_3$	$v_2\ v_4$	12	3
$v_1\ v_4$	$v_2\ v_3$	8	2

- Optimal $\frac{x^\top Lx}{x^\top x} = 1$, when $x = [1 \quad 1 \quad 1 \quad -1]$ or $[-1 \quad -1 \quad -1 \quad 1]$
- This optimal x can be approximately obtained…

Finding x that minimizes $x^\top Lx$

- Let $\lambda_1,\ldots,\lambda_k$ where $\lambda_1 \geq \ldots \geq \lambda_k$ be the eigenvalues of L, and $\mu_1,\ldots,\mu_k$ the respective eigenvectors
- By the min-max theorem of Rayleigh quotient,

$$\min_x \frac{x^\top Lx}{x^\top x} = \lambda_k$$

- However, μ_k is the trivial ($\lambda_k = 0$) solution
  - Compromise and use the second best solution μ_{k-1} (of the 2nd smallest eigenvalue λ_{k-1})
    - Historically μ_{k-1} received more attention than the other eigenvectors, but this is no longer true (will be discussed later)

Eigendecomposition example

- Eigenvalues

λ_1	λ_2	λ_3	λ_4
4.0000	3.0000	1.0000	0.0000

- Eigenvectors

More precisely, -9.51E-17

μ_1	μ_2	μ_3	μ_4
0.8660	0.0000	0.0000	-0.5000
-0.2887	0.7071	-0.4082	-0.5000
-0.2887	-0.7071	-0.4082	-0.5000
-0.2887	0.0000	0.8165	-0.5000

- $\lambda_3 = 1 =$ optimal value for $\frac{1}{2}\sum_{1\le i,j\le m} a_{ij}\left(x_i - x_j\right)^2$
- If group by the ($\pm$) sign, μ_3 correctly places v_1, v_2, v_3 in one group ($-$) and v_4 in another ($+$)

Compromise in +1/-1 restriction

- By relaxing the restriction of +1 and -1 in x to allow any real number, an $x^\top L x$ smaller than the optimal under the restriction is often achieved
  - The improvement can be guaranteed if x is orthogonal to $\mathbf{1}$ (or $-\mathbf{1}$) since by the min-max theorem, $\frac{\mu_{k-1}{}^\top L \mu_{k-1}}{\mu_{k-1}{}^\top \mu_{k-1}}$ is minimal among all $\frac{x^\top L x}{x^\top x}$ that are orthogonal to μ_k
    - However, in the present case, $x = [1\ 1\ 1\ -1]$ and not orthogonal to $\mu_4 = [1\ 1\ 1\ 1]$
    - Still, $\frac{\mu_3{}^\top L \mu_3}{\mu_3{}^\top \mu_3} = \lambda_3 = 1 = \min_{\boldsymbol{x} \in \{1,-1\}^4} \frac{x^\top L x}{x^\top x}$
    - Though no guarantee, improvements are usual

Historical use of μ_{k-1}

- Historically μ_{k-1} received more attention than the other eigenvectors
  - Shi and Malik (2000) started using multiple eigenvectors for clustering (see Part 3)
- μ_{k-1} is called the Fiedler vector
- λ_{k-1} is called the Fiedler value
  - The multiplicity of λ_{k-1} is always 1
  - Also called the algebraic connectivity
    - The further λ_{k-1} is from 0, the more highly connected is the graph (hard to separate)