# Impact of Regularization on Spectral Clustering

Antony Joseph* and Bin Yu#

** Walmart Research Lab in San Francisco*
*(formerly UCB and LBNL)*

*# Departments of Statistics and EECS*
*UC Berkeley*

*Workshop on Spectral Algorithms, Simons Inst, Oct. , 2014*

# Overview

spectral clustering
in
graphs

Berkeley Drosophila Genome Project (BDGP)
(The fruit fly project)

collaborators :

- Siqi Wu, UC Berkeley
- Erwin Frise, Lawrence Berkeley Lab
- Ann Hammonds, Lawrence Berkeley Lab
- Sue Celniker, Lawrence Berkeley Lab

# Overview

spectral clustering
in
graphs

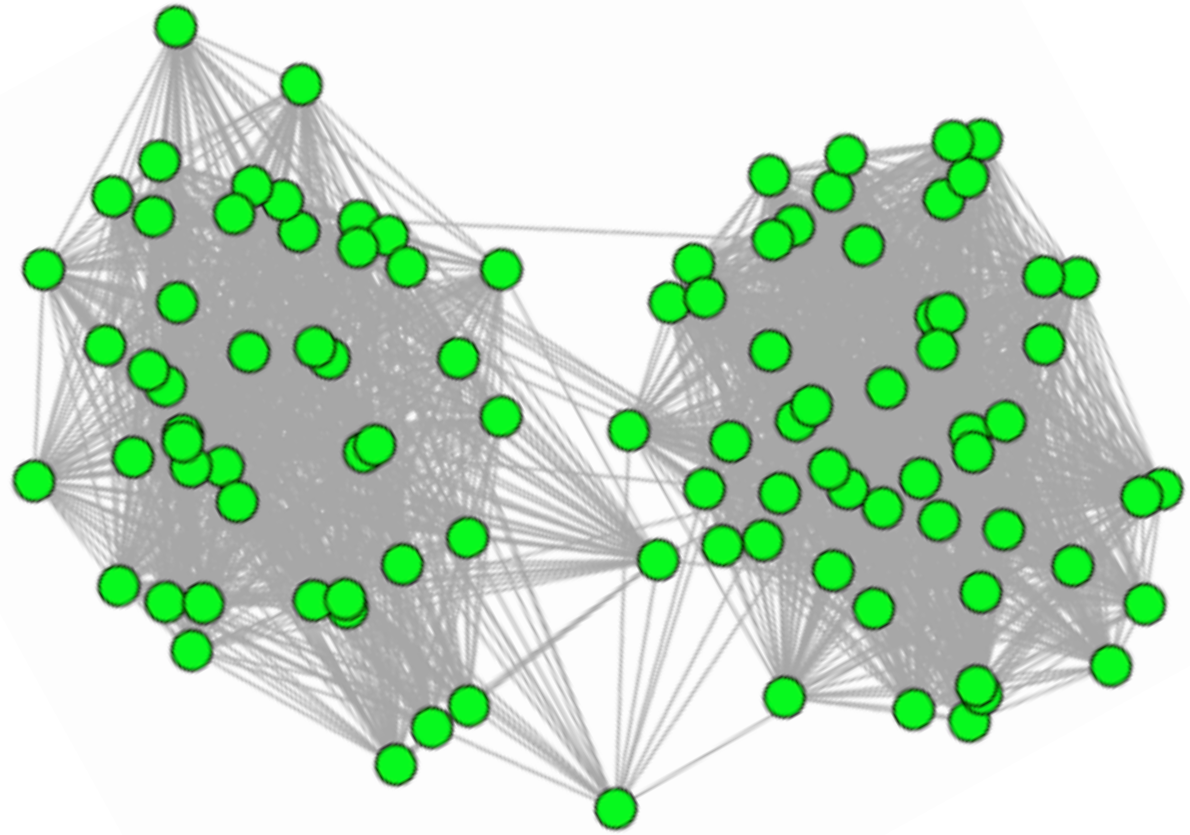
↓

Berkeley Drosophila Genome Project (BDGP)
(The fruit fly project)

collaborators :

- Siqi Wu, UC Berkeley
- Erwin Frise, Lawrence Berkeley Lab
- Ann Hammonds, Lawrence Berkeley Lab
- Sue Celniker, Lawrence Berkeley Lab

# A Graph



| Context | Nodes |
| --- | --- |
| The fruit fly project | *pixels/points in the embryo template* |
| Social network | *people* |

...

The fruit fly project (Berkeley Drosophila Genome Project)

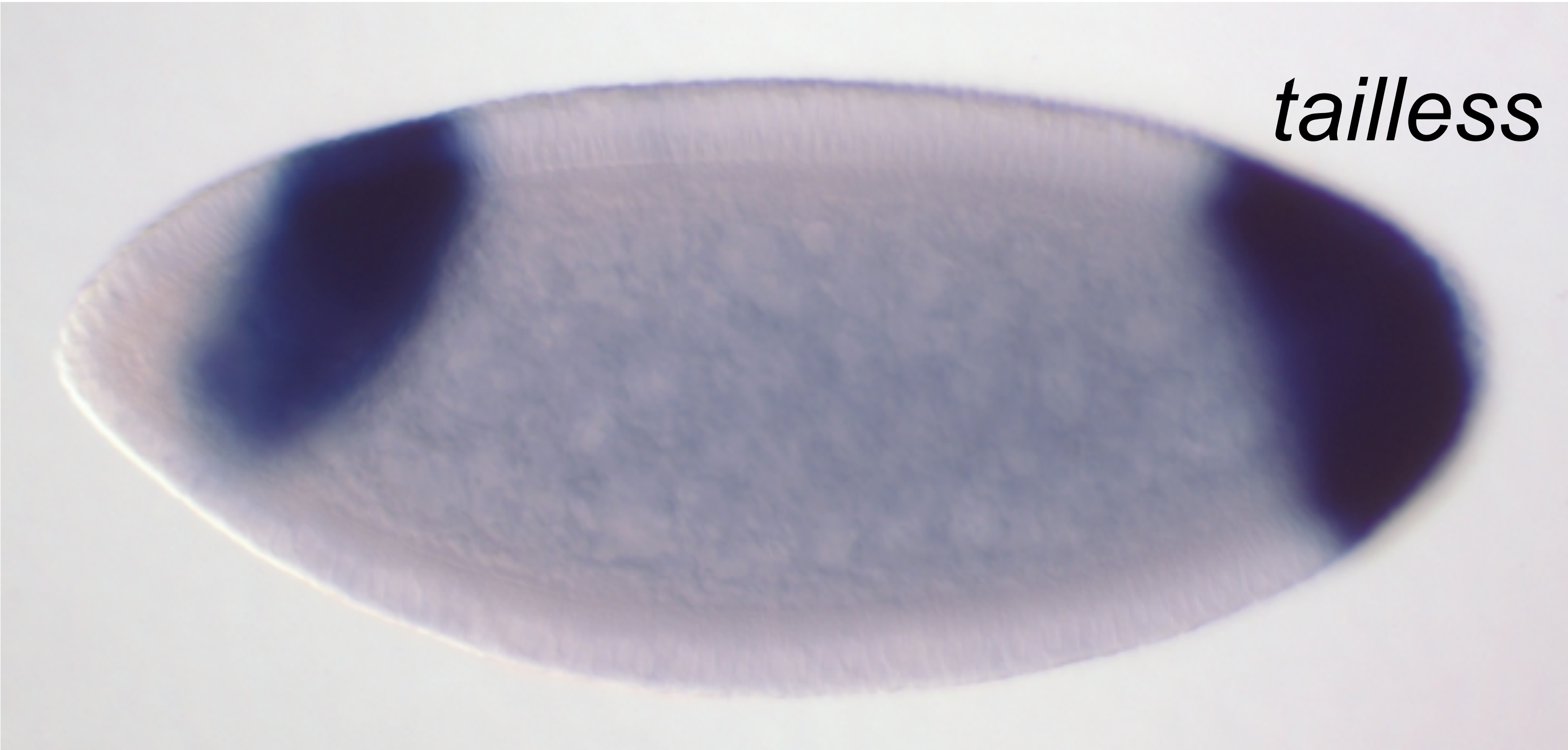
Drosophila
(fruit fly)

Widely studied :

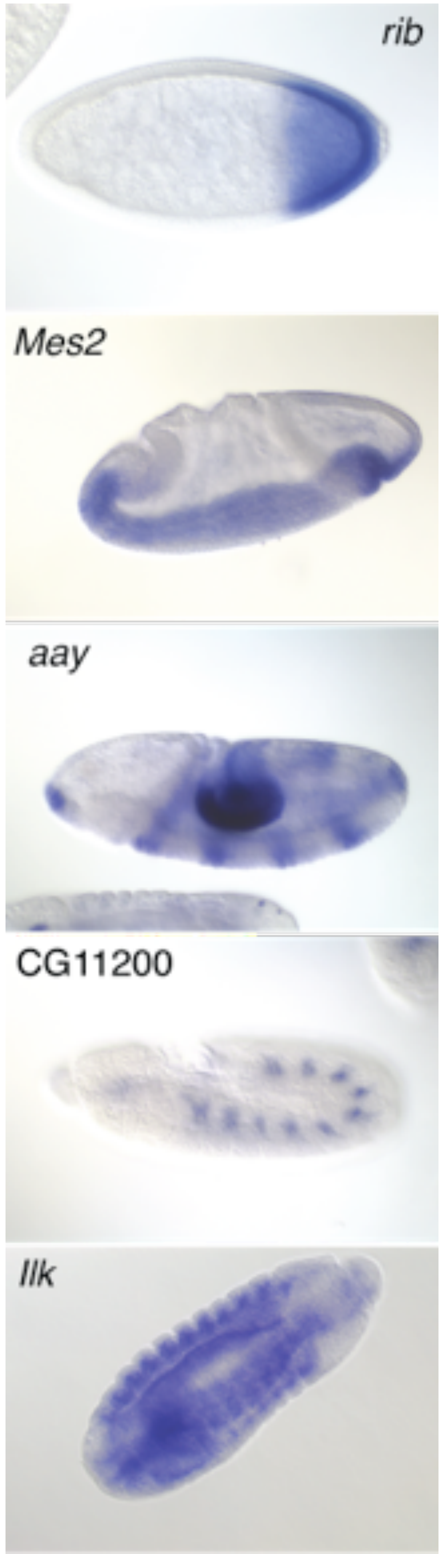
- genetic mechanism similar to humans
- easy to maintain in the lab
- short life cycle
- ...

# Image dataset from the fruit fly project

# Image dataset from the fruit fly project

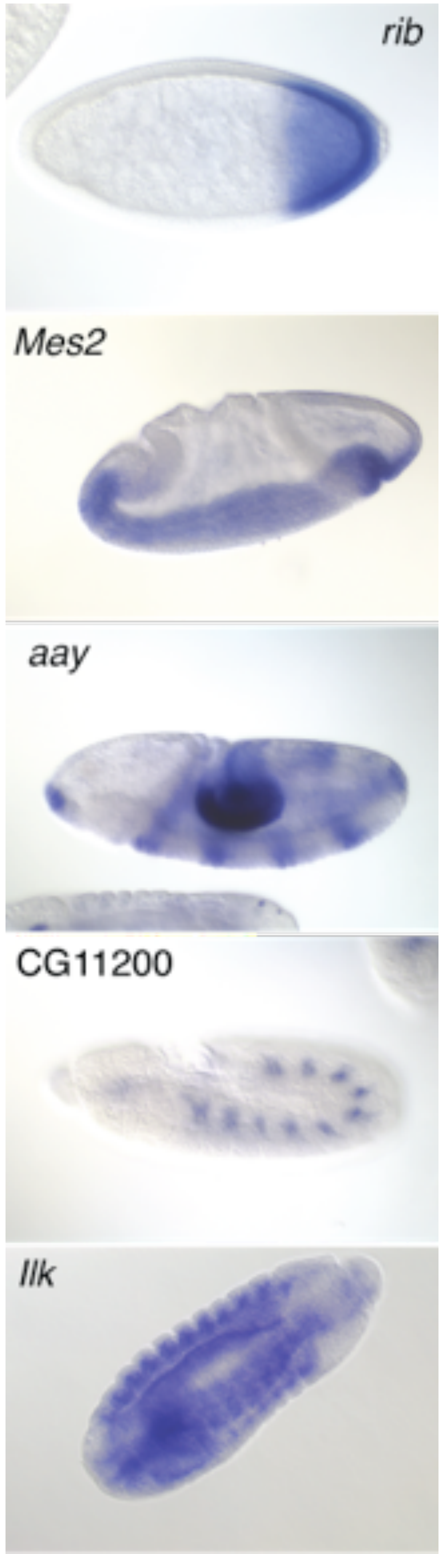


# Image dataset from the fruit fly project




- Over 100,000 stained embryo images (over 7000 genes)

# Image dataset from the fruit fly project

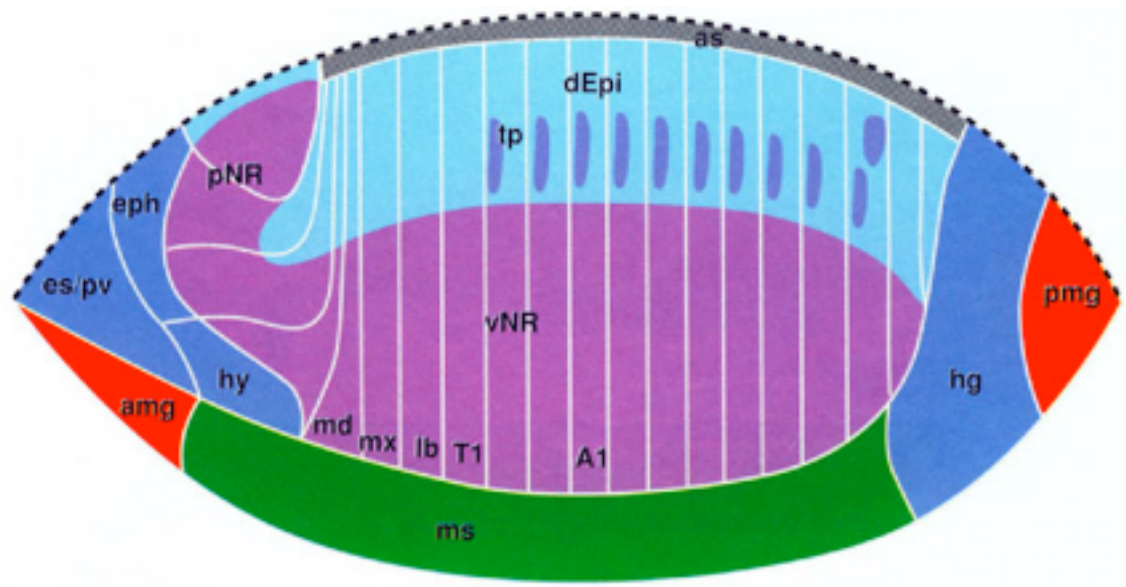



- Over 100,000 stained embryo images (over 7000 genes)

Goals: Contribute to the understanding of ...

- the interaction between different genes
- the genes required for development of various organs.

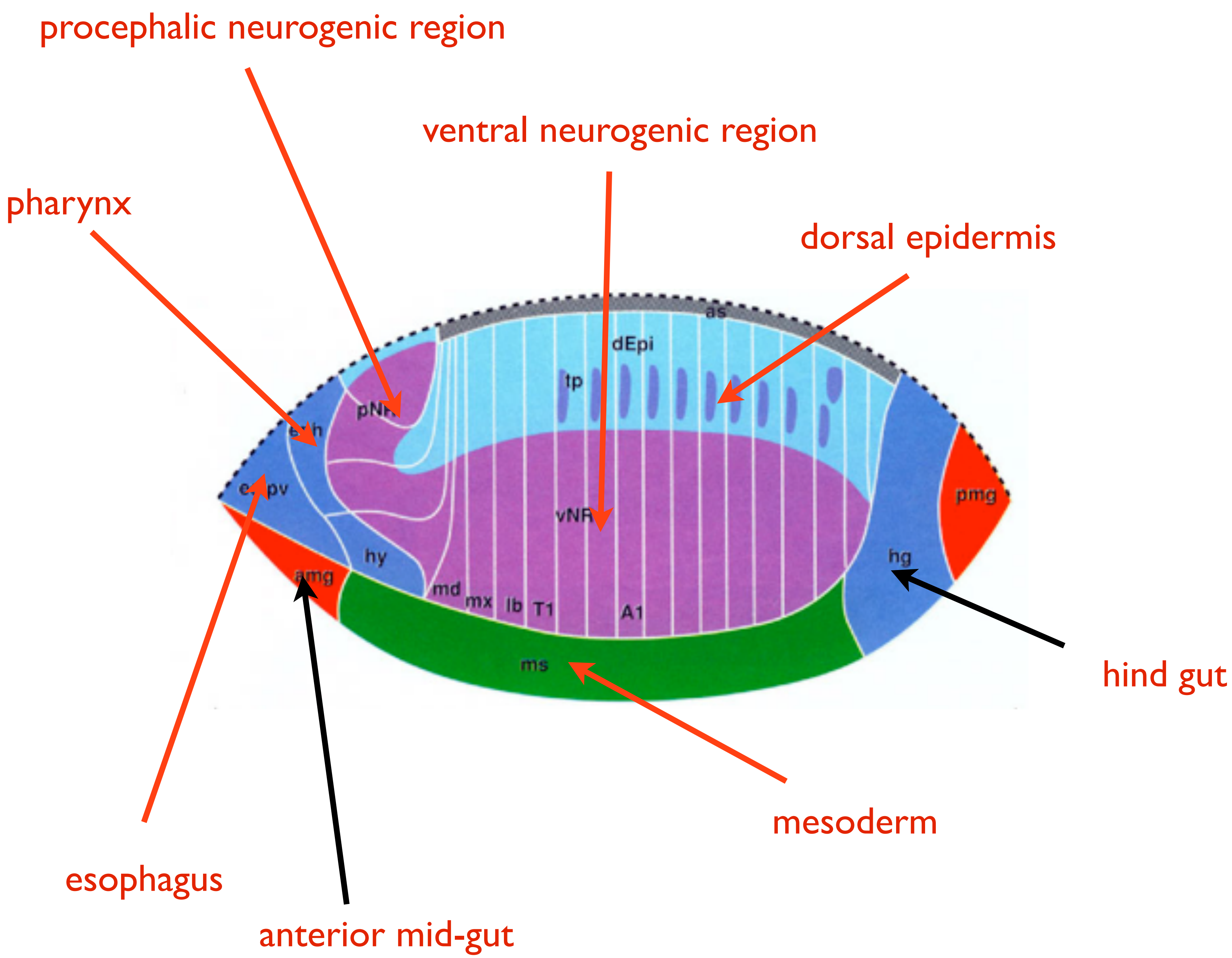
# ‘Fate’ map in early embryos




Laser ablation experiments in embryos in early stages of development

*Lohs-Schardin et. al (’70), Hartenstein et. al. (‘85)*

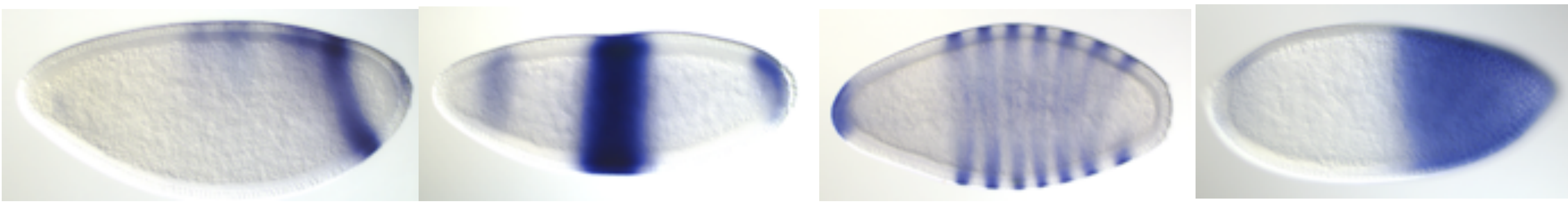
# ‘Fate’ map in early embryos



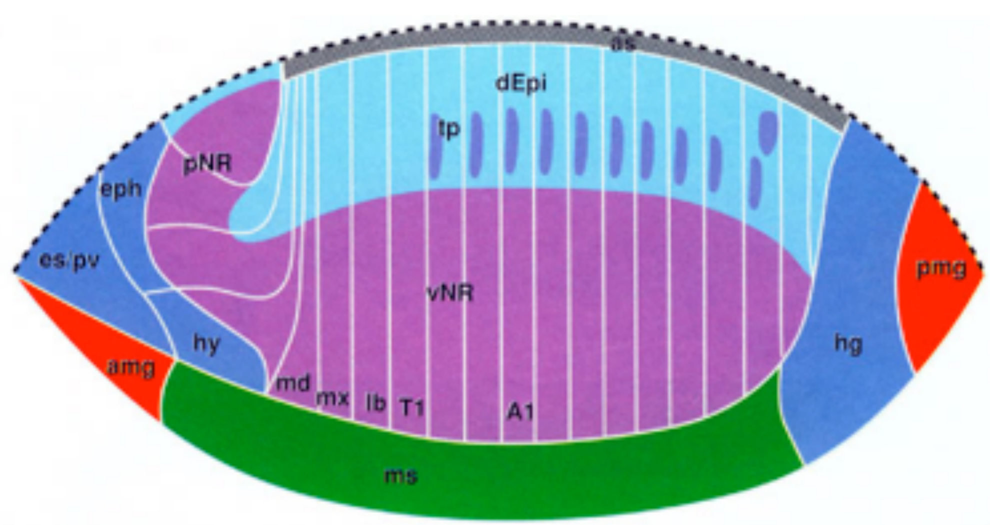

Laser ablation experiments in embryos in early stages of development

*Lohs-Schardin et. al (’70), Hartenstein et. al. (‘85)*

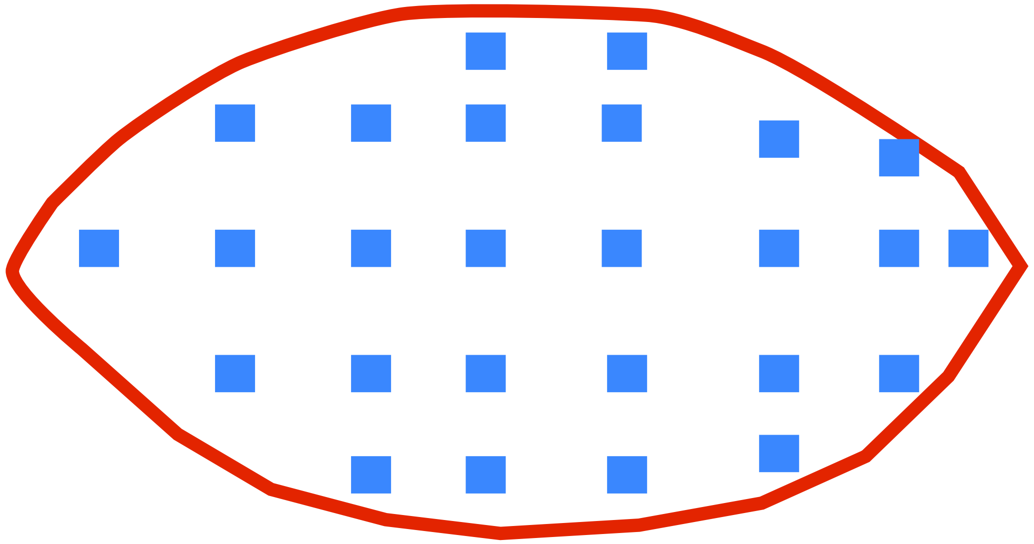
# Do genes explain the `fate' map?



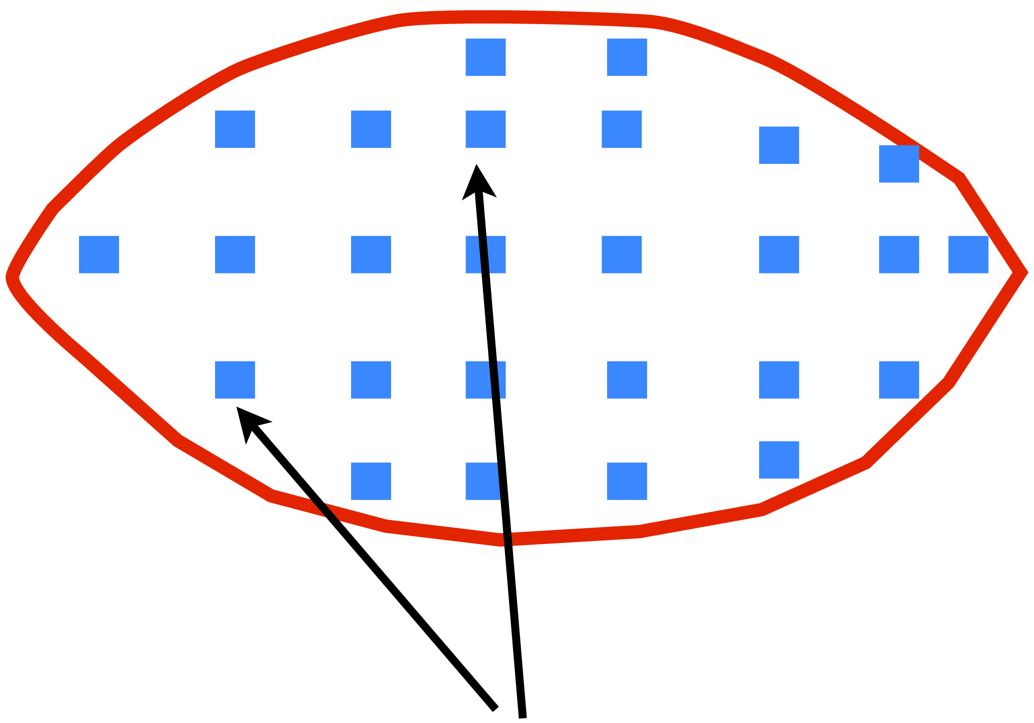
.... early stage gene expression images



# Discovery of fate map
# &
# communities on graphs

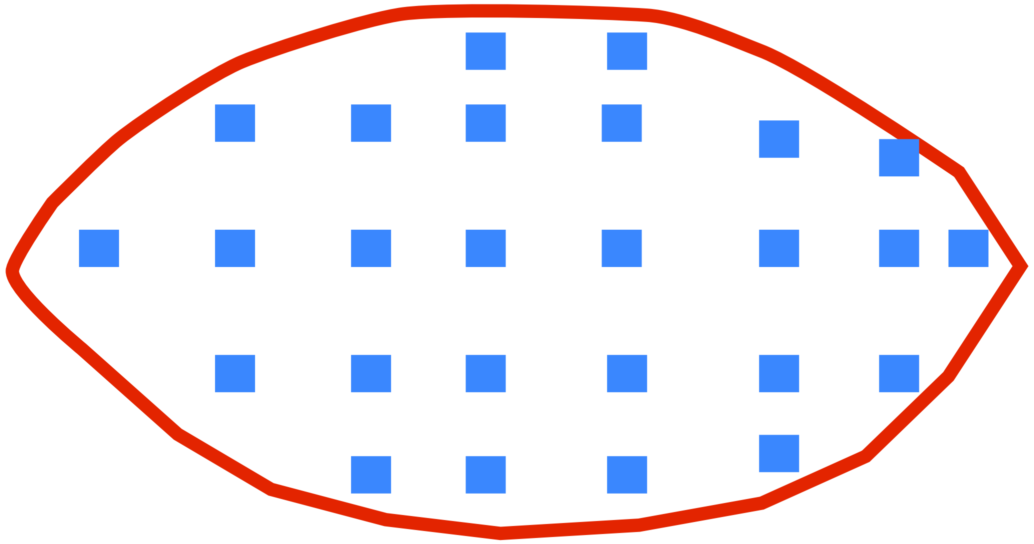


# Discovery of fate map
# &
# communities on graphs

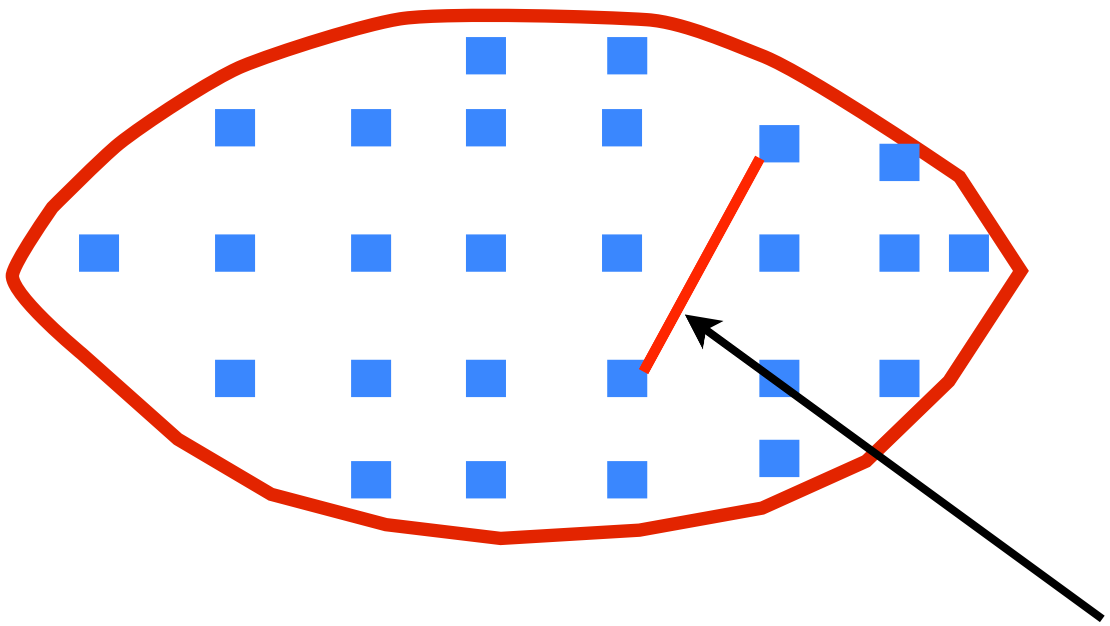


Nodes : pixels/points in the embryo

# Discovery of fate map
# &
# communities on graphs

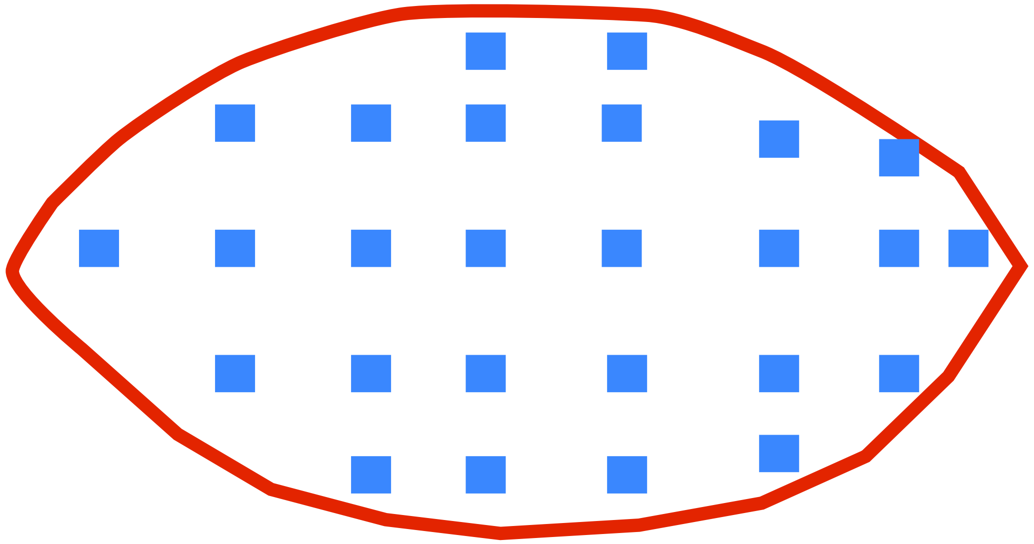


# Discovery of fate map
# &
# communities on graphs

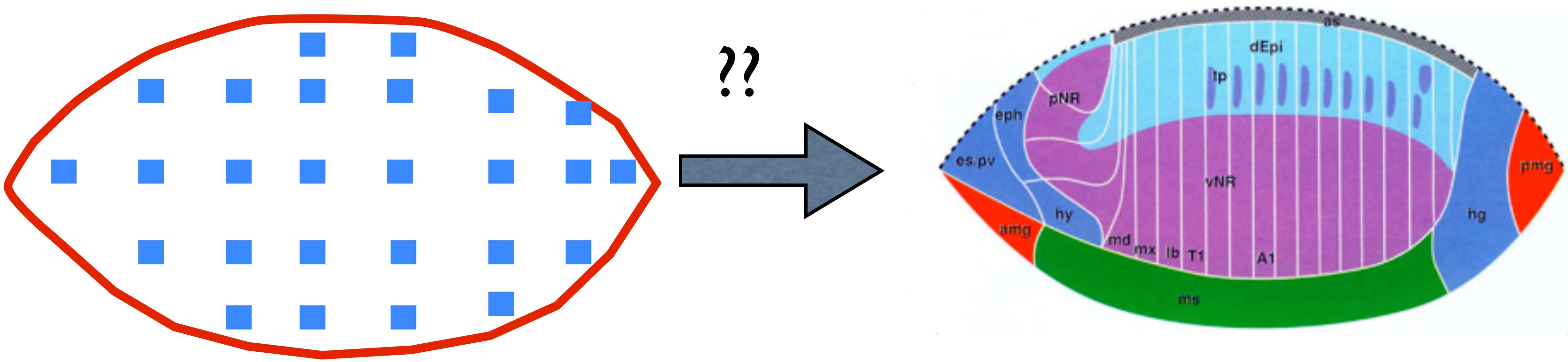


Edge if lot of genes are co-expressed at the two nodes

# Discovery of fate map
# &
# communities on graphs



# Discovery of fate map
# &
# communities on graphs




fate map

# Edge between node *i* and node *j*

# Edge between node *i* and node *j*

$X_i =$ at the *i*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

# Edge between node *i* and node *j*

$X_i$ = at the *i*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

$X_j$ = at the *j*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

# Edge between node *i* and node *j*

$X_i$ = at the *i*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

$X_j$ = at the *j*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

Edge between node *i* and node *j* if $X_i X_j^T > t$, for some $t > 0$.

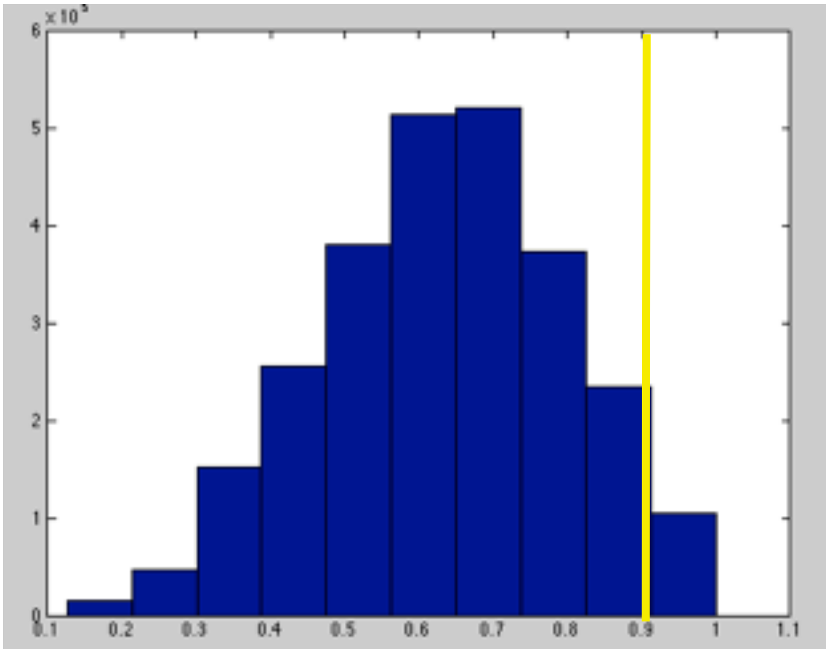
# Edge between node *i* and node *j*

$X_i =$ at the *i*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

$X_j =$ at the *j*-th pixel ( $gene_1$ expression, . . ., . . ., $gene_{1640}$ expression)

Edge between node *i* and node *j* if $X_i X_j^T > t$, for some $t > 0$.

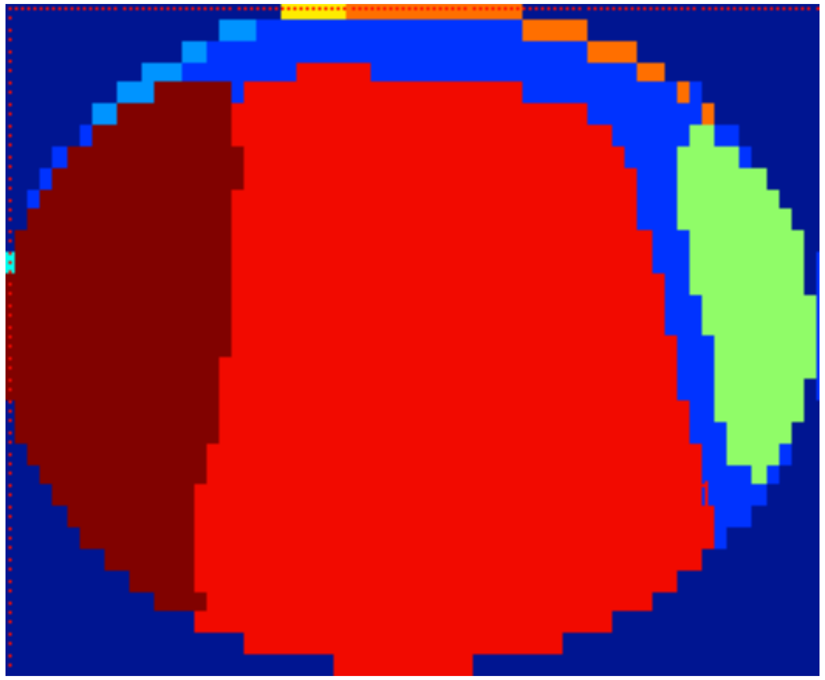
Histogram of $X_i X_j^T$



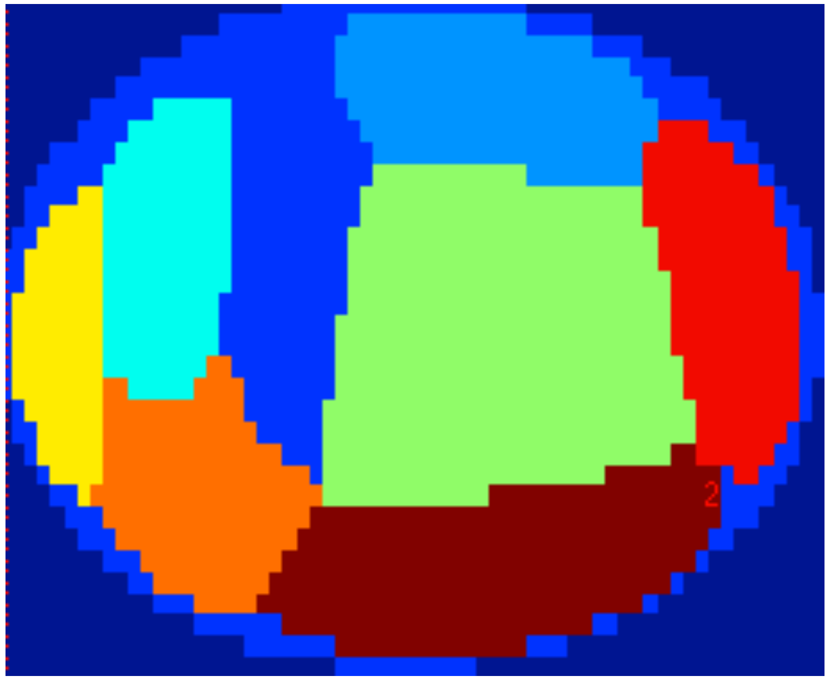

90-th percentile

# Comparing unregularized vs. regularized SC

Take $K = 8$



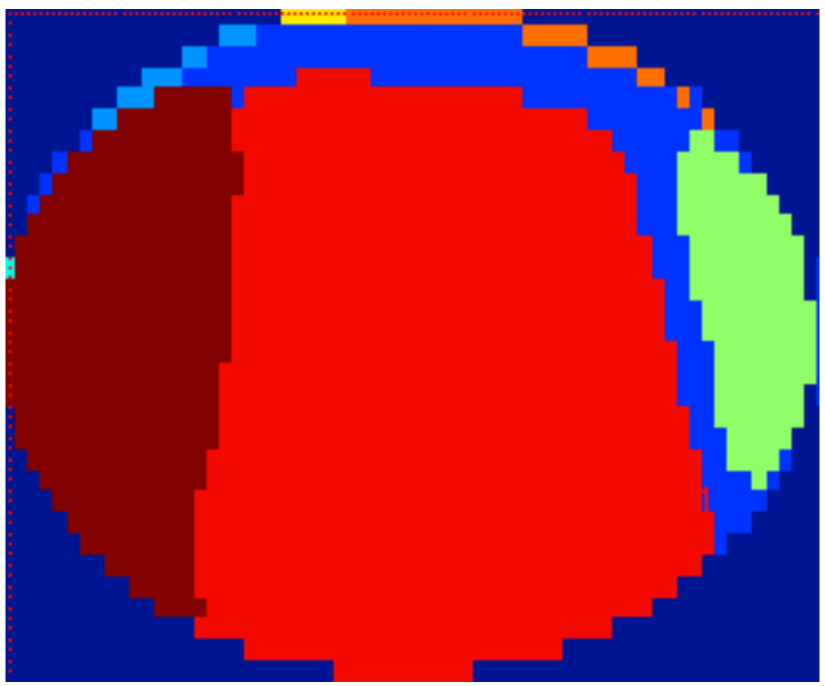
$\tau = 0$



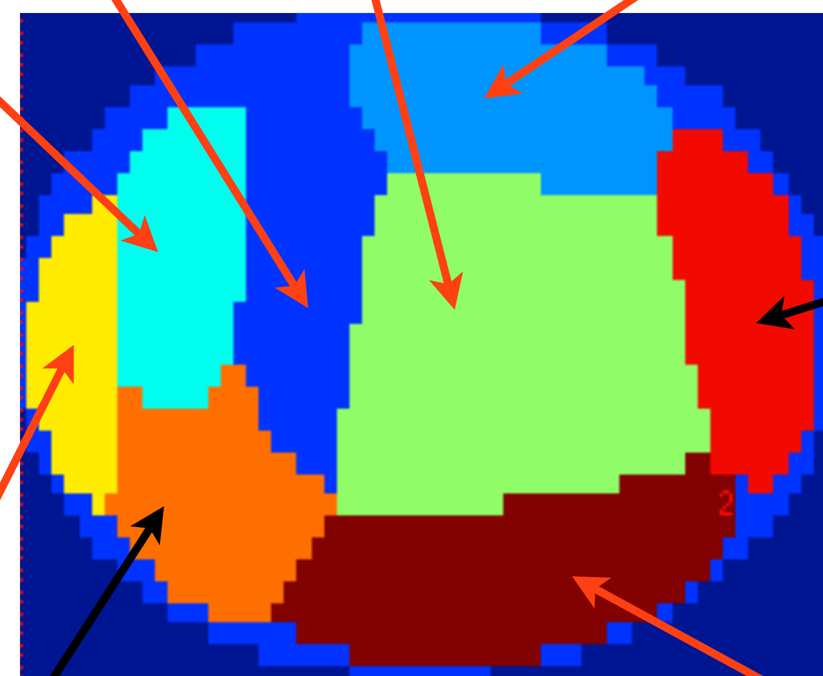
$\tau = 24$

# Comparing unregularized vs. regularized SC

Take $K = 8$

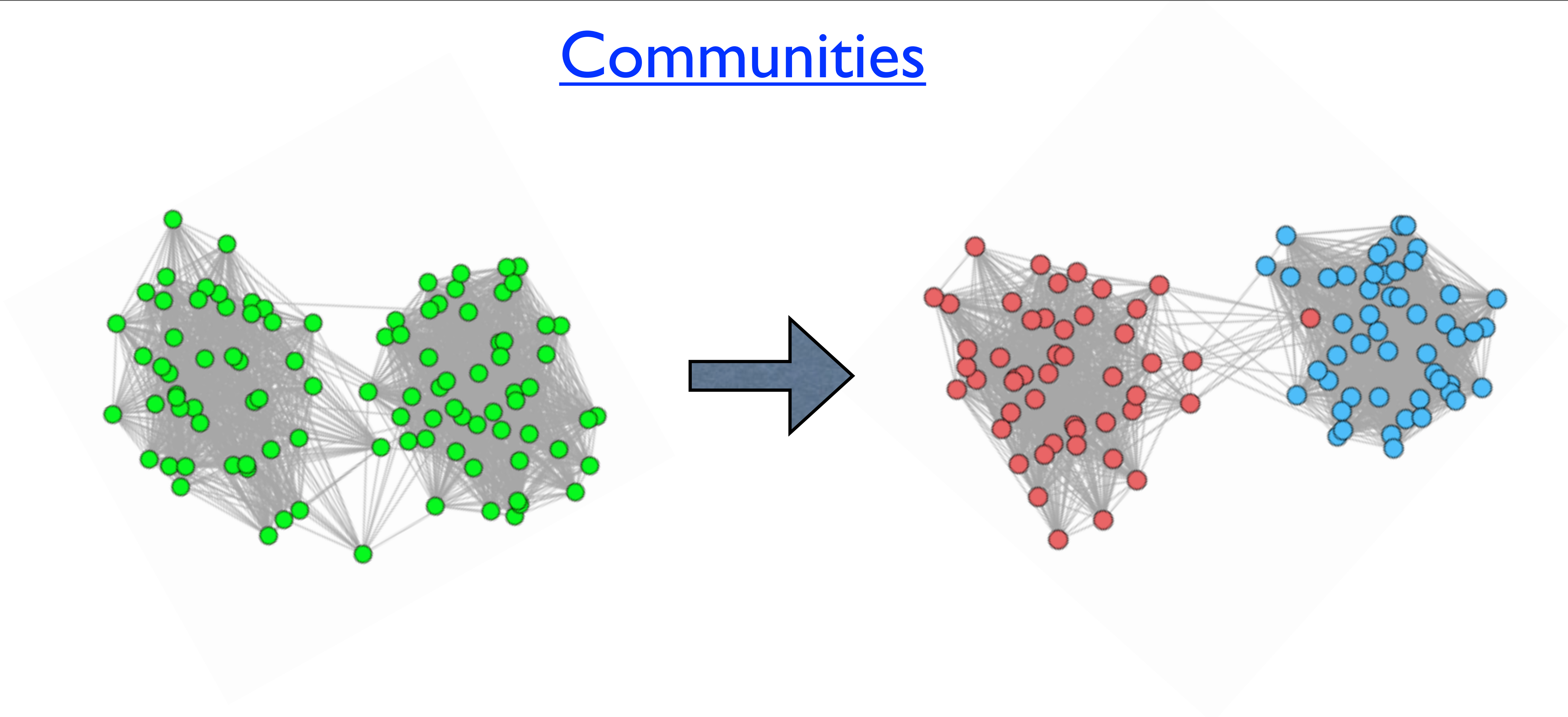


$\tau = 0$




$\tau = 24$

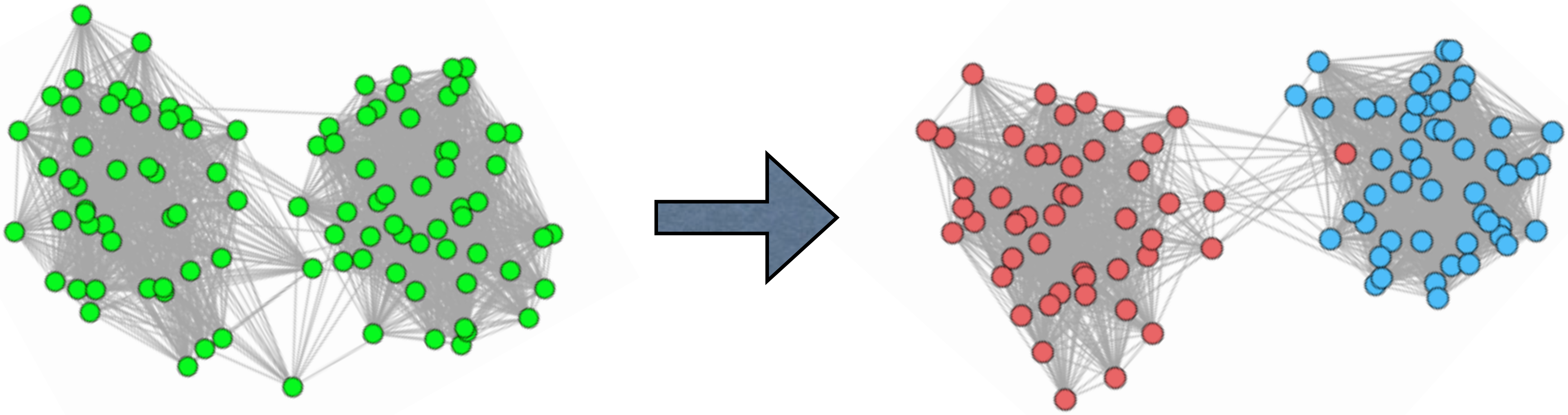
# Communities



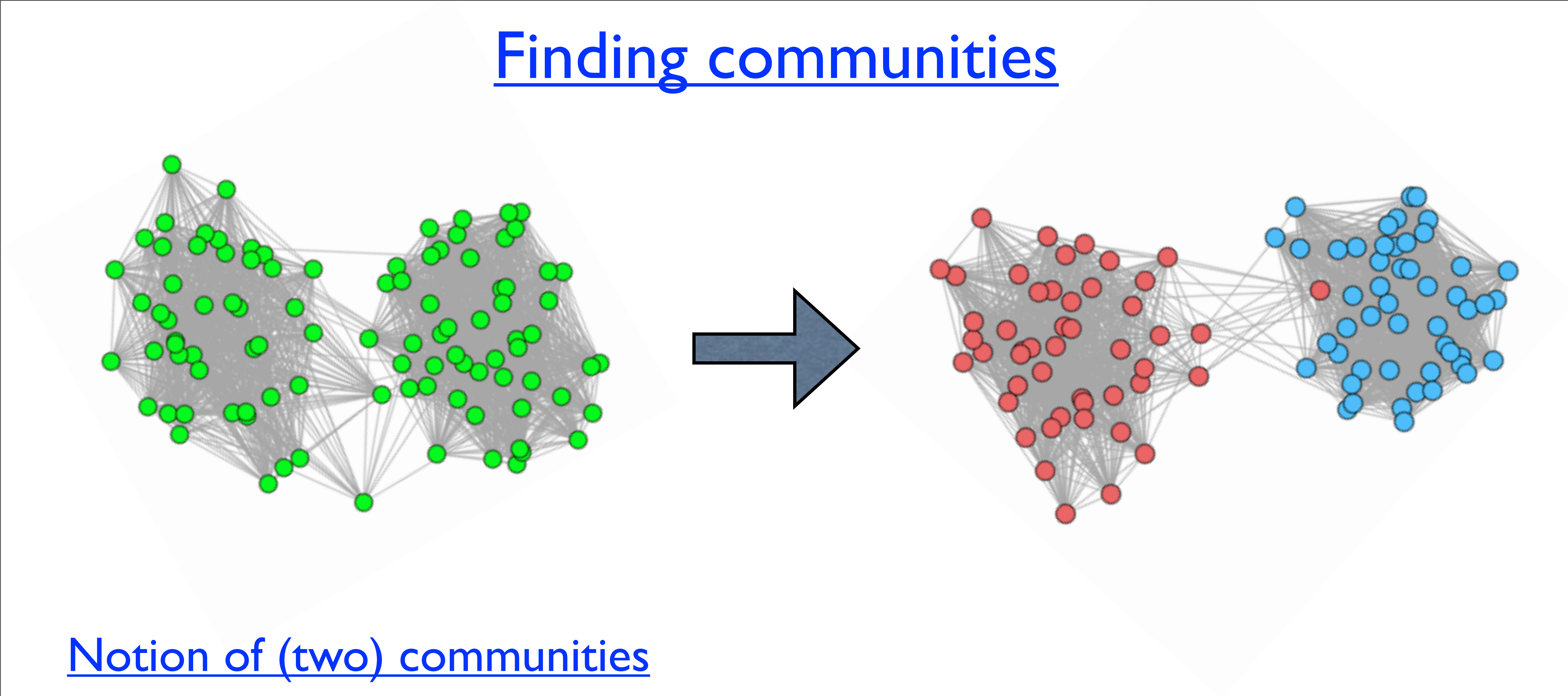
| Nodes | Communities |
|---|---|
| pixels/points in embryo | *area of future organs* |
| people | *like minded people* |

...

# Finding communities



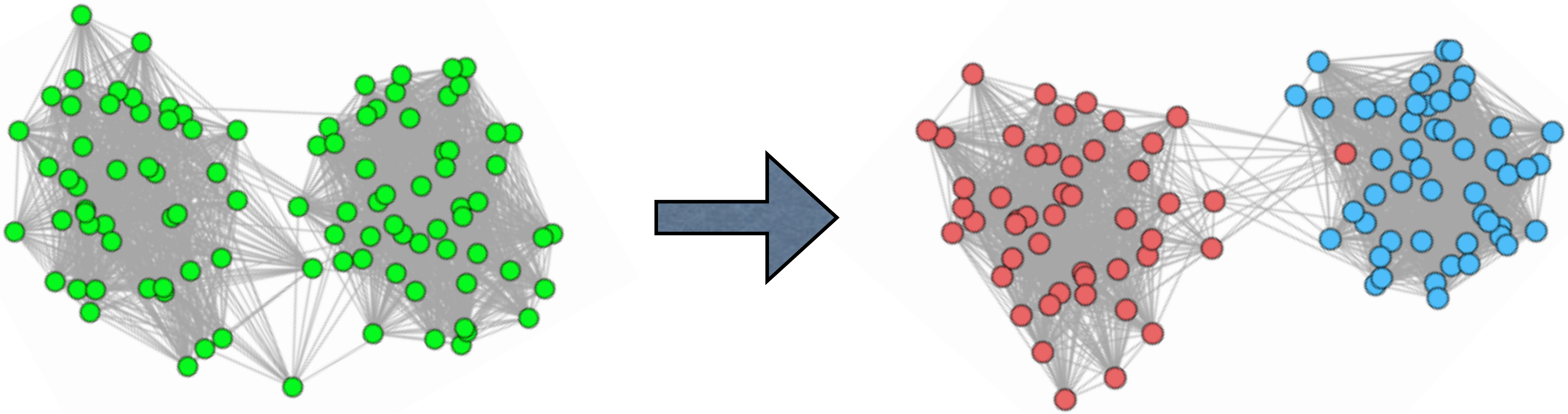
# Finding communities



## Notion of (two) communities

``Partition of nodes into sets $C_1$ and $C_2$,
so that there are very few edges between the nodes in $C_1$ and $C_2$"

# Finding communities



## Notion of (two) communities

``Partition of nodes into sets $C_1$ and $C_2$,
so that there are very few edges between the nodes in $C_1$ and $C_2$"

## Methods

Spectral clustering (Fiedler ('73), Donath & Hoffman ('73), ...)

Modularity (Newman & Girvan ('03)),
Latent space methods (Hoff et. al. ('02))
Profile-likelihood (Bickel & Chen ('09)), Pseudo-Likelihood (Amini et. al. ('13)),

# Spectral Clustering

# Notation

Number of nodes: $n$

Adjacency matrix:
(symmetric binary)

$$A \in \mathbb{R}^{n \times n}$$

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{if } (i,\ j) \text{ is an edge} \\ 0, & \text{otherwise} \end{cases}$$

# Notation

Number of nodes: $n$

Adjacency matrix: (symmetric binary)

$$A \in \mathbb{R}^{n \times n}$$

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{if } (i, j) \text{ is an edge} \\ 0, & \text{otherwise} \end{cases}$$

Each row/column of *A* associated with a node

# Notation

Number of nodes: $n$

Adjacency matrix:
(symmetric binary)

$$A \in \mathbb{R}^{n \times n}$$

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{if } (i,\ j) \text{ is an edge} \\ 0, & \text{otherwise} \end{cases}$$

Degree matrix:
(diagonal)

$$D \in \mathbb{R}^{n \times n}$$

$$D_{ii} = \sum_j A_{ij}$$

# Spectral Clustering

Spectral clustering deals with the eigenvectors of the matrix :

$$L = D^{-1/2}\, A\, D^{-1/2}$$

(Normalized symmetric Laplacian matrix)

# Spectral Clustering

Spectral clustering deals with the eigenvectors of the matrix :

$$L = D^{-1/2}\, A\, D^{-1/2}$$

(Normalized symmetric Laplacian matrix)

Other matrices used ...

$$D^{-1}\, A$$

( Normalized random walk Laplacian)

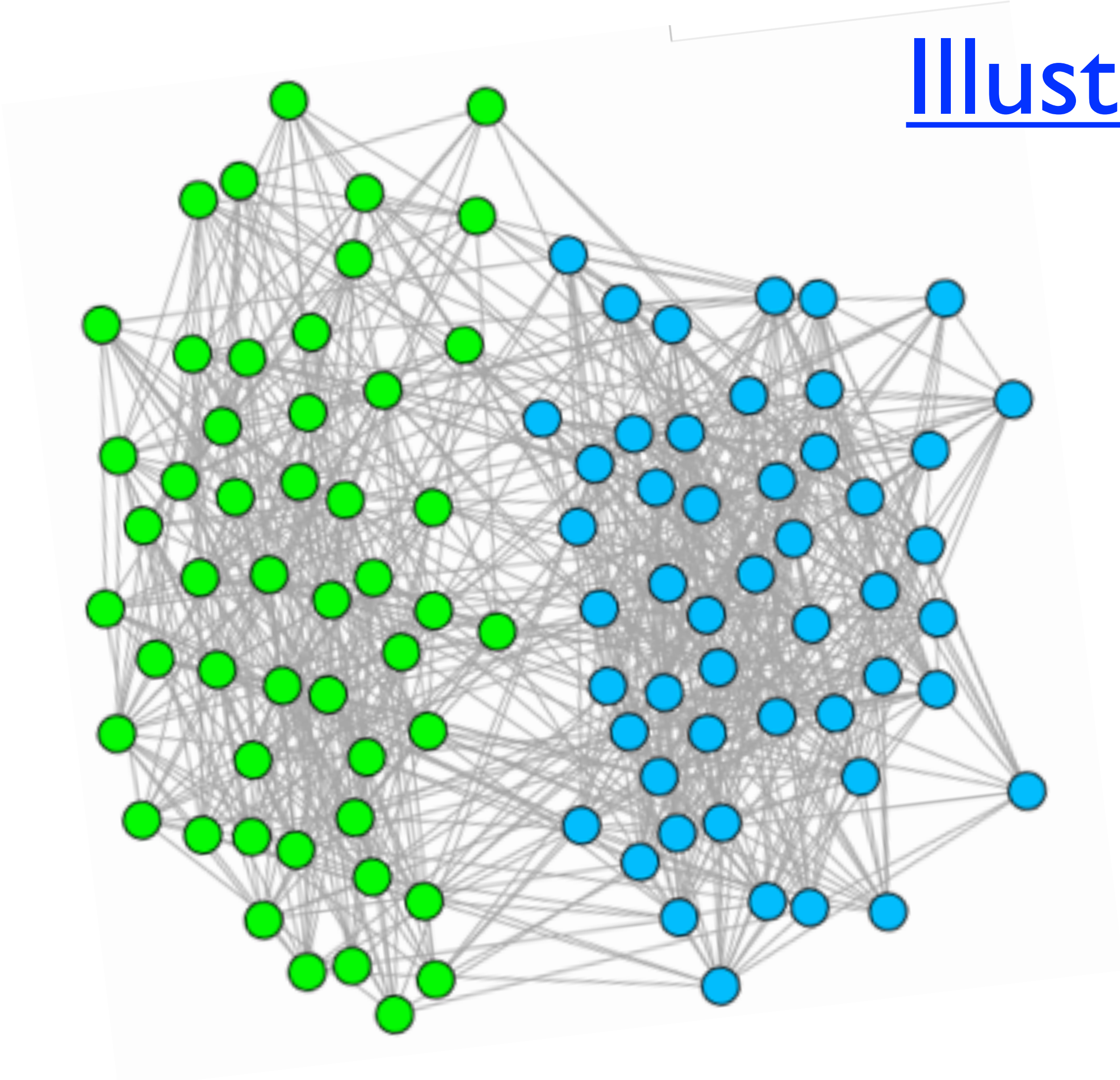
$$D - A$$

(Unnormalized Laplacian)

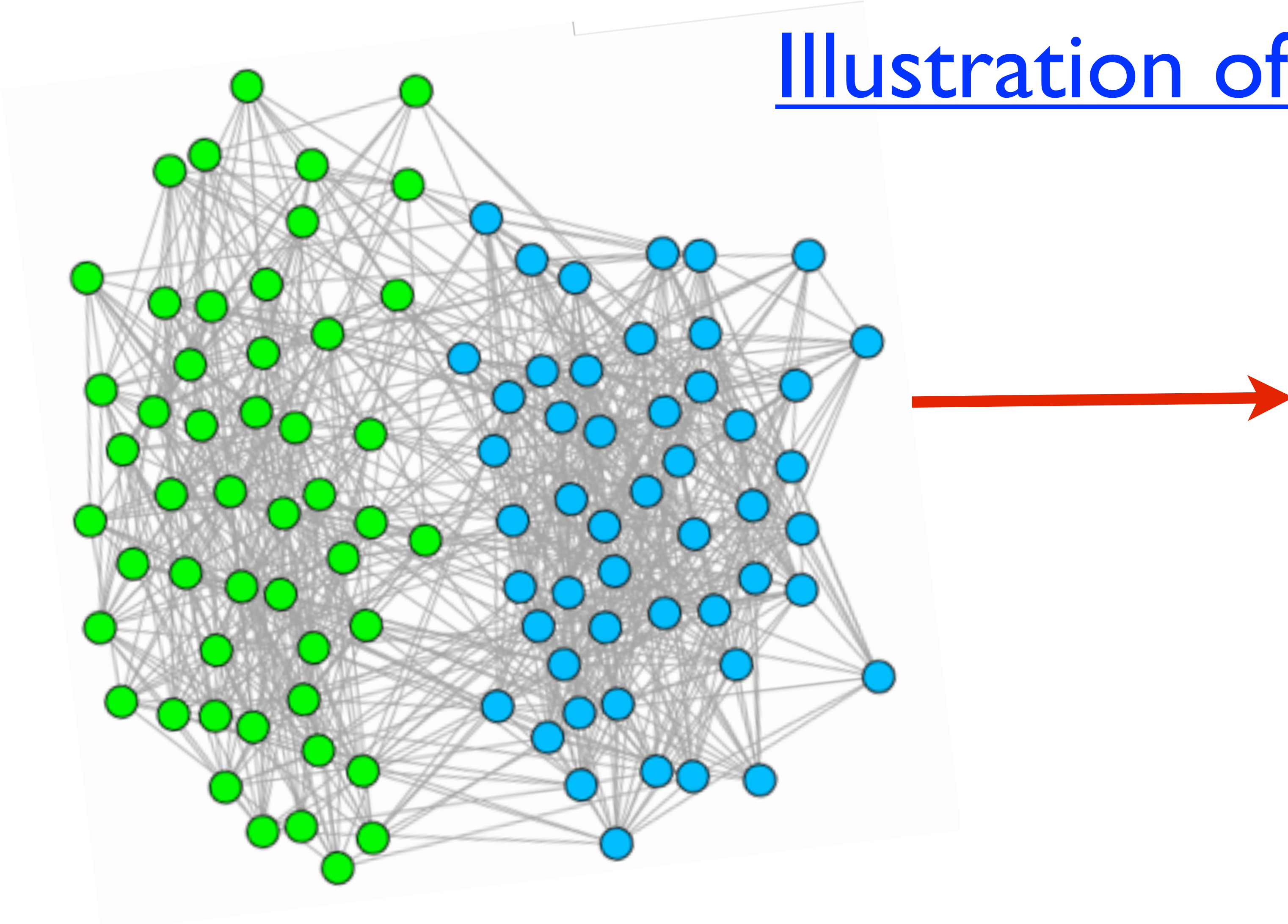
$$A$$

(Adjacency matrix)

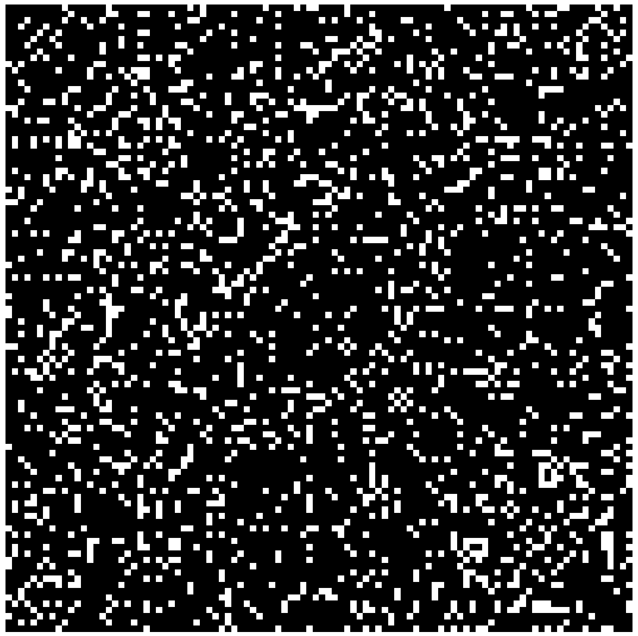
# Illustration of SC



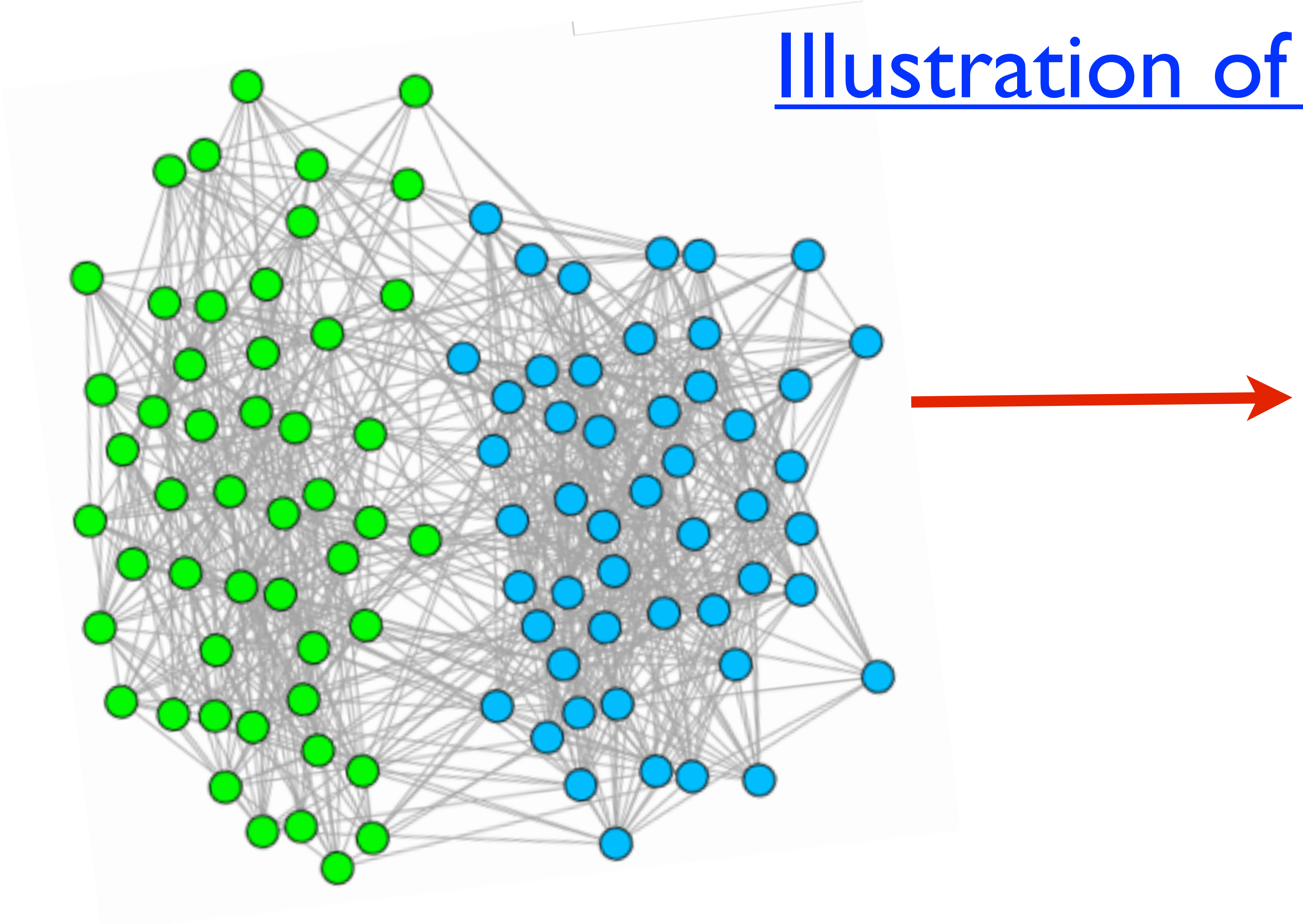
# Illustration of SC

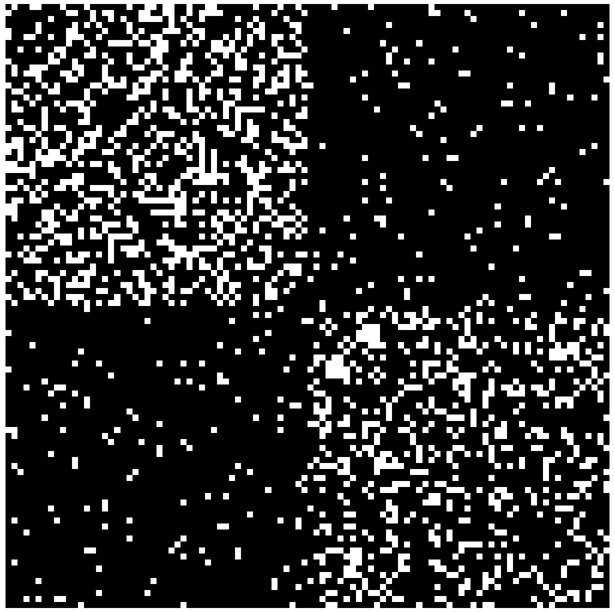


$A =$

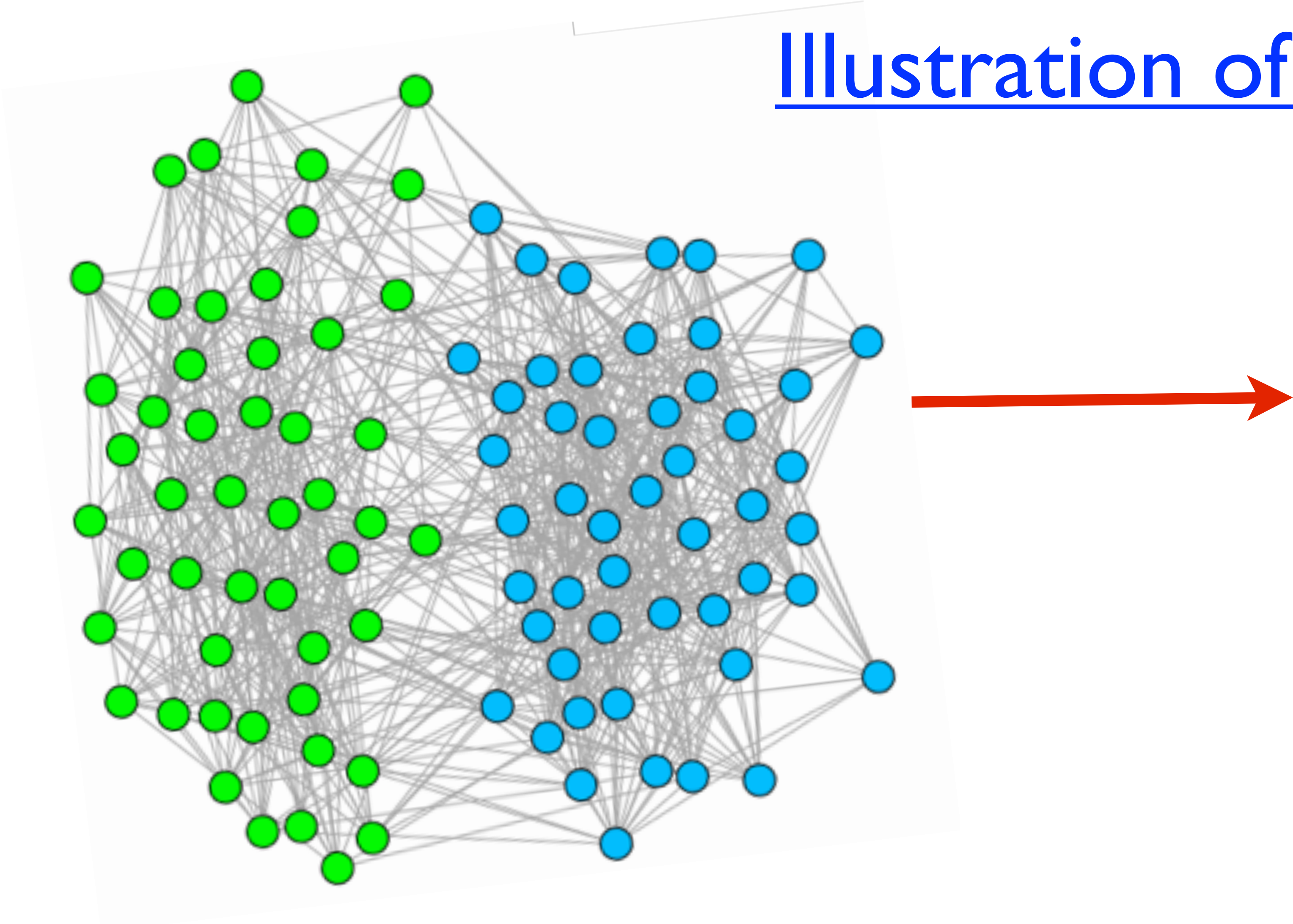


# Illustration of SC

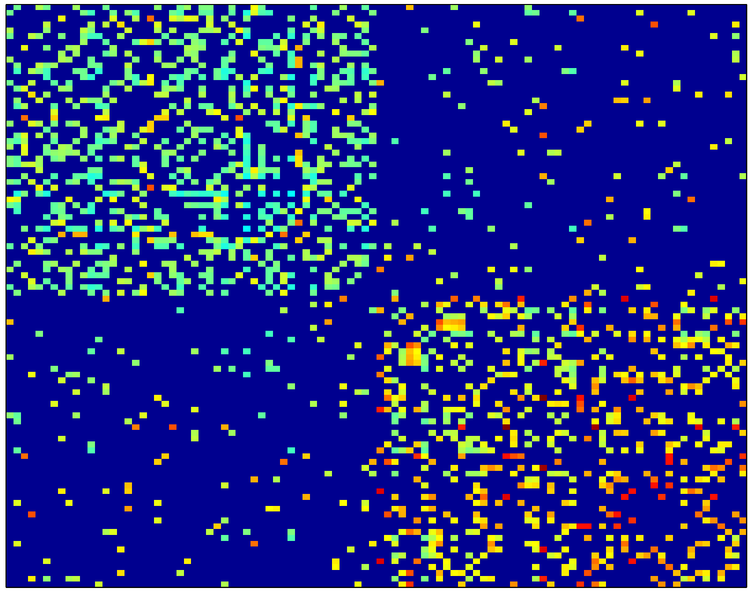


$A =$ 

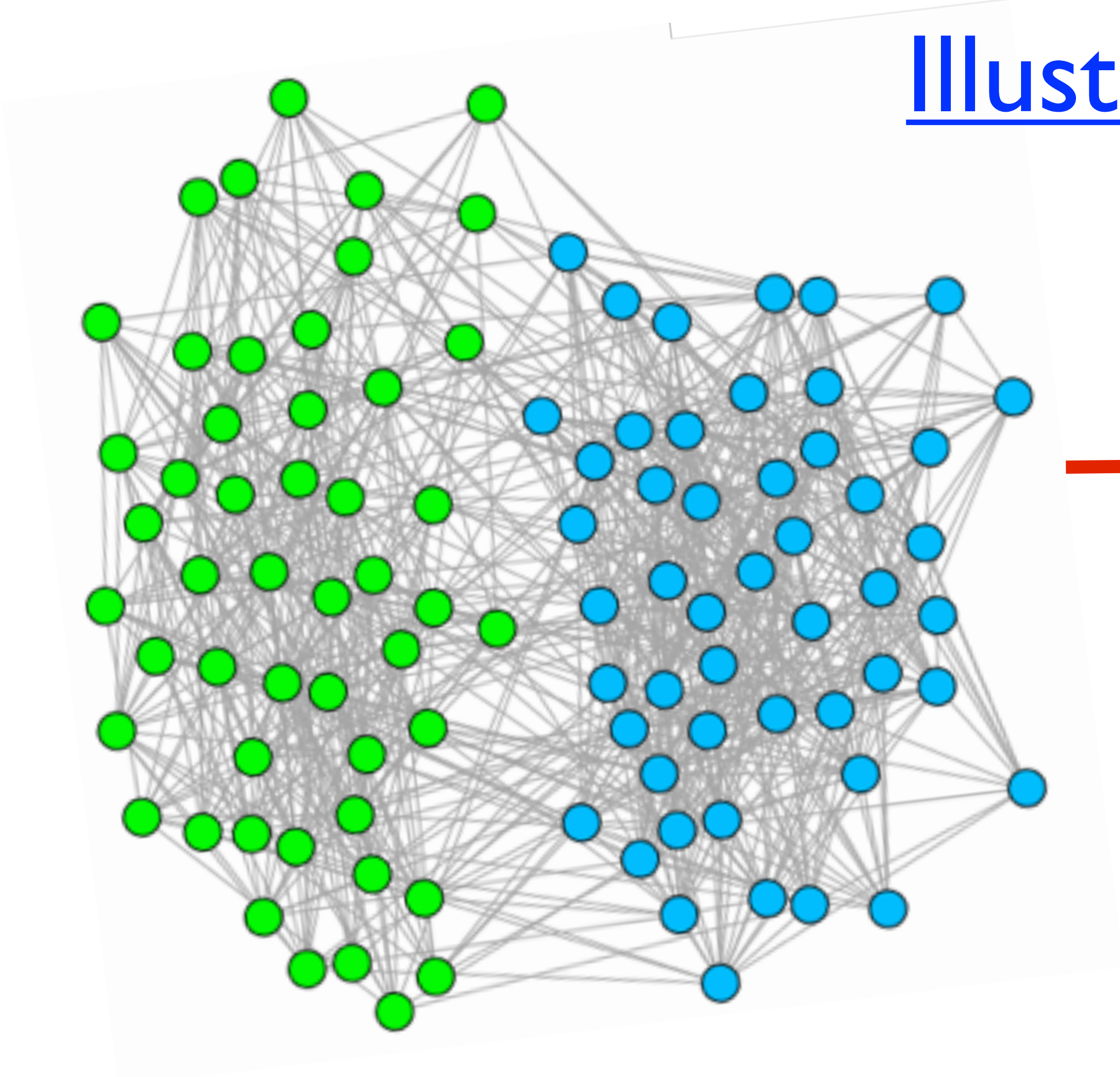
# Illustration of SC



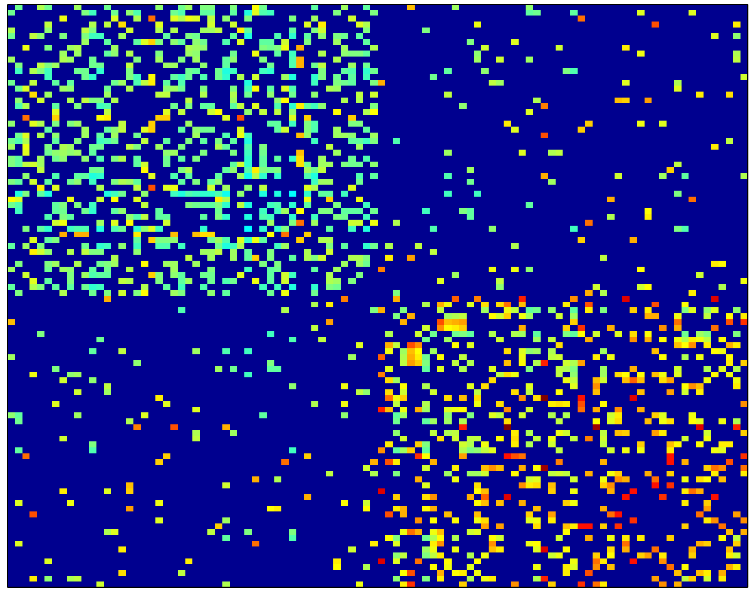
$L =$

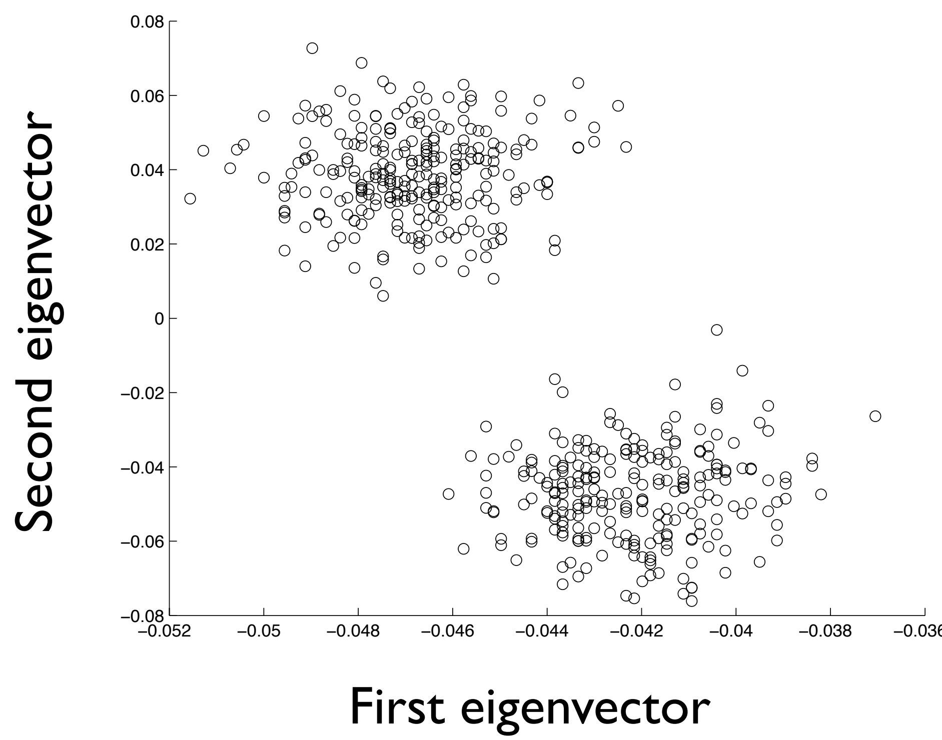


# Illustration of SC

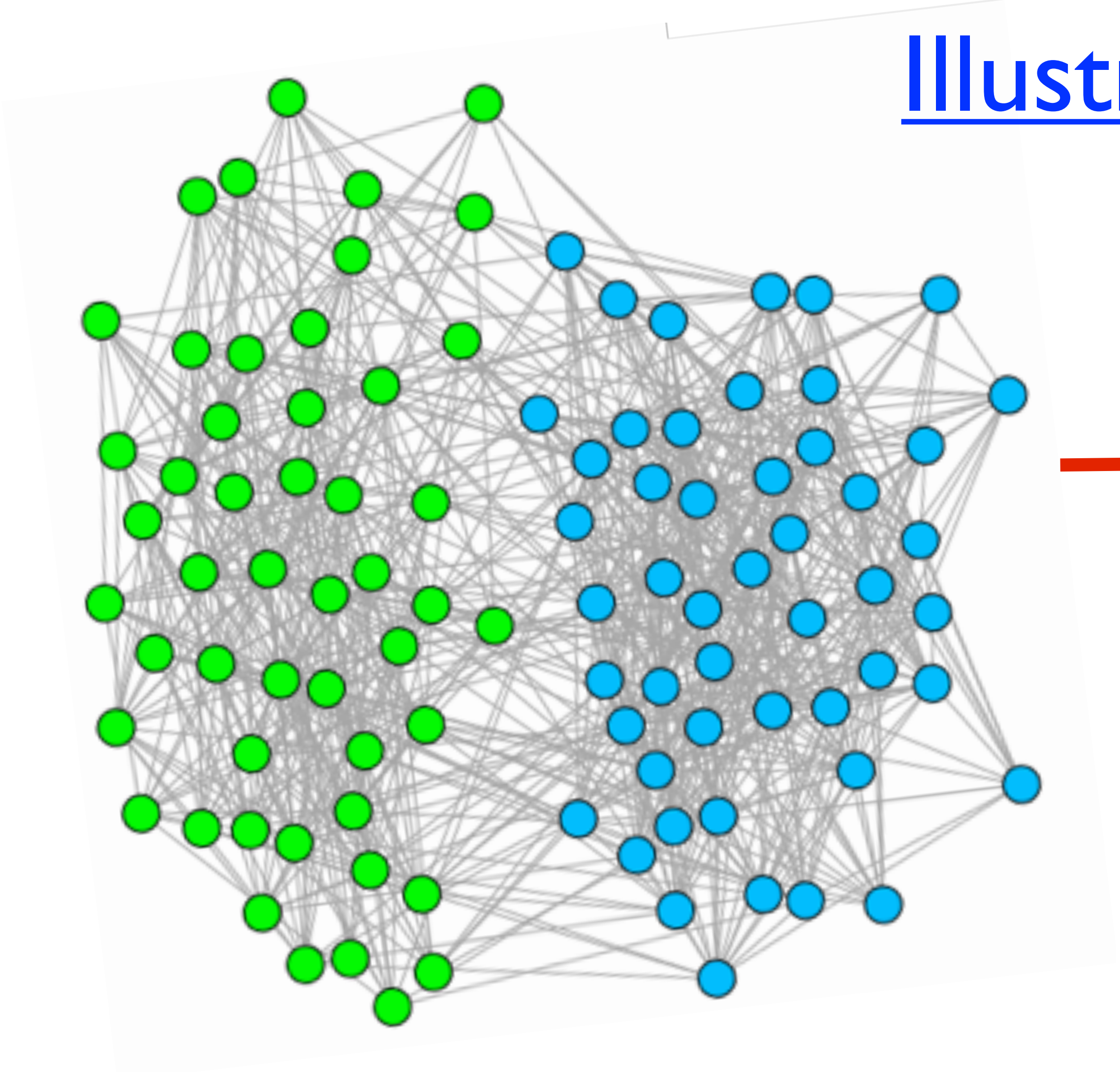


$L =$

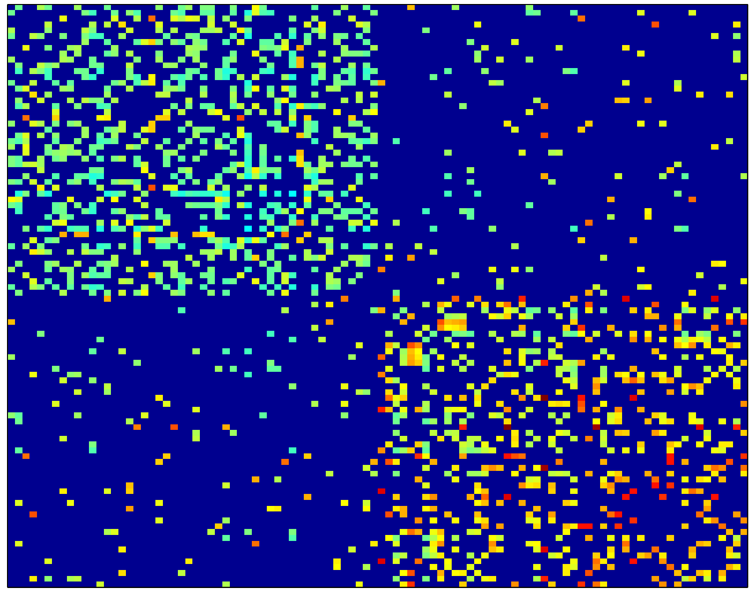


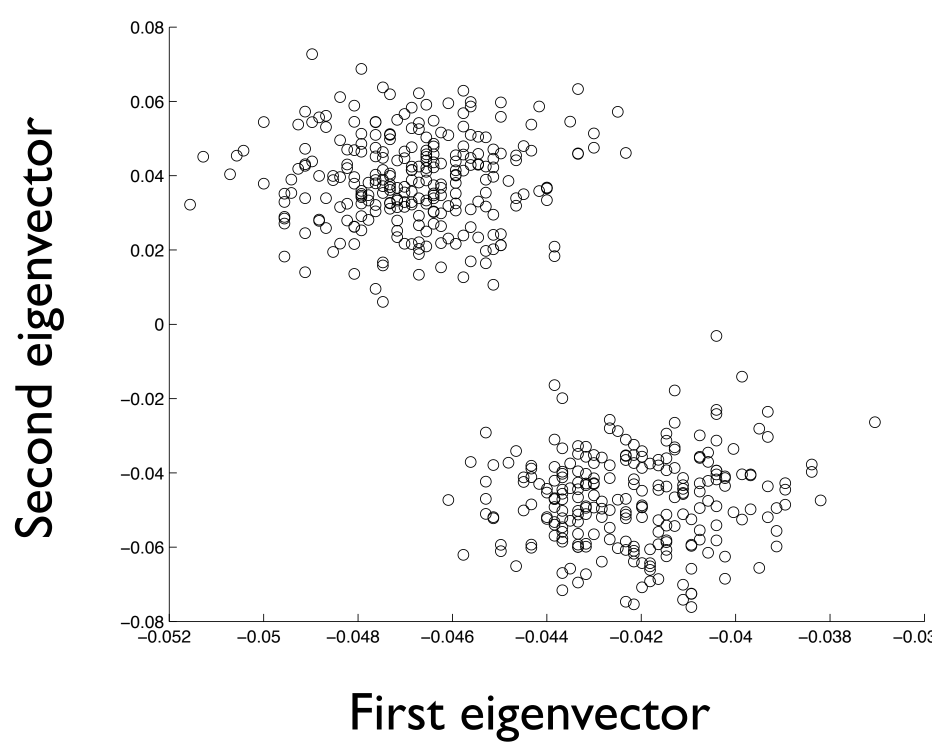


# Illustration of SC

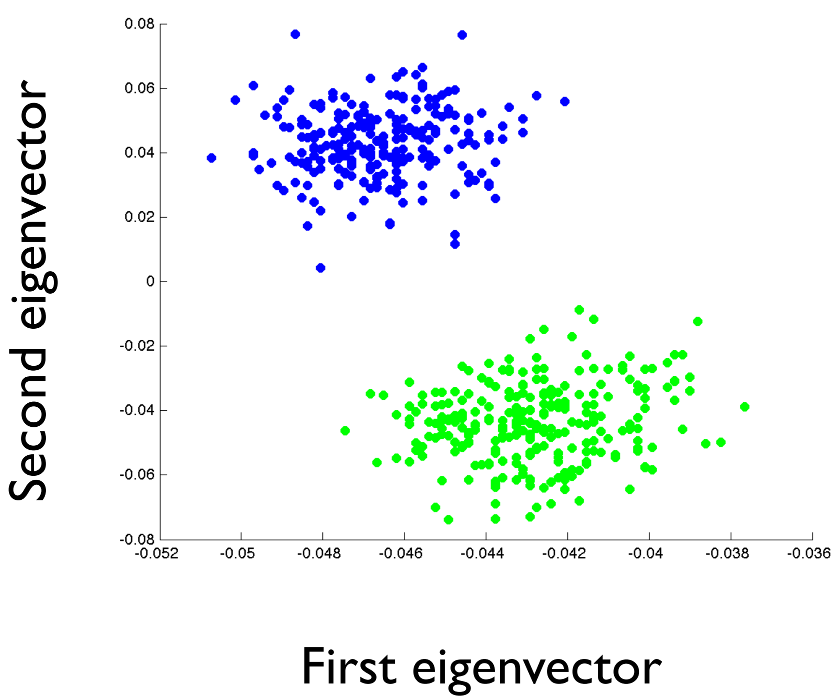


$L =$






cluster



## SC for finding $K$ clusters (Shi and Malik (00), Ng et. al ('02))

- Compute the $n \times K$ matrix $V$ of top $K$ eigenvectors of $L$.
- Cluster the rows of $V$ into $K$ clusters. (eg. using k-means)

# SC for finding $K$ clusters (Shi and Malik (00), Ng et. al ('02))

- Compute the $n \times K$ matrix $V$ of top $K$ eigenvectors of $L$.
- Cluster the rows of $V$ into $K$ clusters. (eg. using k-means)

row of $V$ represents node in the graph

# Popularity of spectral clustering

- Computational advantage :

  -requires eigenvector decomposition which is very fast

  *Theoretical backing :*

  - relaxation of various *cut*-based measures

    (Hagen & Kahng ('92), Shi & Malik ('00), Ng et al, ('02))

  - Stochastic Block Model and its extensions

    (McSherry ('01), Rohe. et. al ('11), Chaudhari et. al. ('12), Sussman ('12), Fishkind ('11))

Regularization proposed by Amini, Chen, Bickel and Levina (*AoS, 2013*)

Performance of spectral clustering improves greatly through regularization

Regularization proposed by Amini, Chen, Bickel and Levina (*AoS, 2013*)

Performance of spectral clustering improves greatly through regularization

- Add a constant matrix to the adjacency matrix $A$.

$$A_\tau = A + \frac{\tau}{n}\mathbf{1}\mathbf{1}', \qquad \tau > 0.$$

- Construct the Laplacian $L_\tau$ from $A_\tau$.
- Cluster the rows of $V_\tau$ into $K$ clusters.

$$V_\tau = \text{matrix of top } K \text{ eigenvectors of } L_\tau$$

Regularization proposed by Amini, Chen, Bickel and Levina (*AoS, 2013*)

Performance of spectral clustering improves greatly through regularization

- Add a constant matrix to the adjacency matrix $A$.

$$A_\tau = A + \frac{\tau}{n}\mathbf{1}\mathbf{1}', \qquad \tau > 0.$$

- Construct the Laplacian $L_\tau$ from $A_\tau$.
- Cluster the rows of $V_\tau$ into $K$ clusters.

$$V_\tau = \text{matrix of top } K \text{ eigenvectors of } L_\tau$$

Alternative forms of regularization proposed and analyzed in Chaudhuri et. al (2012), Qin & Rohe ('13)

## Stochastic Block Model

# Stochastic Block Model (SBM) (Holland et. al ('83))

Given a set of $n$ nodes,

edge $(i, j)$, drawn independently with probability $P_{ij}$

# Stochastic Block Model (SBM) (Holland et. al ('83))

Given a set of $n$ nodes,

edge $(i, j)$, drawn independently with probability $P_{ij}$

$$P = (P_{ij}) = \begin{bmatrix} p_1 & q \\ q & p_2 \end{bmatrix}_{n \times n}$$

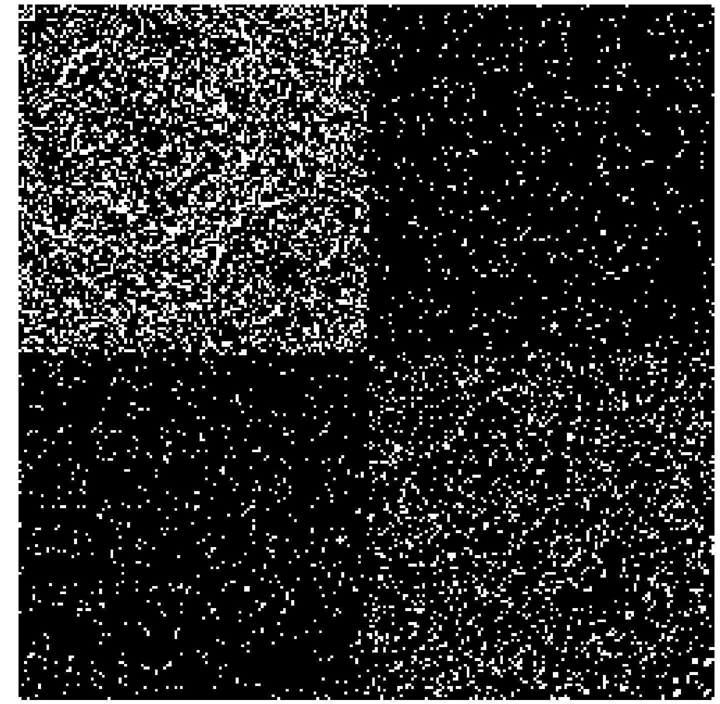
SBM with two blocks

## Edge probability matrix *P*

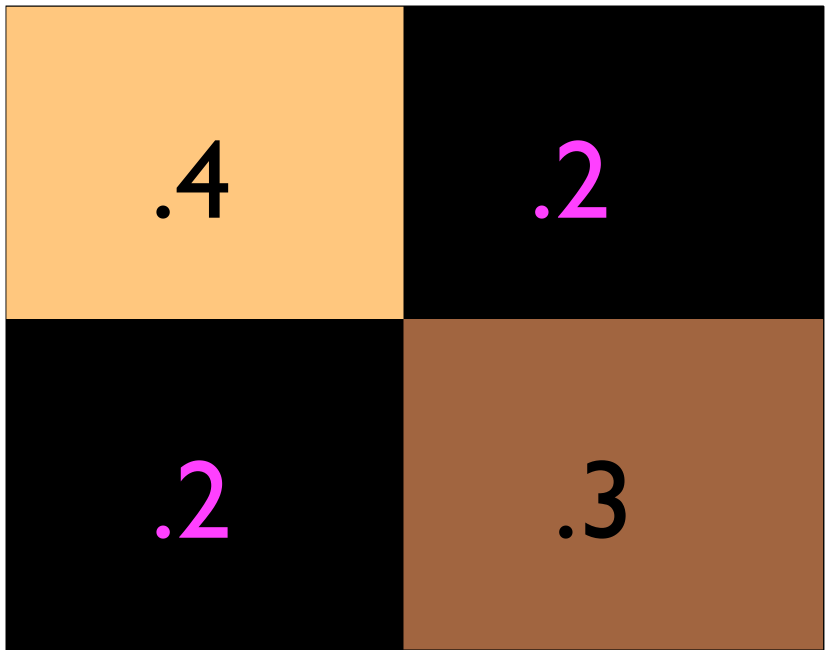
| | |
|---|---|
| .4 | .2 |
| .2 | .3 |

sample

## Adjacency matrix *A*

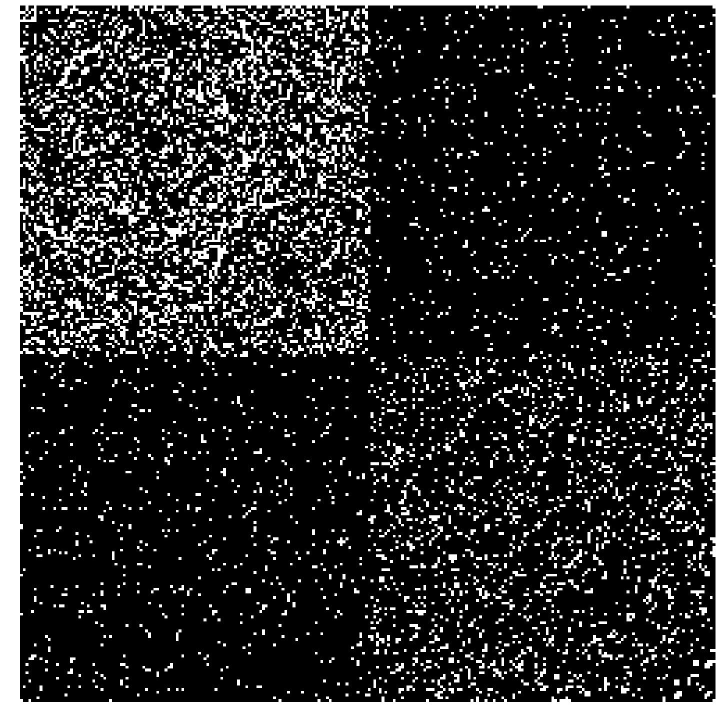


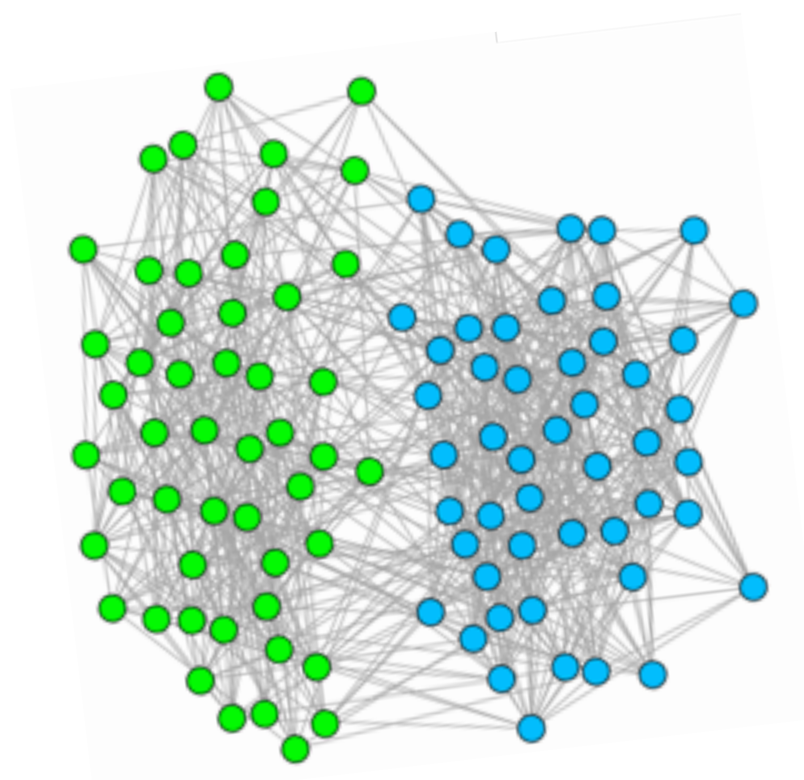
## Edge probability matrix *P*




sample

## Adjacency matrix *A*





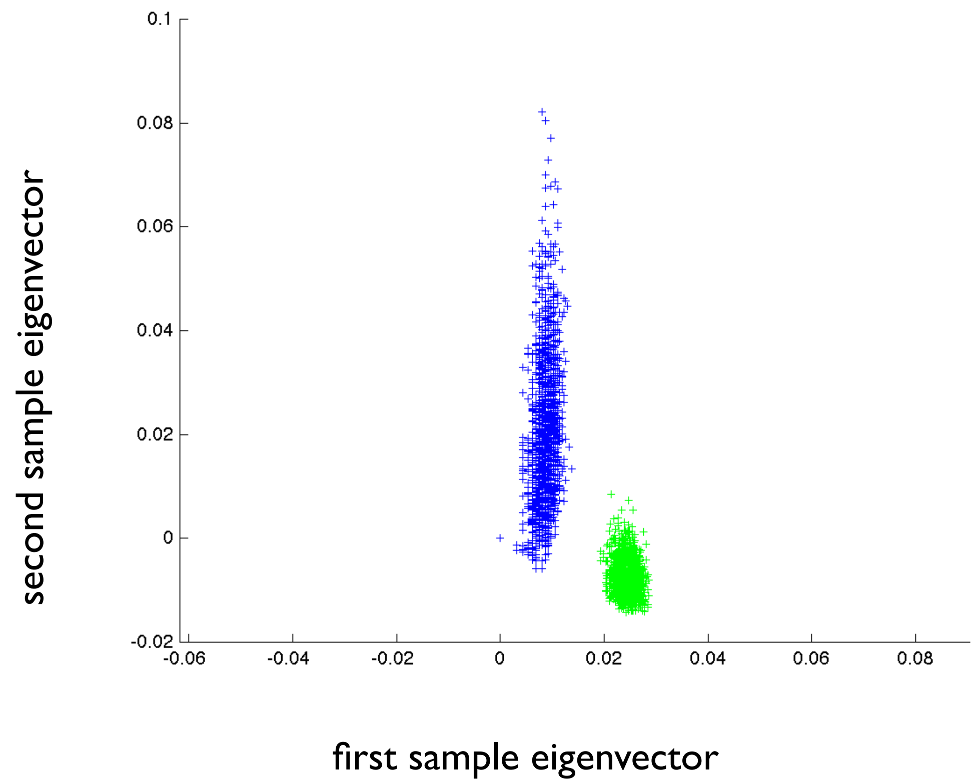
# Analysis of regularization for the SBM
(Focus on $K = 2$)

# Comparing unregularized vs. regularized SC

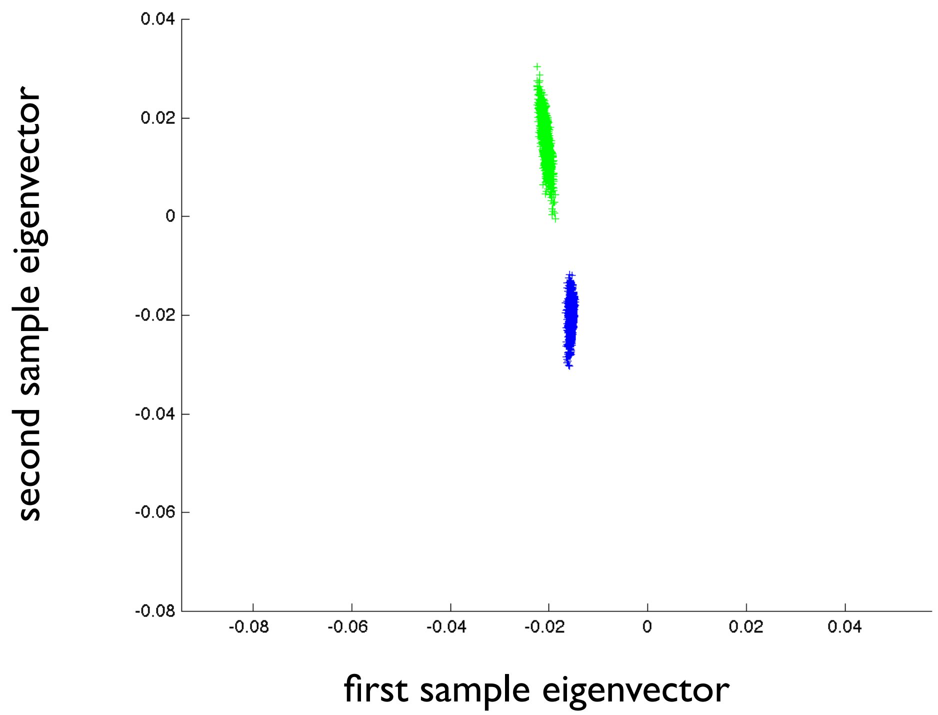
$$P = \begin{pmatrix} .003 & .0025 \\ .0025 & .04 \end{pmatrix}$$

$n = 3000$

$\tau = 0$



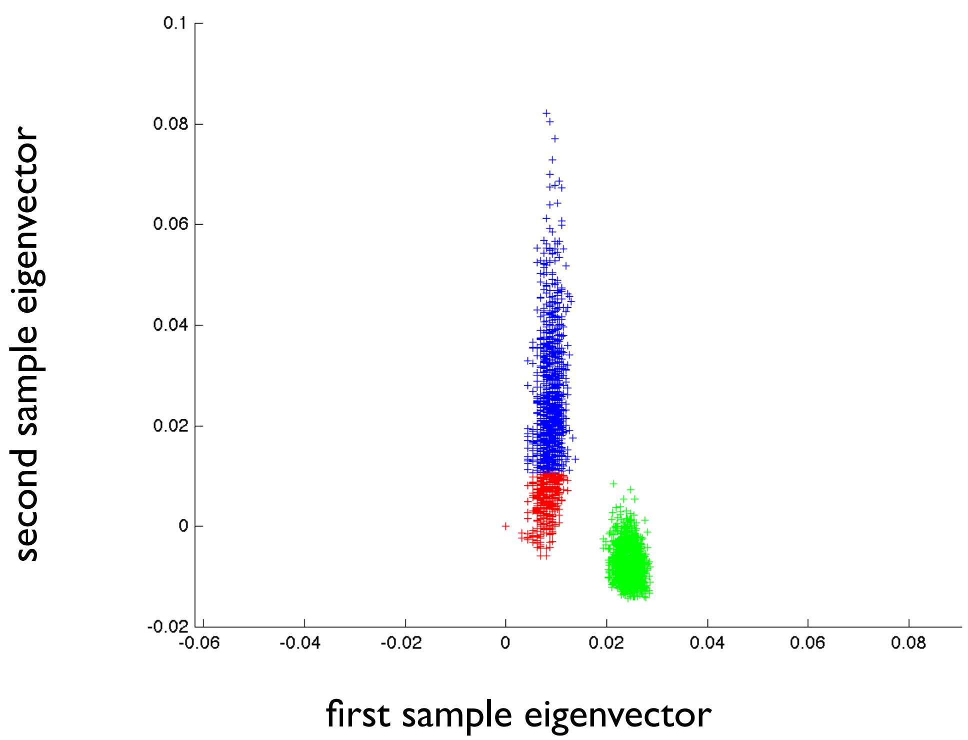

$\tau = 20$



# Comparing unregularized vs. regularized SC

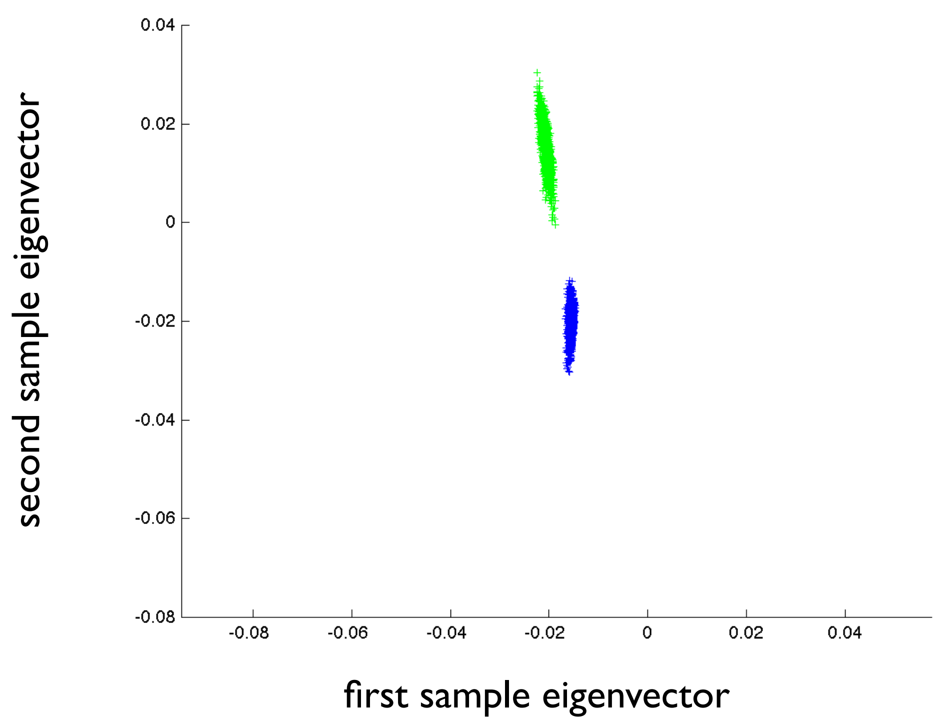
$$P = \begin{bmatrix} .003 & .0025 \\ .0025 & .04 \end{bmatrix} \qquad n = 3000$$

$\tau = 0$




k-means success : 87%

$\tau = 20$




k-means success : 100%

# Recap : Regularized spectral clustering

- Construct,

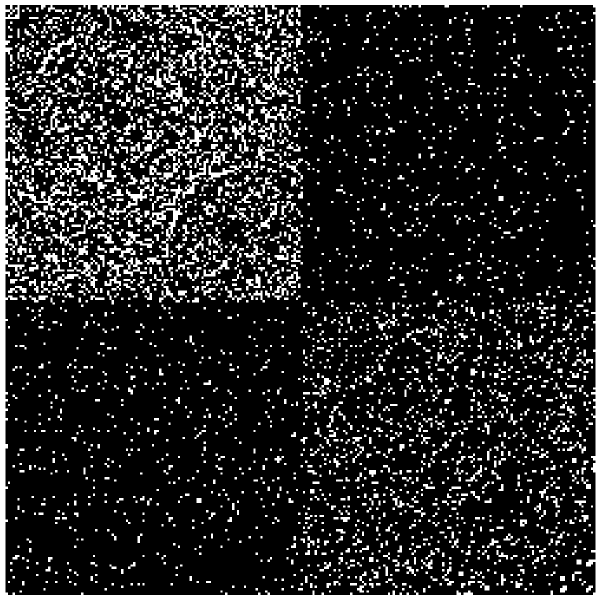
$$A_\tau = A + \frac{\tau}{n}\mathbf{1}\mathbf{1}', \qquad \tau > 0.$$

$$L_\tau = D_\tau^{-1/2} A_\tau D_\tau^{-1/2}$$

- Cluster the rows of $V_\tau$ into two clusters.

$V_\tau =$ matrix of top two eigenvectors of $L_\tau$

# Population level quantities

$A =$ 

$$P = \begin{pmatrix} p_1 & q \\ q & p_2 \end{pmatrix}$$

# Population level quantities

Adjacency matrix $A_\tau$ : $$P_\tau = P + \frac{\tau}{n}\mathbf{1}\mathbf{1}'$$

Laplacian matrix $L_\tau$ : $$L_\tau^{pop}$$

# Population level quantities

Adjacency matrix $A_\tau$ : $$P_\tau = P + \frac{\tau}{n}\mathbf{1}\mathbf{1}'$$

Laplacian matrix $L_\tau$ : $$L_\tau^{pop}$$

## Recall:

- $V_\tau$ is the $n \times 2$ sample eigenvector matrix.
- Rows of $V_\tau$ corresponds to nodes in the graph.

# Population level quantities

Adjacency matrix $A_\tau$ : $$P_\tau = P + \frac{\tau}{n}\mathbf{1}\mathbf{1}'$$

Laplacian matrix $L_\tau$ : $$L_\tau^{pop}$$

- The population version of $V_\tau$ ($V_\tau^{pop}$) has two distinct rows.
- Distinct rows corresponds to nodes in the two communities

# Population level quantities

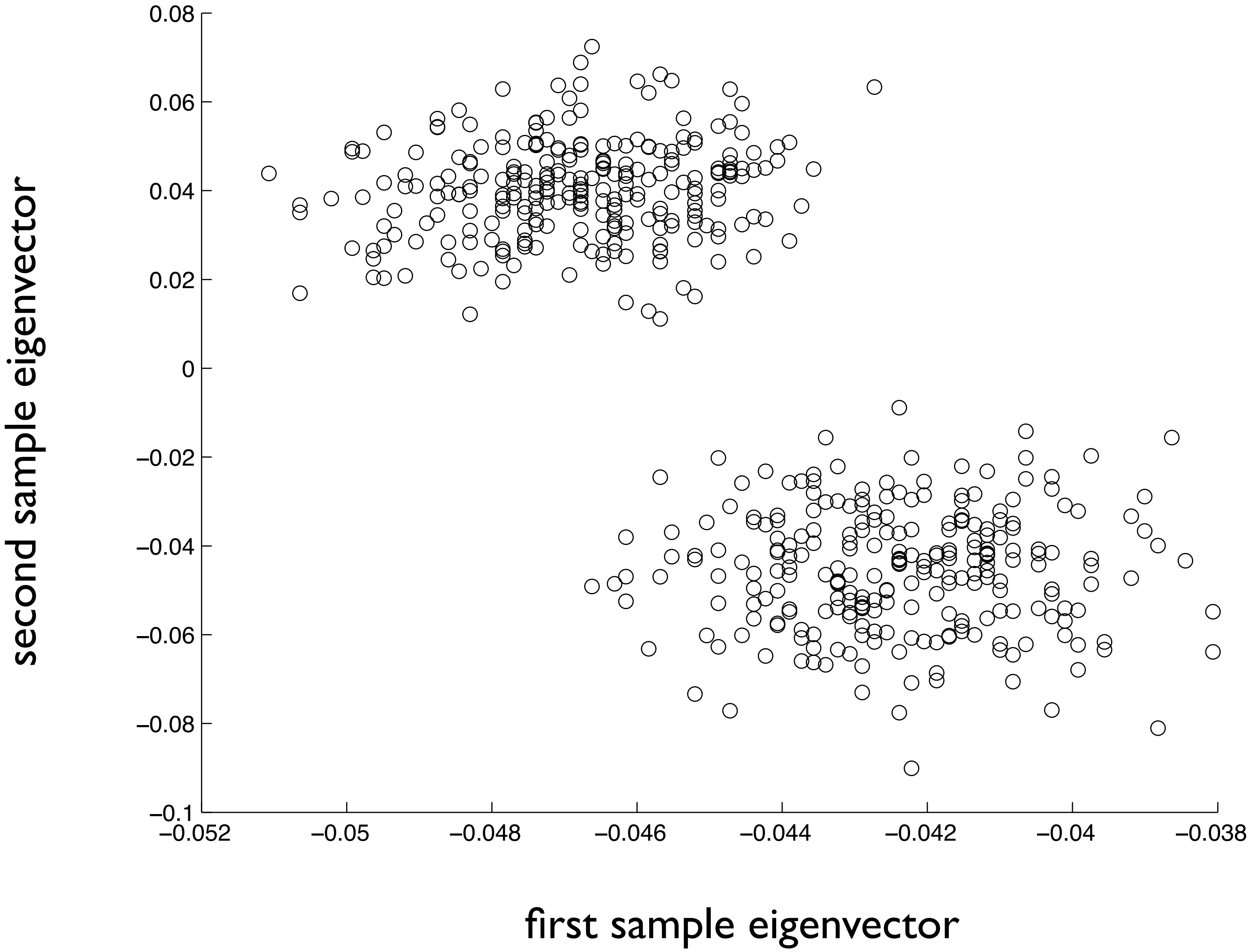
Adjacency matrix $A_\tau$ : $P_\tau = P + \frac{\tau}{n}\mathbf{1}\mathbf{1}'$

Laplacian matrix $L_\tau$ : $L_\tau^{pop}$

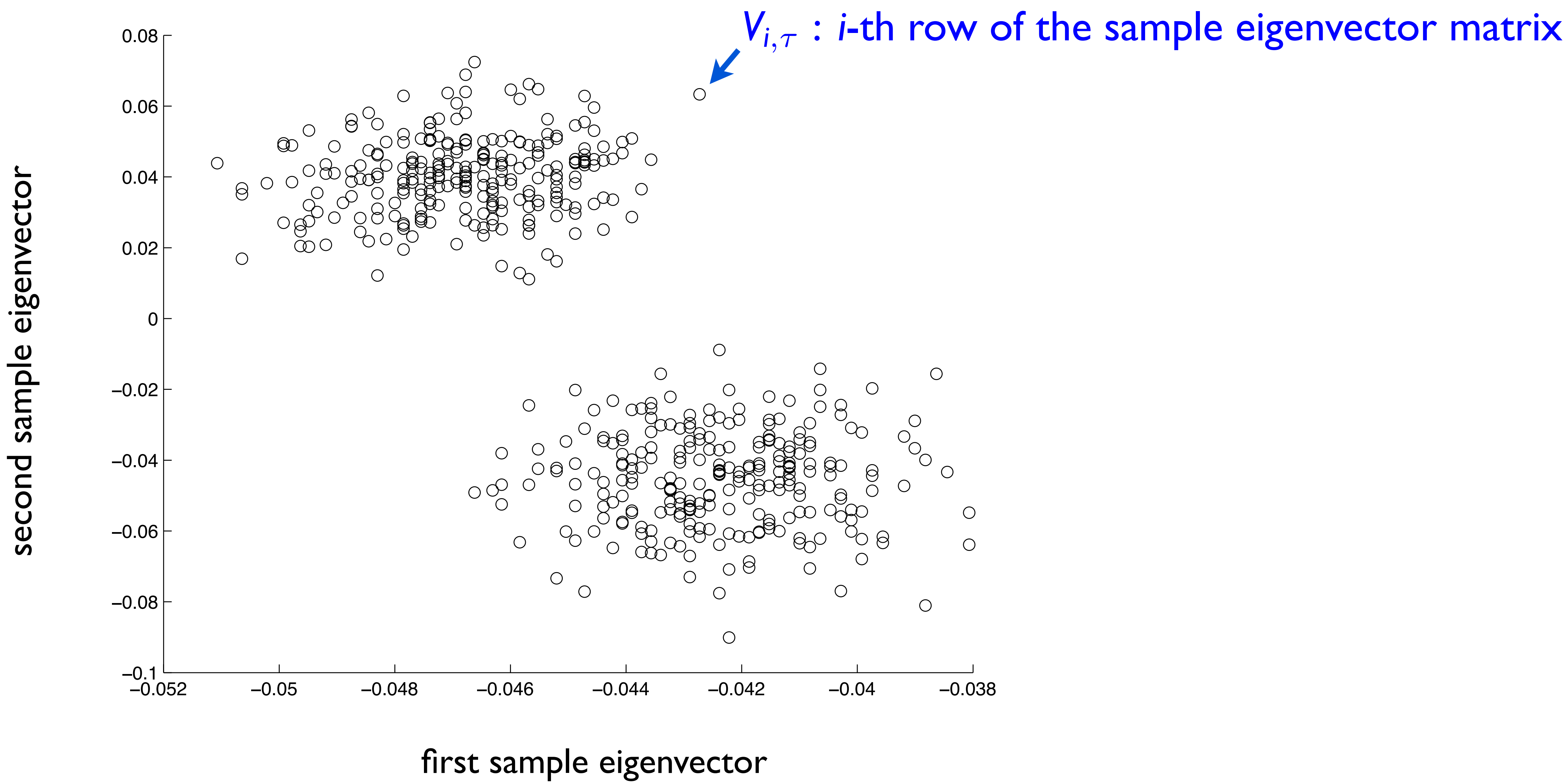
- The population version of $V_\tau$ ($V_\tau^{pop}$) has two distinct rows.
- Distinct rows corresponds to nodes in the two communities

Denote these by $\text{center}_{1,\tau}$, $\text{center}_{2,\tau}$

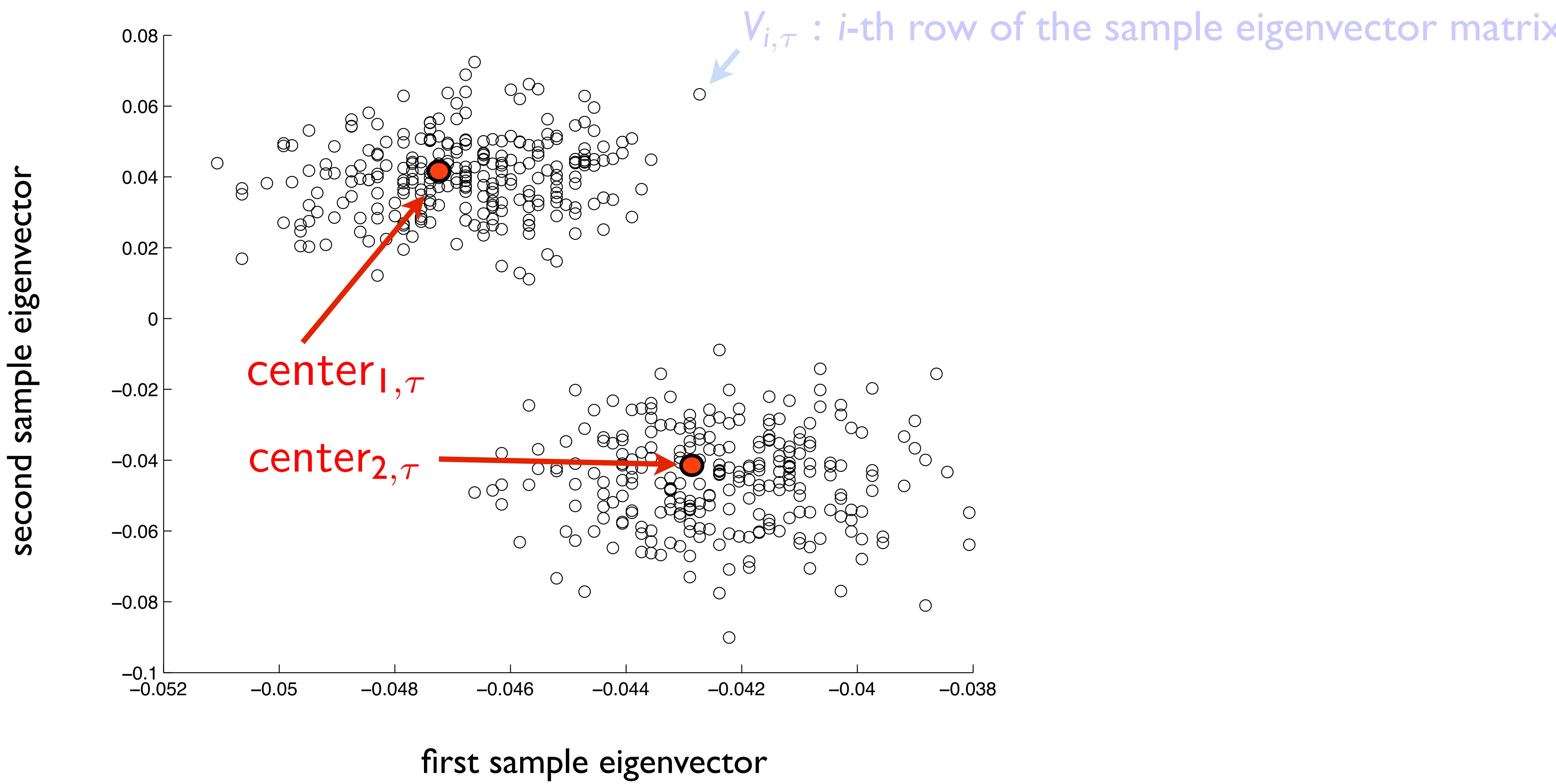
# Scatter plot for a particular $\tau$



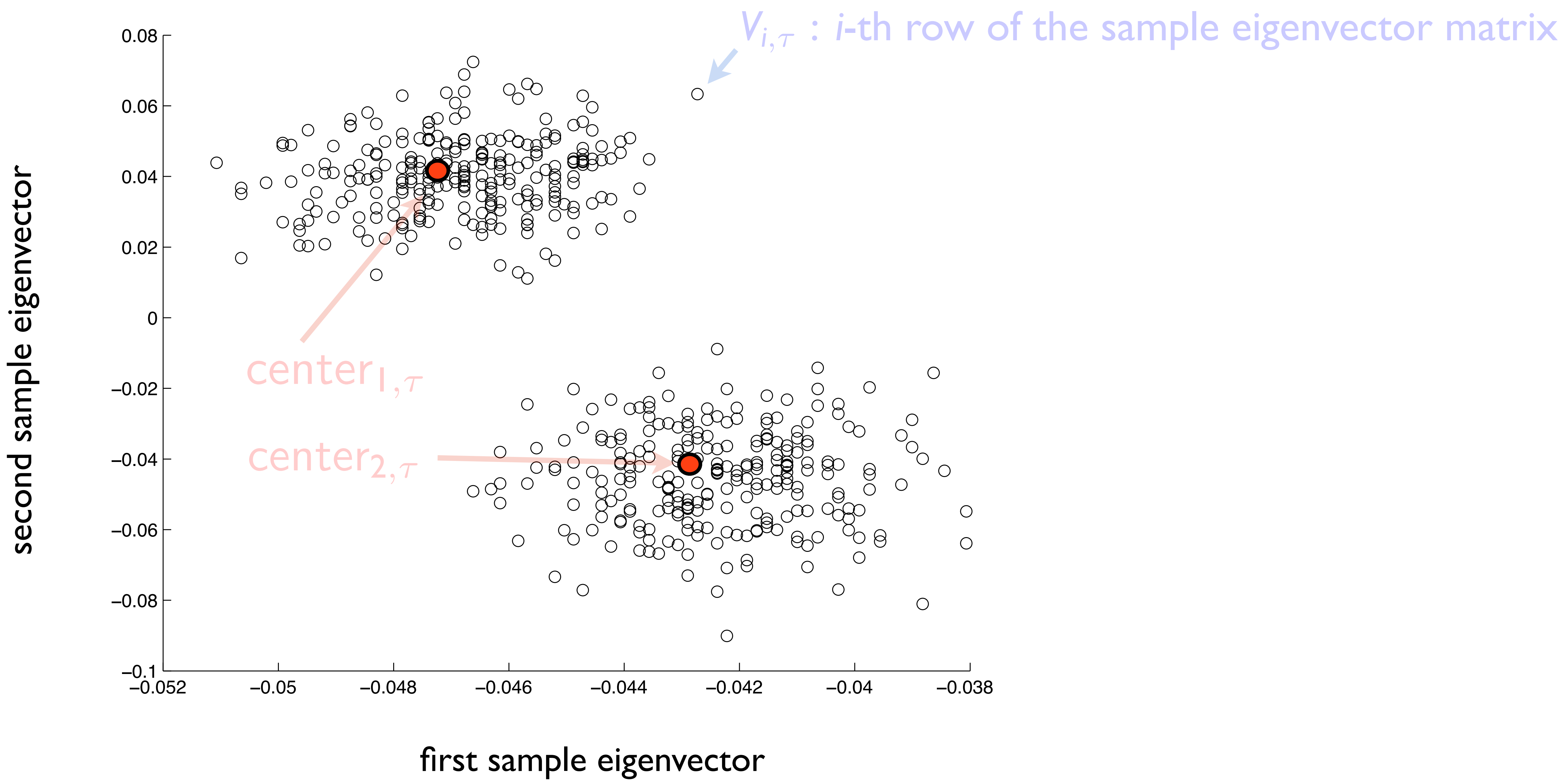
# Scatter plot for a particular $\tau$



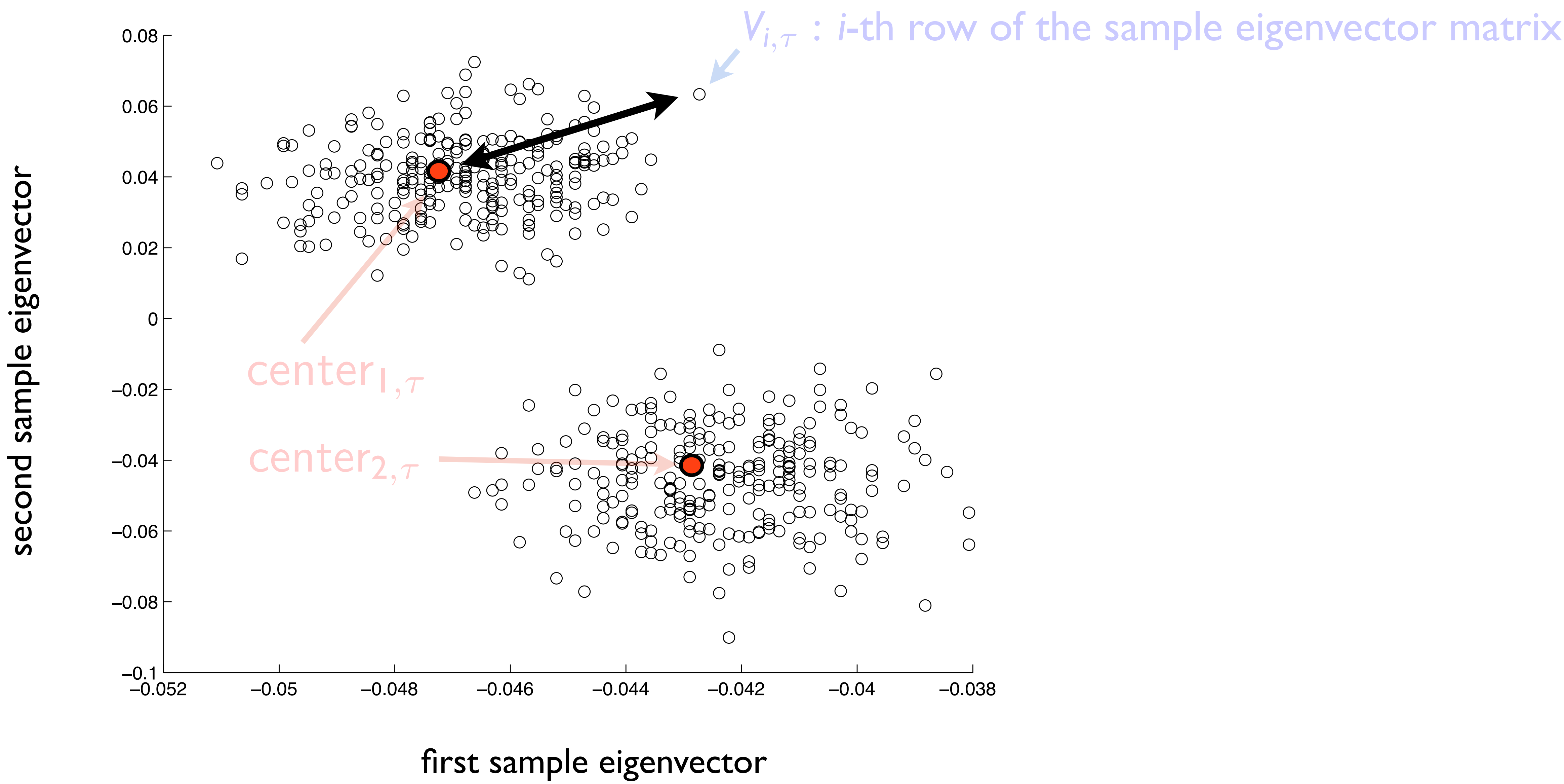
# Scatter plot for a particular $\tau$



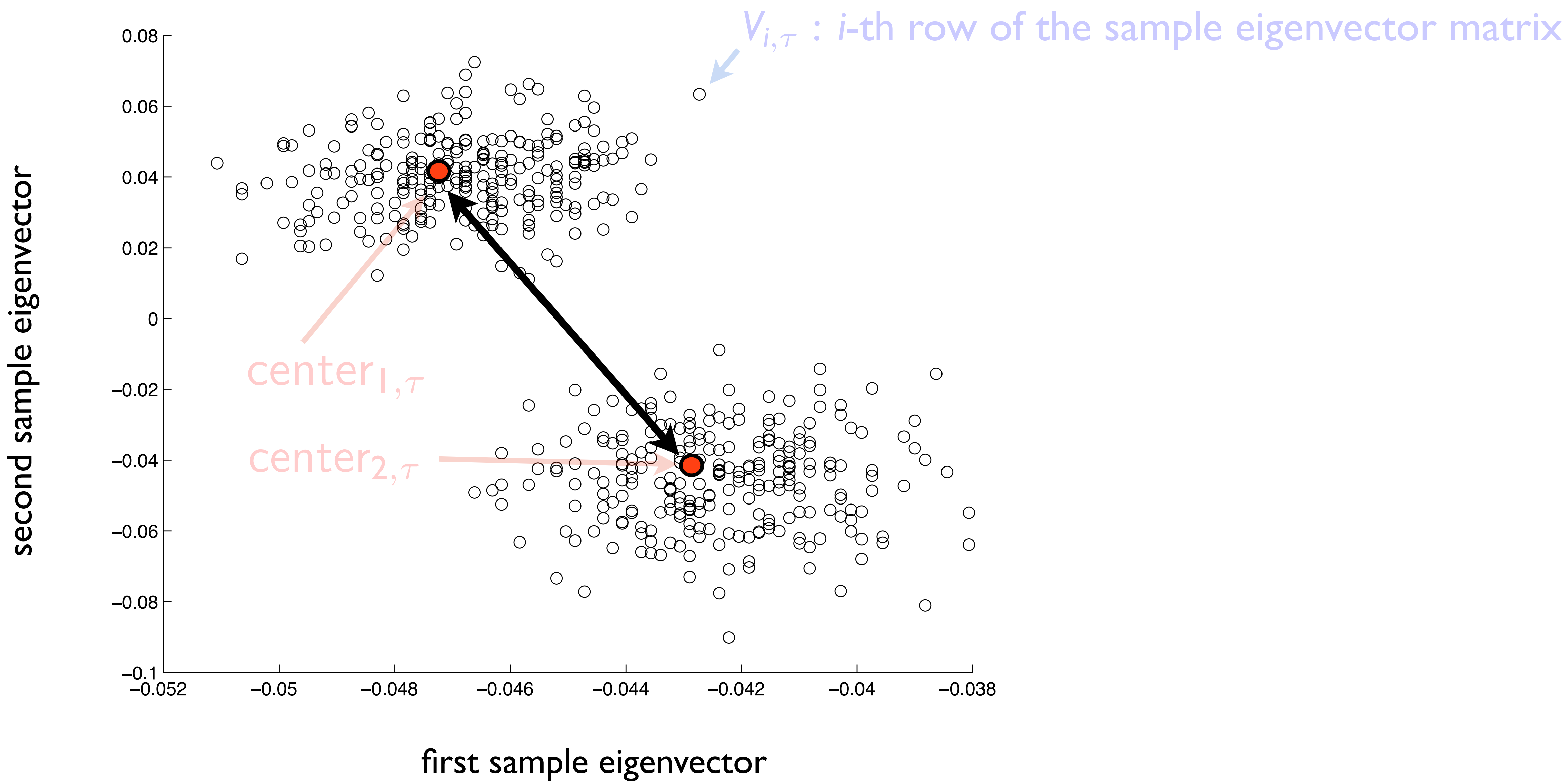
# Scatter plot for a particular $\tau$



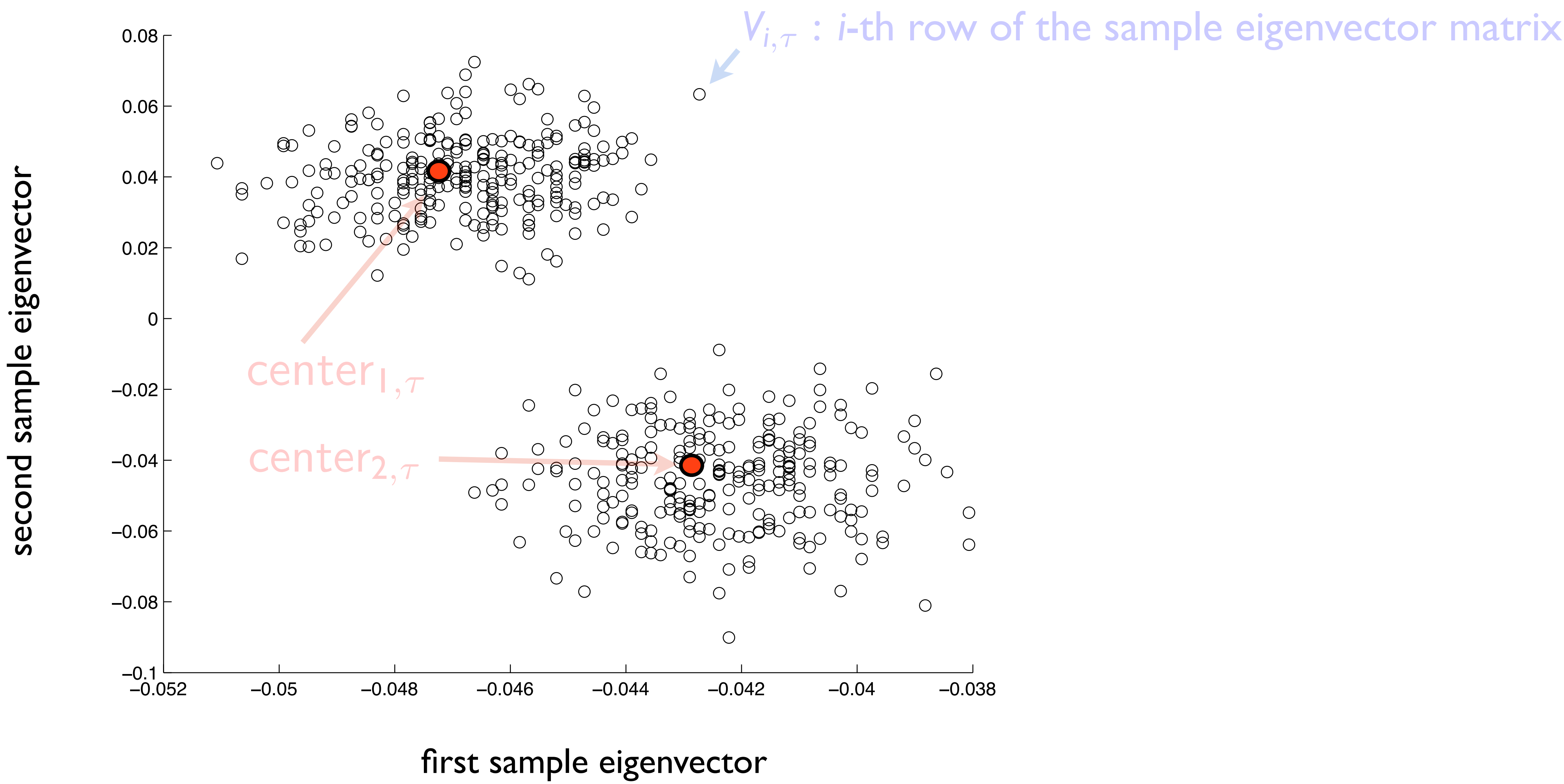
# Scatter plot for a particular $\tau$



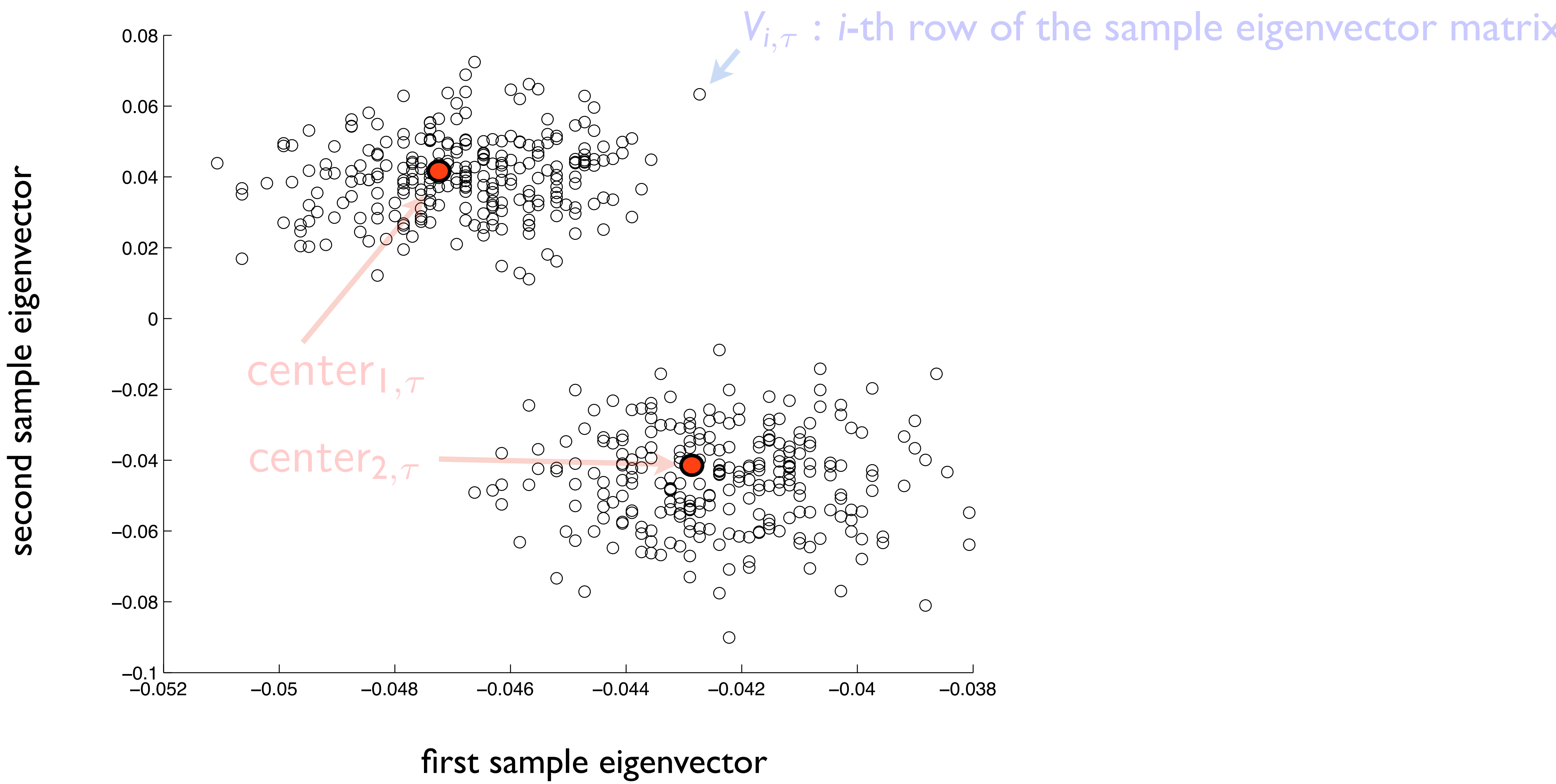
# Scatter plot for a particular $\tau$



# Scatter plot for a particular $\tau$



# Scatter plot for a particular $\tau$




$$\text{pert}_\tau = \frac{\max_{k=1,2} \ \max_{i \in \text{cluster } k} \|V_{i,\tau} - \text{center}_{k,\tau}\|}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\max_{k=1,2} \max_{i \in \text{cluster } k} \|V_{i,\tau} - \text{center}_{k,\tau}\|}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

↑ does not depend on $\tau$

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

Implication of matrix perturbation theory (Davis - Kahan) :

$$\text{pert}_\tau \lesssim \sqrt{n}\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

Implication of matrix perturbation theory (Davis - Kahan) :

$$\text{pert}_\tau \lesssim \sqrt{n}\,\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

second eigenvalue of $L_\tau^{pop}$

($\mu_{2,\tau}$ *decreases with* $\tau$)

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

Implication of matrix perturbation theory (Davis - Kahan) :

$$\text{pert}_\tau \lesssim \sqrt{n}\,\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

Implication of concentration of Laplacian (Oliveira ('10)):

If $\tau \gtrsim \log n$,

$$\|L_\tau - L_\tau^{pop}\| \lesssim \min\left\{\frac{1}{\sqrt{c_{1,n}+\tau}}, \frac{c_{2,n}}{(c_{1,n}+\tau)}\right\}\sqrt{\log n}$$ with high probability

# Understanding $\text{pert}_\tau$

$$\text{pert}_\tau = \frac{\text{"Distance" between eigenvector matrices of } L_\tau \text{ and } L_\tau^{pop}}{\|\text{center}_{1,\tau} - \text{center}_{2,\tau}\|}$$

Implication of matrix perturbation theory (Davis - Kahan) :

$$\text{pert}_\tau \lesssim \sqrt{n}\,\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

Improvements using extension of techniques in Balakrishnan et. al. ('11).

Let,

$$d_n := \text{average expected degree of the nodes}$$

Set,

$$\tau = d_n$$

Let,

$$d_n := \text{average expected degree of the nodes}$$

Set,

$$\tau = d_n$$

Result (SBM with two blocks):

If

$$d_n \gtrsim \frac{\sqrt{n \log n}}{\mu_{2,0}}$$

then regularized SC recovers the clusters with high probability.

Let,

$$d_n := \text{average expected degree of the nodes}$$

Set,

$$\tau = d_n$$

Result (SBM with two blocks):

If

$$d_n \gtrsim \frac{\sqrt{n \log n}}{\mu_{2,0}}$$

then regularized SC recovers the clusters with high probability.

Summary:

Unlike McSherry ('01), Rohe et. al. ('11), Chaudhuri et. al ('12), the results don't depend on the minimum degree.

# Choice of regularization parameter

Recall: trade-offs dictated by

$$\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

Recall: trade-offs dictated by

$$\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

- Consider,

$$\frac{\|L_\tau - \hat{L}_\tau^{pop}\|}{\hat{\mu}_{2,\tau}}$$

- Choose $\tau$ that minimizes the statistic, over a grid of values.

Recall: trade-offs dictated by

$$\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

- Consider,

$$\frac{\|L_\tau - \hat{L}_\tau^{pop}\|}{\hat{\mu}_{2,\tau}}$$

Estimates based on estimated SBM (or degree corrected SBM)

- Choose $\tau$ that minimizes the statistic, over a grid of values.

Recall: trade-offs dictated by

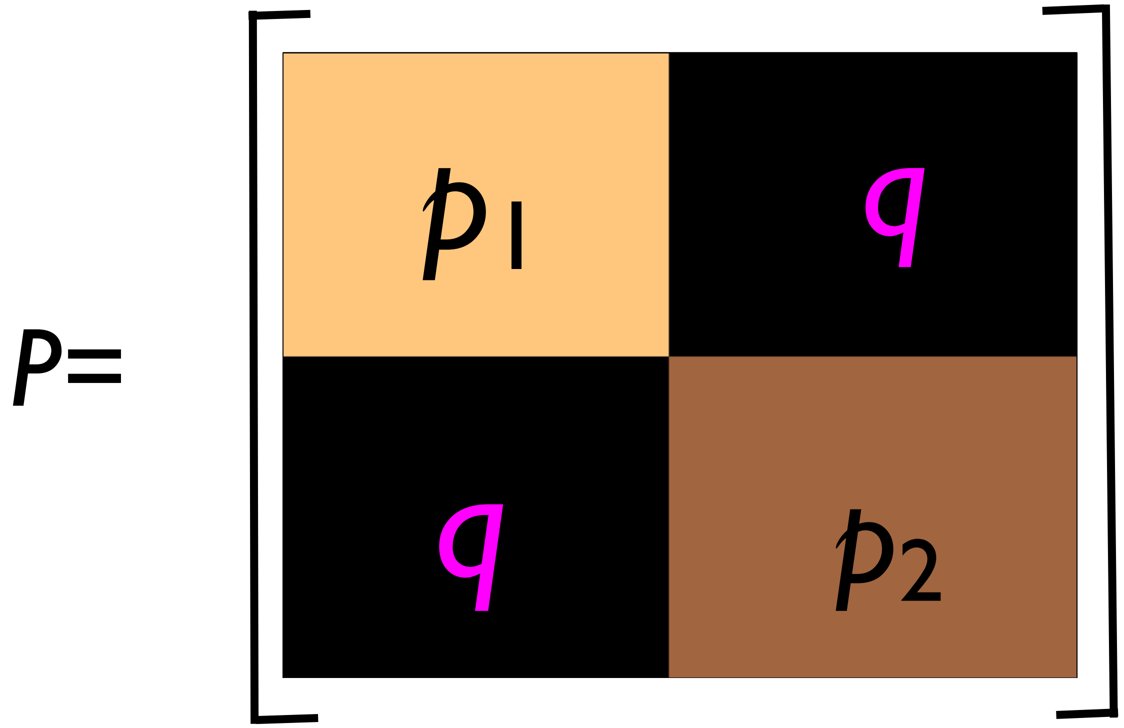
$$\frac{\|L_\tau - L_\tau^{pop}\|}{\mu_{2,\tau}}$$

- Consider,

$$\frac{\|L_\tau - \hat{L}_\tau^{pop}\|}{\hat{\mu}_{2,\tau}}$$

- Choose $\tau$ that minimizes the statistic, over a grid of values.

# The estimates $\hat{L}_{\tau}^{pop}, \ \hat{\mu}_{2,\tau}$

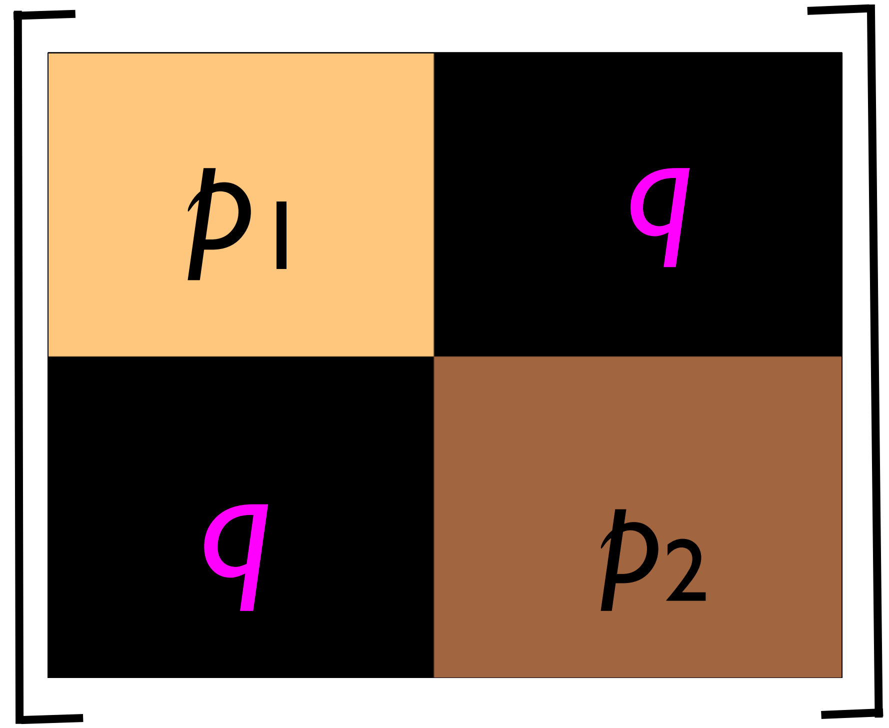



For a particular $\tau$,

- Let $C_1, \ C_2$ be clusters outputted from regularized SC algorithm.

# The estimates $\hat{L}_{\tau}^{pop}$, $\hat{\mu}_{2,\tau}$

$P=$




For a particular $\tau$,

- Let $C_1$, $C_2$ be clusters outputted from regularized SC algorithm.
- Estimate $p_1$, $p_2$ and $q$ from $C_1$ and $C_2$

$e.g.$ $\hat{p}_1 =$ fraction of edges for nodes in $C_1$

# The estimates $\hat{L}_{\tau}^{pop}$, $\hat{\mu}_{2,\tau}$

$$\hat{P} = \begin{bmatrix} \hat{p}_1 & \hat{q} \\ \hat{q} & \hat{p}_2 \end{bmatrix}$$

For a particular $\tau$,

- Let $C_1$, $C_2$ be clusters outputted from regularized SC algorithm.
- Estimate $p_1$, $p_2$ and $q$ from $C_1$ and $C_2$

$e.g.$ $\hat{p}_1 =$ fraction of edges for nodes in $C_1$

## The estimates $\hat{L}_\tau^{pop}$, $\hat{\mu}_{2,\tau}$

$$\hat{P} = \begin{bmatrix} \hat{p}_1 & \hat{q} \\ \hat{q} & \hat{p}_2 \end{bmatrix}$$

For a particular $\tau$,

- Let $C_1, C_2$ be clusters outputted from regularized SC algorithm.
- Estimate $p_1, p_2$ and $q$ from $C_1$ and $C_2$

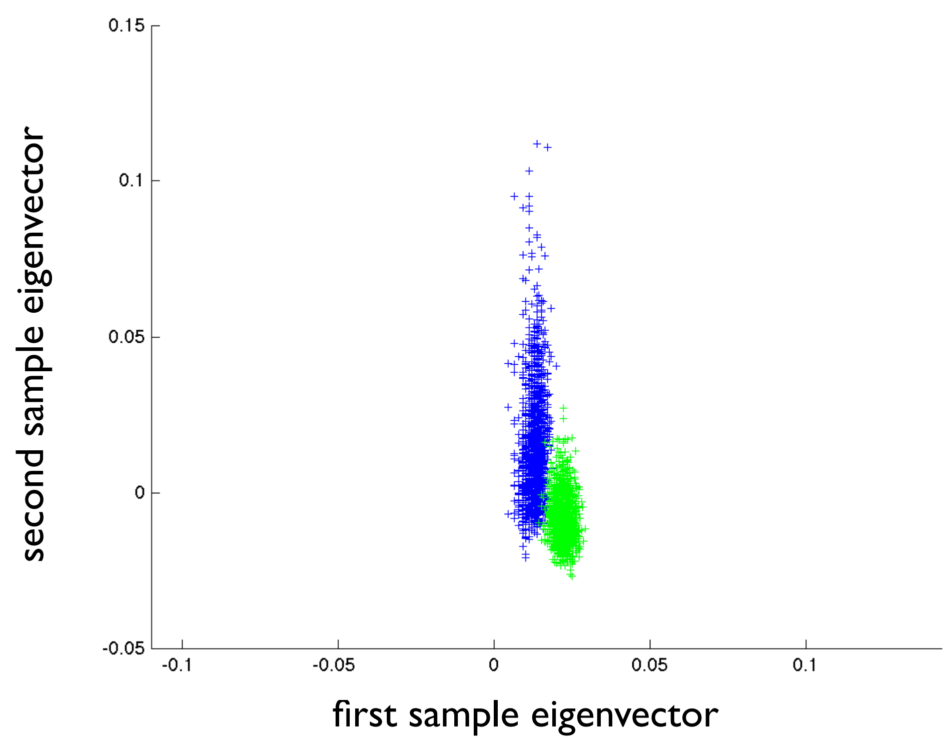
$e.g.\ \ \hat{p}_1 =$ fraction of edges for nodes in $C_1$

- Use $\hat{P}$ to calculate $\hat{L}_\tau^{pop}$. Take $\hat{\mu}_{2,\tau}$ to be the second eigenvalue of $\hat{L}_\tau^{pop}$.

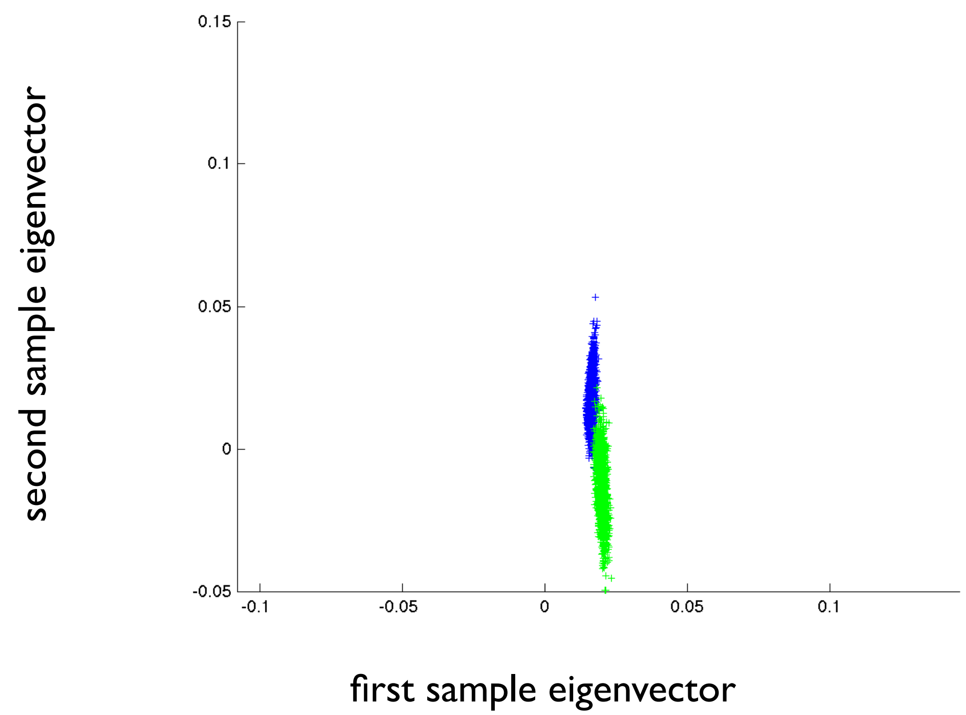
# Example

$$P = \begin{pmatrix} .003 & .0025 \\ .0025 & .01 \end{pmatrix} \qquad n = 3000$$

$\tau = 0$



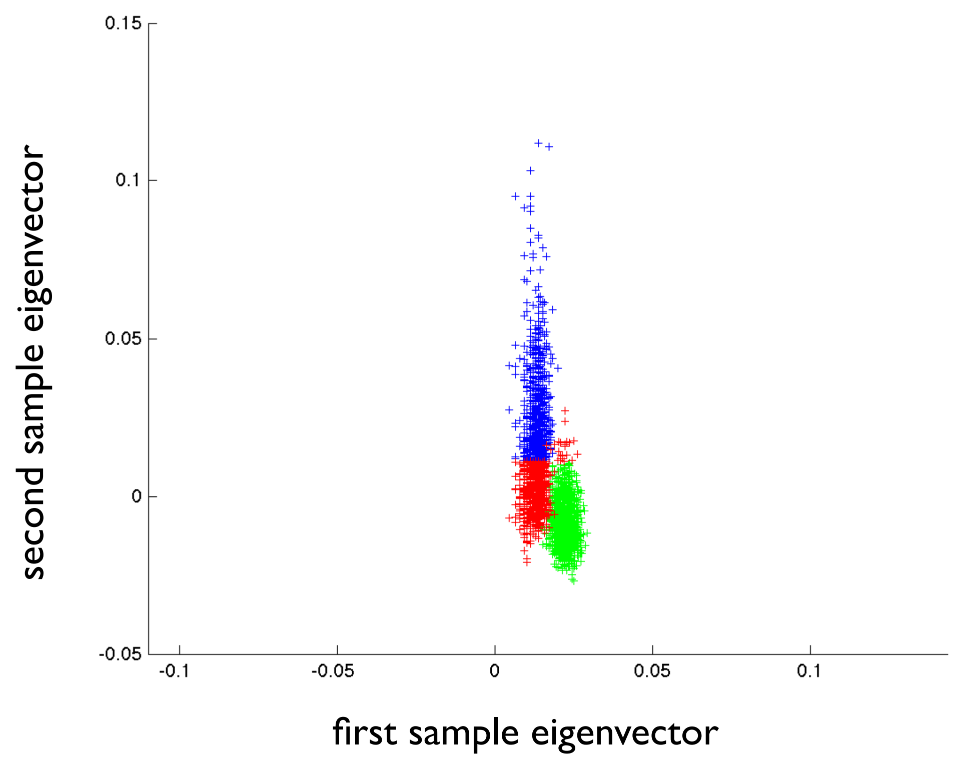

$\tau = 18$



# Example

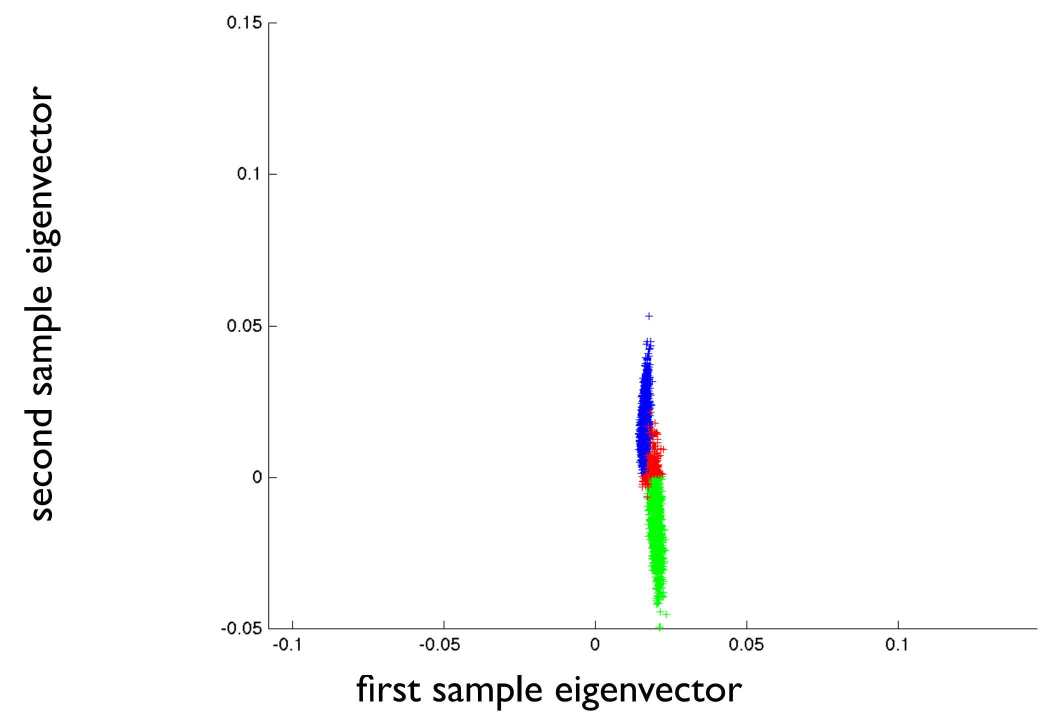
$$P = \begin{pmatrix} .003 & .0025 \\ .0025 & .01 \end{pmatrix} \qquad n = 3000$$

$\tau = 0$



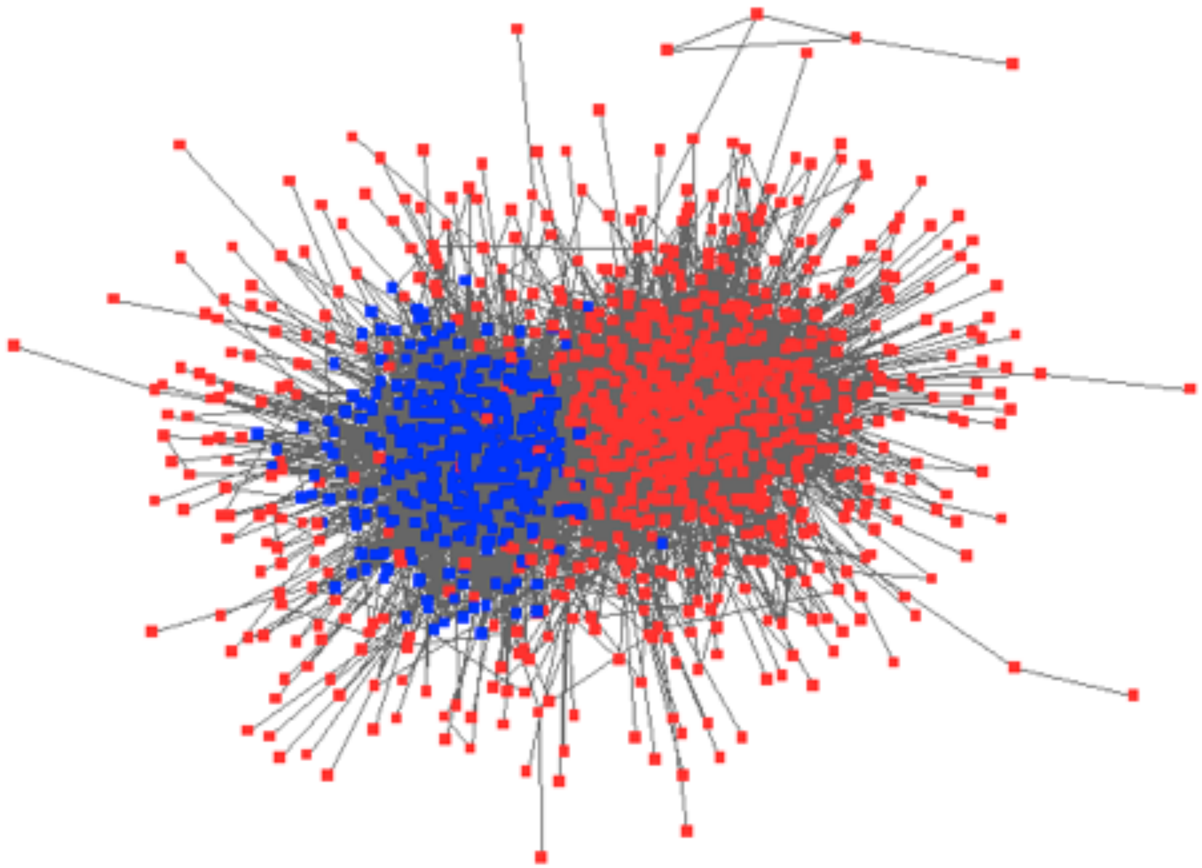

k-means success : 75%

$\tau = 18$




k-means success : 94%

[Political blog data]



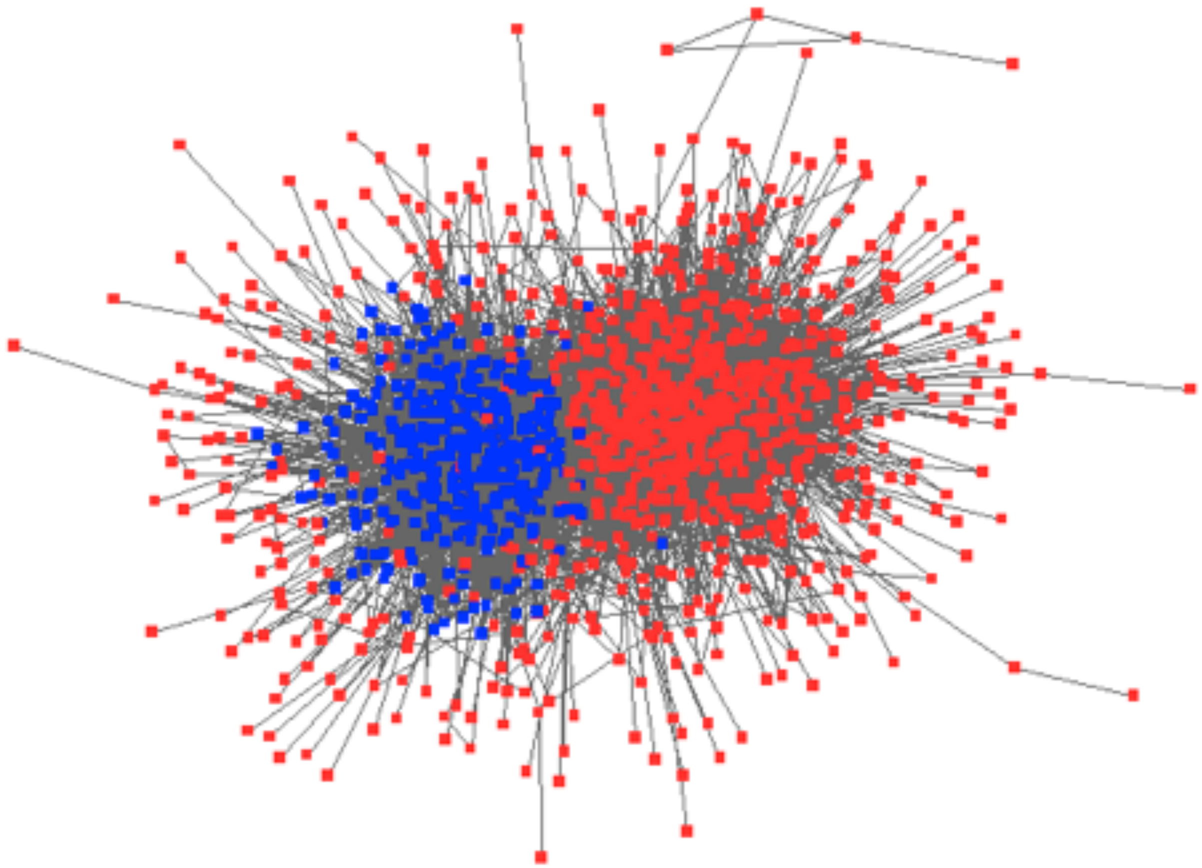
- Nodes are political blog sites. ($n = 1222$)

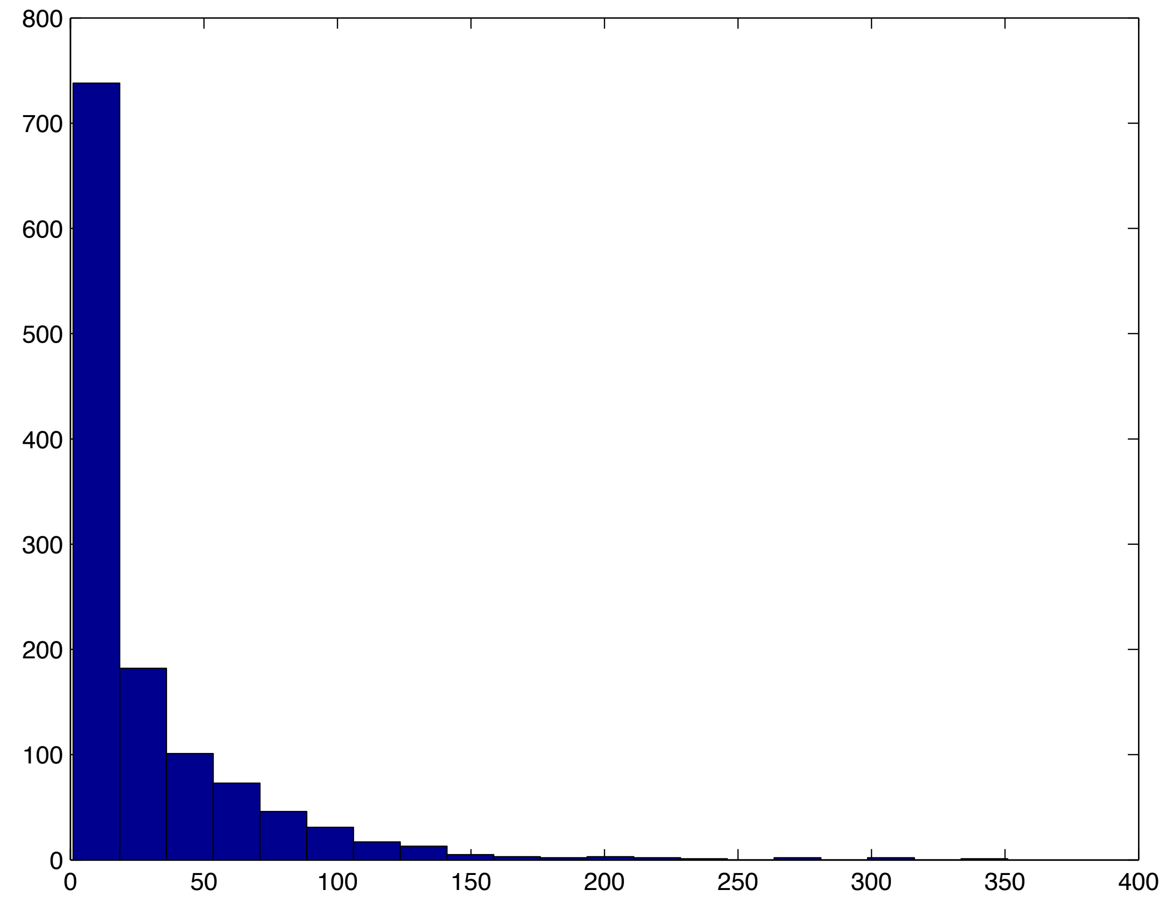
  red nodes : conservative blogs

  blue nodes : liberal blogs

- Edge between two nodes if either website has a link to the other.

*source : Adamic & Glance ('05)*






Histogram of degrees

- Nodes are political blog sites. ($n = 1222$)

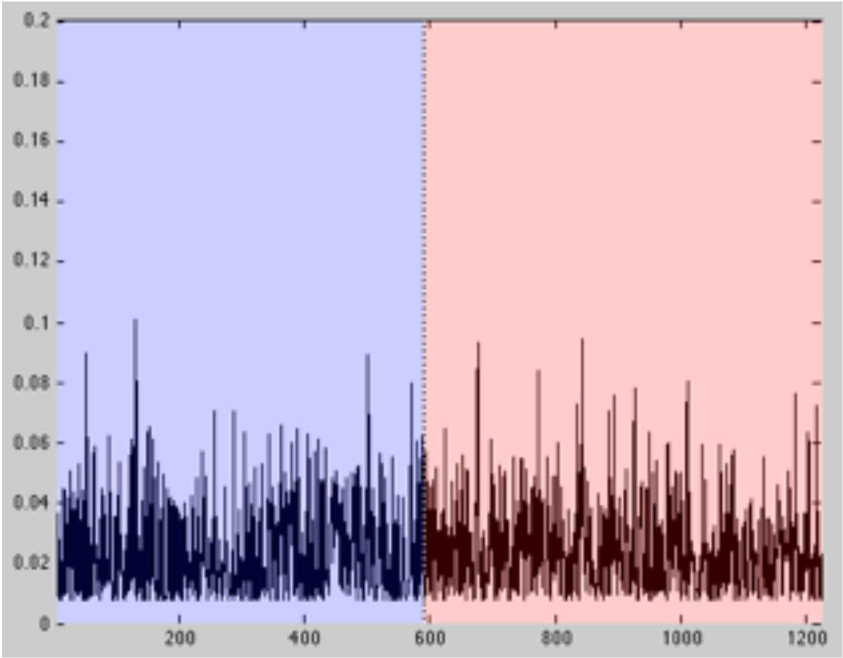
  red nodes : conservative blogs

  blue nodes : liberal blogs

- Edge between two nodes if either website has a link to the other.

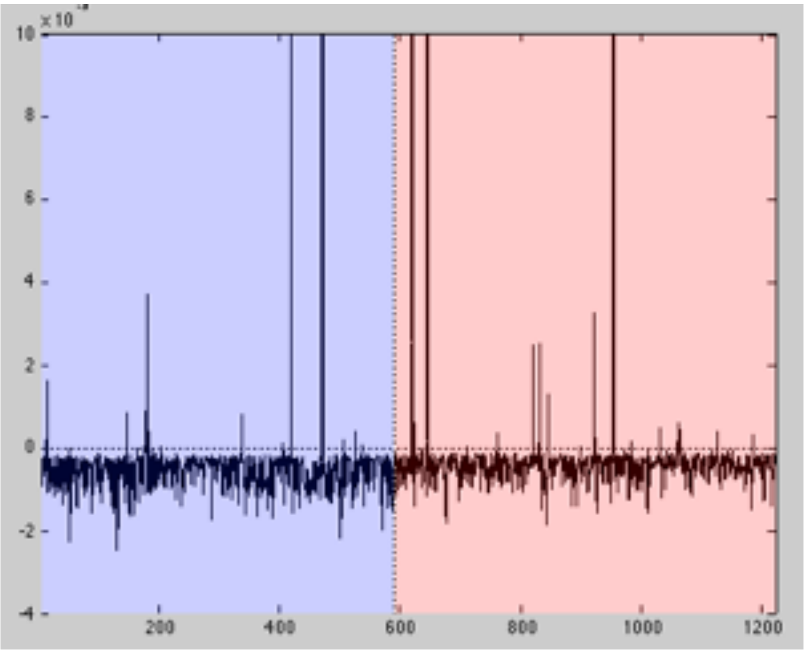
*source : Adamic & Glance ('05)*

# Political blogs data set




first eigenvector

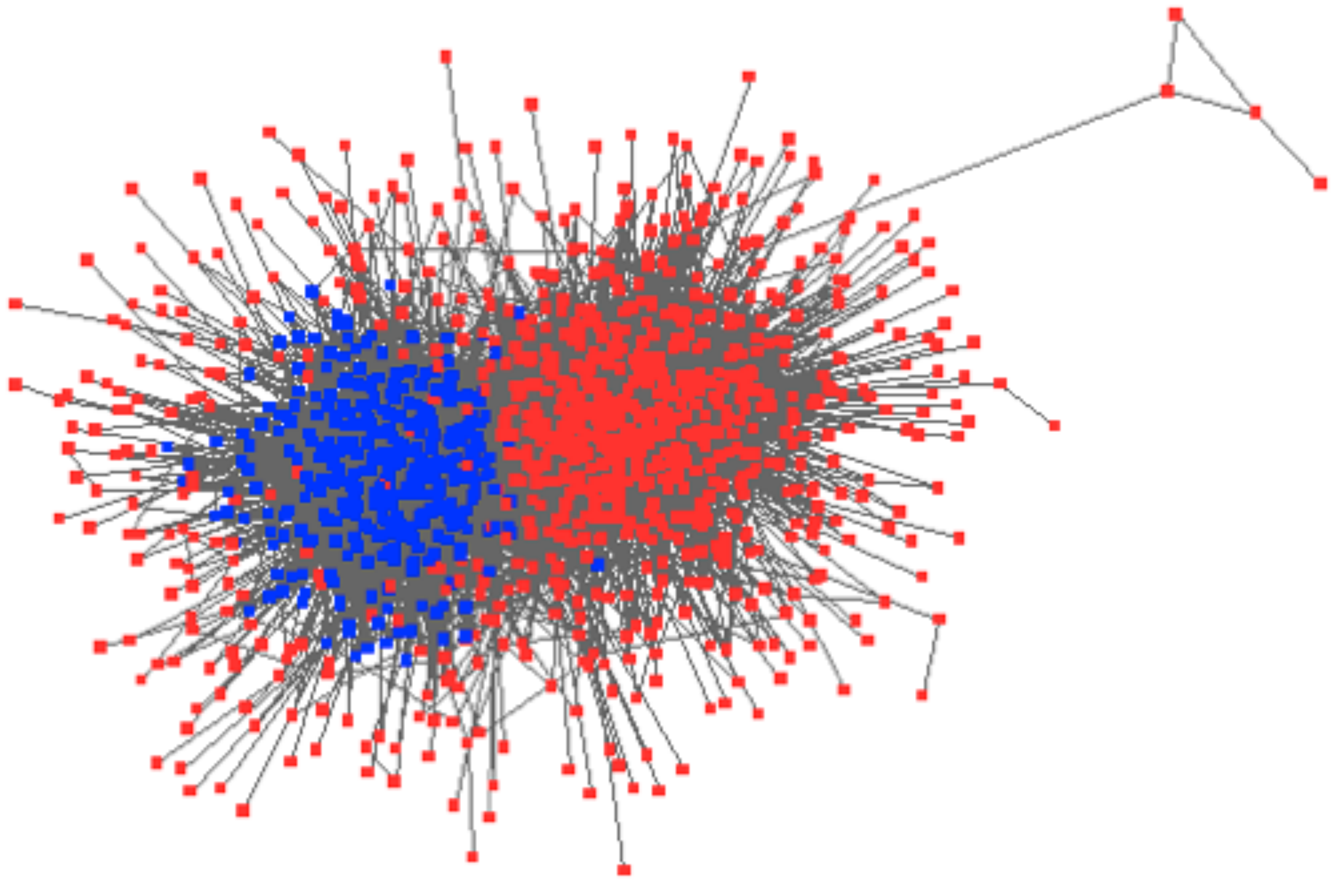
nodes

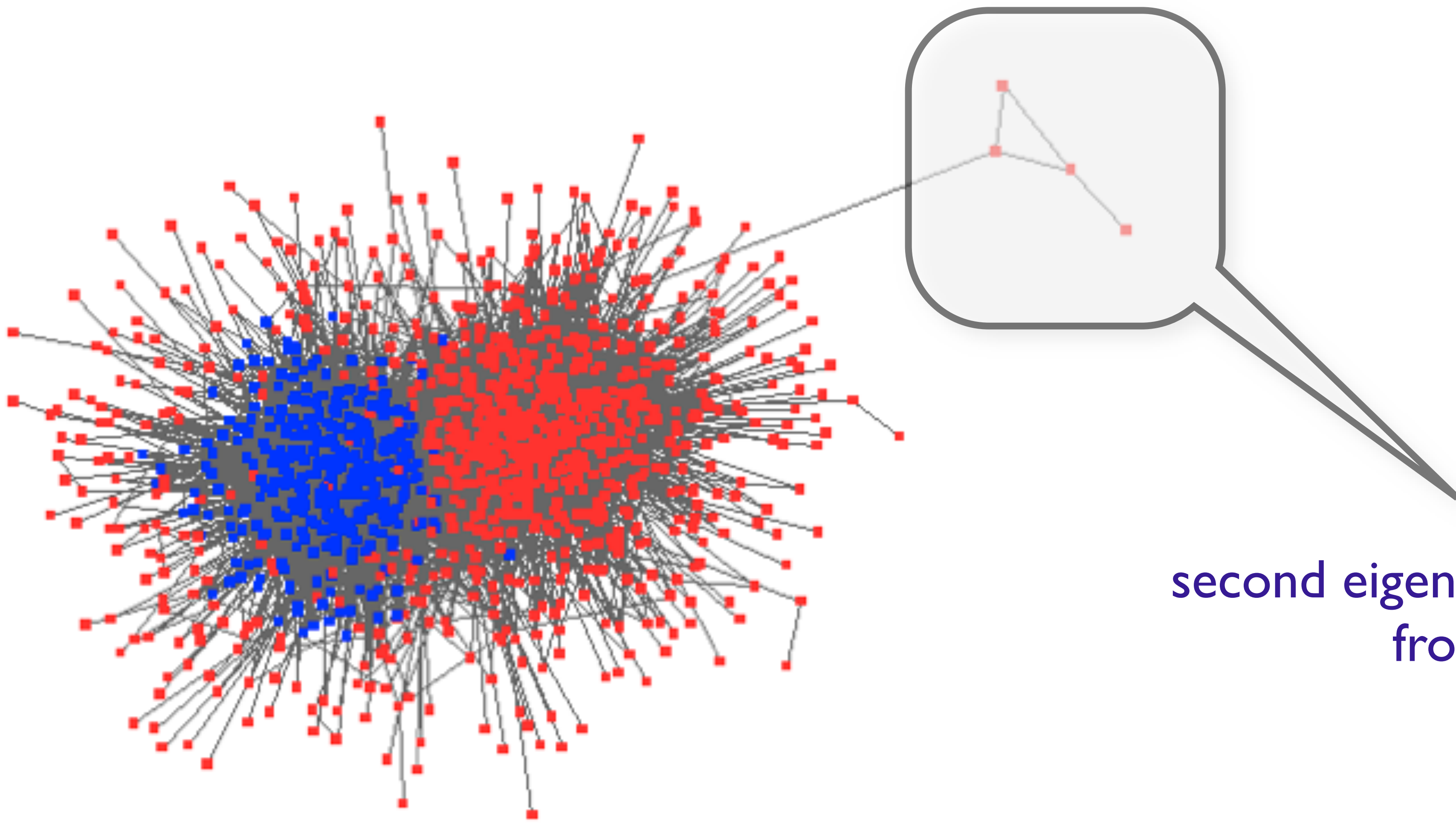



second eigenvector

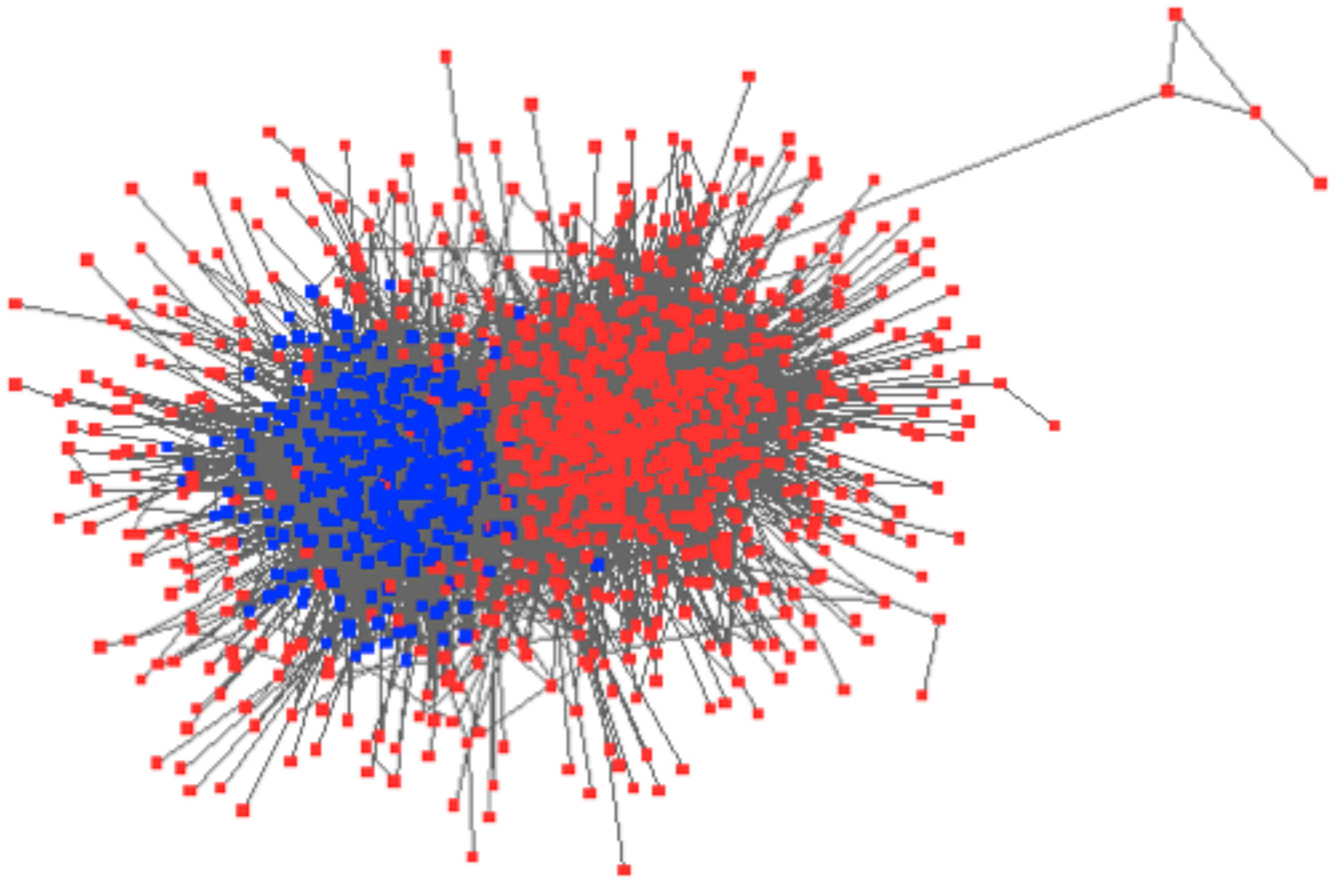
nodes

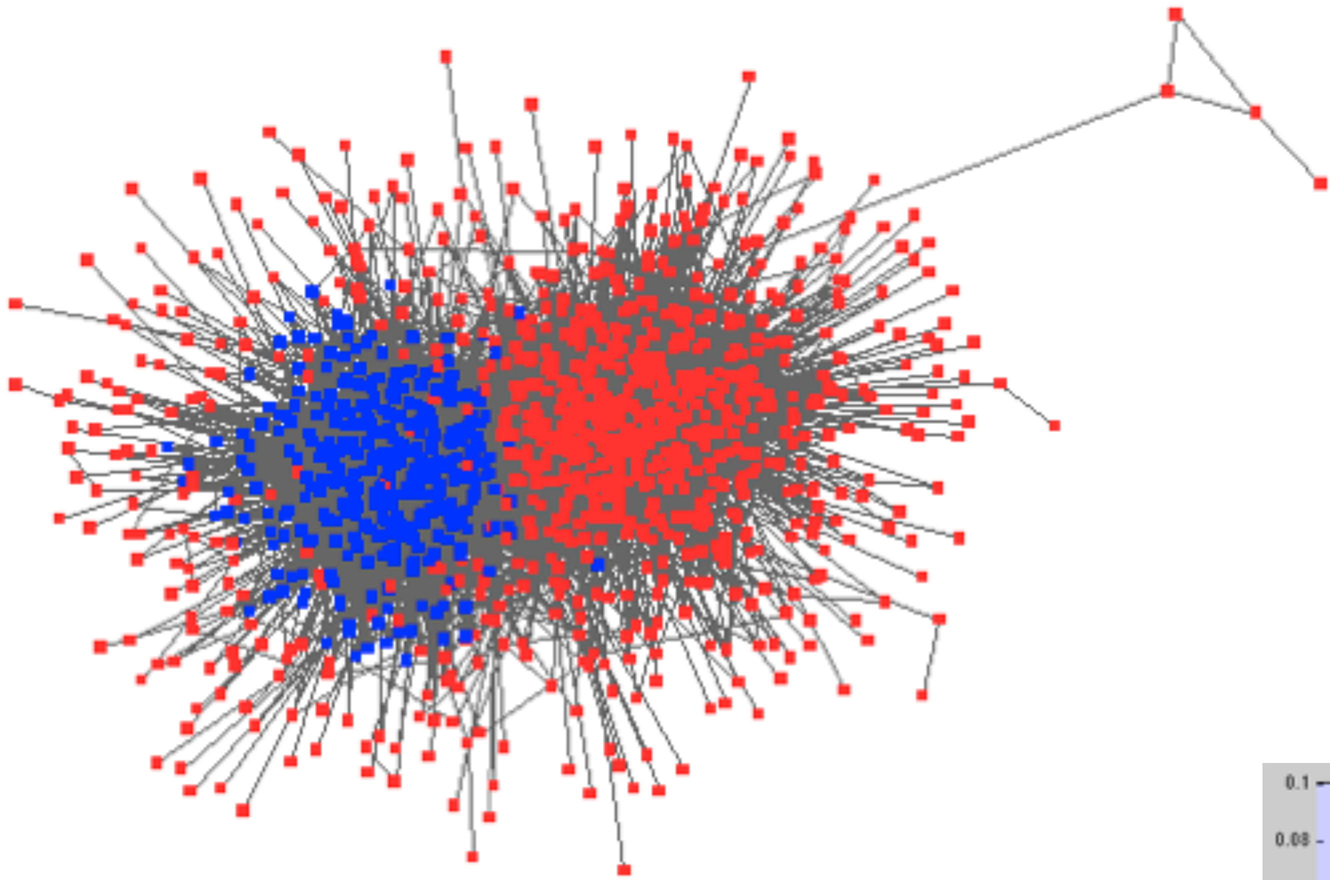
Unregularized
Spectral Clustering

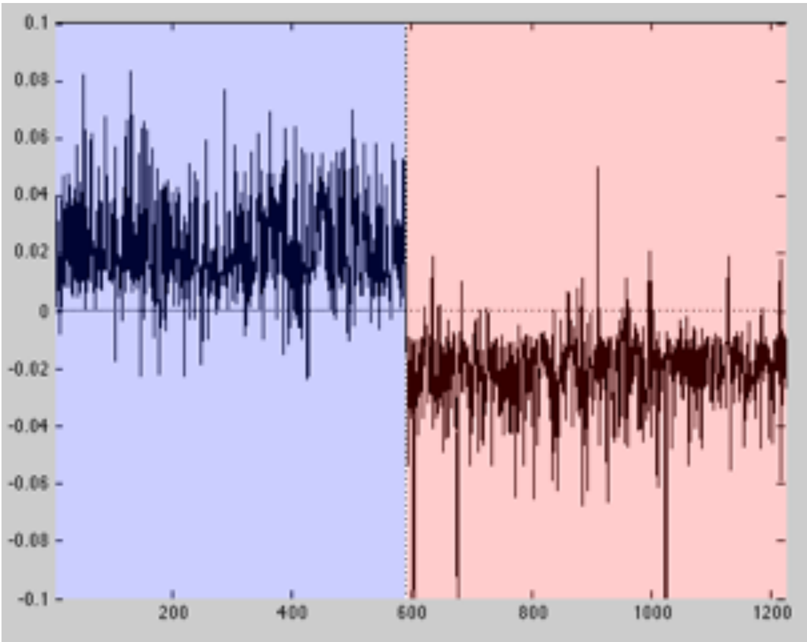




second eigenvector discriminates these
from the remaining



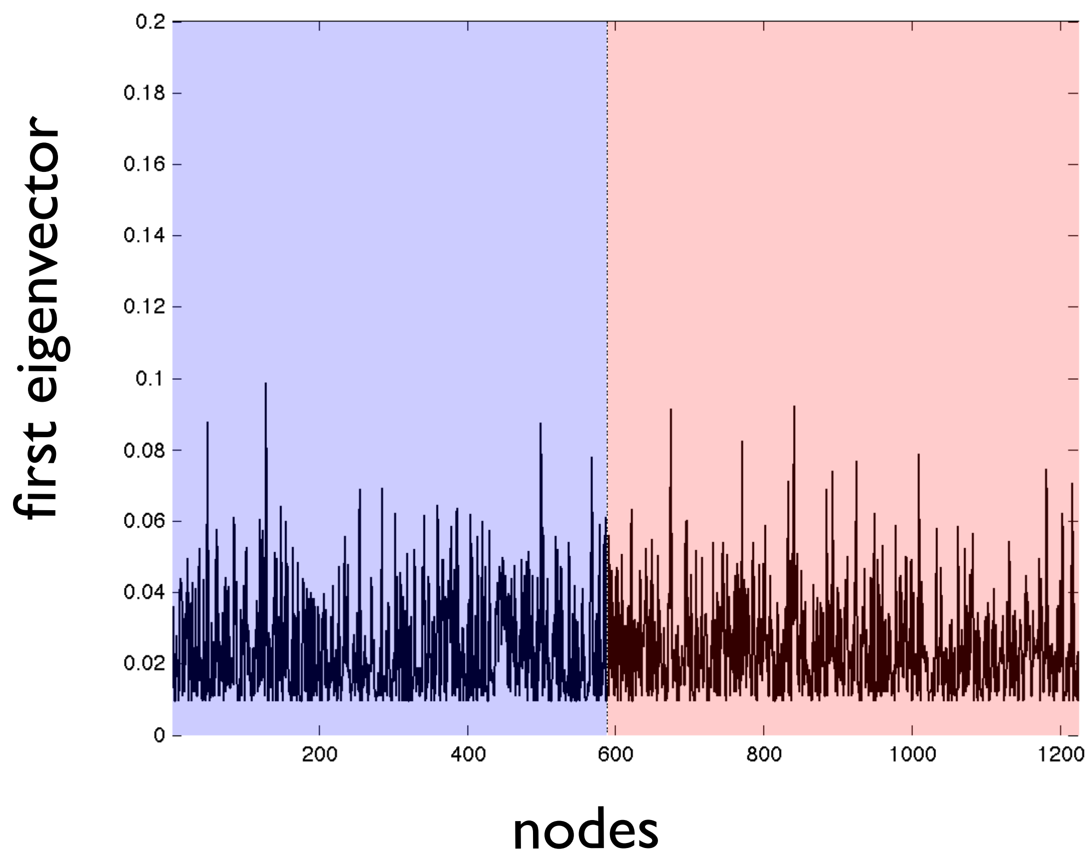


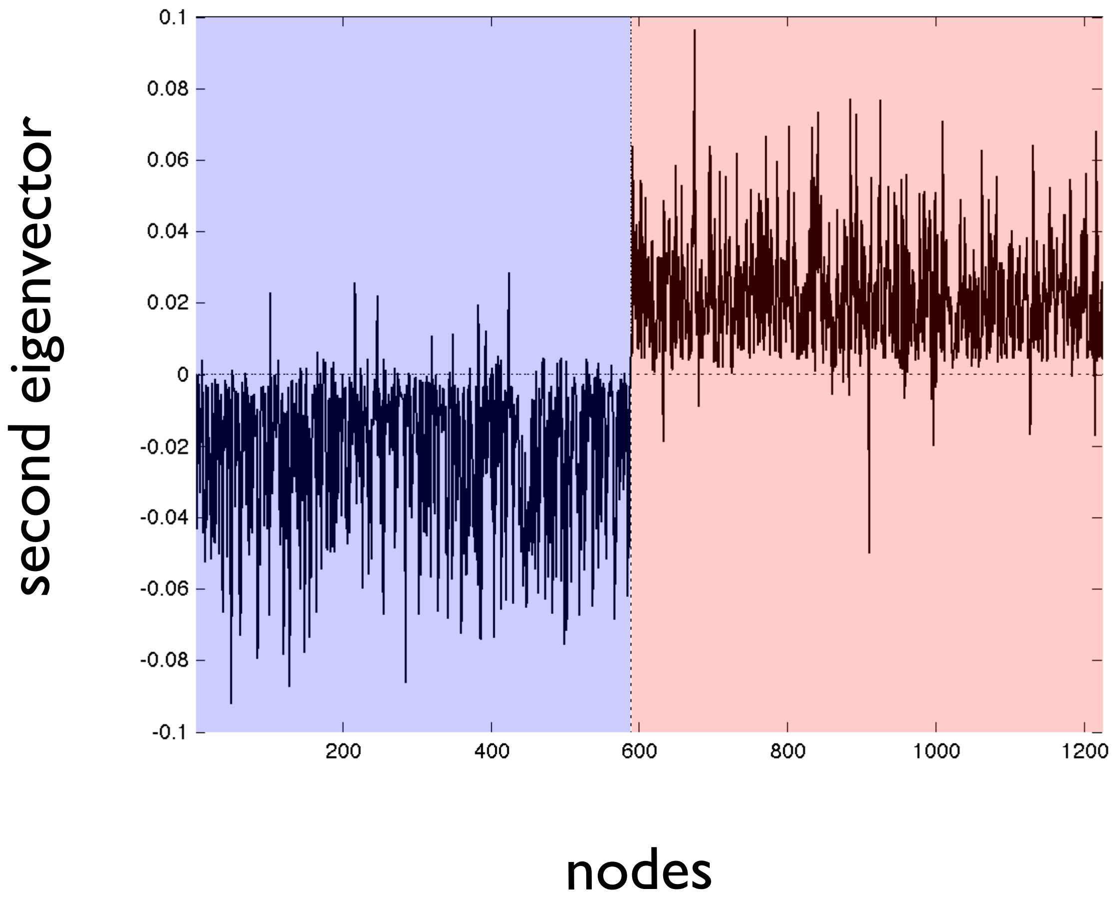



third eigenvector

# Regularized SC for political blogs dataset

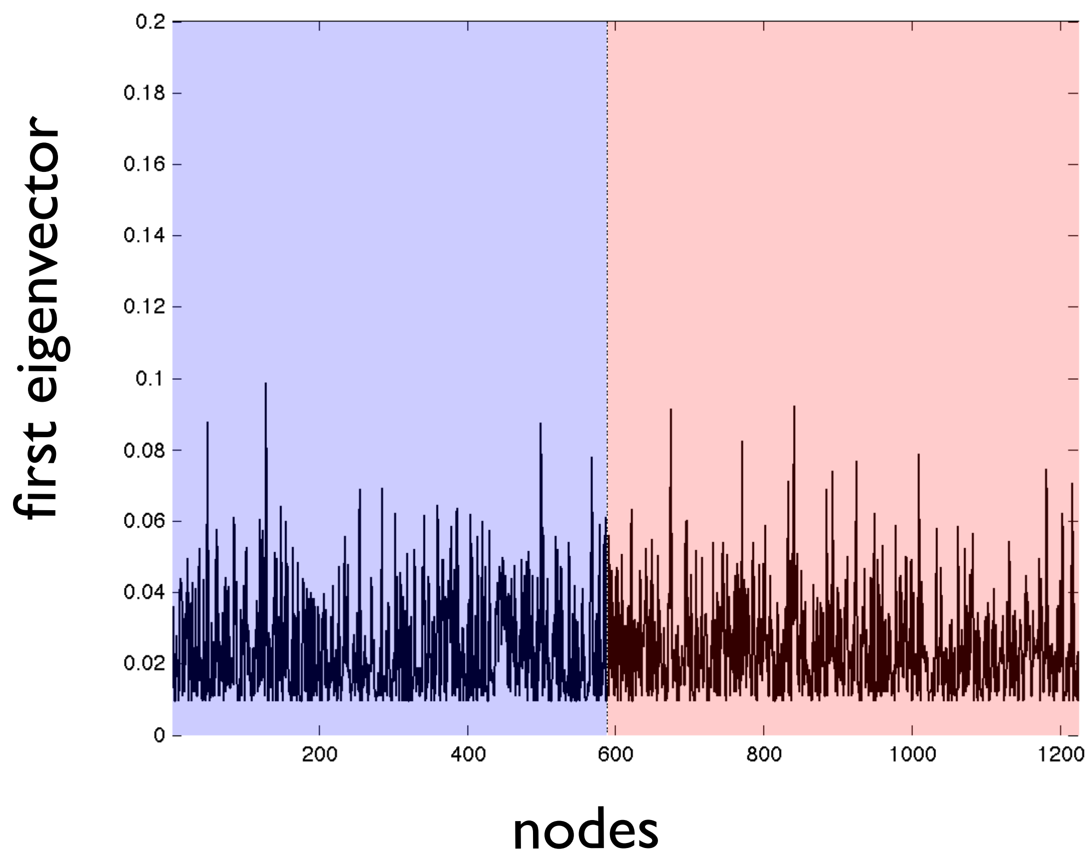
$\tau = 2.5$

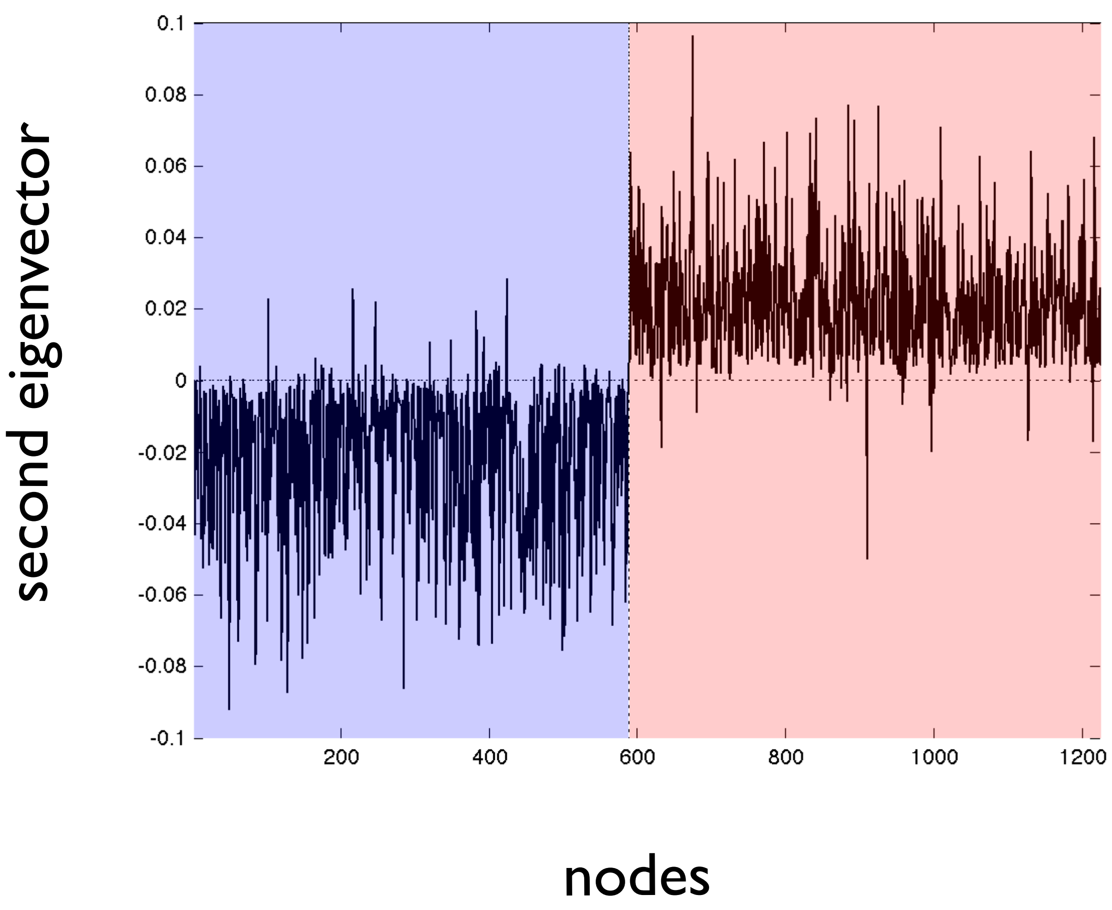


# Regularized SC for political blogs dataset

$\tau = 2.5$

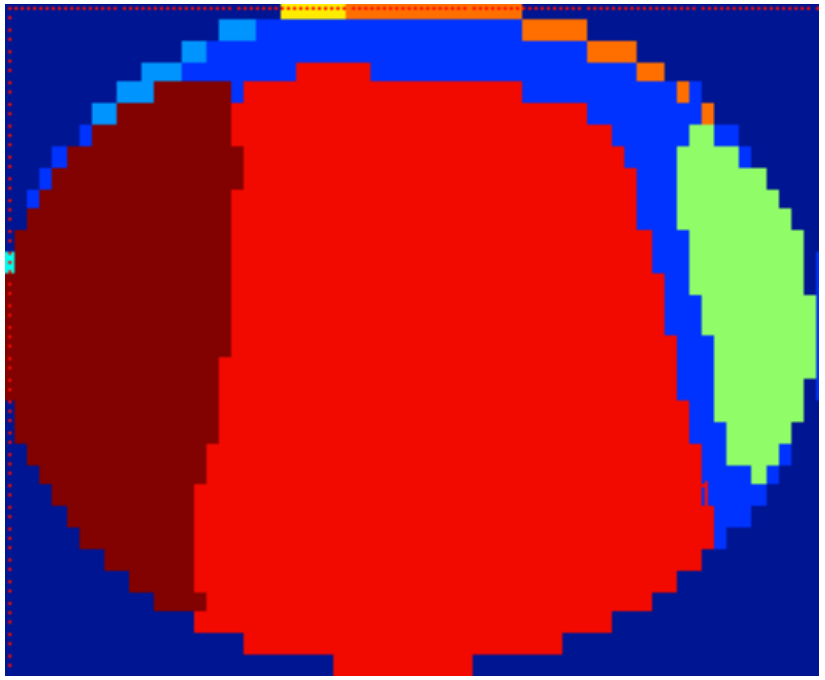


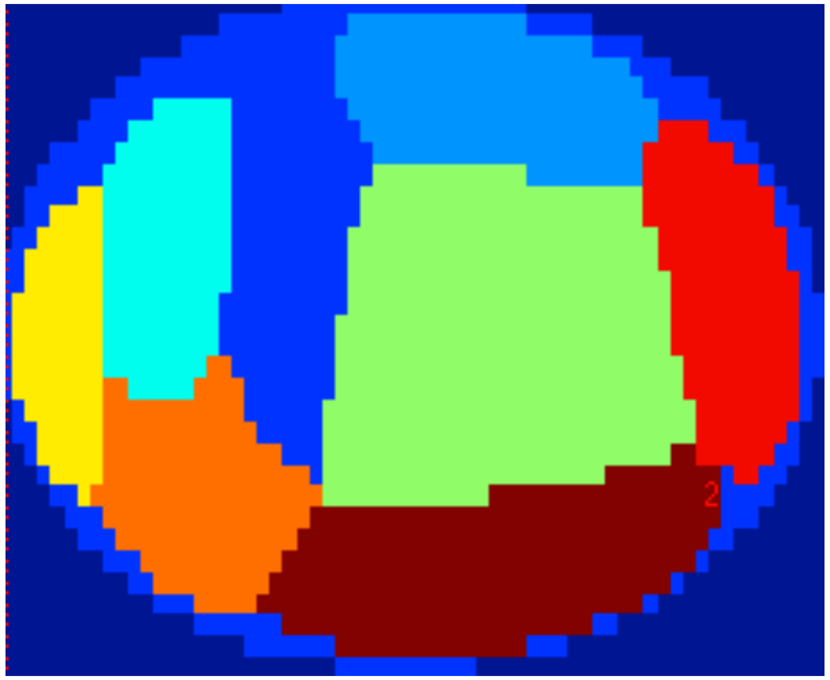

13% of misclassified nodes for regularized compared to 48% for unregularized

# Comparing unregularized vs. regularized Spectral Clustering (SC)

Take $K = 8$



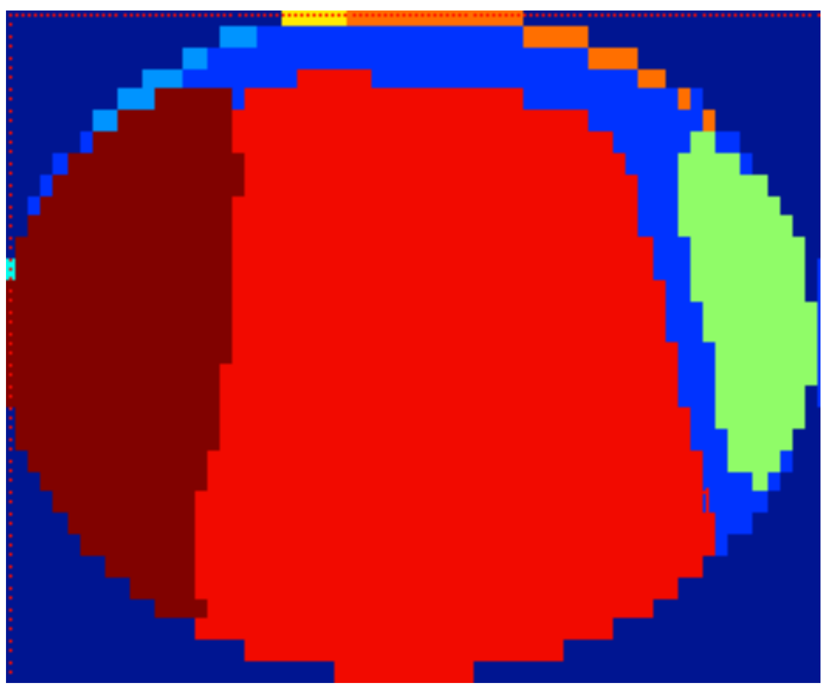
$\tau = 0$



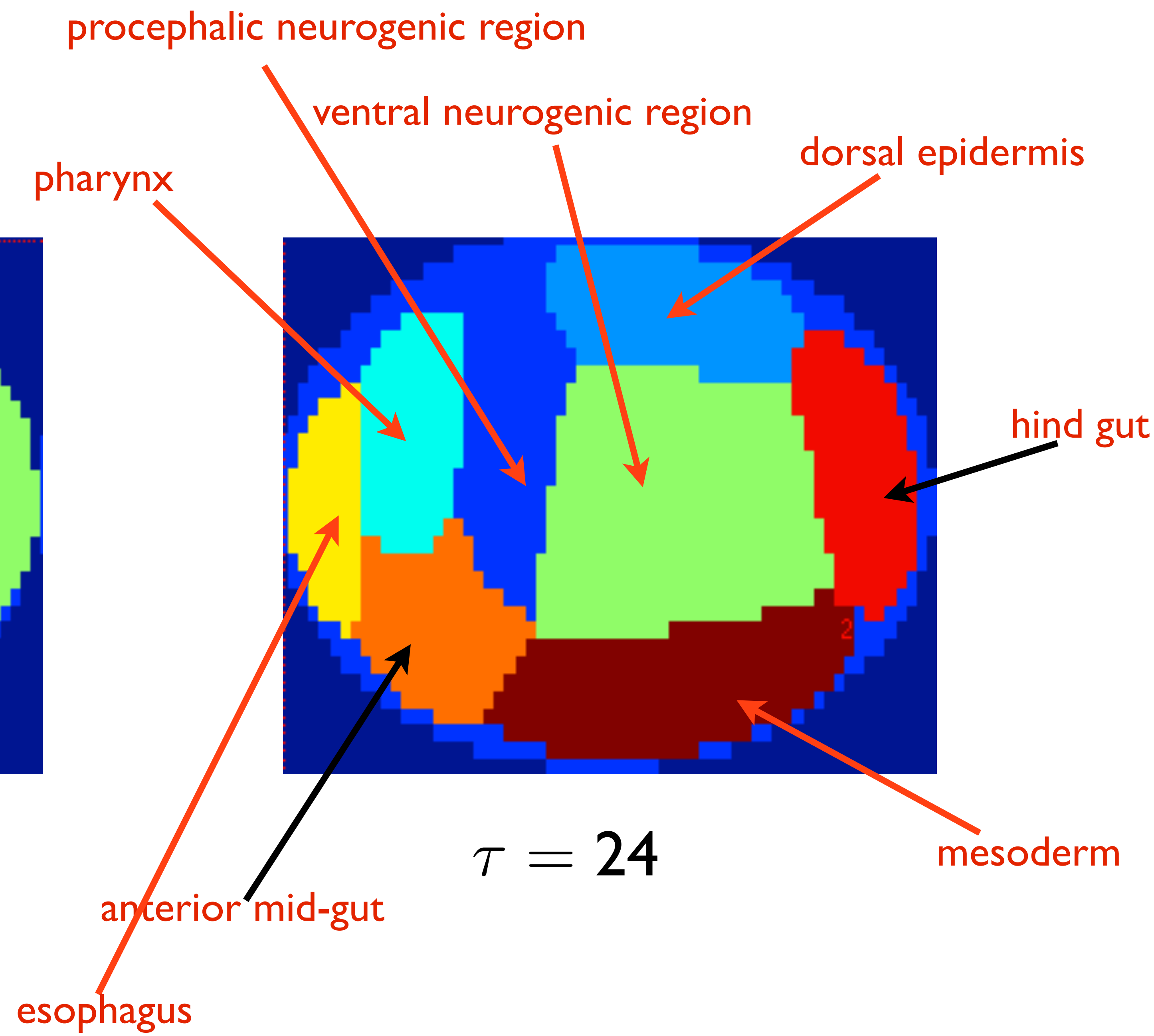
$\tau = 24$

# Comparing unregularized vs. regularized Spectral Clustering (SC)

Take $K = 8$



$\tau = 0$




$\tau = 24$

# Summary

- Theoretical upper bound under SBM shows "bias-variance"-like trade-off while the amount of regularization increases in SC

- Theoretical analysis motivates practically useful scheme (using SBM or degree-corrected SBM) to select regularization parameter in RSC.

Promising results in fruitfly image segmentation

Paper at (2014 rev):
http://arxiv.org/pdf/1312.1733.pdf

# Ongoing/future directions

## The BDGP project

(with Antony Joseph, Siqi Wu, Ann Hammonds, Sue Celniker, Erwin Frise)

- Analysis of gene interactions in different regions of early stage embryos
- Extension of analysis to later stage embryos

## Spectral Clustering (with *Antony Joseph*)

- Fast algorithm for computing the data-driven choice of regularization parameter
- Role of regularization in other scenarios, such as hierarchical clusters
- Regularization parameter choice for continuous data