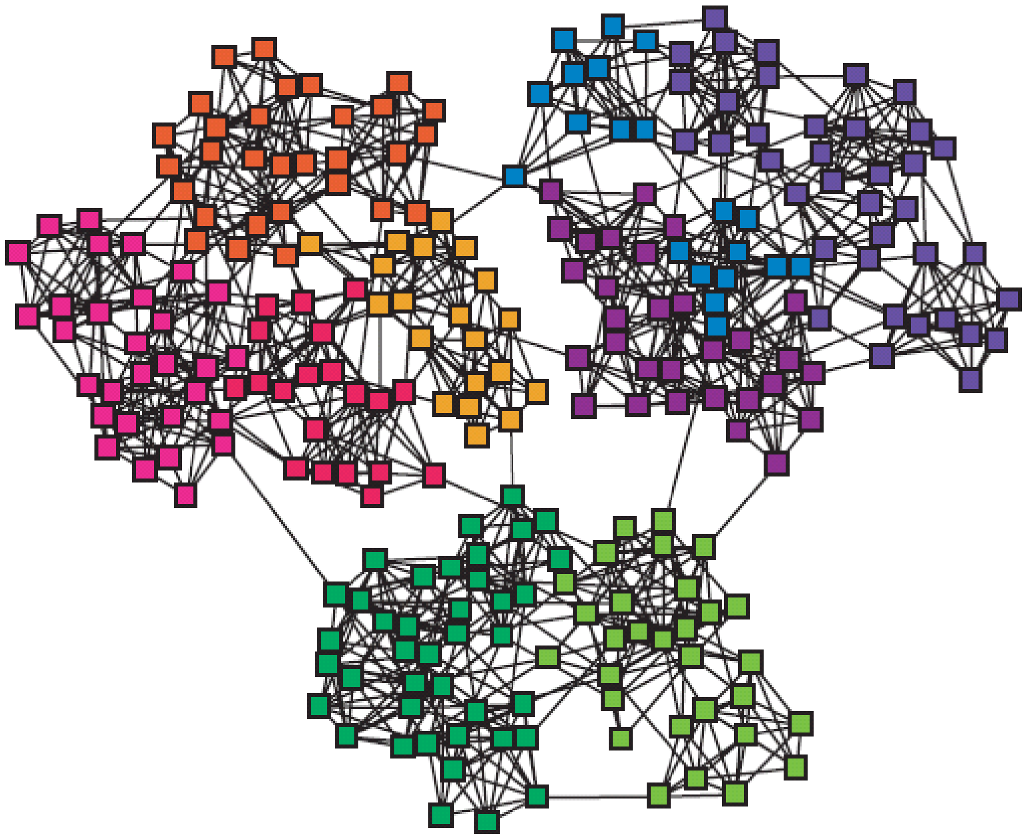
Analysis of Large Graphs: Community Detection

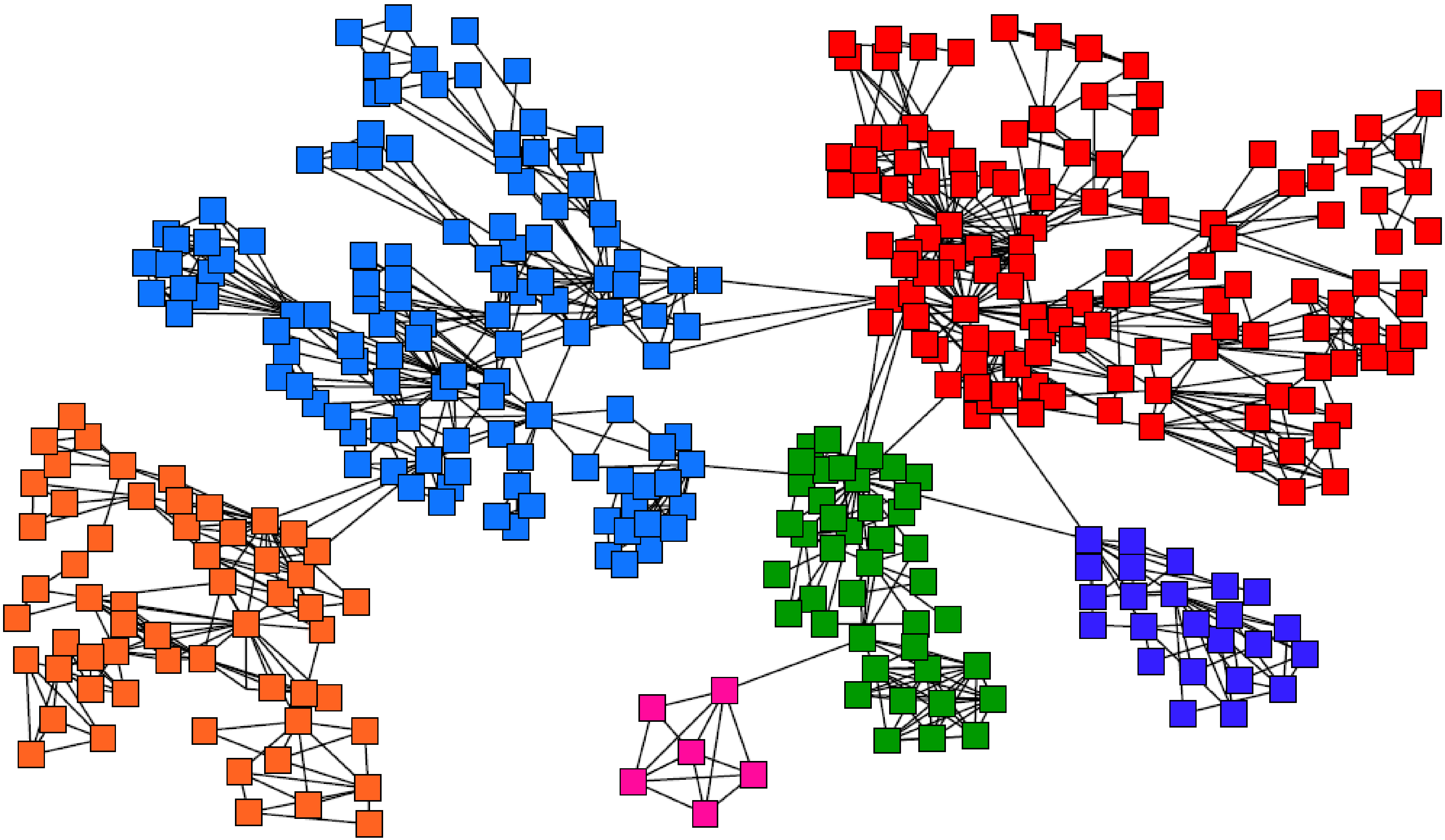
Mining of Massive Datasets
Jure Leskovec, and Rajaraman, Jeff Ullman Stanford University
http://www.mmds.org

Networks & Communities

- **We often think of a network being organized into modules, clusters, communities, groups:**

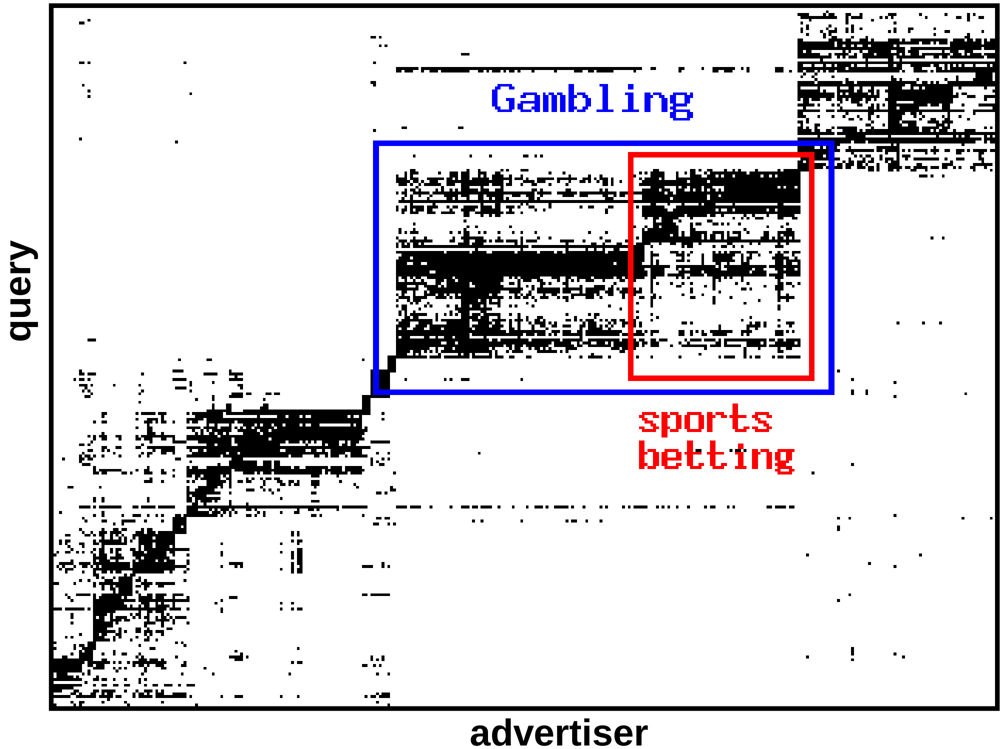


Goal: Find Densely Linked Clusters



Micro-Markets in Sponsored Search

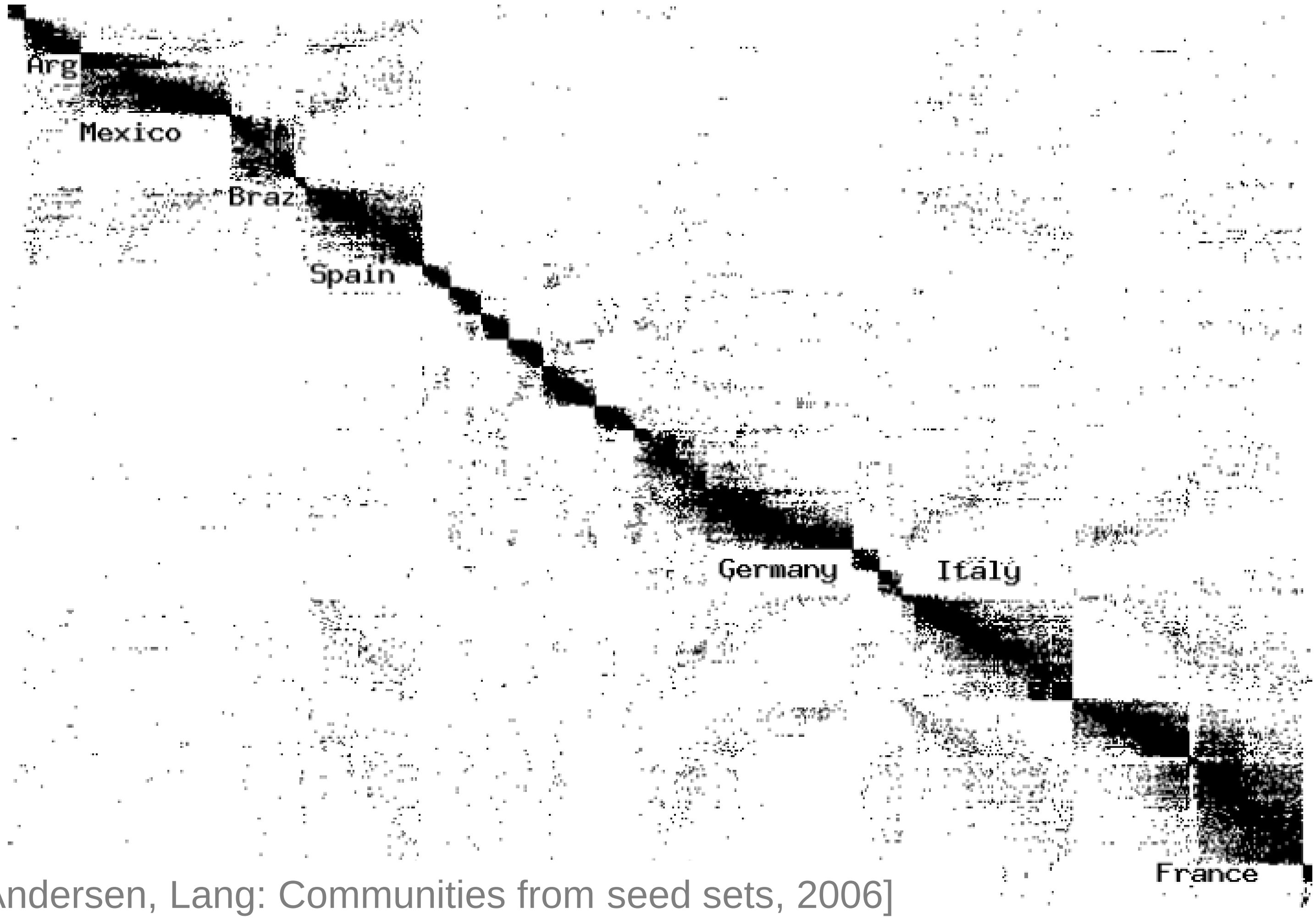
- **Find micro-markets by partitioning the query-to-advertiser graph:**



[Andersen, Lang: Communities from seed sets, 2006]

Movies and Actors

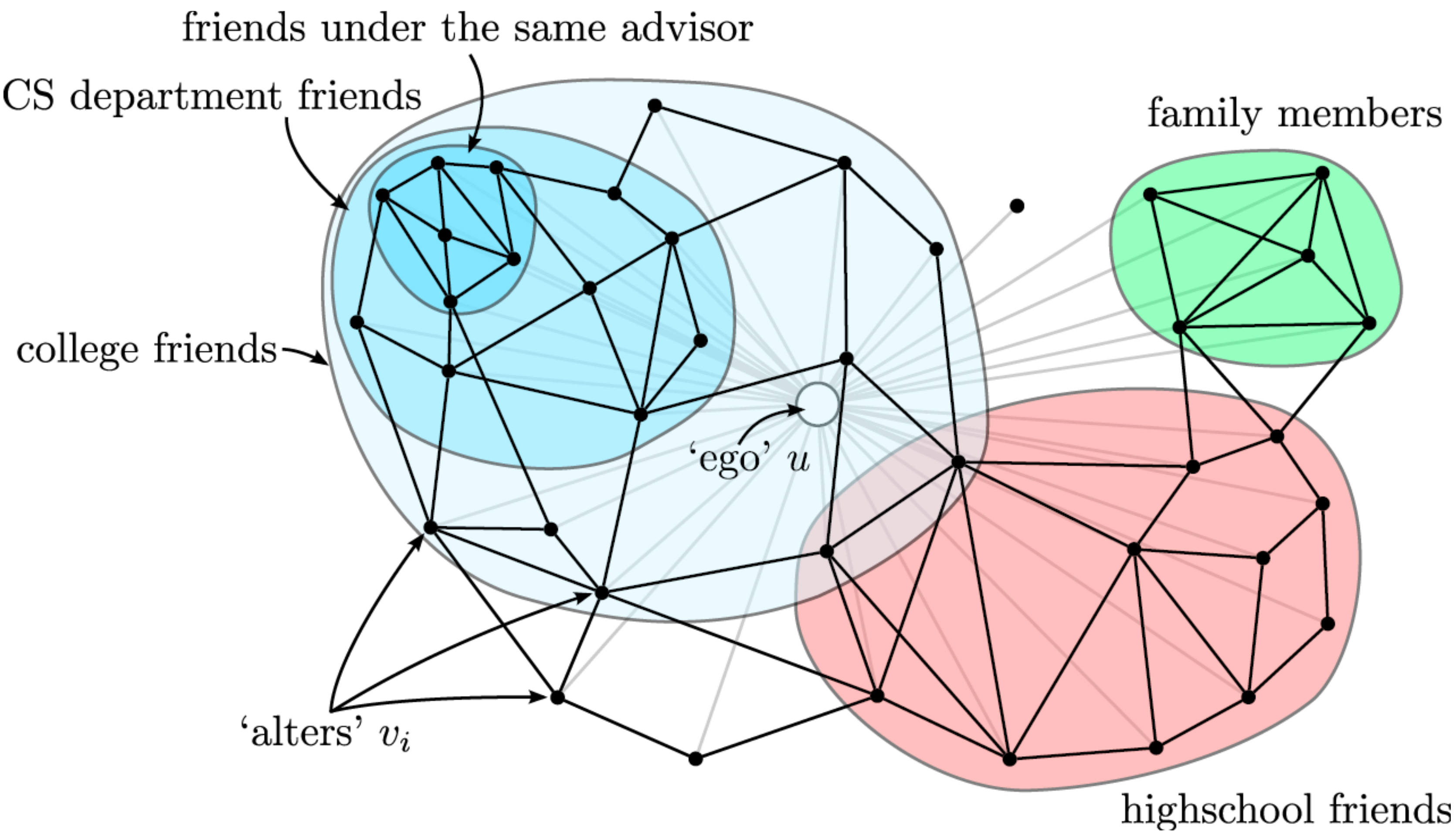
- **Clusters in Movies-to-Actors graph:**



[Andersen, Lang: Communities from seed sets, 2006]

Twitter & Facebook

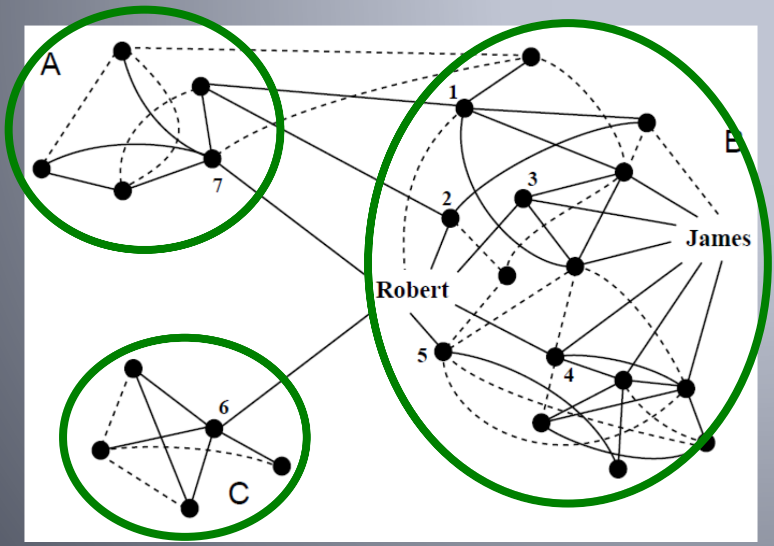
- **Discovering social circles, circles of trust:**

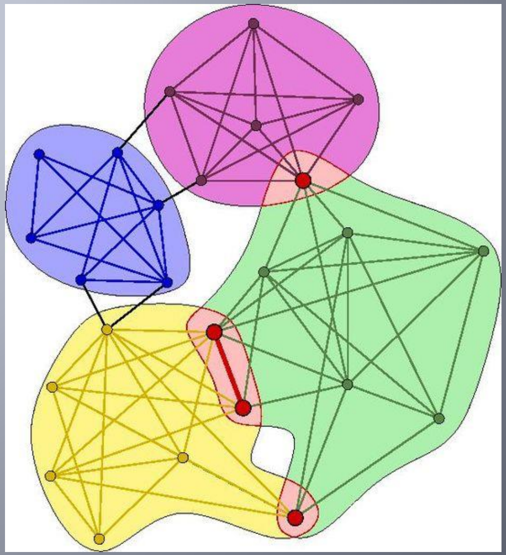


[McAuley, Leskovec: Discovering social circles in ego networks, 2012]

Community Detection

How to find communities?

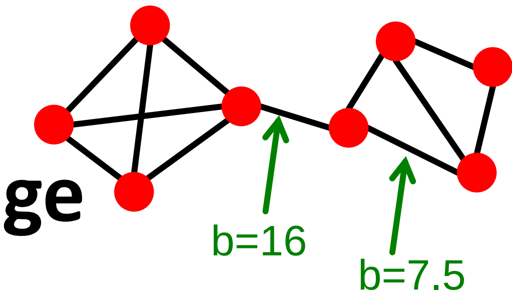


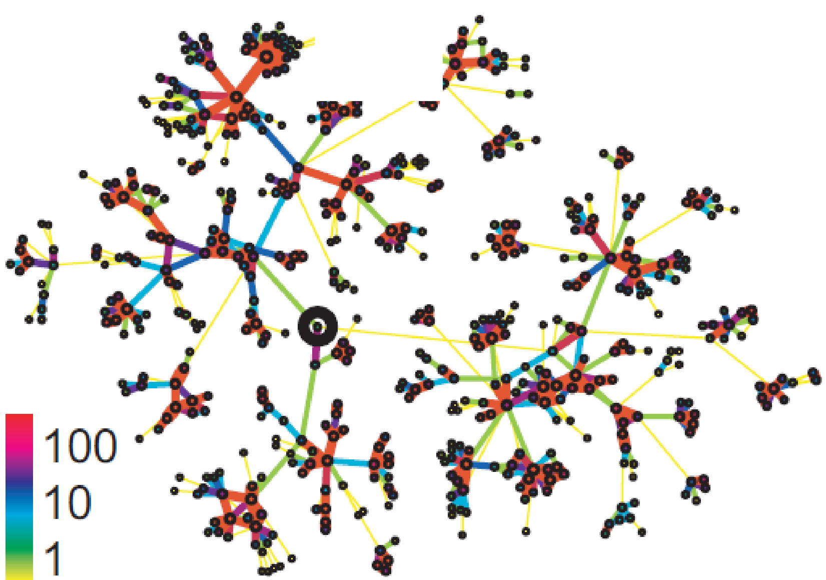


We will work with **undirected** (unweighted) networks

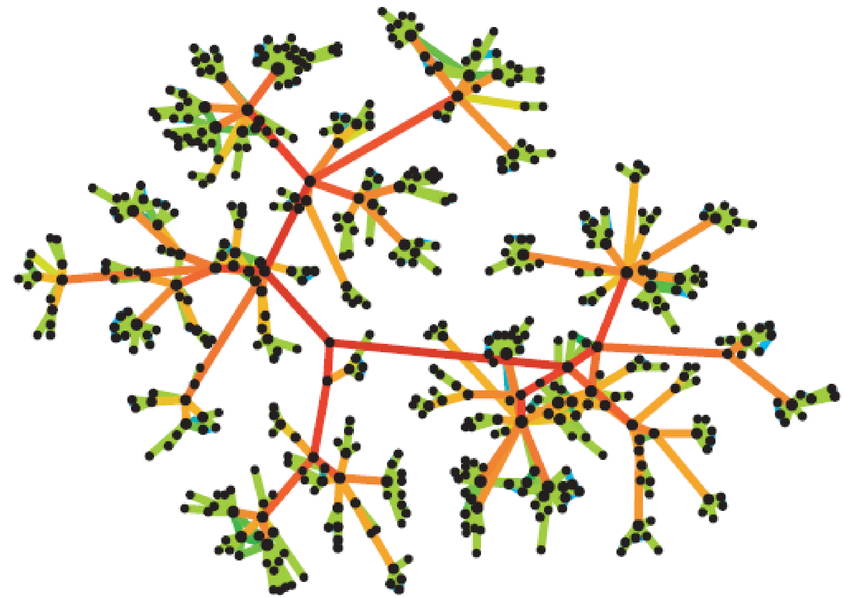
Method 1: Strength of Weak Ties

- **Edge betweenness:** Number of shortest paths passing over the edge
- **Intuition:**





Edge strengths (call volume) in a real network



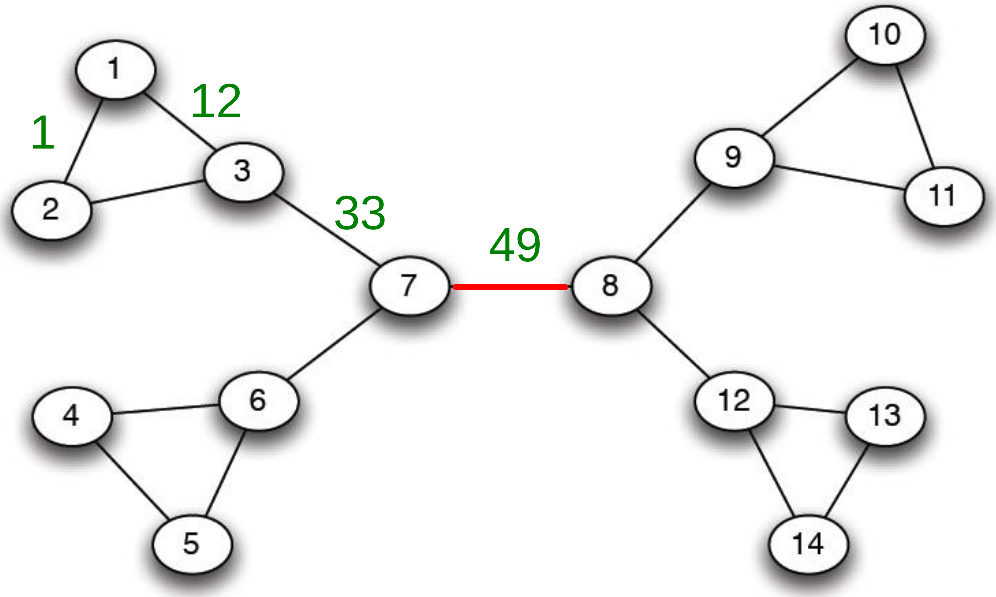
Edge betweenness in a real network

Method 1: Girvan-Newman

- Divisive hierarchical clustering based on the notion of edge **betweenness:**

  Number of shortest paths passing through the edge
- **Girvan-Newman Algorithm:**
  - **Undirected unweighted networks**
  - **Repeat until no edges are left:**
    - Calculate betweenness of edges
    - Remove edges with highest betweenness
  - Connected components are communities
  - Gives a hierarchical decomposition of the network

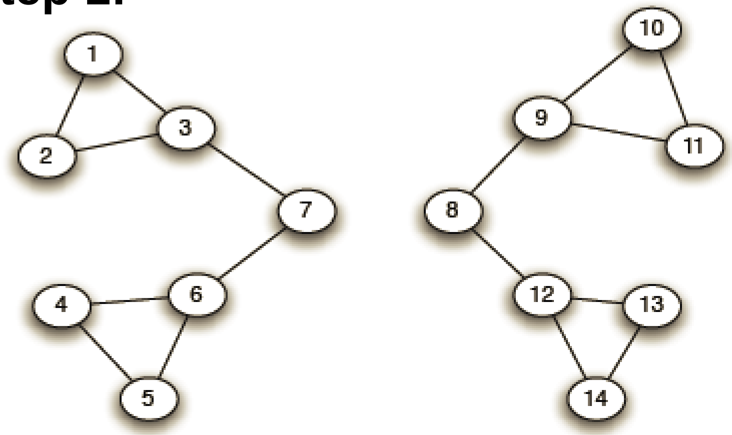
Girvan-Newman: Example



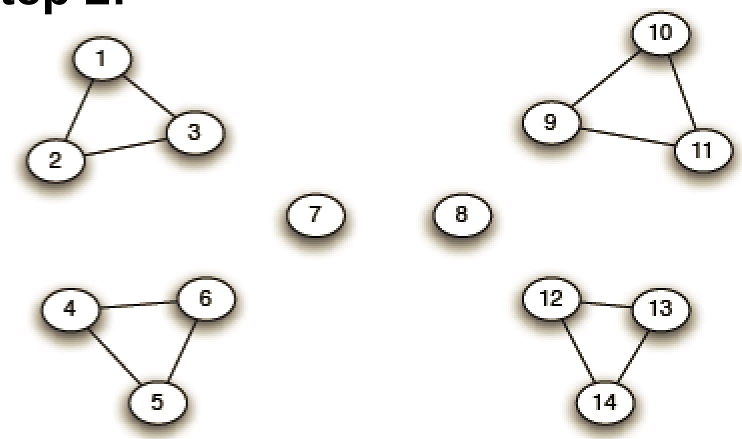
Need to re-compute betweenness at every step

Girvan-Newman: Example

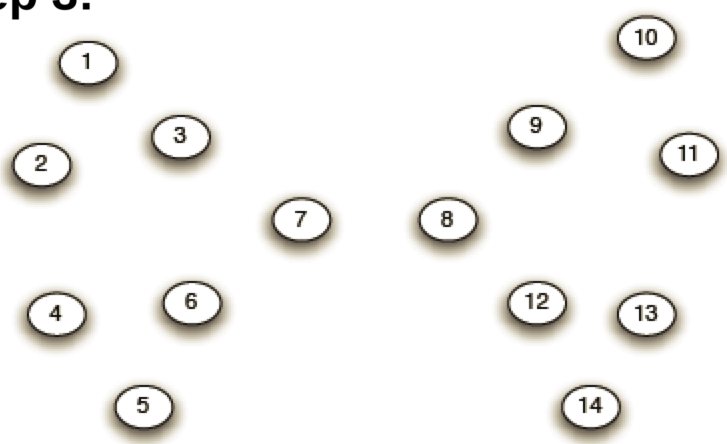
Step 1:



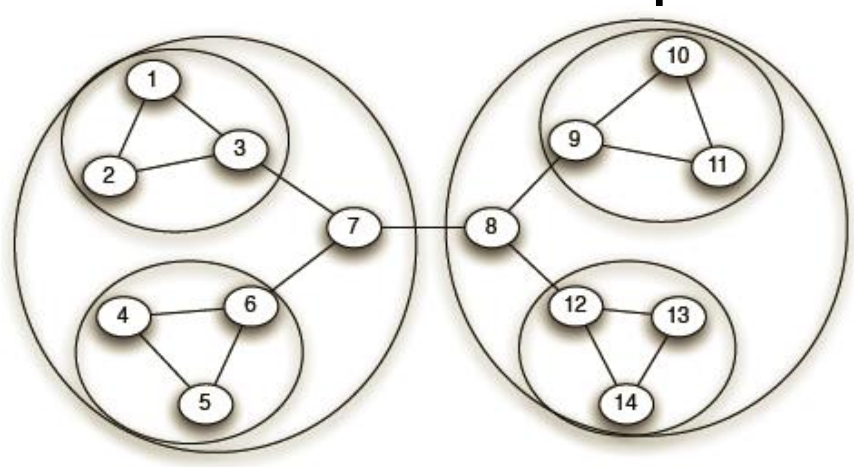
Step 2:



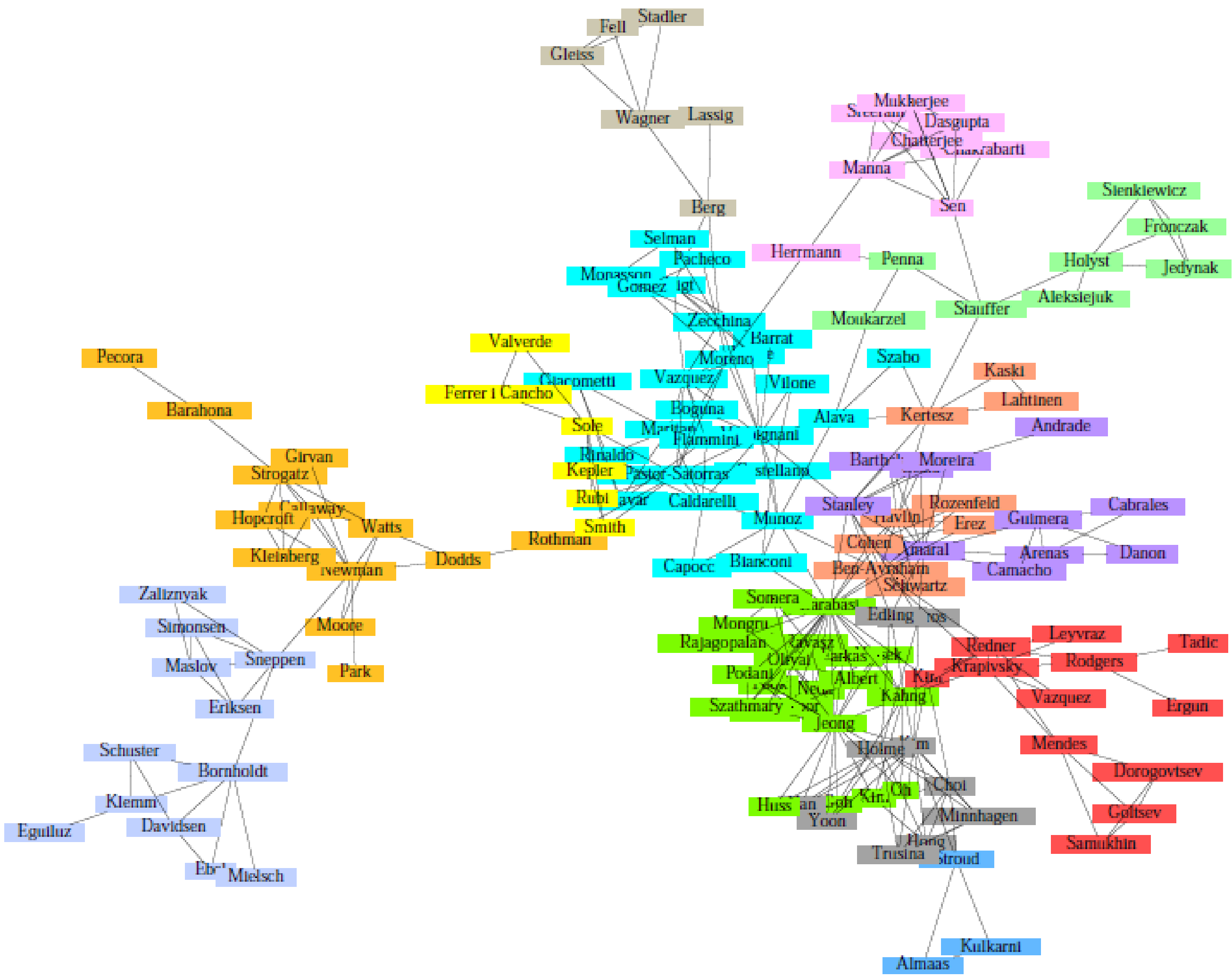
Step 3:



Hierarchical network decomposition:



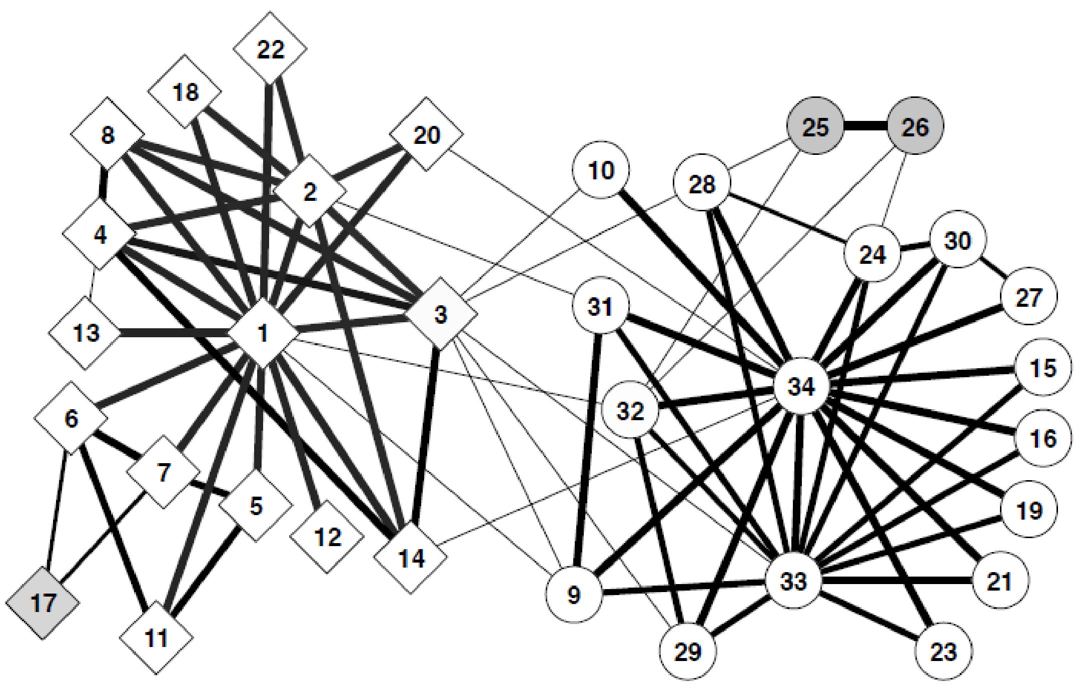
Girvan-Newman: Results

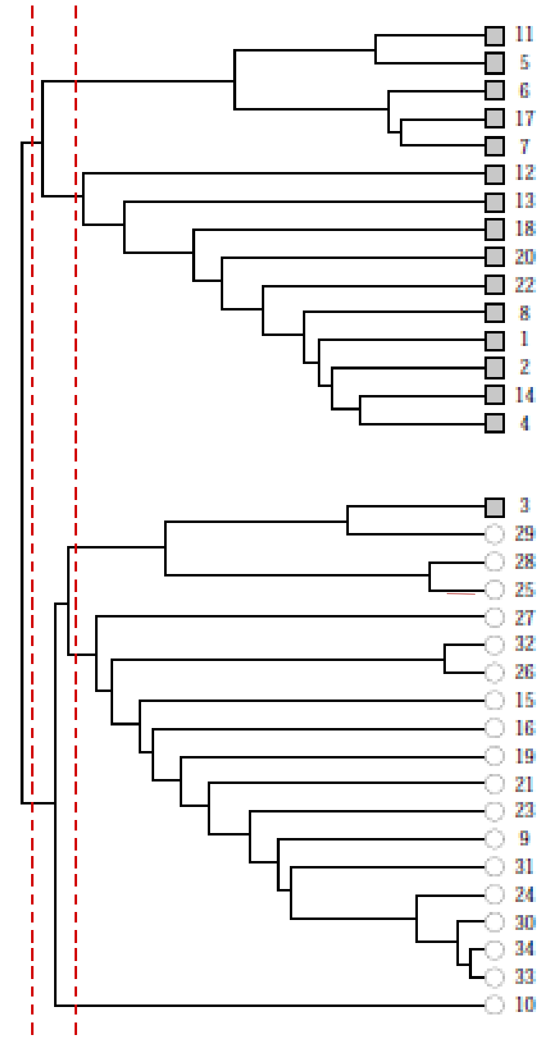


Communities in physics collaborations

Girvan-Newman: Results

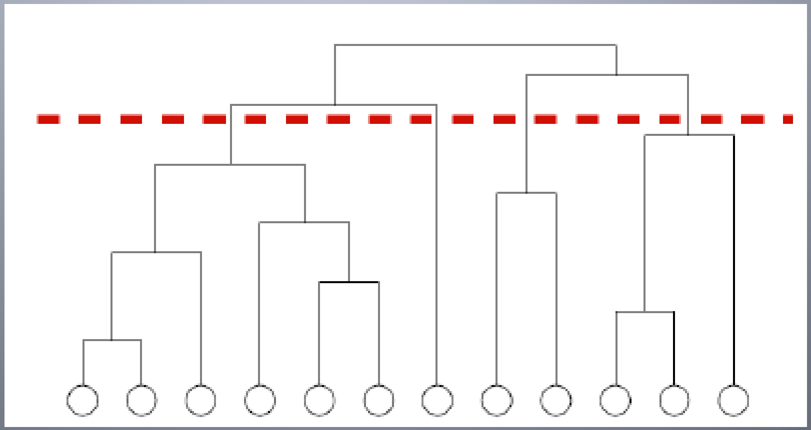
- **Zachary's Karate club:** Hierarchical decomposition





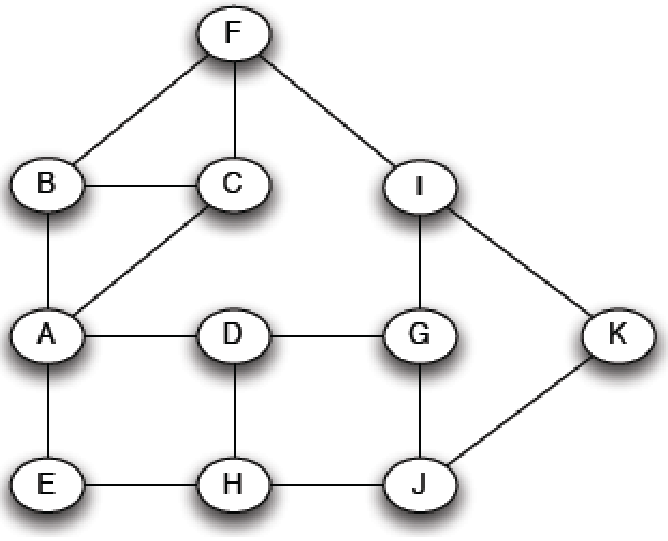
We need to resolve 2 questions

1. **How to compute betweenness?**
2. **How to select the number of clusters?**

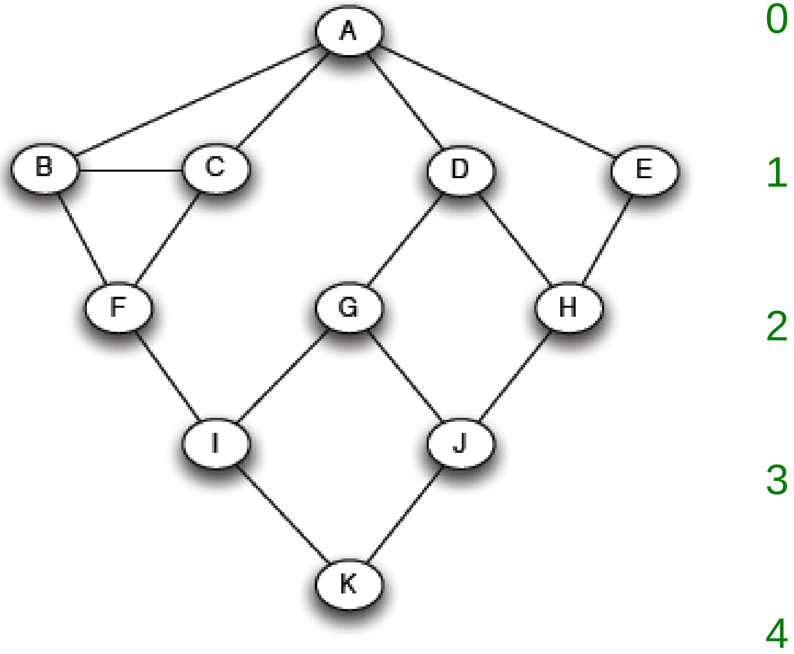


How to Compute Betweenness?

- **Want to compute betweenness of paths starting at node A**

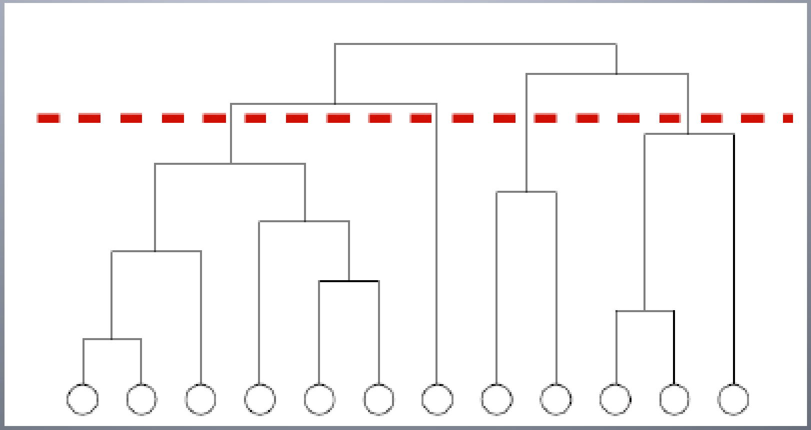


- **Breath first search starting from A:**



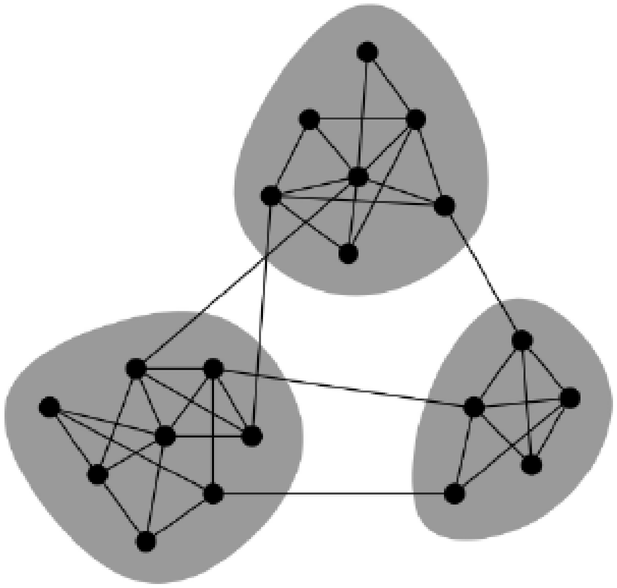
We need to resolve 2 questions

1. How to compute betweenness?
2. How to select the number of clusters?



Network Communities

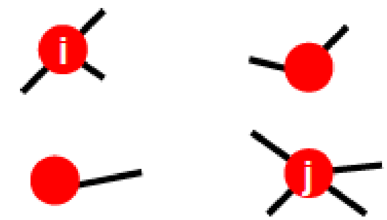
- **Communities:** sets of **tightly connected nodes**
- Define: **Modularity Q**
  - A measure of how well a network is partitioned into communities
  - Given a partitioning of the network into groups $s \in S$:



$$Q \propto \sum_{s \in S} [\ (\#\text{ edges within group } s) - \underbrace{(\text{expected }\#\text{ edges within group } s)\ ]}_{\text{Need a null model!}}$$

Null Model: Configuration Model

- **Given real G on n nodes and m edges, construct rewired network G'**
  - Same degree distribution but random connections
  - Consider G' as a multigraph
  - **The expected number of edges between nodes i and j of degrees k_i and k_j equals to:** $k_i \cdot \frac{k_j}{2m} = \frac{k_i k_j}{2m}$
    - The expected number of edges in (multigraph) **G'**:
      - $= \frac{1}{2}\sum_{i \in N}\sum_{j \in N}\frac{k_i k_j}{2m} = \frac{1}{2} \cdot \frac{1}{2m}\sum_{i \in N} k_i \left(\sum_{j \in N} k_j\right) =$
      - $= \frac{1}{4m} 2m \cdot 2m = m$



Note:
$$\sum_{u \in N} k_u = 2m$$

Modularity

- Modularity of partitioning S of graph G:
  - Q ∝ $\sum_{s \in S}$ [(# edges within group *s*) – (expected # edges within group *s*)]
  - $Q(G,S) = \underbrace{\frac{1}{2m}}_{\text{Normalizing cost.: } -1<Q<1} \sum_{s \in S} \sum_{i \in s} \sum_{j \in s} \left(A_{ij} - \frac{k_i k_j}{2m} \right)$

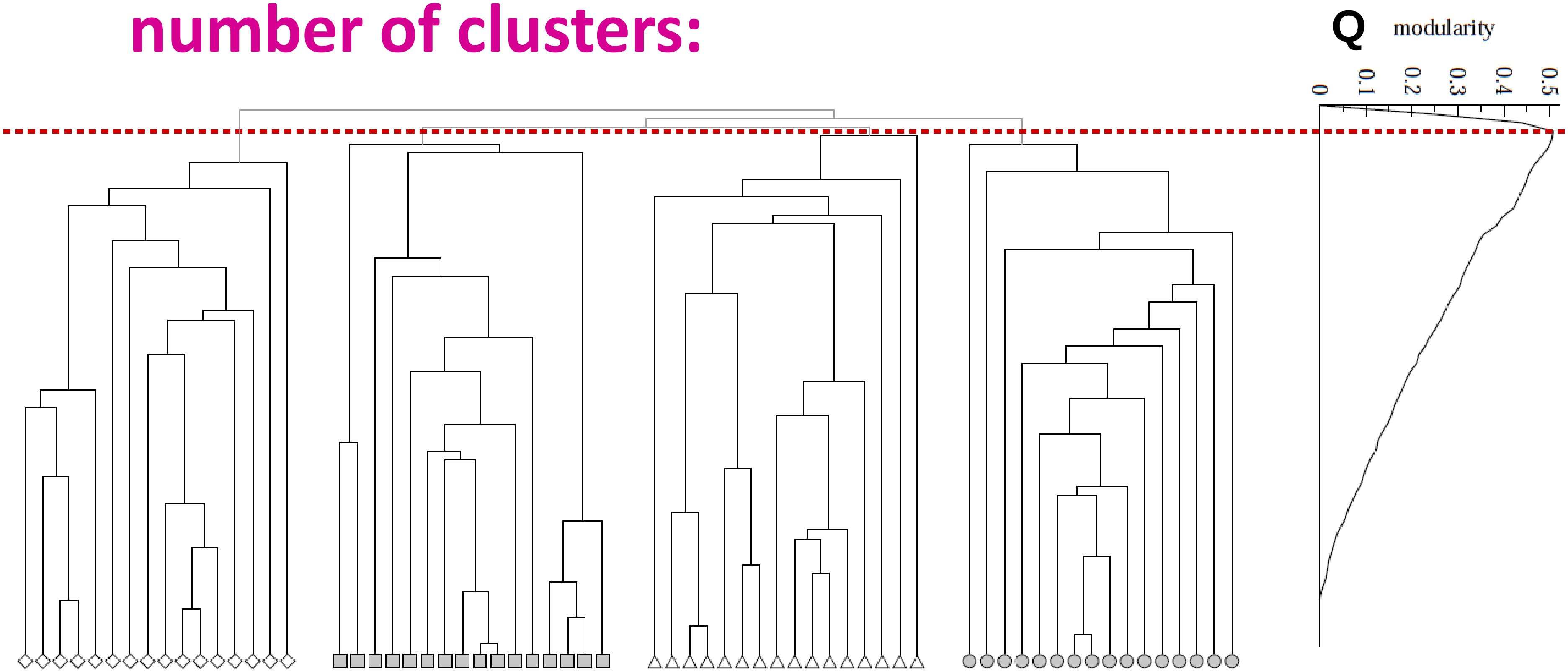
A_{ij} = 1 if i→j,
0 else

$$Q(G,\{g_1,\ldots,g_n\}) = \frac{1}{2m} \sum_{i,j=1}^{n} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(g_i, g_j)$$

- Modularity values take range [−1,1]
  - It is positive if the number of edges within groups exceeds the expected number
  - **0.3~0.7<Q** means significant community structure

Modularity: Number of clusters

- **Modularity is useful for selecting the number of clusters:**



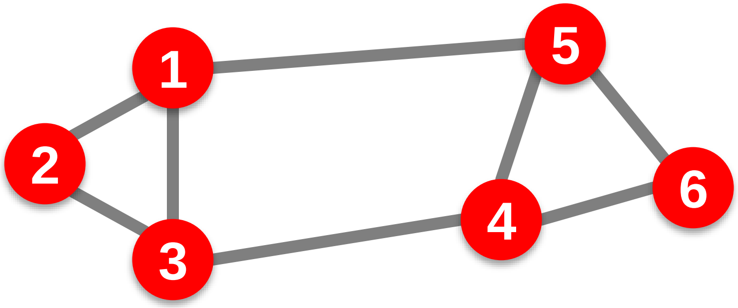
Next time: Why not optimize Modularity directly?

http://netwiki.amath.unc.edu/GenLouvain/GenLouvain

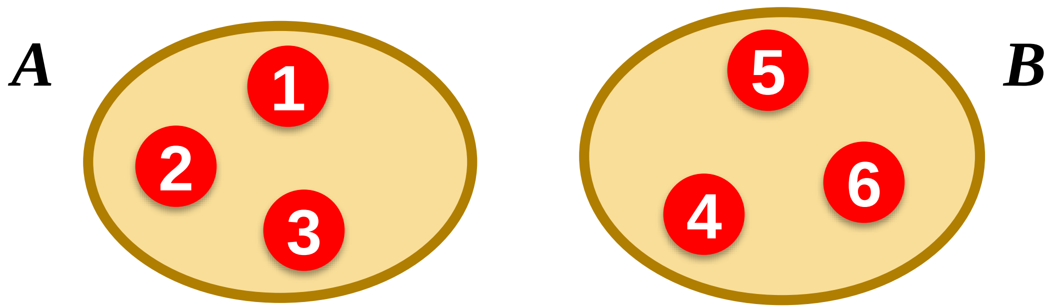
Spectral Clustering

Graph Partitioning

- **Undirected graph $G(V, E)$:**



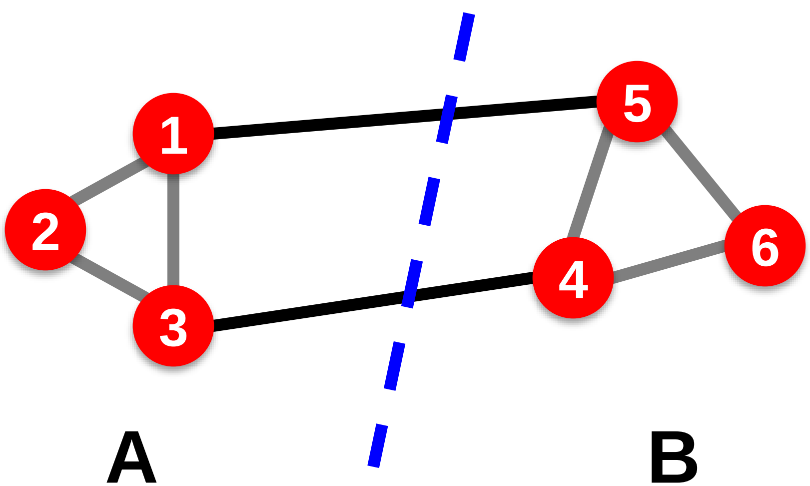
- **Bi-partitioning task:**
  - Divide vertices into two disjoint groups A, B



- **Questions:**
  - How can we define a "good" partition of G?
  - How can we efficiently identify such a partition?

Graph Partitioning

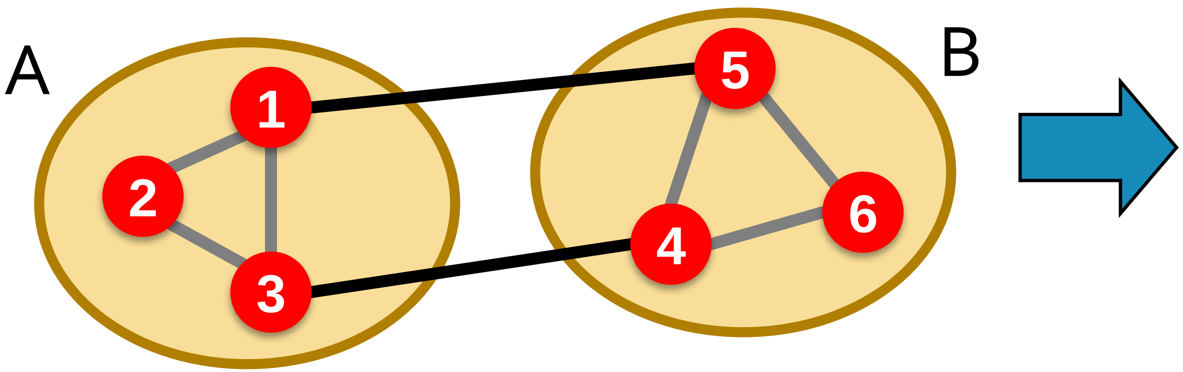
- **What makes a good partition?**
  - Maximize the number of within-group connections
  - Minimize the number of between-group connections



Graph Cuts

- **Express partitioning objectives as a function of the "edge cut" of the partition**
- **Cut:** Set of edges with only one vertex (node) in a group:

$$cut(A,B) = \sum_{i \in A, j \in B} w_{ij}$$



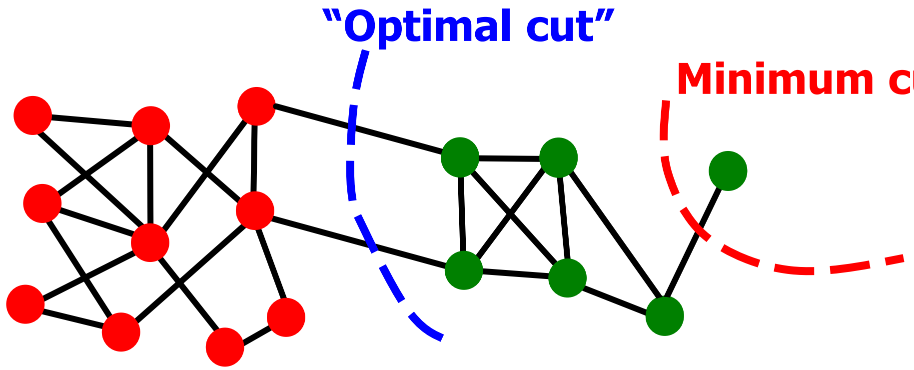
$cut(A,B) = 2$

Graph Cut Criterion

- **Criterion: Minimum-cut**
  - Minimize weight of connections between groups

$$\arg\min_{A,B} cut(A,B)$$

- **Degenerate case:**



- **Problem:**
  - Only considers external cluster connections
  - Does not consider internal cluster connectivity

Graph Cut Criteria

- **Criterion: Normalized-cut** [Shi-Malik, '97]
  - Connectivity between groups relative to the density of each group

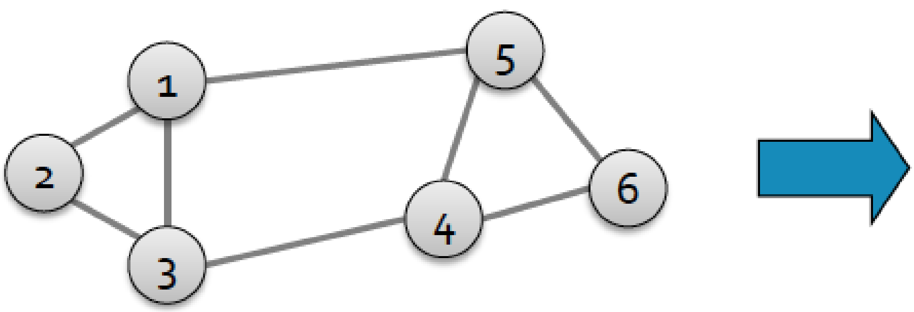
$$ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

$\boldsymbol{vol(A)}$: total weight of the edges with at least one endpoint in $\boldsymbol{A}$: $\boldsymbol{vol(A) = \sum_{i \in A} k_i}$

- **Why use this criterion?**
  - Produces more balanced partitions
- **How do we efficiently find a good partition?**
  - **Problem:** Computing optimal cut is NP-hard

Matrix Representations

- **Adjacency matrix (A):**
  - $n \times n$ matrix
  - $A=[a_{ij}]$, $a_{ij}=1$ if edge between node i and j

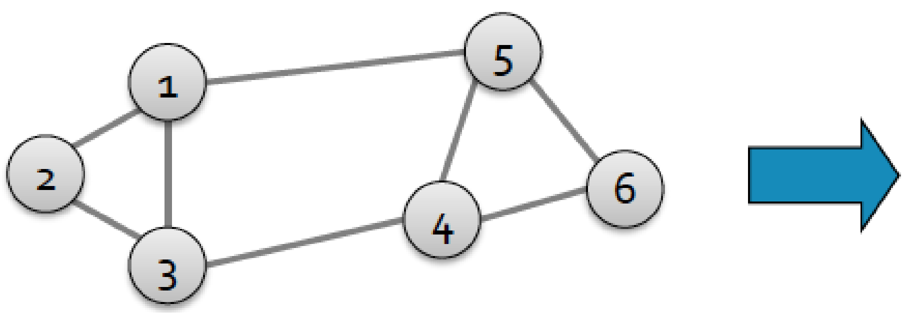


	1	2	3	4	5	6
1	0	1	1	0	1	0
2	1	0	1	0	0	0
3	1	1	0	1	0	0
4	0	0	1	0	1	1
5	1	0	0	1	0	1
6	0	0	0	1	1	0

- **Important properties:**
  - Symmetric matrix
  - Eigenvectors are real and orthogonal

Matrix Representations

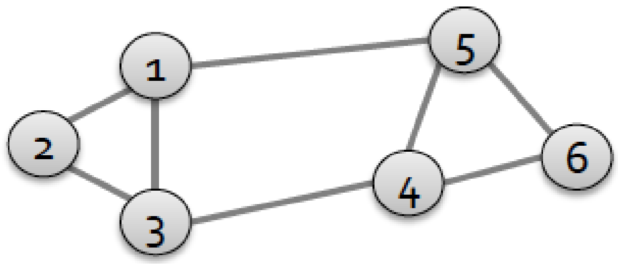
- **Degree matrix (D):**
  - $n \times n$ diagonal matrix
  - $D=[d_{ii}]$, $d_{ii}=$ degree of node i



	1	2	3	4	5	6
1	3	0	0	0	0	0
2	0	2	0	0	0	0
3	0	0	3	0	0	0
4	0	0	0	3	0	0
5	0	0	0	0	3	0
6	0	0	0	0	0	2

Matrix Representations

- Laplacian matrix (L):
  - $n \times n$ symmetric matrix



	1	2	3	4	5	6
1	3	-1	-1	0	-1	0
2	-1	2	-1	0	0	0
3	-1	-1	3	-1	0	0
4	0	0	-1	3	-1	-1
5	-1	0	0	-1	3	-1
6	0	0	0	-1	-1	2

$$L = D - A$$

- What is trivial eigenpair?
  - $x = (1, \dots, 1)$ then $L \cdot x = 0$ and so $\lambda = \lambda_1 = 0$
- Important properties:
  - **Eigenvalues** are non-negative real numbers
  - **Eigenvectors** are real and orthogonal

Facts about the Laplacian L

Details!

(a) All eigenvalues are ≥ 0

(b) $x^T L x = \sum_{ij} L_{ij} x_i x_j \geq 0$ for every x

(c) $L = N^T \cdot N$

- That is, L is positive semi-definite

- **Proof:**
  - **(c)⇒(b):** $x^T L x = x^T N^T N x = (xN)^T (Nx) \geq 0$
    - As it is just the square of length of Nx
  - **(b)⇒(a):** Let λ be an eigenvalue of $\boldsymbol{L}$. Then by **(b)** $x^T L x \geq 0$ so $x^T L x = x^T \lambda x = \lambda x^T x \Rightarrow \boldsymbol{\lambda} \geq \mathbf{0}$
  - **(a)⇒(c):** is also easy! Do it yourself.

λ_2 as optimization problem

- **Fact:** **For symmetric matrix M:**

$$\lambda_2 = \min_x \frac{x^T M x}{x^T x}$$

- **What is the meaning of min $x^T L x$ on G?**
  - $x^T L x = \sum_{i,j=1}^{n} L_{ij} x_i x_j = \sum_{i,j=1}^{n} (D_{ij} - A_{ij}) x_i x_j$
  - $= \sum_i D_{ii} x_i^2 - \sum_{(i,j)\in E} 2 x_i x_j$
  - $= \sum_{(i,j)\in E} (\underbrace{x_i^2 + x_j^2} - 2 x_i x_j) = \sum_{(i,j)\in E} (x_i - x_j)^2$

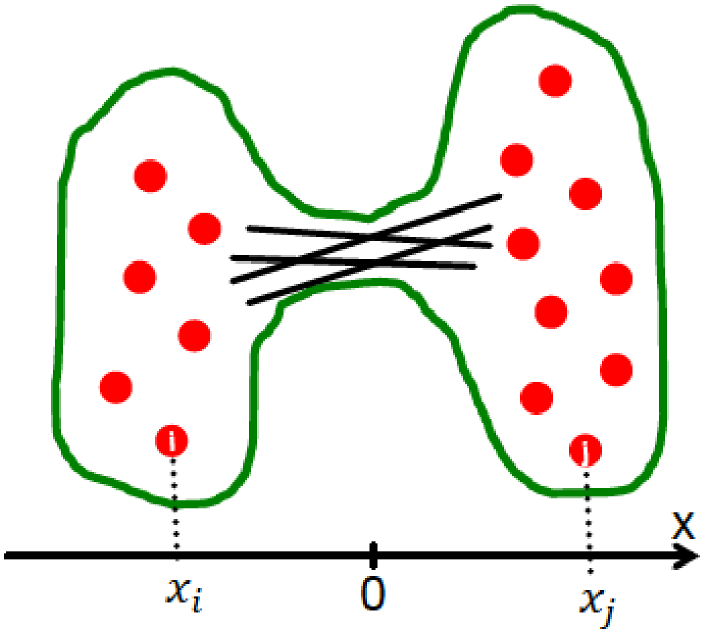
Node i has degree d_i. So, value x_i^2 needs to be summed up d_i times.
But each edge (i, j) has two endpoints so we need $x_i^2 + x_j^2$

λ_2 as optimization problem

- **What else do we know about x?**
  - $\boldsymbol{x}$ is unit vector: $\sum_i x_i^2 = 1$
  - $\boldsymbol{x}$ is orthogonal to 1st eigenvector $(\mathbf{1}, \dots, \mathbf{1})$ thus: $\sum_i x_i \cdot 1 = \sum_i x_i = 0$
- **Remember:**

$$\lambda_2 = \min_{\substack{\text{All labelings} \\ \text{of nodes } i \text{ so} \\ \text{that } \sum x_i = 0}} \frac{\sum_{(i,j) \in E} (x_i - x_j)^2}{\sum_i x_i^2}$$

We want to assign values x_i to nodes i such that few edges cross 0.
(we want x_i and x_j to subtract each other)



Balance to minimize

Find Optimal Cut [Fiedler'73]

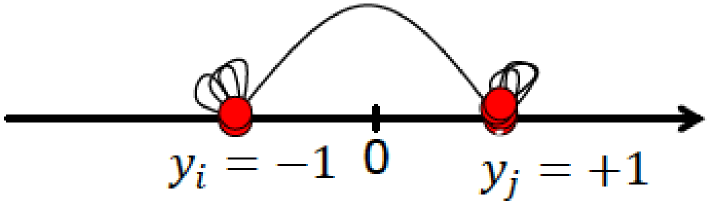
- **Back to finding the optimal cut**
- **Express partition (A,B) as a vector**

$$y_i = \begin{cases} +1 & if\ i \in A \\ -1 & if\ i \in B \end{cases}$$

- We can minimize the cut of the partition by finding a non-trivial vector x that **minimizes**:

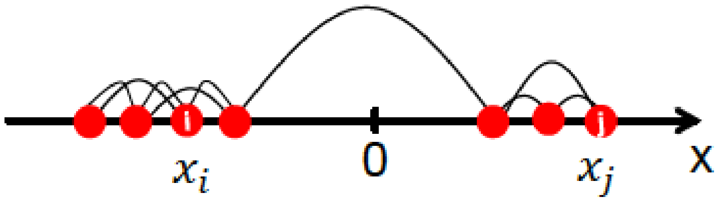
$$\arg\min_{y \in [-1,+1]^n} f(y) = \sum_{(i,j) \in E} (y_i - y_j)^2$$

Can't solve exactly. Let's relax y and allow it to take any real value.



Rayleigh Theorem

$$\min_{y \in \Re^n} f(y) = \sum_{(i,j)\in E} (y_i - y_j)^2 = y^T L y$$



- $\lambda_2 = \min_y f(y)$: The minimum value of $f(y)$ is given by the 2nd smallest eigenvalue λ_2 of the Laplacian matrix L
- $\mathbf{x} = \arg\min_y f(y)$: The optimal solution for y is given by the corresponding eigenvector x, referred as the **Fiedler vector**

So far…

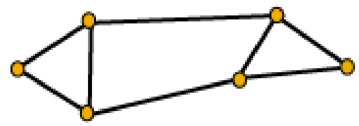
- **How to define a "good" partition of a graph?**
  - Minimize a given graph cut criterion
- **How to efficiently identify such a partition?**
  - Approximate using information provided by the eigenvalues and eigenvectors of a graph
- **Spectral Clustering**

Spectral Clustering Algorithms

- **Three basic stages:**
  - **1) Pre-processing**
    - Construct a matrix representation of the graph
  - **2) Decomposition**
    - Compute eigenvalues and eigenvectors of the matrix
    - Map each point to a lower-dimensional representation based on one or more eigenvectors
  - **3) Grouping**
    - Assign points to two or more clusters, based on the new representation

Spectral Partitioning Algorithm

- 1) **Pre-processing:**
  - Build Laplacian matrix L of the graph



	1	2	3	4	5	6
1	3	-1	-1	0	-1	0
2	-1	2	-1	0	0	0
3	-1	-1	3	-1	0	0
4	0	0	-1	3	-1	-1
5	-1	0	0	-1	3	-1
6	0	0	0	-1	-1	2

- 2) **Decomposition:**
  - Find eigenvalues λ and eigenvectors x of the matrix L

λ=

0.0
1.0
3.0
3.0
4.0
5.0

X =

0.4	0.3	-0.5	-0.2	-0.4	-0.5
0.4	0.6	0.4	-0.4	0.4	0.0
0.4	0.3	0.1	0.6	-0.4	0.5
0.4	-0.3	0.1	0.6	0.4	-0.5
0.4	-0.3	-0.5	-0.2	0.4	0.5
0.4	-0.6	0.4	-0.4	-0.4	0.0

  - Map vertices to corresponding components of λ_2

1	0.3
2	0.6
3	0.3
4	-0.3
5	-0.3
6	-0.6

How do we now find the clusters?

Spectral Partitioning

- 3) **Grouping:**
  - Sort components of reduced 1-dimensional vector
  - Identify clusters by splitting the sorted vector in two
- **How to choose a splitting point?**
  - Naïve approaches:
    - Split at **0** or median value
  - More expensive approaches:
    - Attempt to minimize normalized cut in 1-dimension (sweep over ordering of nodes induced by the eigenvector)

1	0.3
2	0.6
3	0.3
4	-0.3
5	-0.3
6	-0.6

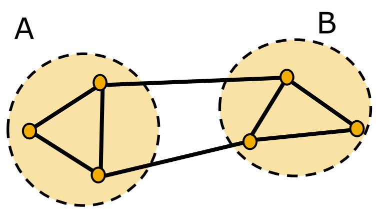
Split at 0:

Cluster A: Positive points

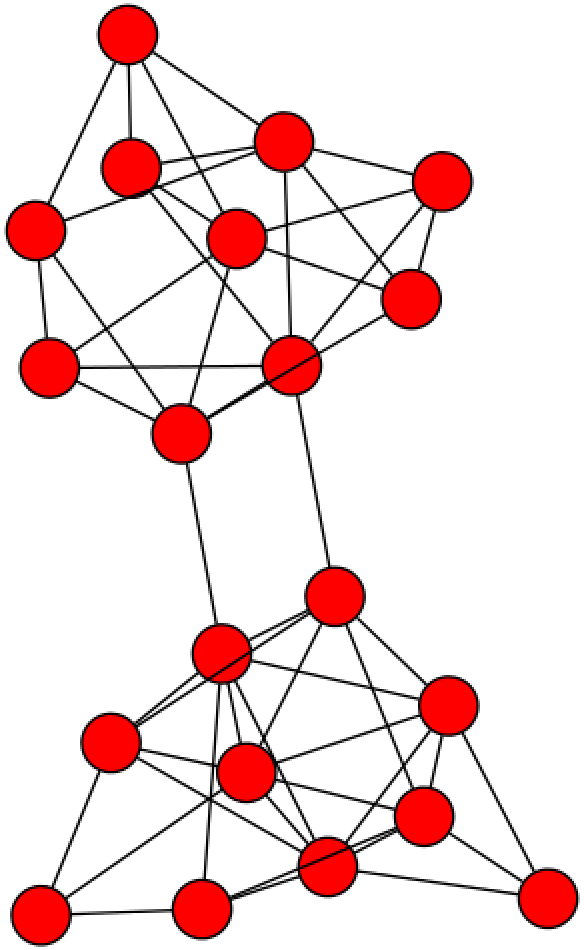
Cluster B: Negative points

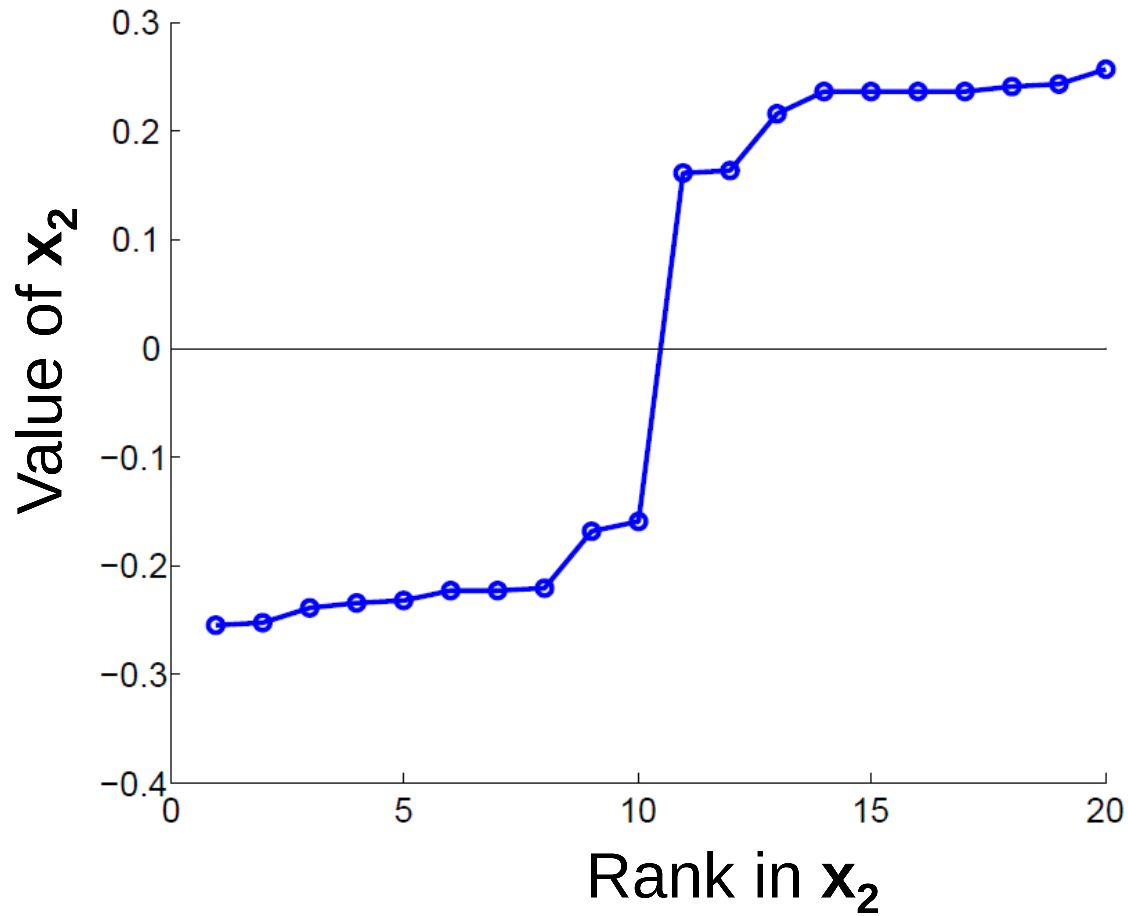
1	0.3
2	0.6
3	0.3

4	-0.3
5	-0.3
6	-0.6

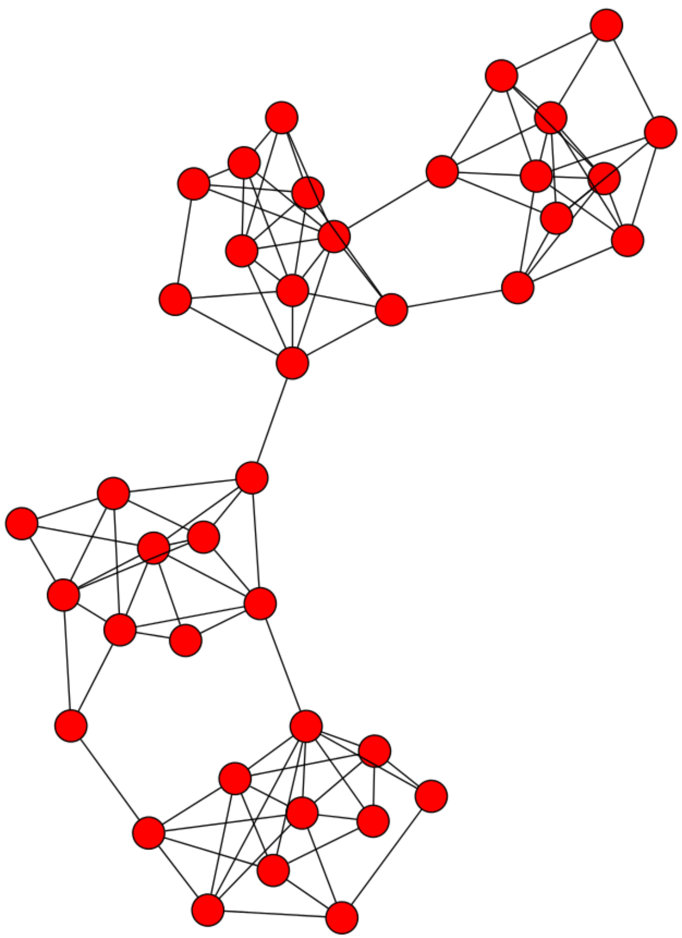


Example: Spectral Partitioning

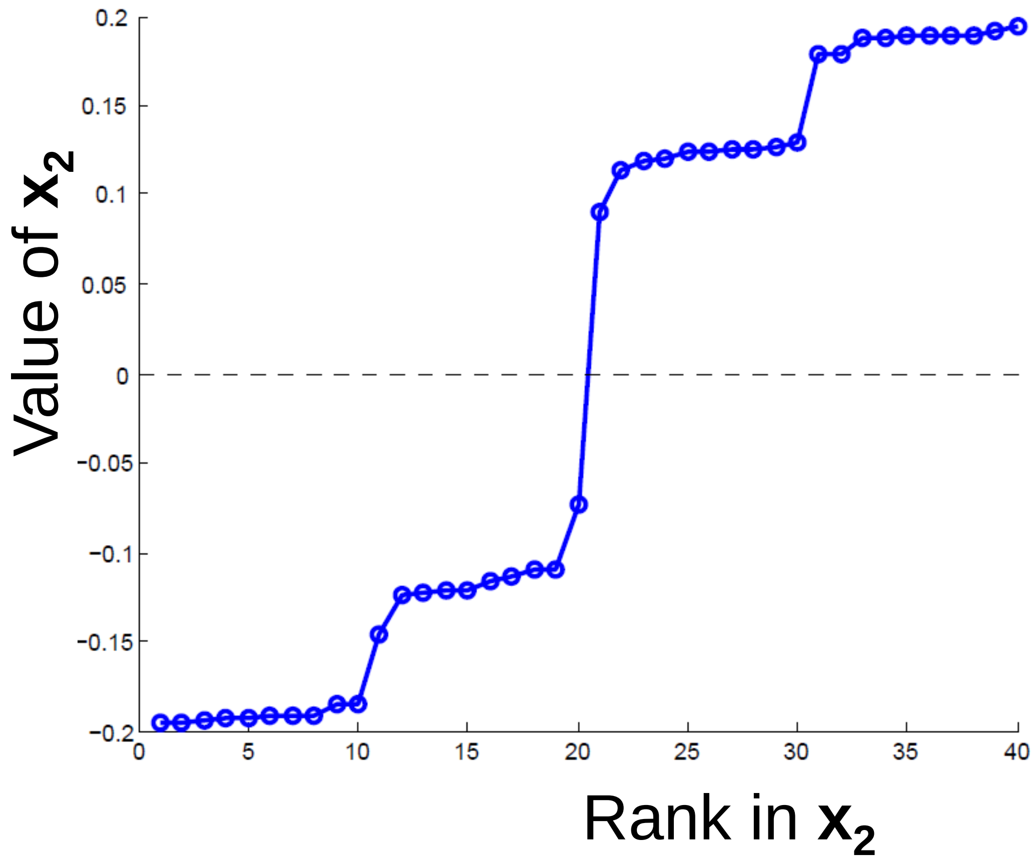




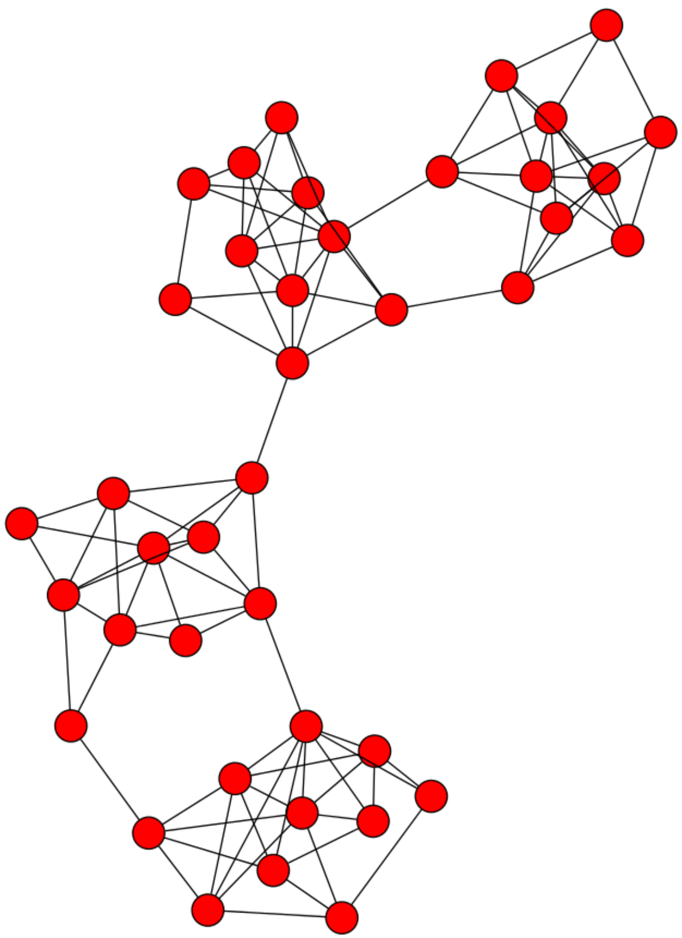
Example: Spectral Partitioning

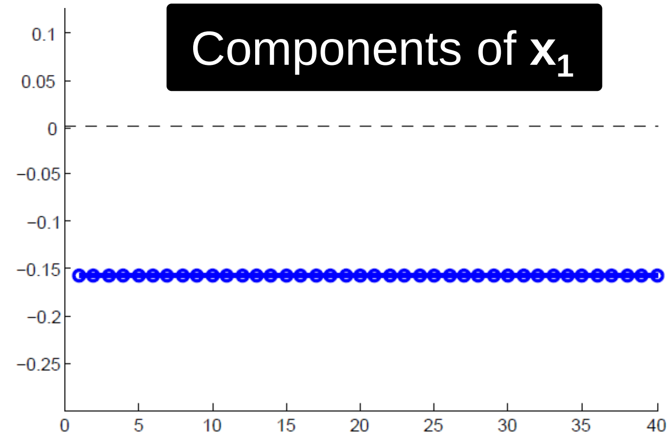


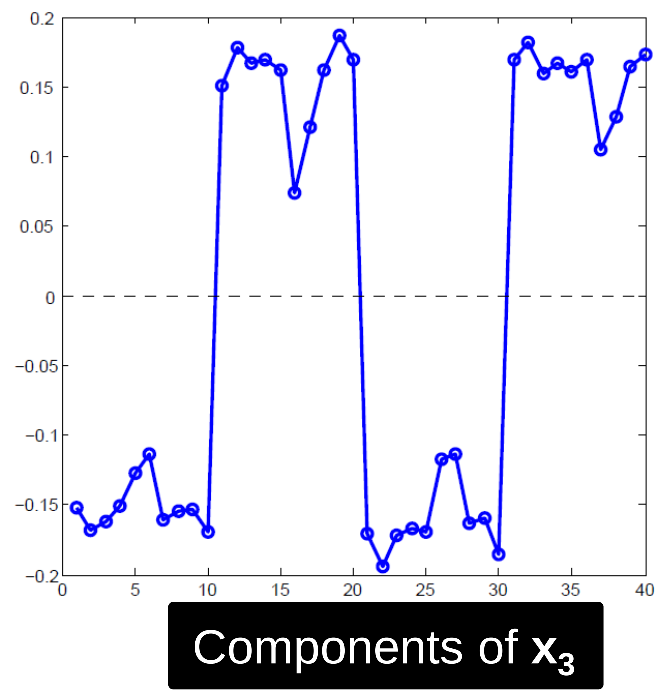
Components of $\mathbf{x_2}$



Example: Spectral partitioning







k-Way Spectral Clustering

- **How do we partition a graph into *k* clusters?**
- **Two basic approaches:**
  - **Recursive bi-partitioning** [Hagen et al., '92]
    - Recursively apply bi-partitioning algorithm in a hierarchical divisive manner
    - Disadvantages: Inefficient, unstable
  - **Cluster multiple eigenvectors** [Shi-Malik, '00]
    - Build a reduced space from multiple eigenvectors
    - Commonly used in recent papers
    - A preferable approach…

Why use multiple eigenvectors?

- **Approximates the optimal cut** [Shi-Malik, ’00]
  - Can be used to approximate optimal k-way normalized cut
- **Emphasizes cohesive clusters**
  - Increases the unevenness in the distribution of the data
  - Associations between similar points are amplified, associations between dissimilar points are attenuated
  - The data begins to “approximate a clustering”
- **Well-separated space**
  - Transforms data to a new “embedded space”, consisting of k orthogonal basis vectors
- Multiple eigenvectors prevent instability due to information loss

Some publications

- http://cse.sysu.edu.cn/node/2465

Homework 3

作业题目：实现并测试Modularity算法

作业要求:

1. 实现Modularity算法，采用Fast unfolding of communities in large networks（查资料）实现modularity的优化。
2. 在斯坦福大学网络数据集网站https://snap.stanford.edu/data/或者其他网站（e.g. https://www.scholat.com/research/opendata/)找到自己认为合适的5个数据集，进行如下分析:
   - 自己算法的社区发现结果，用Normalized Mutual Information度量效果，见ANMI_analytical_11.m，非Matlab版本的度量方法同学们可以网上找到。
   - GenLouvain算法的社区发现结果，代码下载链接：http://netwiki.amath.unc.edu/GenLouvain/GenLouvain），用Normalized Mutual Information度量效果。
   - 分析两者的差异并发现自己代码的问题。
3. 提交代码+数据集+详细实验报告及分析（编程语言不限、报告字数不限，需要透彻分析），压缩包提交：学号+姓名。
4. 提交日期：6月1日。提交邮箱：sysumldm2022@163.com