Script generated by TTT

Title: groh: profile1 (19.05.2015)

Date: Tue May 19 14:58:23 CEST 2015

Duration: 91:37 min

Pages: 91

Social Network Analysis: Graph Clustering

Lecture follows [1]. Citations of [1] are mostly omitted because of simplicity

Graph Clustering

- Given directed, weighted graph G=(V,E,w); A graph clustering **C**={C_1, C_2, ..., C_k} is a partition of V into non-empty subsets C_i;
- Notations:
  - E(C_i,C_j): Set of edges in G from C_i to C_j;
  - $E(\mathbf{C}) = \cup_{i=1,\dots,k} E(C\_i)$: Set of intra-cluster edges;
  - $\overline{E(\mathbf{C})} = E \setminus E(\mathbf{C})$: Set of inter-cluster edges;
  - $m(\mathbf{C}) = |E(\mathbf{C})|$; $\overline{m}(\mathbf{C}) = |\overline{E(\mathbf{C})}|$;
  - G([C_i]): subgraph induced by C_i;
  - **C** with k=1: *1-clustering*; **C** with k=|V|: *singletons*; (both: *Trivial clustering);* **C** with k=2: *cut;*
  - **A**(G): Set of all possible clusterings on G;

Graph Clustering

- Given directed, weighted graph G=(V,E,w); A graph clustering **C**={C_1, C_2, ..., C_k} is a partition of V into non-empty subsets C_i;
- Notations:
  - E(C_i,C_j): Set of edges in G from C_i to C_j;
  - $E(\mathbf{C}) = \cup_{i=1,\dots,k} E(C\_i)$: Set of intra-cluster edges;
  - $\overline{E(\mathbf{C})} = E \setminus E(\mathbf{C})$: Set of inter-cluster edges;
  - $m(\mathbf{C}) = |E(\mathbf{C})|$; $\overline{m}(\mathbf{C}) = |\overline{E(\mathbf{C})}|$;
  - G([C_i]): subgraph induced by C_i;
  - **C** with k=1: *1-clustering*; **C** with k=|V|: *singletons*; (both: *Trivial clustering);* **C** with k=2: *cut;*
  - **A**(G): Set of all possible clusterings on G;

Graph Clustering

- Given directed, weighted graph G=(V,E,w); A graph clustering **C**={C_1, C_2, ..., C_k} is a partition of V into non-empty subsets C_i;
- Notations:
  - E(C_i,C_j): Set of edges in G from C_i to C_j;
  - E(**C**) = $\bigcup_{i=1,\dots,k}$E(C_i): Set of intra-cluster edges;
  - $\overline{E(\mathbf{C})}$ = E \ E(**C**) : Set of inter-cluster edges;
  - m(**C**) = |E(**C**)|; $\overline{m}$(**C**) = |$\overline{E(\mathbf{C})}$|;
  - G([C_i]): subgraph induced by C_i;
  - **C** with k=1: *1-clustering*; **C** with k=|V|: *singletons*;
    (both: *Trivial clustering);*
    C with k=2: *cut;*
  - **A**(G): Set of all possible clusterings on G;

Graph Clustering

- Given directed, weighted graph G=(V,E,w); A graph clustering **C**={C_1, C_2, ..., C_k} is a partition of V into non-empty subsets C_i;
- Notations:
  - E(C_i,C_j): Set of edges in G from C_i to C_j;
  - E(**C**) = $\bigcup_{i=1,\dots,k}$E(C_i): Set of intra-cluster edges;
  - $\overline{E(\mathbf{C})}$ = E \ E(**C**) : Set of inter-cluster edges;
  - m(**C**) = |E(**C**)|; $\overline{m}$(**C**) = |$\overline{E(\mathbf{C})}$|;
  - G([C_i]): subgraph induced by C_i;
  - **C** with k=1: *1-clustering*; **C** with k=|V|: *singletons*;
    (both: *Trivial clustering);*
    C with k=2: *cut;*
  - **A**(G): Set of all possible clusterings on G;

Graph Clustering

- Notations (continued):
  - Let **C_1**={C_1, C_2, ..., C_k} and **C_2**={C'_1, C'_2, ..., C'_l}:
    C_1 ≤ **C_2** ↔ ∀i ∃j : C_i ⊆C'_j;
    C_1: *refinement of* **C_2; C_2:** *coarsening of* **C_1;**
  - *Chain* (comparable set) of clusterings: *hierarchy;*
  - Hierarchy is *total* ←→ Both trivial clusterings are contained;
  - Hierarchy that contains one clustering for each {1,2,...,|V|}: *complete;*
  - S(V): *(Proper) Cut function*: Cutting the node-set in two (non-empty) subsets: S(V) and V \ S(V);

Graph Clustering

- Notations (continued):
  - Let **C_1**={C_1, C_2, ..., C_k} and **C_2**={C'_1, C'_2, ..., C'_l}:
    C_1 ≤ **C_2** ↔ ∀i ∃j : C_i ⊆C'_j;
    C_1: *refinement of* **C_2; C_2:** *coarsening of* **C_1;**
  - *Chain* (comparable set) of clusterings: *hierarchy;*
  - Hierarchy is *total* ←→ Both trivial clusterings are contained;
  - Hierarchy that contains one clustering for each {1,2,...,|V|}: *complete;*
  - S(V): *(Proper) Cut function*: Cutting the node-set in two (non-empty) subsets: S(V) and V \ S(V);

Graph Clustering

- Notations (continued):
  - Let **C_1**={C_1, C_2, ..., C_k} and **C_2**={C'_1, C'_2, ..., C'_l}:
    C_1 ≤ **C_2** ↔ ∀i ∃j : C_i ⊆C'_j;
    C_1: *refinement of* **C_2; C_2:** *coarsening of* **C_1;**
  - *Chain* (comparable set) of clusterings: *hierarchy;*
  - Hierarchy is *total* ←→ Both trivial clusterings are contained;
  - Hierarchy that contains one clustering for each {1,2,...,|V|}: *complete;*
  - S(V): *(Proper) Cut function*: Cutting the node-set in two (non-empty) subsets: S(V) and V \ S(V);

Graph Clustering

- Notations (continued):
  - Let **C_1**={C_1, C_2, ..., C_k} and **C_2**={C'_1, C'_2, ..., C'_l}:
    C_1 ≤ **C_2** ↔ ∀i ∃j : C_i ⊆C'_j;
    C_1: *refinement of* **C_2; C_2:** *coarsening of* **C_1;**
  - *Chain* (comparable set) of clusterings: *hierarchy;*
  - Hierarchy is *total* ←→ Both trivial clusterings are contained;
  - Hierarchy that contains one clustering for each {1,2,...,|V|}: *complete;*
  - S(V): *(Proper) Cut function*: Cutting the node-set in two (non-empty) subsets: S(V) and V \ S(V);

Quality Measures for Clusterings

- Quality measure: Objective function **A**(G) → $\mathbb{R}$ that formalizes the clustering paradigm in a special way

- G= (V,E,w): Weight function w: E→ $\mathbb{R}^+$ is interpreted as "similarity" (higher weights correspond to more intense tie); also possible: negative weights = dissimilarity; or w: E → [0,1] or w: E → [-1,1] etc.

- Distinguish between no edge and edge with weight zero;

- Notation: $w(E) = \sum_{e \in E} w(e)$

Quality Measures for Clusterings

- Quality measure: Objective function **A**(G) → $\mathbb{R}$ that formalizes the clustering paradigm in a special way

- G= (V,E,w): Weight function w: E→ $\mathbb{R}^+$ is interpreted as "similarity" (higher weights correspond to more intense tie); also possible: negative weights = dissimilarity; or w: E → [0,1] or w: E → [-1,1] etc.

- Distinguish between no edge and edge with weight zero;

- Notation: $w(E) = \sum_{e \in E} w(e)$

6: Graph Clustering

- General framework for a quality index of a clustering:

$$index(\mathbf{C}) = \frac{f(\mathbf{C}) + g(\mathbf{C})}{\max\{f(\mathbf{C}') + g(\mathbf{C}') : \mathbf{C}' \in \mathbf{A}(G)\}}$$

- $f : \mathbf{A}(G) \rightarrow \mathbb{R}^+$ measures intra cluster density (coherence);
  $g : \mathbf{A}(G) \rightarrow \mathbb{R}^+$ measures inter cluster sparseness (decoherence);

6: Graph Clustering

- General framework for a quality index of a clustering:

$$index(\mathbf{C}) = \frac{f(\mathbf{C}) + g(\mathbf{C})}{\max\{f(\mathbf{C}') + g(\mathbf{C}') : \mathbf{C}' \in \mathbf{A}(G)\}}$$

- $f : \mathbf{A}(G) \rightarrow \mathbb{R}^+$ measures intra cluster density (coherence);
  $g : \mathbf{A}(G) \rightarrow \mathbb{R}^+$ measures inter cluster sparseness (decoherence);

Coverage

- First quality measure: Coverage

$$\gamma(\mathbf{C}) = \frac{w(E(\mathbf{C}))}{w(E)} = \frac{\sum_{e \in E(\mathbf{C})} w(e)}{\sum_{e \in E} w(e)}$$

- Thus: f = w(E(**C**)) and g = 0; → only accumulated intra cluster density is measured
- Maximum value 1 achieved for **C**={V} (1-clustering)
- A clustering has coverage $\gamma(\mathbf{C})=1$ iff $\overline{E}(\mathbf{C})=\varnothing$ (clustering is composed of connected components of G) or $w(\overline{E}(\mathbf{C}))=0$
- **C** with k>1 can be transformed into **C´** with k´<k and γ(**C**) ≤ γ(**C´**) by merging two clusters in **C** → optimal non-trivial **C** is a minimum cut
- → „Monotonic“ behavior of coverage → coverage is not a good sole quality index

Coverage

- First quality measure: Coverage

$$\gamma(\mathbf{C}) = \frac{w(E(\mathbf{C}))}{w(E)} = \frac{\sum_{e \in E(\mathbf{C})} w(e)}{\sum_{e \in E} w(e)}$$

- Thus: f = w(E(**C**)) and g = 0; → only accumulated intra cluster density is measured
- Maximum value 1 achieved for **C**={V} (1-clustering)
- A clustering has coverage $\gamma(\mathbf{C})=1$ iff $\overline{E}(\mathbf{C})=\varnothing$ (clustering is composed of connected components of G) or $w(\overline{E}(\mathbf{C}))=0$
- **C** with k>1 can be transformed into **C´** with k´<k and γ(**C**) ≤ γ(**C´**) by merging two clusters in **C** → optimal non-trivial **C** is a minimum cut
- → „Monotonic“ behavior of coverage → coverage is not a good sole quality index

Coverage

- First quality measure: Coverage

$$\gamma(\mathbf{C}) = \frac{w(E(\mathbf{C}))}{w(E)} = \frac{\sum_{e \in E(\mathbf{C})} w(e)}{\sum_{e \in E} w(e)}$$

- Thus: f = w(E(**C**)) and g = 0; → only accumulated intra cluster density is measured
- Maximum value 1 achieved for **C**={V} (1-clustering)
- A clustering has coverage γ(**C**)=1 iff $\overline{E(\mathbf{C})}=\varnothing$ (clustering is composed of connected components of G) or $w(\overline{E(\mathbf{C})})=0$
- **C** with k>1 can be transformed into **C´** with k´<k and γ(**C**) ≤ γ(**C´**) by merging two clusters in **C** → optimal non-trivial **C** is a minimum cut
- → „Monotonic“ behavior of coverage → coverage is not a good sole quality index

Coverage

- First quality measure: Coverage

$$\gamma(\mathbf{C}) = \frac{w(E(\mathbf{C}))}{w(E)} = \frac{\sum_{e \in E(\mathbf{C})} w(e)}{\sum_{e \in E} w(e)}$$

- Thus: f = w(E(**C**)) and g = 0; → only accumulated intra cluster density is measured
- Maximum value 1 achieved for **C**={V} (1-clustering)
- A clustering has coverage γ(**C**)=1 iff $\overline{E(\mathbf{C})}=\varnothing$ (clustering is composed of connected components of G) or $w(\overline{E(\mathbf{C})})=0$
- **C** with k>1 can be transformed into **C´** with k´<k and γ(**C**) ≤ γ(**C´**) by merging two clusters in **C** → optimal non-trivial **C** is a minimum cut
- → „Monotonic“ behavior of coverage → coverage is not a good sole quality index

Conductance

- Clustering paradigm reformulated: Clusters should be well connected (many edges need to be removed to make it unconnected); few inter cluster edges (ideally none)
- Conductance: Measure for bottlenecks (Bottleneck: Cut that separates V into roughly same size halves and “cuts across” relatively few edges)
- Let **C**={C_1, V \ C_1} be a cut. Conductance φ of **C** is defined as

$$\varphi(\mathbf{C}) = \begin{cases} 1 & \text{if } C\_1 \in \{\varnothing, V\} \\ 0 & \text{if } C\_1 \notin \{\varnothing, V\},\ w(\overline{E(\mathbf{C})})=0 \\ \dfrac{w(\overline{E(\mathbf{C})})}{\min\left(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e)\right)} & \text{otherwise} \end{cases}$$

the smaller φ(**C**), the more „bottlenecky“ is **C**

Conductance

- Clustering paradigm reformulated: Clusters should be well connected (many edges need to be removed to make it unconnected); few inter cluster edges (ideally none)
- Conductance: Measure for bottlenecks (Bottleneck: Cut that separates V into roughly same size halves and “cuts across” relatively few edges)
- Let **C**={C_1, V \ C_1} be a cut. Conductance φ of **C** is defined as

$$\varphi(\mathbf{C}) = \begin{cases} 1 & \text{if } C\_1 \in \{\varnothing, V\} \\ 0 & \text{if } C\_1 \notin \{\varnothing, V\},\ w(\overline{E(\mathbf{C})})=0 \\ \dfrac{w(\overline{E(\mathbf{C})})}{\min\left(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e)\right)} & \text{otherwise} \end{cases}$$

the smaller φ(**C**), the more „bottlenecky“ is **C**

Conductance

- Clustering paradigm reformulated: Clusters should be well connected (many edges need to be removed to make it unconnected); few inter cluster edges (ideally none)
- Conductance: Measure for bottlenecks (Bottleneck: Cut that separates V into roughly same size halves and “cuts across” relatively few edges)
- Let **C**={C_1, V \ C_1} be a cut. Conductance φ of **C** is defined as

$$\varphi(\mathbf{C}) = \begin{cases} 1 & \text{if C\_1} \in \{\varnothing, V\} \\ 0 & \text{if C\_1} \notin \{\varnothing, V\},\ w(\overline{E(\mathbf{C})})=0 \\ \dfrac{w(\overline{E(\mathbf{C})})}{\min(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e))} & \text{otherwise} \end{cases}$$

the smaller φ(**C**), the more „bottlenecky“ is **C**

Conductance

- Clustering paradigm reformulated: Clusters should be well connected (many edges need to be removed to make it unconnected); few inter cluster edges (ideally none)
- Conductance: Measure for bottlenecks (Bottleneck: Cut that separates V into roughly same size halves and “cuts across” relatively few edges)
- Let **C**={C_1, V \ C_1} be a cut. Conductance φ of **C** is defined as

$$\varphi(\mathbf{C}) = \begin{cases} 1 & \text{if C\_1} \in \{\varnothing, V\} \\ 0 & \text{if C\_1} \notin \{\varnothing, V\},\ w(\overline{E(\mathbf{C})})=0 \\ \dfrac{w(\overline{E(\mathbf{C})})}{\min(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e))} & \text{otherwise} \end{cases}$$

the smaller φ(**C**), the more „bottlenecky“ is **C**

Conductance

- Clustering paradigm reformulated: Clusters should be well connected (many edges need to be removed to make it unconnected); few inter cluster edges (ideally none)
- Conductance: Measure for bottlenecks (Bottleneck: Cut that separates V into roughly same size halves and “cuts across” relatively few edges)
- Let **C**={C_1, V \ C_1} be a cut. Conductance φ of **C** is defined as

$$\varphi(\mathbf{C}) = \begin{cases} 1 & \text{if C\_1} \in \{\varnothing, V\} \\ 0 & \text{if C\_1} \notin \{\varnothing, V\},\ w(\overline{E(\mathbf{C})})=0 \\ \dfrac{w(\overline{E(\mathbf{C})})}{\min(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e))} & \text{otherwise} \end{cases}$$

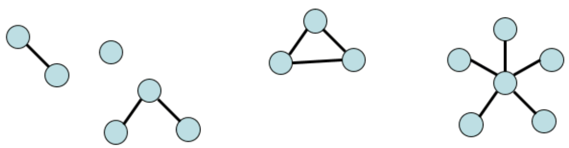
the smaller φ(**C**), the more „bottlenecky“ is **C**

Conductance

- Conductance φ of G is defined as

$$\varphi(G) = \min_{C\_1 \subseteq V} \varphi(C\_1, V \setminus C\_1)$$

- Small conductance ←→”good cut possible”
- All unconnected graphs have conductance 0
- Theorem: If G is undirected and positively weighted, G has maximum conductance φ(G)=1 iff G is connected and has at most three nodes or is a star.

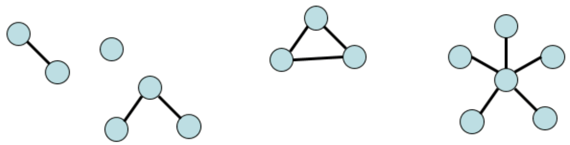


Conductance

- Conductance φ of G is defined as

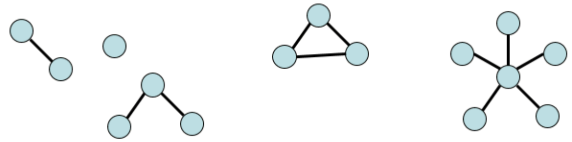
$$\varphi(G) = \min_{C\_1 \subseteq V} \varphi(C\_1,\ V \setminus C\_1)$$

- Small conductance ←→"good cut possible"
- All unconnected graphs have conductance 0
- Theorem: If G is undirected and positively weighted, G has maximum conductance φ(G)=1 iff G is connected and has at most three nodes or is a star.



Conductance

- Theorem: If G is undirected and positively weighted, G has maximum conductance φ(G)=1 iff G is connected and has at most three nodes or is a star. (Proof: see [1])



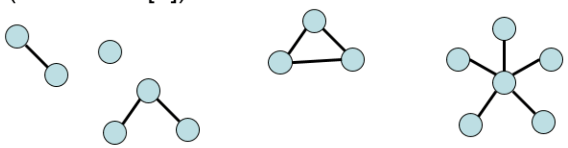
- Proof `←' $\sum_{e \in E(C\_1, V)} w(e) = w(E(C\_1)) + w(\overline{E(\mathbf{C})}) \rightarrow$

$$\frac{w(\overline{E(\mathbf{C})})}{\min\left(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e)\right)} =$$

$$\frac{w(\overline{E(\mathbf{C})})}{w(\overline{E(\mathbf{C})}) + \min(\underbrace{w(E(C\_1)), w(E(V \setminus C\_1))}_{0 \text{ if star or at most 3 nodes}})} = 1$$

Conductance

- Theorem: If G is undirected and positively weighted, G has maximum conductance φ(G)=1 iff G is connected and has at most three nodes or is a star. (Proof: see [1])



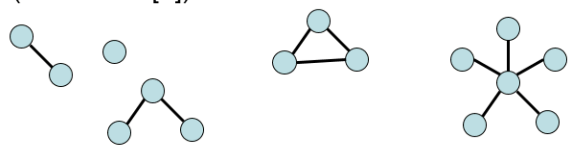
- Proof `←' $\sum_{e \in E(C\_1, V)} w(e) = w(E(C\_1)) + w(\overline{E(\mathbf{C})}) \rightarrow$

$$\frac{w(\overline{E(\mathbf{C})})}{\min\left(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e)\right)} =$$

$$\frac{w(\overline{E(\mathbf{C})})}{w(\overline{E(\mathbf{C})}) + \min(\underbrace{w(E(C\_1)), w(E(V \setminus C\_1))}_{0 \text{ if star or at most 3 nodes}})} = 1$$

Conductance

- Theorem: If G is undirected and positively weighted, G has maximum conductance φ(G)=1 iff G is connected and has at most three nodes or is a star. (Proof: see [1])



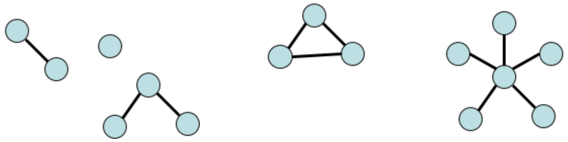
- Proof `←' $\sum_{e \in E(C\_1, V)} w(e) = w(E(C\_1)) + w(\overline{E(\mathbf{C})}) \rightarrow$

$$\frac{w(\overline{E(\mathbf{C})})}{\min\left(\sum_{e \in E(C\_1, V)} w(e), \sum_{e \in E(V \setminus C\_1, V)} w(e)\right)} =$$

$$\frac{w(\overline{E(\mathbf{C})})}{w(\overline{E(\mathbf{C})}) + \min(\underbrace{w(E(C\_1)), w(E(V \setminus C\_1))}_{0 \text{ if star or at most 3 nodes}})} = 1$$

Conductance

- Theorem: If G is undirected and positively weighted, G has maximum conductance φ(G)=1 iff G is connected and has at most three nodes or is a star. (Proof: see [1])



- Proof `←' $\sum_{e\in E(C\_1,V)} w(e) = w(E(C\_1)) + w(\overline{E}(\mathbf{C}))$ →

$$\frac{w(\overline{E}(\mathbf{C}))}{\min\left(\sum_{e\in E(C\_1,V)} w(e), \sum_{e\in E(V\setminus C\_1,V)} w(e)\right)} =$$

$$\frac{w(\overline{E}(\mathbf{C}))}{w(\overline{E}(\mathbf{C})) + \min(\underbrace{w(E(C\_1))}_{\text{0 if star or at most 3 nodes}}, w(E(V\setminus C\_1)))} = 1$$

Conductance

- Conductance appears to be a good element for a quality measure, but: calculating it is NP hard ☹ but: it can be approximated with guarantee $O((\log |V|)^{1/2})$ (see [1]) (This means: approximation * $O((\log |V|)^{1/2})$ = true value)

- The conductance of a complete graph is asymptotically 0.5: Let n be an integer:

$$\varphi(K_n) = \begin{cases} 0.5\frac{n}{n-1} & \text{if } n \text{ is even} \\ 0.5 + \frac{1}{n-1} & \text{if } n \text{ is odd} \end{cases}$$

Conductance

- Conductance appears to be a good element for a quality measure, but: calculating it is NP hard ☹ but: it can be approximated with guarantee $O((\log |V|)^{1/2})$ (see [1]) (This means: approximation * $O((\log |V|)^{1/2})$ = true value)

- The conductance of a complete graph is asymptotically 0.5: Let n be an integer:

$$\varphi(K_n) = \begin{cases} 0.5\frac{n}{n-1} & \text{if } n \text{ is even} \\ 0.5 + \frac{1}{n-1} & \text{if } n \text{ is odd} \end{cases}$$

Conductance

- With conductance we can define two appropriate quality measures for clusterings:

- First measure: g=0 and $f(\mathbf{C}) = \min_{1\le i\le k} \varphi(G[C\_i])$

- If the measure is small: At least one of the clusters (more precisely: the induced subgraph) contains at least one bottleneck → This cluster is too coarse →Use minimum conductance cut to cut this cluster in "halves"

- From theorem before: Only clusterings where the clusters induce subgraphs that are stars or have size at most three have f=1 (f is called *intra cluster conductance*)

Conductance

- With conductance we can define two appropriate quality measures for clusterings:
- First measure: g=0 and $f(\mathbf{C}) = \min_{1 \le i \le k} \varphi(G[\mathrm{C\_i}])$
- If the measure is small: At least one of the clusters (more precisely: the induced subgraph) contains at least one bottleneck → This cluster is too coarse →Use minimum conductance cut to cut this cluster in “halves”
- From theorem before: Only clusterings where the clusters induce subgraphs that are stars or have size at most three have f=1 (f is called *intra cluster conductance*)

Conductance

- Second measure: f=0 and

$$g(\mathbf{C}) = \begin{cases} 1 & \text{if } \mathbf{C} = \{V\} \\ 1 - \max_{1 \le i \le k} \varphi(\mathrm{C\_i},\ V \setminus \mathrm{C\_i}) & \text{otherwise} \end{cases}$$

- If the measure is small: At least one of the clusters (more precisely: the induced subgraph) has many connections to outside → The clustering is too fine →Merge clusters
- From theorem before: Only clusterings that have inter cluster edge weight zero have g=1 (g is called *inter cluster conductance*)

Conductance

- Second measure: f=0 and

$$g(\mathbf{C}) = \begin{cases} 1 & \text{if } \mathbf{C} = \{V\} \\ 1 - \max_{1 \le i \le k} \varphi(\mathrm{C\_i},\ V \setminus \mathrm{C\_i}) & \text{otherwise} \end{cases}$$

- If the measure is small: At least one of the clusters (more precisely: the induced subgraph) has many connections to outside → The clustering is too fine →Merge clusters
- From theorem before: Only clusterings that have inter cluster edge weight zero have g=1 (g is called *inter cluster conductance*)

Performance

- Main idea: Clustering paradigm → Count “correctly classified pairs of nodes”. A pair of nodes is correctly classified if:
  - It is in the same cluster AND connected by an edge → f counts the number of edges within clusters
  - If it is not in the same cluster AND not connected by an edge →g counts the number of non-existent edges between clusters

$$f(\mathbf{C}) = \sum_{i=1}^{k} |E(\mathrm{C\_i})|$$

$$g(\mathbf{C}) = \sum_{u,v \in V} [(u,v) \notin E] * [u \in \mathrm{C\_i}, v \in \mathrm{C\_j}, i \neq j]$$

Iverson-notation: [L]=1 if L is true

Performance

- Main idea: Clustering paradigm → Count “correctly classified pairs of nodes”. A pair of nodes is correctly classified if:
  - It is in the same cluster AND connected by an edge → f counts the number of edges within clusters
  - If it is not in the same cluster AND not connected by an edge → g counts the number of non-existent edges between clusters

$$f(\mathbf{C}) = \sum_{i=1}^{k} |E(\mathrm{C\_i})|$$

$$g(\mathbf{C}) = \sum_{u,v \in V} [(u,v) \notin E] * [u \in \mathrm{C\_i}, v \in \mathrm{C\_j}, i \neq j]$$

Iverson-notation: [L]=1 if L is true

Performance

- Main idea: Clustering paradigm → Count “correctly classified pairs of nodes”. A pair of nodes is correctly classified if:
  - It is in the same cluster AND connected by an edge → f counts the number of edges within clusters
  - If it is not in the same cluster AND not connected by an edge → g counts the number of non-existent edges between clusters

$$f(\mathbf{C}) = \sum_{i=1}^{k} |E(\mathrm{C\_i})|$$

$$g(\mathbf{C}) = \sum_{u,v \in V} [(u,v) \notin E] * [u \in \mathrm{C\_i}, v \in \mathrm{C\_j}, i \neq j]$$

Iverson-notation: [L]=1 if L is true

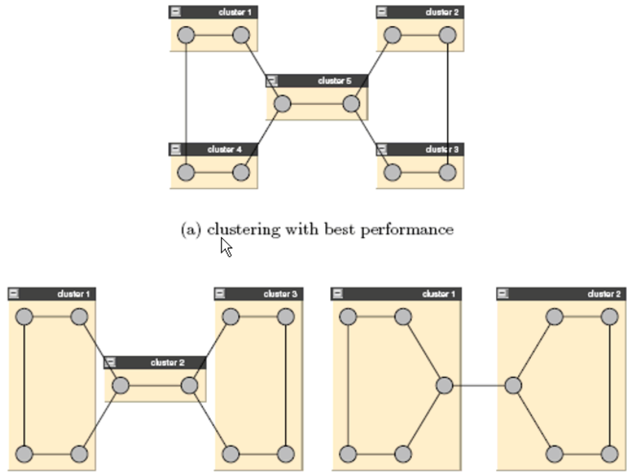
Performance

- Calculating the maximum of f+g is NP-hard (In fact calculating the maximum of f+g would in essence be calculating the optimal clustering) →use |V| (|V|-1) as normalization for quality measure
- The performance index is then:

$$\mathrm{perf}(\mathbf{C}) = \frac{f(\mathbf{C}) + g(\mathbf{C})}{|V|(|V|-1)}$$

- Problems with Performance: when graph is sparse (example: planar graphs: |E| is linear in |V|). Tendency: Performance delivers many small clusters

Performance



(a) clustering with best performance

(b) intuitive clustering (c) another intuitive clustering

Fig. 8.4. A situation where the clustering with optimal performance is a refinement (Figure 8.4(b)) of an intuitive clustering and is skew (Figure 8.4(c)) to another intuitive clustering

Performance

- If using weighted edges → some modifications:
  - use weights normalized to 1 → Max weight M = 1

$$f(\mathbf{C}) = \sum_{i=1}^{k} w(E(C\_i))$$

$$g(\mathbf{C}) = \sum_{u,v \in V} M * [(u,v) \notin E] * [u \in C\_i, v \in C\_j, i \neq j]$$

Performance

- If using weighted edges → some modifications:
  - use weights normalized to 1 → Max weight M = 1

$$f(\mathbf{C}) = \sum_{i=1}^{k} w(E(C\_i))$$

$$g(\mathbf{C}) = \sum_{u,v \in V} M * [(u,v) \notin E] * [u \in C\_i, v \in C\_j, i \neq j]$$

Performance

- If using weighted edges → some modifications:
  - use weights normalized to 1 → Max weight M = 1

$$f(\mathbf{C}) = \sum_{i=1}^{k} w(E(C\_i))$$

$$g(\mathbf{C}) = \sum_{u,v \in V} M * [(u,v) \notin E] * [u \in C\_i, v \in C\_j, i \neq j]$$

Performance

- In that version g neglects the individual inter-cluster edges → Introduce g_w

$$g'(\mathbf{C}) = g(\mathbf{C}) + \underbrace{M\,|\overline{E(\mathbf{C})}| - w(\overline{E(\mathbf{C})})}_{g_w(\mathbf{C})}$$

- Overall index is then:

$$\text{perf}_w(\mathbf{C}) = \frac{f(\mathbf{C}) + g(\mathbf{C}) + \vartheta\, g_w(\mathbf{C})}{M(|V|(|V|-1))}$$

- other possibility: minimize incorrectly classified edges (dual problem)

Performance

- In that version g neglects the individual inter-cluster edges → Introduce g_w

$$g'(\mathbf{C}) = g(\mathbf{C}) + \underbrace{M\,|\overline{E(\mathbf{C})}| - w(\overline{E(\mathbf{C})})}_{g_w(\mathbf{C})}$$

- Overall index is then:

$$\text{perf}_w(\mathbf{C}) = \frac{f(\mathbf{C}) + g(\mathbf{C}) + \vartheta\, g_w(\mathbf{C})}{M(|V|(|V|-1)}$$

- other possibility: minimize incorrectly classified edges (dual problem)

Performance

- In that version g neglects the individual inter-cluster edges → Introduce g_w

$$g'(\mathbf{C}) = g(\mathbf{C}) + \underbrace{M\,|\overline{E(\mathbf{C})}| - w(\overline{E(\mathbf{C})})}_{g_w(\mathbf{C})}$$

- Overall index is then:

$$\text{perf}_w(\mathbf{C}) = \frac{f(\mathbf{C}) + g(\mathbf{C}) + \vartheta\, g_w(\mathbf{C})}{M(|V|(|V|-1)}$$

- other possibility: minimize incorrectly classified edges (dual problem)

Other Indices

- If density measure π on graphs is available:

$$\begin{aligned} \text{worst case: } & \min_i\{\pi(G[C_1]), \ldots, \pi(G[C_k])\} \\ \text{average case: } & \frac{1}{k}\sum_i \pi(G[C_i]) \\ \text{best case: } & \max_i\{\pi(G[C_1]), \ldots, \pi(G[C_k])\} \end{aligned}$$

- (especially suitable in metric spaces)

Other Indices

- If density measure π on graphs is available:

$$\begin{aligned} \text{worst case: } & \min_i\{\pi(G[C_1]), \ldots, \pi(G[C_k])\} \\ \text{average case: } & \frac{1}{k}\sum_i \pi(G[C_i]) \\ \text{best case: } & \max_i\{\pi(G[C_1]), \ldots, \pi(G[C_k])\} \end{aligned}$$

- (especially suitable in metric spaces)

- What have we seen so far? Measures for cluster quality
- But how do we compute such clusters?
- First group of methods: Greedy approaches

greedy minimization:

```
let L0 be a feasible solution;
i ←0;
while({L | L∈N(Li), c(L)<c(Li)} ≠ Ø) {
        Li+1 ← argminL∈N(Li) c(L);
        i ← i+1;
}
```

Space of all solutions L that can be constructed from solution L_i

c(L) is the cost of solution L

Graph Clustering Algorithms

- What have we seen so far? Measures for cluster quality
- But how do we compute such clusters?
- First group of methods: Greedy approaches

greedy minimization:

```
let L0 be a feasible solution;
i ←0;
while({L | L∈N(Li), c(L)<c(Li)} ≠ Ø) {
        Li+1 ← argminL∈N(Li) c(L);
        i ← i+1;
}
```

Space of all solutions L that can be constructed from solution L_i

c(L) is the cost of solution L

Graph Clustering Algorithms

- What have we seen so far? Measures for cluster quality
- But how do we compute such clusters?
- First group of methods: Greedy approaches

greedy minimization:

```
let L0 be a feasible solution;
i ←0;
while({L | L∈N(Li), c(L)<c(Li)} ≠ Ø) {
        Li+1 ← argminL∈N(Li) c(L);
        i ← i+1;
}
```

Space of all solutions L that can be constructed from solution L_i

c(L) is the cost of solution L

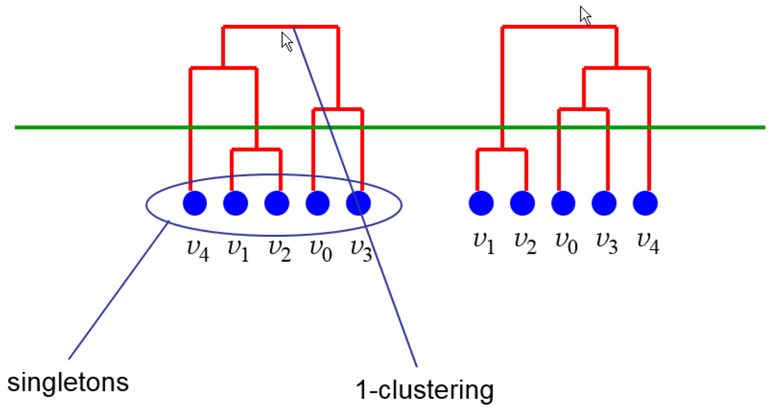
Graph Clustering Algorithms

greedy minimization:

```
let L0 be a feasible solution;
i ←0;
while({L | L∈N(Li), c(L)<c(Li)} ≠ Ø) {
        Li+1 ← argminL∈N(Li) c(L);
        i ← i+1;
}
```

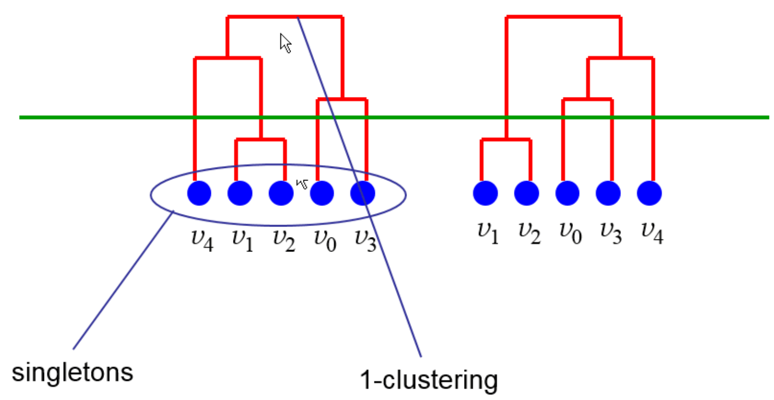
- For greedy maximization substitute argmax and >
- What function do we use as c(L) → cluster quality measures modeling the clustering paradigm
- How do we construct solutions (clusterings) from other solutions: Merging or splitting of clusters → a hierarchy of clusters results → "Dendrogram"

- Advantage of Dendrograms: Can be "**cut**" at any desired number of clusters.



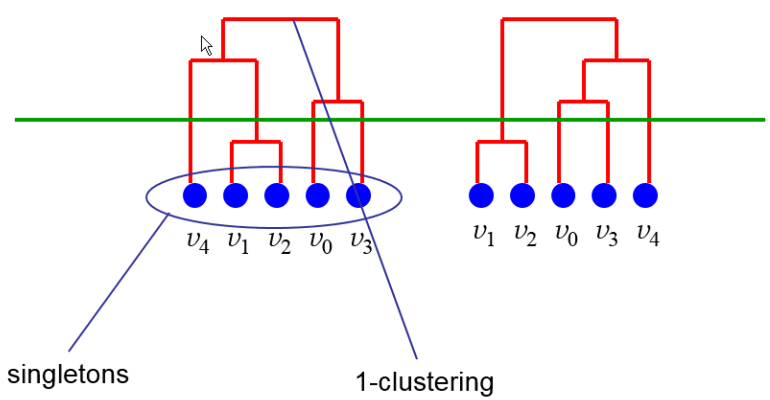
Graph Clustering Algorithms

- Advantage of Dendrograms: Can be "**cut**" at any desired number of clusters.



Graph Clustering Algorithms

- Advantage of Dendrograms: Can be "**cut**" at any desired number of clusters.



Graph Clustering Algorithms

- Linkage (Agglomeration): Iteratively coarsens a given clustering by merging two clusters until 1-clustering is reached ("bottom up")
- Splitting (Division): Iteratively refines a given clustering by splitting one cluster until slingleton clustering is reached ("top down").

- Linkage:

  P(V) := 2^V =: power-set

  - Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
  - Given: Either c_{global}: A(G) → $\mathbb{R}^+$ or c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for merging operations)
  - i→i+1: Either merge those two clusters where resulting clustering yields the minimum global cost
    or merge those two clusters with the minimum local merging cost

Graph Clustering Algorithms

- Linkage (Agglomeration): Iteratively coarsens a given clustering by merging two clusters until 1-clustering is reached ("bottom up")
- Splitting (Division): Iteratively refines a given clustering by splitting one cluster until slingleton clustering is reached ("top down").

- Linkage:

P(V) := 2^V =: power-set

  - Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
  - Given: Either c_{global}: A(G) → $\mathbb{R}^+$ or c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for merging operations)
  - i→i+1: Either merge those two clusters where resulting clustering yields the minimum global cost or merge those two clusters with the minimum local merging cost

Graph Clustering Algorithms

- Linkage (Agglomeration): Iteratively coarsens a given clustering by merging two clusters until 1-clustering is reached ("bottom up")
- Splitting (Division): Iteratively refines a given clustering by splitting one cluster until slingleton clustering is reached ("top down").

- Linkage:

P(V) := 2^V =: power-set

  - Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
  - Given: Either c_{global}: A(G) → $\mathbb{R}^+$ or c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for merging operations)
  - i→i+1: Either merge those two clusters where resulting clustering yields the minimum global cost or merge those two clusters with the minimum local merging cost

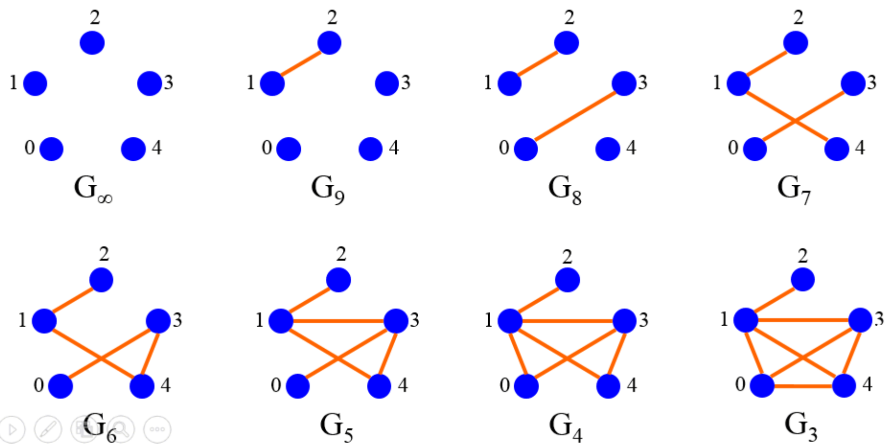
Linkage

- Example

weight matrix:

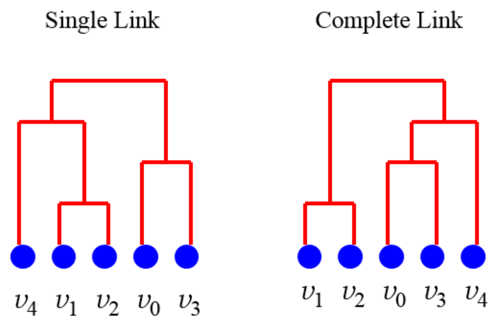
	v_0	v_1	v_2	v_3	v_4
v_0	∞	4	2	8	3
v_1	4	∞	9	5	7
v_2	2	9	∞	0	1
v_3	8	5	0	∞	6
v_4	3	7	1	6	∞

Threshold graphs:



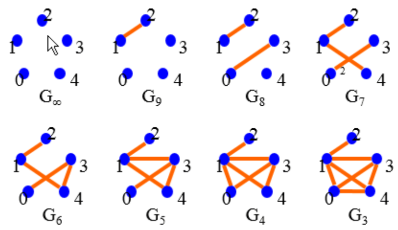
Linkage

Resulting dendrograms:



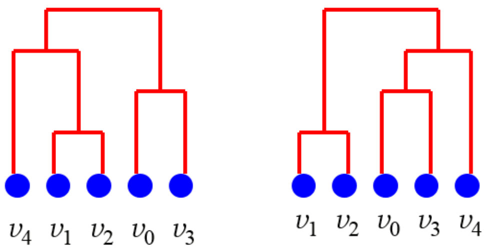
weight matrix:

	v_0	v_1	v_2	v_3	v_4
v_0	∞	4	2	8	3
v_1	4	∞	9	5	7
v_2	2	9	∞	0	1
v_3	8	5	0	∞	6
v_4	3	7	1	6	∞



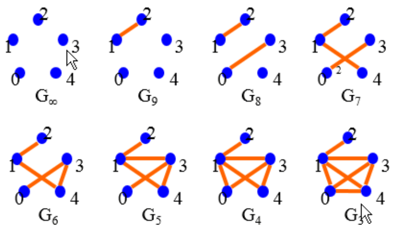
Resulting dendrograms:

Single Link   Complete Link



weight matrix:

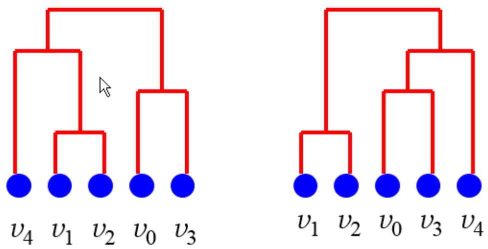
	v_0	v_1	v_2	v_3	v_4
v_0	∞	4	2	8	3
v_1	4	∞	9	5	7
v_2	2	9	∞	0	1
v_3	8	5	0	∞	6
v_4	3	7	1	6	∞



Linkage

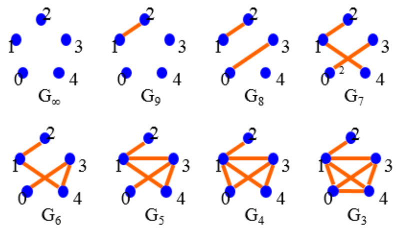
Resulting dendrograms:

Single Link   Complete Link



weight matrix:

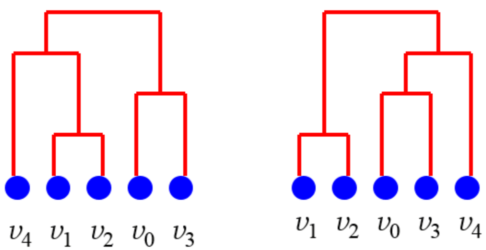
	v_0	v_1	v_2	v_3	v_4
v_0	∞	4	2	8	3
v_1	4	∞	9	5	7
v_2	2	9	∞	0	1
v_3	8	5	0	∞	6
v_4	3	7	1	6	∞



Linkage

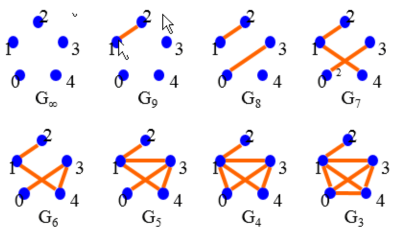
Resulting dendrograms:

Single Link   Complete Link



weight matrix:

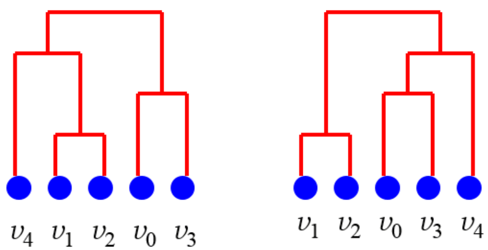
	v_0	v_1	v_2	v_3	v_4
v_0	∞	4	2	8	3
v_1	4	∞	9	5	7
v_2	2	9	∞	0	1
v_3	8	5	0	∞	6
v_4	3	7	1	6	∞



Linkage

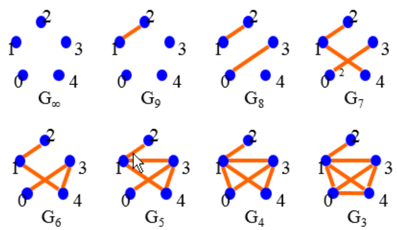
Resulting dendrograms:

Single Link   Complete Link



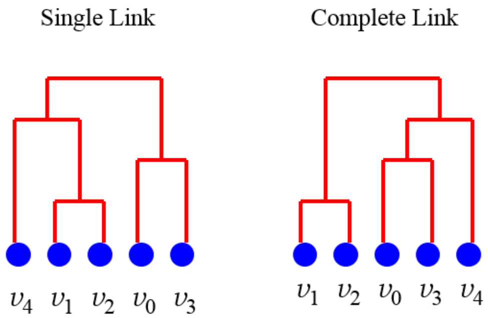
weight matrix:

	v_0	v_1	v_2	v_3	v_4
v_0	∞	4	2	8	3
v_1	4	∞	9	5	7
v_2	2	9	∞	0	1
v_3	8	5	0	∞	6
v_4	3	7	1	6	∞



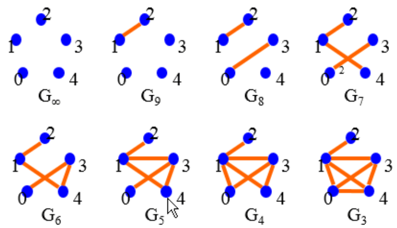
Linkage

Resulting dendrograms:



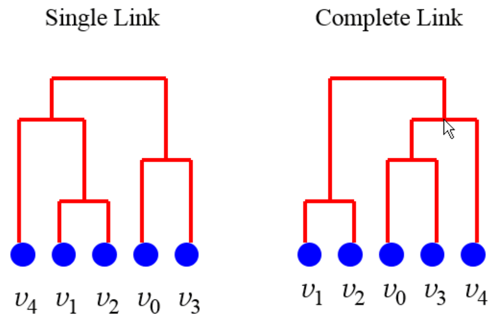
weight matrix:

$$\begin{array}{c} \begin{matrix} v_0 & v_1 & v_2 & v_3 & v_4 \end{matrix} \\ \begin{pmatrix} \infty & 4 & 2 & 8 & 3 \\ 4 & \infty & 9 & 5 & 7 \\ 2 & 9 & \infty & 0 & 1 \\ 8 & 5 & 0 & \infty & 6 \\ 3 & 7 & 1 & 6 & \infty \end{pmatrix} \end{array} \begin{matrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{matrix}$$



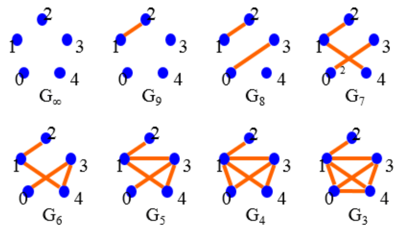
Linkage

Resulting dendrograms:



weight matrix:

$$\begin{array}{c} \begin{matrix} v_0 & v_1 & v_2 & v_3 & v_4 \end{matrix} \\ \begin{pmatrix} \infty & 4 & 2 & 8 & 3 \\ 4 & \infty & 9 & 5 & 7 \\ 2 & 9 & \infty & 0 & 1 \\ 8 & 5 & 0 & \infty & 6 \\ 3 & 7 & 1 & 6 & \infty \end{pmatrix} \end{array} \begin{matrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{matrix}$$



Linkage

- Instead of the weight matrix

weight matrix:

$$\begin{array}{c} \begin{matrix} v_0 & v_1 & v_2 & v_3 & v_4 \end{matrix} \\ \begin{pmatrix} \infty & 4 & 2 & 8 & 3 \\ 4 & \infty & 9 & 5 & 7 \\ 2 & 9 & \infty & 0 & 1 \\ 8 & 5 & 0 & \infty & 6 \\ 3 & 7 & 1 & 6 & \infty \end{pmatrix} \end{array} \begin{matrix} v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{matrix}$$

we may as well use an "equivalent" distance matrix $d(i,j) = d_{ij}$, e.g.

$$\begin{pmatrix} 0 & 6 & 8 & 2 & 7 \\ 6 & 0 & 1 & 5 & 3 \\ 8 & 1 & 0 & 10 & 9 \\ 2 & 5 & 10 & 0 & 4 \\ 7 & 3 & 9 & 4 & 0 \end{pmatrix}$$

and would have to modify the threshold graph based algorithm(replace < with > and ≥ with ≤ and ∞ with 0) (or set weight = 1/distance). Then this algorithm implements precisely the aforementioned cost function

$$c_{local}(C_i, C_j) = \begin{matrix} \max \\ \min \end{matrix} \{ d(u,v) \mid u \in C_i, v \in C_j \}$$

Complete Linkage   Single Linkage

Splitting

- Given: $G=(V,E,w)$; initial clustering $\mathbf{C}_1$;
- Given: Either c_{global}: $A(G) \rightarrow \mathbb{R}^+$

  c_{local}: $P(V) \times P(V) \rightarrow \mathbb{R}^+$ (for splitting operations) or

  c_{global}: $A(G) \rightarrow \mathbb{R}^+$ and cut function $S: P(V) \rightarrow P(V)$ or

  c_{local}: $P(V) \times P(V) \rightarrow \mathbb{R}^+$ and cut function $S: P(V) \rightarrow P(V)$
- $i \rightarrow i+1$: Split that cluster where the resulting clustering yields the minimum global cost or

  split the cluster with the minimum local splitting cost or

  split that cluster (according to cut S) where the resulting clustering yields the minimum global cost or

  split the cluster (according to cut S) with the minimum local splitting cost

Splitting

- Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
- Given: Either c_{global}: A(G) → $\mathbb{R}^+$

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for splitting operations) or

  c_{global}: A(G) → $\mathbb{R}^+$ and cut function S: P(V) → P(V) or

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ and cut function S: P(V) → P(V)

- i→i+1: Split that cluster where the resulting clustering yields the minimum global cost or

  split the cluster with the minimum local splitting cost or

  split that cluster (according to cut S) where the resulting clustering yields the minimum global cost or

  split the cluster (according to cut S) with the minimum local splitting cost

Splitting

- Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
- Given: Either c_{global}: A(G) → $\mathbb{R}^+$

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for splitting operations) or

  c_{global}: A(G) → $\mathbb{R}^+$ and cut function S: P(V) → P(V) or

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ and cut function S: P(V) → P(V)

- i→i+1: Split that cluster where the resulting clustering yields the minimum global cost or

  split the cluster with the minimum local splitting cost or

  split that cluster (according to cut S) where the resulting clustering yields the minimum global cost or

  split the cluster (according to cut S) with the minimum local splitting cost

Splitting

C_x: one of the clusters

one “half” of the split of C_x

strictly global

strictly local

semi global

semi local

one “half” of the cut (split) of some C_x, defining the cut

- Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
- Given: Either c_{global}: A(G) → $\mathbb{R}^+$

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for splitting operations) or

  c_{global}: A(G) → $\mathbb{R}^+$ and cut function S: P(V) → P(V) or

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ and cut function S: P(V) → P(V)

- i→i+1: Split that cluster where the resulting clustering yields the minimum global cost or

  split the cluster with the minimum local splitting cost or

  split that cluster (according to cut S) where the resulting clustering yields the minimum global cost or

  split the cluster (according to cut S) with the minimum local splitting cost

Splitting

C_x: one of the clusters

one “half” of the split of C_x

strictly global

strictly local

semi global

semi local

one “half” of the cut (split) of some C_x, defining the cut

- Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
- Given: Either c_{global}: A(G) → $\mathbb{R}^+$

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for splitting operations) or

  c_{global}: A(G) → $\mathbb{R}^+$ and cut function S: P(V) → P(V) or

  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ and cut function S: P(V) → P(V)

- i→i+1: Split that cluster where the resulting clustering yields the minimum global cost or

  split the cluster with the minimum local splitting cost or

  split that cluster (according to cut S) where the resulting clustering yields the minimum global cost or

  split the cluster (according to cut S) with the minimum local splitting cost

Splitting

- Cut function avoids having to test all possible splits
  - Variants of Cut functions:

$$S(V) := \underset{\emptyset \neq V' \subset V}{\operatorname{argmin}} \; \omega(E(V', V \setminus V'))$$

$$S_{\text{ratio}}(V) := \underset{\emptyset \neq V' \subset V}{\operatorname{argmin}} \; \frac{\omega(E(V', V \setminus V'))}{|V'| \cdot (|V| - |V'|)}$$ [1]

$$S_{\text{balanced}}(V) := \underset{\emptyset \neq V' \subset V}{\operatorname{argmin}} \; \frac{\omega(E(V', V \setminus V'))}{\min(|V'|, (|V| - |V'|))}$$

$$S_{\text{conductance}}(V) := \underset{\emptyset \neq V' \subset V}{\operatorname{argmax}} \; \delta(V') = \underset{V' \subset V}{\operatorname{argmin}} \; \varphi(V', V \setminus V')$$

inter cluster conductance (slide 14):

$$g(\mathbf{C} = \{V', V \setminus V'\}) = \delta(V') = \begin{cases} 1 & \text{if } \mathbf{C} = \{V, \{\}\} \\ 1 - \varphi(V', V \setminus V') & \text{otherwise} \end{cases}$$

Splitting

- Given: G=(V,E,w); initial clustering $\mathbf{C}_1$;
- Given: Either c_{global}: A(G) → $\mathbb{R}^+$
  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ (for splitting operations) or
  c_{global}: A(G) → $\mathbb{R}^+$ and cut function S: P(V) → P(V) or
  c_{local}: P(V) x P(V) → $\mathbb{R}^+$ and cut function S: P(V) → P(V)
- i→i+1: Split that cluster where the resulting clustering yields the minimum global cost or
  split the cluster with the minimum local splitting cost or
  split that cluster (according to cut S) where the resulting clustering yields the minimum global cost or
  split the cluster (according to cut S) with the minimum local splitting cost

Shifting

shifting minimization:

```
let L0 be a feasible solution;
i ←0;
while({L | L∈N(Li)} ≠ Ø){
      choose Li+1 from N(Li)according to ☺;
      i ← i+1;
}
```

- Choosing schema ☺ can be either based on potential function φ, on random selection or based on genetic algorithms with fitness function etc.
- Potential function φ: A(G) x A(G) → $\mathbb{R}$ based: Chose a new clustering $\mathbf{C}_{i+1}$ so that $\varphi(\mathbf{C}_i, \mathbf{C}_{i+1}) > 0$

Newman Girvan Method: Centrality-based Splitting + Modularity

Last example of this part: bringing it all together (see [3]):

- Observations → critique on agglomerative methods: fail to cluster peripheral nodes correctly [3] → Newman Girvan method: Divisive hierarchical clustering (splitting) + Modularity:

1. Calculate edge betweenness for all edges
2. Remove edge with highest edge betweenness
3. goto 1.

→ dendrogram

- Use Modularity as intra cluster coherence (f) cluster validity measure (g=0) to optimally cut dendrogram:

$$Q = \sum_i (e_{ii} - a_i^2) = \operatorname{Tr} \mathbf{e} - \| \mathbf{e}^2 \|$$

Newman Girvan Method: Centrality-based Splitting + Modularity

Last example of this part: bringing it all together (see [3]):

- Observations → critique on agglomerative methods: fail to cluster peripheral nodes correctly [3] → Newman Girvan method: Divisive hierarchical clustering (splitting) + Modularity:

1. Calculate edge betweenness for all edges
2. Remove edge with highest edge betweenness
3. goto 1.

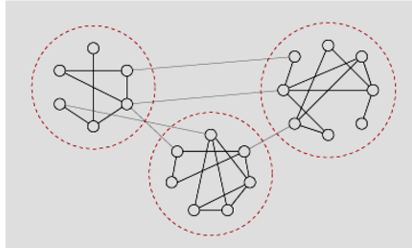
→ dendrogram

- Use Modularity as intra cluster coherence (f) cluster validity measure (g=0) to optimally cut dendrogram:

$$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:

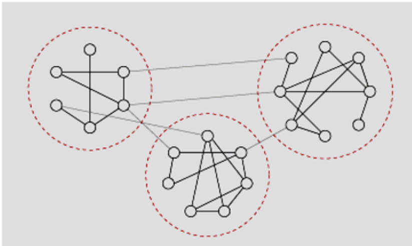


[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{rnd} = a_i a_j \rightarrow e_{ii}^{rnd} = a_i^2$
- f: Compare (→difference) real with rnd → $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:

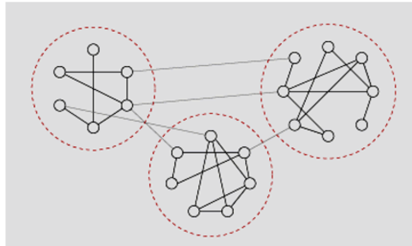


[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{rnd} = a_i a_j \rightarrow e_{ii}^{rnd} = a_i^2$
- f: Compare (→difference) real with rnd → $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:

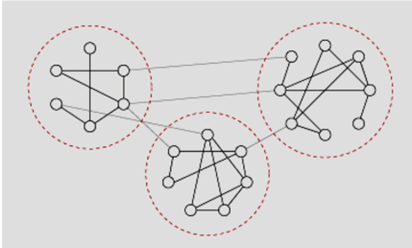


[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{rnd} = a_i a_j \rightarrow e_{ii}^{rnd} = a_i^2$
- f: Compare (→difference) real with rnd → $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:



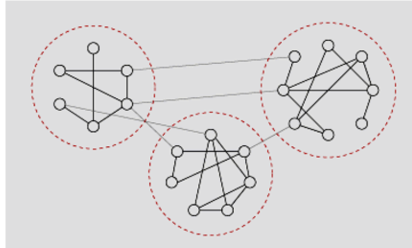
[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{\text{rnd}} = a_i a_j \rightarrow e_{ii}^{\text{rnd}} = a_i^2$
- f: Compare (→difference)

real with rnd → $$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:



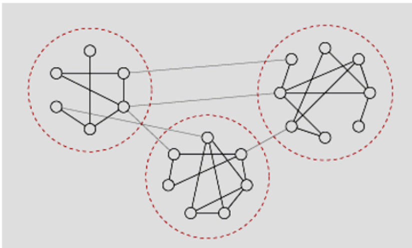
[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{\text{rnd}} = a_i a_j \rightarrow e_{ii}^{\text{rnd}} = a_i^2$
- f: Compare (→difference)

real with rnd → $$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:



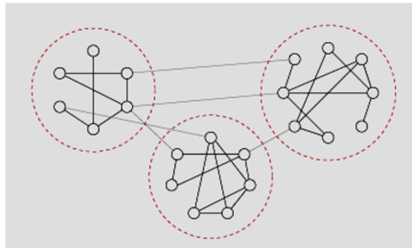
[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{\text{rnd}} = a_i a_j \rightarrow e_{ii}^{\text{rnd}} = a_i^2$
- f: Compare (→difference)

real with rnd → $$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:



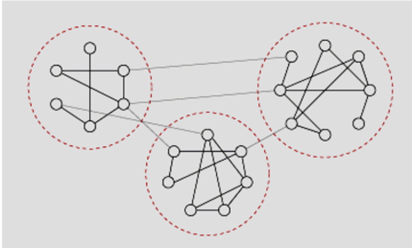
[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i' e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{\text{rnd}} = a_i a_j \rightarrow e_{ii}^{\text{rnd}} = a_i^2$
- f: Compare (→difference)

real with rnd → $$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:

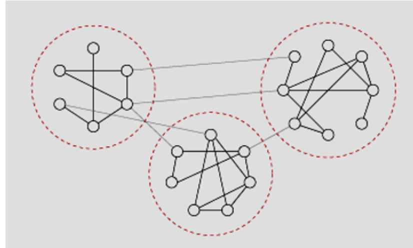


[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{rnd} = a_i a_j$ → $e_{ii}^{rnd} = a_i^2$
- f: Compare (→difference)
  real with rnd → $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:

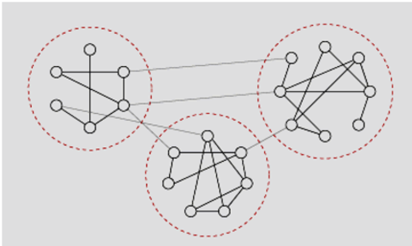


[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{rnd} = a_i a_j$ → $e_{ii}^{rnd} = a_i^2$
- f: Compare (→difference)
  real with rnd → $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:



[3]

- k clusters → k x k symmetric matrix **e**: e_{ij} = |E(C_i,C_j)| / |E| : fraction of edges between communities
- $\mathrm{Tr}\,\mathbf{e} = \sum_i e_{ii}$: fraction of edges within communities
- $a_i = \sum_j e_{ij}$: fraction of edges that connect to cluster C_i
- Random network (keep a_i fixed): $e_{ij}^{rnd} = a_i a_j$ → $e_{ii}^{rnd} = a_i^2$
- f: Compare (→difference)
  real with rnd → $Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$

Newman Girvan Method: Centrality-based Splitting + Modularity

How to compute Modularity with a given (weighted) adjacency matrix?

- Real graph: Fraction of edges within clusters:

  |E(C_i)| / |E| = $\frac{\sum_{ij} A_{ij}\delta(c_i,c_j)}{\sum_{ij} A_{ij}} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,c_j)$ $m = \frac{1}{2}\sum_{ij} A_{ij}$,

- Random graph (keep degrees k_i of vertices fixed): prob of edge between vertices i and j is $k_i k_j/(2m)^2$

  $k_v = \sum_w A_{vw}$.

- → modularity: $Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i,c_j)$

which is equal (as before) $= \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \| \mathbf{e}^2 \|$

- → with the new formulation we can compute modularity for weighted graphs

Newman Girvan Method: Centrality-based Splitting + Modularity

How to compute Modularity with a given (weighted) adjacency matrix?

- Real graph: Fraction of edges within clusters:

  |E(C_i)| / |E| = $\frac{\sum_{ij} A_{ij}\delta(c_i,c_j)}{\sum_{ij} A_{ij}} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,c_j)$ $\quad m = \frac{1}{2}\sum_{ij} A_{ij},$

- Random graph (keep degrees k_i of vertices fixed): prob of edge between vertices i and j is $k_i k_j/(2m)^2$

  $k_v = \sum_w A_{vw}.$

- → modularity: $Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i,c_j)$

which is equal (as before) $= \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$

- → with the new formulation we can compute modularity for weighted graphs

Newman Girvan Method: Centrality-based Splitting + Modularity

How to compute Modularity with a given (weighted) adjacency matrix?

- Real graph: Fraction of edges within clusters:

  |E(C_i)| / |E| = $\frac{\sum_{ij} A_{ij}\delta(c_i,c_j)}{\sum_{ij} A_{ij}} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,c_j)$ $\quad m = \frac{1}{2}\sum_{ij} A_{ij},$

- Random graph (keep degrees k_i of vertices fixed): prob of edge between vertices i and j is $k_i k_j/(2m)^2$

  $k_v = \sum_w A_{vw}.$

- → modularity: $Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i,c_j)$

which is equal (as before) $= \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$

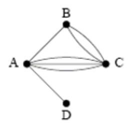
- → with the new formulation we can compute modularity for weighted graphs

Newman Girvan Method: Centrality-based Splitting + Modularity

Modularity:

- In [1]: different notion (not keeping a_i fixed): $\sum_{i=1}^{k}\left(|E(C_i)| - m\frac{|C_i|\cdot(|C_i|-1)}{n\cdot(n-1)}\right)$

- In [4]: Newman's version for weighted graphs:
  idea: use multiple edges to model weights



$$\equiv \begin{array}{c} \begin{matrix} A & B & C & D \end{matrix} \\ \begin{pmatrix} 0 & 1 & 3 & 1 \\ 1 & 0 & 2 & 0 \\ 3 & 2 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

Newman Girvan Method: Centrality-based Splitting + Modularity

How to compute Modularity with a given (weighted) adjacency matrix?

- Real graph: Fraction of edges within clusters:

  |E(C_i)| / |E| = $\frac{\sum_{ij} A_{ij}\delta(c_i,c_j)}{\sum_{ij} A_{ij}} = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,c_j)$ $\quad m = \frac{1}{2}\sum_{ij} A_{ij},$

- Random graph (keep degrees k_i of vertices fixed): prob of edge between vertices i and j is $k_i k_j/(2m)^2$

  $k_v = \sum_w A_{vw}.$

- → modularity: $Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i,c_j)$

which is equal (as before) $= \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\,\mathbf{e} - \|\mathbf{e}^2\|$

- → with the new formulation we can compute modularity for weighted graphs