Graph Theory

Graph features

Sergio Peignier sergio.peignier@insa-lyon.fr

Associate Professor
INSA Lyon
Biosciences department

Table of contents

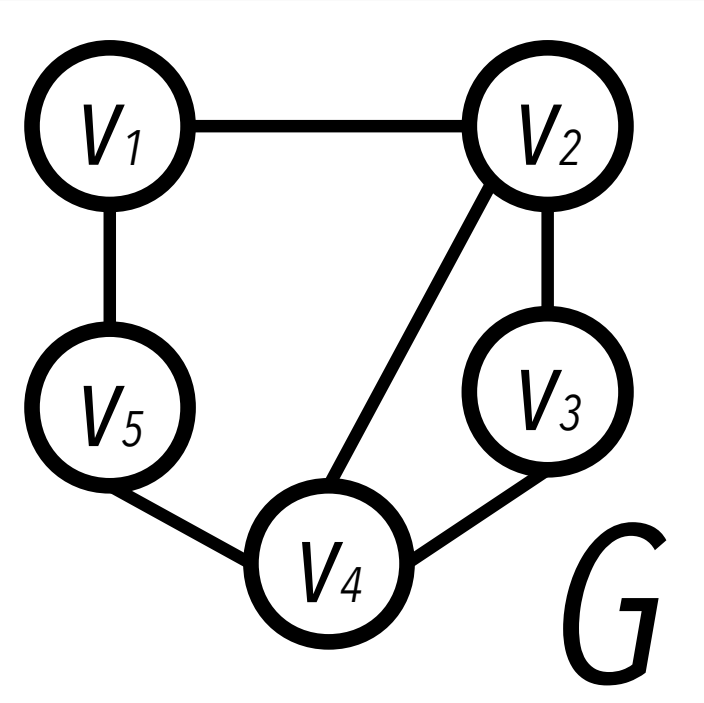
Local features

Node Degree (undirected)

$G = \langle V, E \rangle$, node $v_i \in V$, adjacency matrix A

- $deg(v_i)$: **Nb. of edges** of v_i.
- $deg(v_i) = \sum_j A_{i,j}$

Example: $deg(v_1) = 2$, $deg(v_2) = 3$



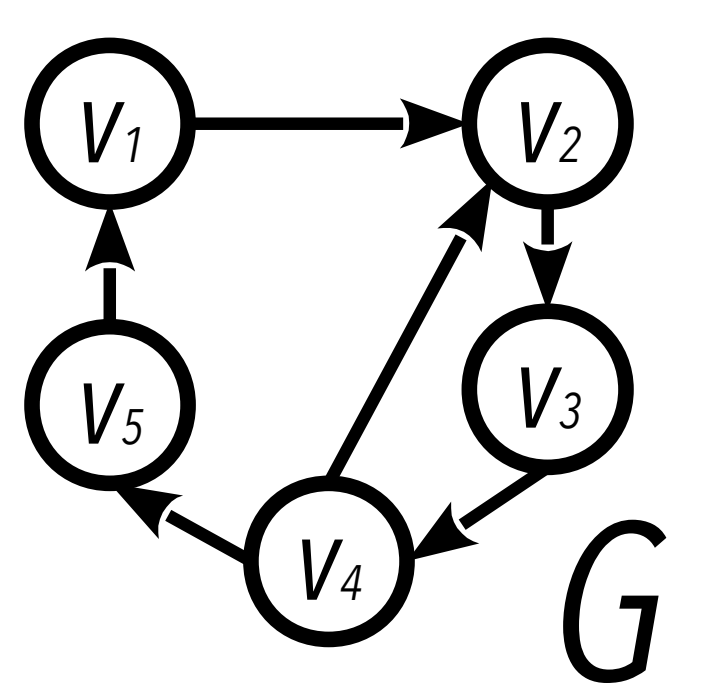
$$\begin{bmatrix} & v_1 & v_2 & v_3 & v_4 & v_5 \\ v_1 & 0 & 1 & 0 & 0 & 1 \\ v_2 & 1 & 0 & 1 & 1 & 0 \\ v_3 & 0 & 1 & 0 & 1 & 0 \\ v_4 & 0 & 1 & 1 & 0 & 1 \\ v_5 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Node Degree (directed)

$G = \langle V, E \rangle$, node $v_i \in V$, adjacency matrix A

- $deg_{out}(v_i)$: **Nb. of outgoing edges**: $deg_{out}(v_i) = \sum_j A_{i,j}$
- $deg_{in}(v_i)$: **Nb. of incoming edges**: $deg_{in}(v_i) = \sum_j A_{j,i}$

Example: $deg_{out}(v_4) = 2$, $deg_{in}(v_4) = 1$, $deg_{in}(v_2) = 2$, $deg_{out}(v_2) = 1$



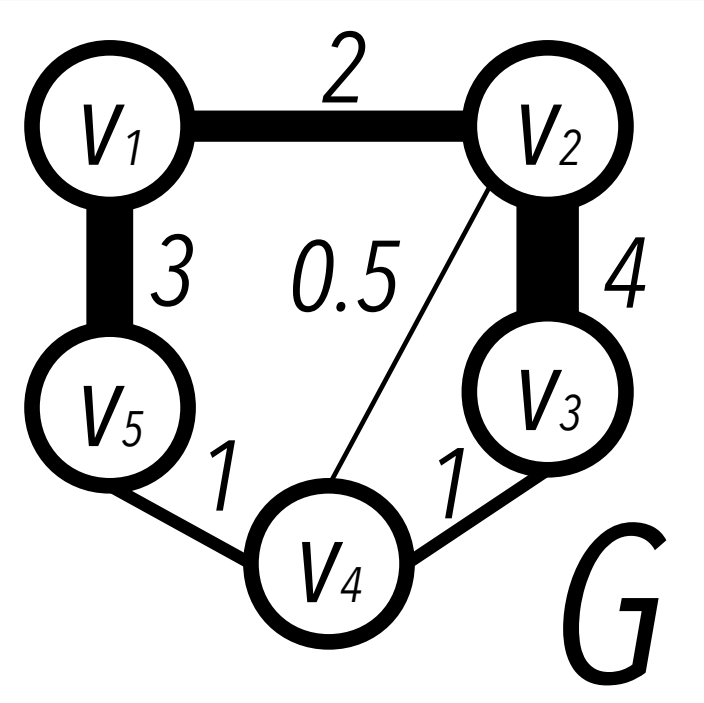
$$\begin{bmatrix} & v_1 & v_2 & v_3 & v_4 & v_5 \\ v_1 & 0 & 1 & 0 & 0 & 0 \\ v_2 & 0 & 0 & 1 & 0 & 0 \\ v_3 & 0 & 0 & 0 & 1 & 0 \\ v_4 & 0 & 1 & 0 & 0 & 1 \\ v_5 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Node strength

$G = \langle V, E, w \rangle$, node $v_i \in V$, adjacency matrix A

- $deg(v)$: **weighted sum of the edges** of v_i.
- $deg(v_i) = \sum_j A_{i,j}$

Example: $deg(v_1) = 5$, $deg(v_2) = 6.5$



$$\begin{bmatrix} & v_1 & v_2 & v_3 & v_4 & v_5 \\ v_1 & 0 & 2 & 0 & 0 & 3 \\ v_2 & 2 & 0 & 4 & 0.5 & 0 \\ v_3 & 0 & 4 & 0 & 1 & 0 \\ v_4 & 0 & 0.5 & 1 & 0 & 1 \\ v_5 & 3 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Clustering coefficient Watts and Strogatz 2002

$G = \langle V, E \rangle$, node $v \in V$

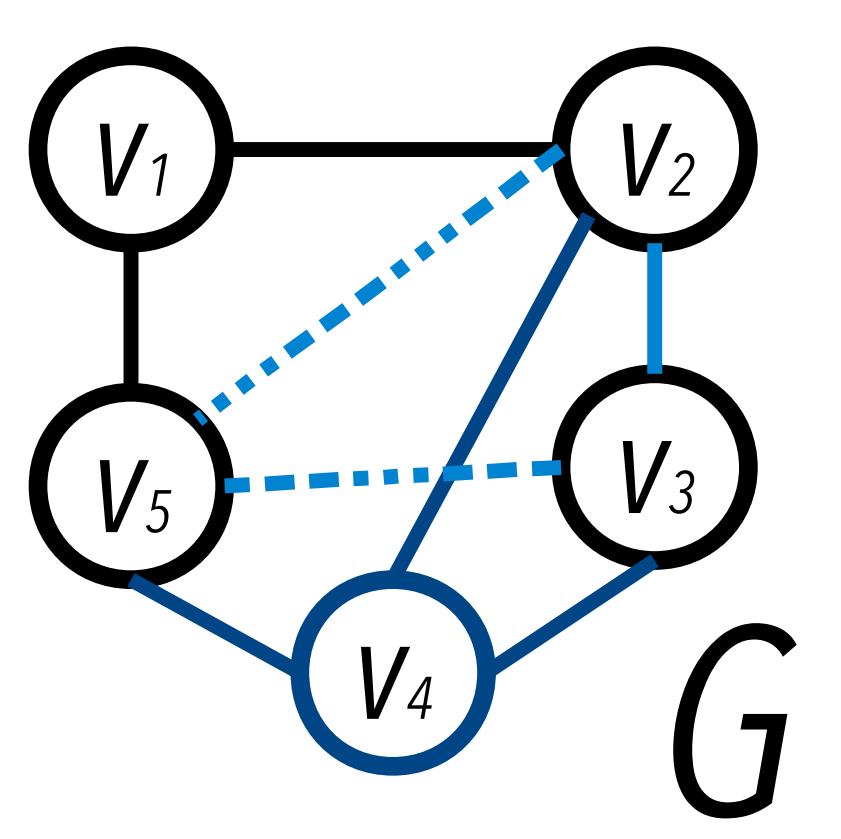
- $C(v)$: Measure of **inter-connectivity** around v_i.
- Portion of **neighbours** of v_i **connected** to each other.

$$C(v_i) = \frac{2e_{v_i}}{k_{v_i}(k_{v_i} - 1)}$$

- e_{v_i}: Nb. of **edges between the neighbours** of v_i
- $k_{v_i} = deg(v_i)$
- Max. nb. of **triangles** formed by v_i and 2 neighbors: $\frac{k_{v_i}(k_{v_i}-1)}{2}$

Clustering coefficient Watts and Strogatz 2002

Example: $C_{v_4} = \frac{2e_{v_i}}{k_{v_i}(k_{v_i}-1)} = \frac{2\times 1}{3\times 2} = \frac{1}{3}$



Overlap

$G = \langle V, E \rangle$, node $v \in V$

- $O(v_i, v_j)$:
  Fraction of **common neighbours** of a **connected pair**.

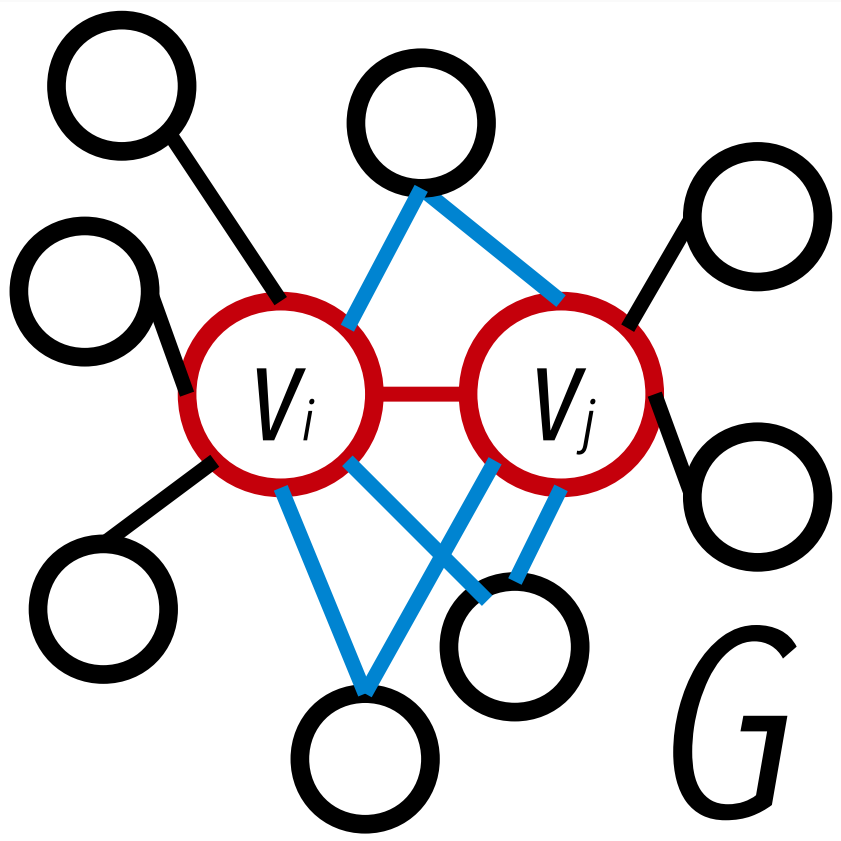
$$O(\langle v_i, v_j \rangle) = \frac{n_{v_i,v_j}}{(k_{v_i} - 1) + (k_{v_j} - 1) - n_{v_i,v_j}}$$

- n_{v_i,v_j}: Nb. **common neighbours** between v_i and v_j
- $k_{v_i} = deg(v_i)$
- $(k_{v_i} - 1) + (k_{v_j} - 1) - n_{v_i,v_j}$:
  Max. nb. **triangles** formed by v_i, v_j and 1 neighbor of v_i or v_j.

Overlap

Example:

$$O(v_i, v_j) = \frac{n_{v_i,v_j}}{(k_{v_i}-1)+(k_{v_j}-1)-n_{v_i,v_j}} = \frac{3}{6+5-3} = \frac{3}{8}$$

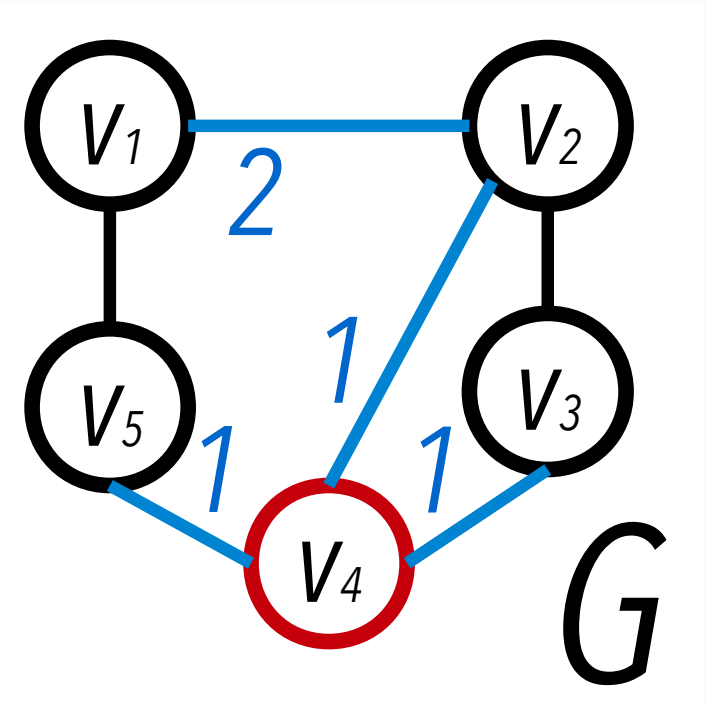


Closeness Centrality

Inverse of **avg. distance** between v_i and all other ones.

$$d_{avg}(v_i) = \frac{\sum_{i \neq j} dist(v_i, v_j)}{|V| - 1}, \quad g_{v_i} = \frac{1}{d_{avg}(v_i)}$$

Not defined for unconnected networks.



Example: $g_{v_4} = \frac{4}{5}$

Betweeness Centrality

Fraction of shortest paths that pass through v.

$$b_v = \sum_{v_s \neq v \neq v_t} \frac{\sigma_{v_s,v_t}(v)}{\sigma_{v_s,v_t}}$$

- $\sigma_{v_s,v_t}(v)$: **Nb. of shortest paths** between v_s and v_t, **passing through** v
- σ_{v_s,v_t}: **Nb. of shortest paths** between v_s and v_t.

$\exists$ Analogous definition for edges

k-shell Index

Identify **central** and the **dense cores** of graphs. **k-core decomposition:**

1. $G = \langle V, E \rangle, \quad k = 1$
2. Remove nodes with degree $\leq k$, and all those that get degree ≤ 1 because of the removal, recursively.
3. $k \leftarrow k + 1$
4. Go to step 2 and 4 until there are no more nodes

- **k-shell** : Nodes removed in the k turn.
- **k-core**: Remaining nodes.

Write the previous instructions in clean pseudo-code

Global features

Density

Ratio of the **nb. of edges** and **max. nb. of possible edges**

$$\rho = \frac{2|E|}{|V|(|V|-1)}$$

- Fully connected graph: $\rho = 1$
- Real world networks: sparse

Diameter

Maximum distance between 2 nodes in G

$$d_{max}(G) = \max_{v_i, v_j} dist(v_i, v_j)$$

Average Distance / Average Path Length

Connected directed graph $G = \langle V, E \rangle$

$$d_{avg} = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} dist(v_i, v_j)$$

Connected undirected graph $G = \langle V, E \rangle$

$$d_{avg} = \frac{2}{|V|(|V|-1)} \sum_{i,j>i} dist(v_i, v_j)$$

Clustering Coefficient

Average Clustering Coefficient

$$C(G) = \frac{\sum_i C_{v_i}}{|V|}$$

Rich-Club Coefficient

How **well connected** are the **well connected nodes between themselves**.

$$\phi(k) = \frac{2|E_{\geq k}|}{|V_{\geq k}| \times (|V_{\geq k}| - 1)}$$

- $V_{\geq k}$: Nodes $v \in V$ s.t. $deg(v) \geq k$
- $E_{\geq k}$: Edges between nodes in $V_{\geq k}$

Degree Distribution

Fraction of nodes in G **with degree** k.

Let $V_k = \{v \in V | deg(v) = k\}$

$$P(k) = \frac{|V_k|}{|V|}$$

Pearson Degree-Degree Correlation

$$r = \frac{1}{\sigma_q^2} \sum_{k,l} k \times l \times (e_{kl} - q_k q_l)$$

Let $\langle v_i, v_j \rangle \in E$ a random edge.

e_{kl}: Prob. that the ends of $\langle v_i, v_j \rangle$ have degrees k and l.

q_k: Prob. that one end of $\langle v_i, v_j \rangle$ has degree k

σ_q^2: VAR of the distribution of q_k.

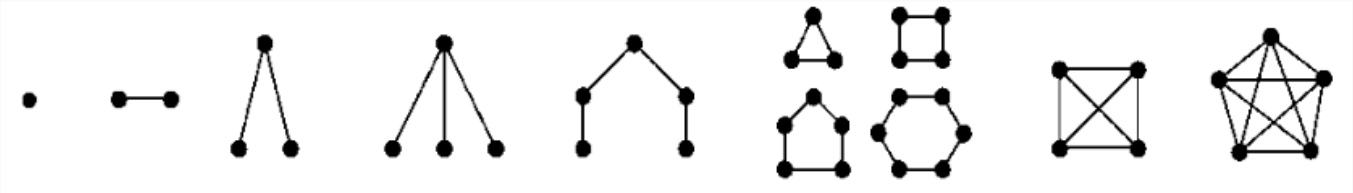
Null Hypothesis: No degree correlation $e_{kl} = q_k q_l$

- $r < 0, \quad \Longrightarrow$ G is Disassortative
- $r = 0, \quad \Longrightarrow$ G is Neutral
- $r > 0, \quad \Longrightarrow$ G is Assortative

Motifs

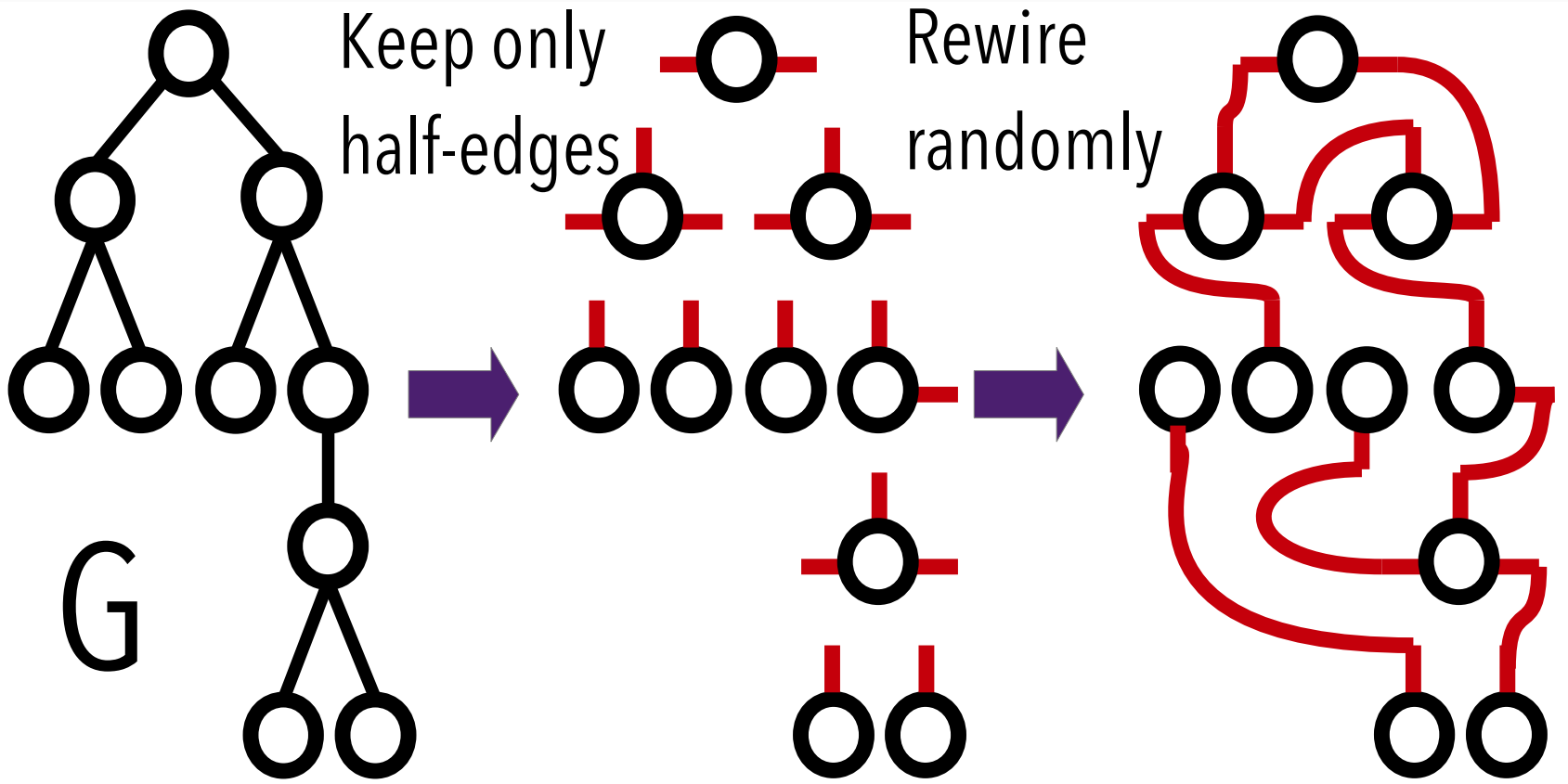
Motifs

Sub-graph having a **significantly higher frequency** w.r.t. **configuration model (null hypothesis)**[1].



[1] **Configuration model**: Ensemble of random networks preserving the degree distribution of the original graph.

Motifs | Configuration Model



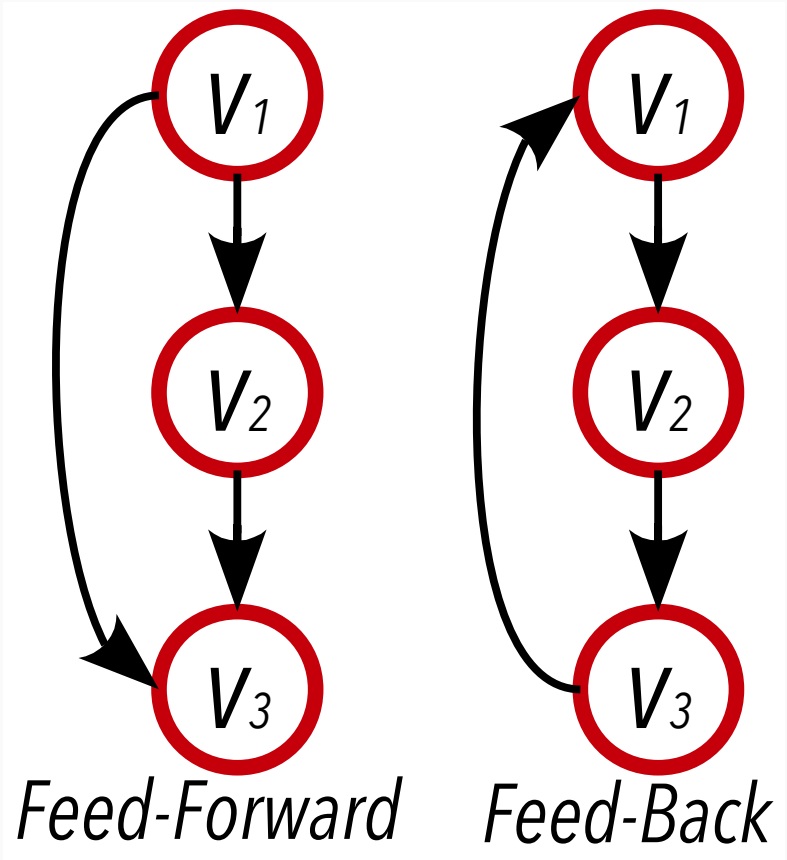
Motifs | Algorithm

- Count sub-graphs n_m in the original graph G.
- Repeat
  - Rewire randomly G: configuration model G_{rand}
  - Count sub-graphs n_m^{rand} in G_{rand}
- Compute $Z - score$:

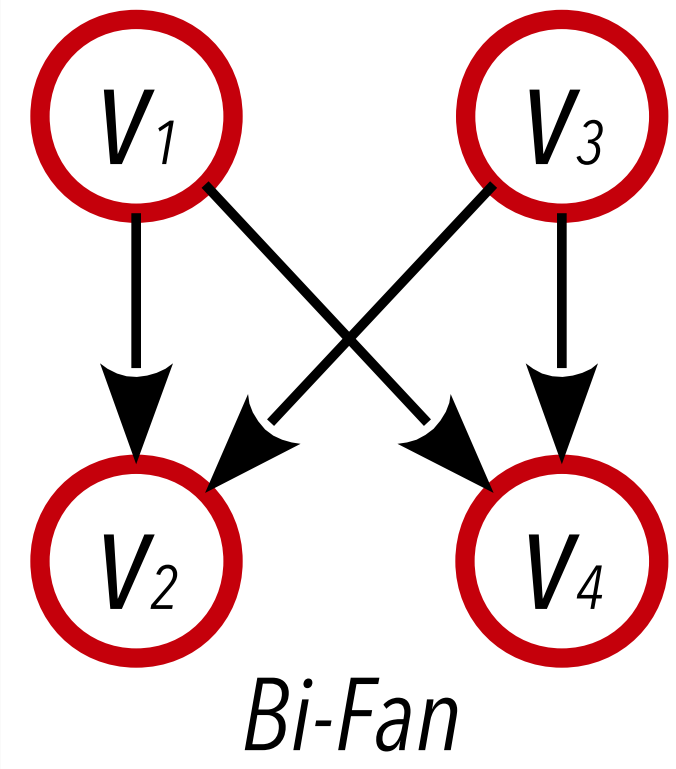
$$Z_m = \frac{n_m - E[n_m^{rand}]}{\sigma_{n_m}^{rand}}$$

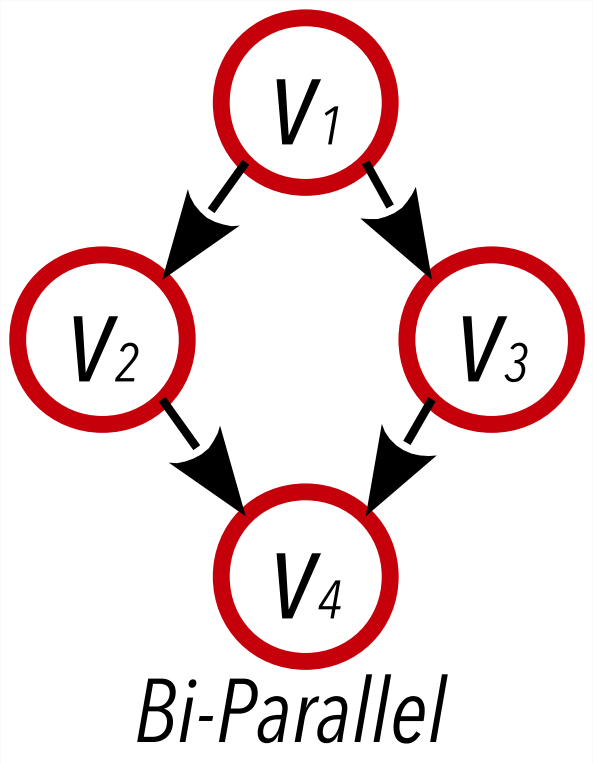
Motif search algorithms: mfinder, PFP, ESU, ...

Feed-Forward/Feed-Back Loop



Bi-Fan/Bi-Parallel





Communities

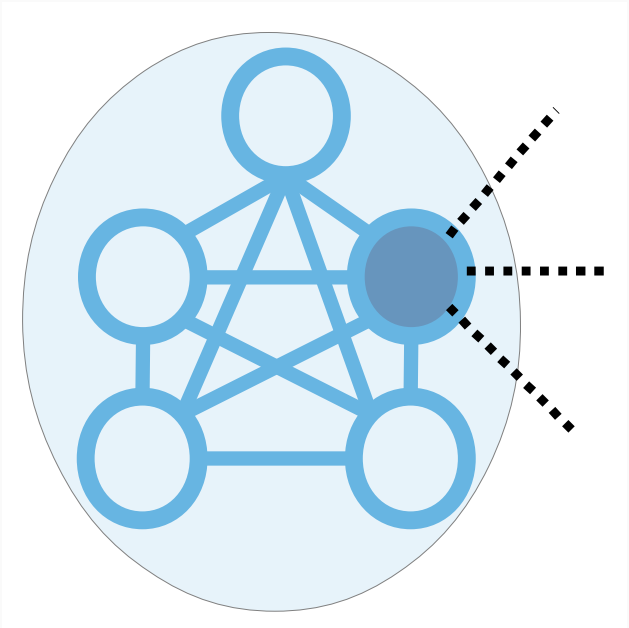
Basic Definitions

Community V^c in a graph $G = \langle V, E \rangle$ is:

$V^c \subseteq V$ s.t. nodes in V^c are **densely connected internally**.

$G^{int} = \langle V^c, E^c \rangle$: **induced subgraph** of G

- **Internal degree**: k_v^{int}:
  Degree of v in G^{int}
- **Community internal degree**:
  $k_{V^c}^{int} = \sum_{v \in V^c} k_v^{int}$
- **External degree**:
  $k_v^{ext} = deg(v) - k_v^{int}$
- **Community external degree**:
  $k_{V^c}^{ext} = \sum_{v \in C} k_v^{ext}$.



e.g. $k_v^{int} = 4$, $k_v^{ext} = 3$,
$k_{V^c}^{int} = 20$, $k_{V^c}^{ext} = 3$

Community Detection Problem

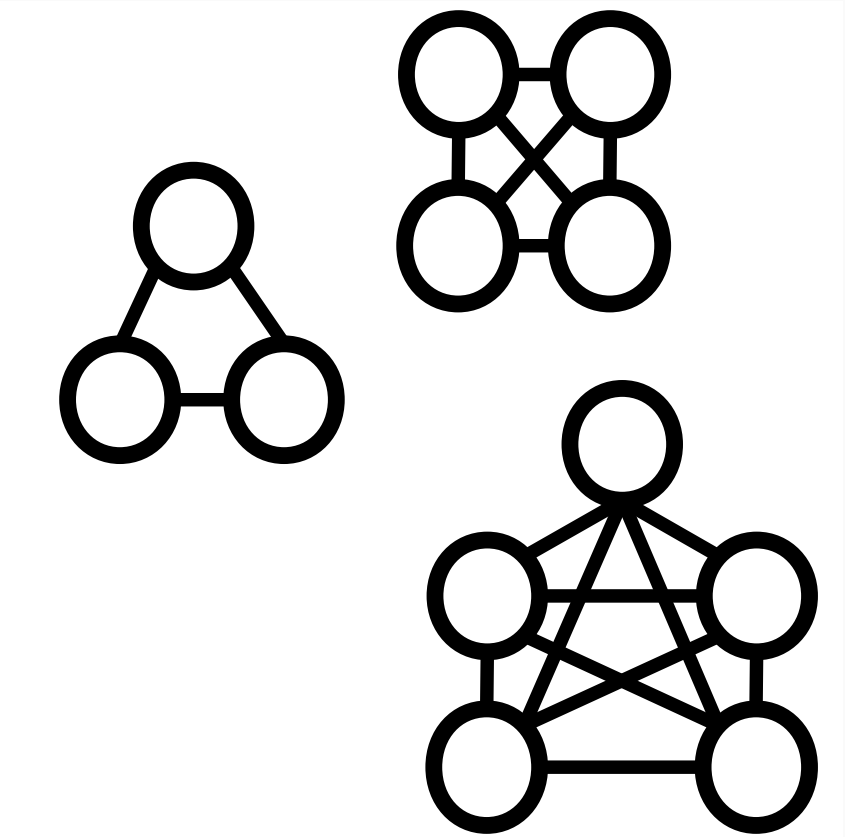
- Communities **inferred** only from the **graph structure.**
- Communities definition (Ill-posed problem):
  Each **community detection algorithm** has its own **implicit definition**.
- **Choice depends** on the specific system/application

Cliques

k-Clique: Subgraph $G^c = \langle V^c, E^c \rangle$ of G s.t.:

$$|V^c| = k \quad \text{and} \quad |E^c| = \frac{k(k-1)}{2}$$

Problem: Too strict condition



Strong/Weak Communities

Strong community:

Community V^c s.t. $\forall v \in V^c \quad k_v^{int} > k_v^{ext}$

Weak communities:

Community V^c s.t. $k_{V^c}^{int} > k_{V^c}^{ext}$

Problems:

- Too strict condition
- Unrealistic in practical cases

k-plex

Maximal subgraph $G^c = \langle V^c, E^c \rangle$ s.t.

$$\forall v \in V^c, \quad k_v^{int} \geq (|E^c| - 1) - k, \quad \text{with} \quad k \geq 0$$

Each node v is **adjacent to all other** nodes of G^c **except at most** k of them.

k-core

Maximal subgraph $G^c = \langle V^c, E^c \rangle$ s.t.

$$\forall v \in V^c, \quad k_v^{int} \geq k, \quad \text{with} \quad k \geq 0$$

Each node v is **adjacent to at least** k other nodes in V^c

p-quasi complete subgraph

Maximal subgraph $G^c = \langle V^c, E^c \rangle$ s.t.

$$\forall v \in V^c, \quad k_v^{int} \geq p(|E^c| - 1), \quad \text{with} \quad p \in [0, 1]$$

Each node v is **adjacent** to **at least** $p(|E^c| - 1)$ other nodes.

Similarity Based Community Detection

Community: **subgraph** of **"similar" nodes**

Let $G = \langle V, E \rangle$ be a graph with adjacency matrix A

Some Similarity measures:

- **Structural equivalence dissimilarity:**

$$dist(v_i, v_j) = \sqrt{\sum_{k \neq i,j} (A_{i,k} - A_{j,k})^2}$$

- **Neighborhoods overlap:**

$$w_{v_i,v_j} = \frac{Neighbors(v_j) \cap Neighbors(v_i)}{Neighbors(v_j) \cup Neighbors(v_i)}$$

- **Pearson correlation coefficient**:

$$Corr_{i,j} = \frac{\sum_k (A_{i,k} - \mu_i)(A_{j,k} - \mu_j)}{\sigma_i \sigma_j}$$

Partitioning | Girvan-Newman method

Girvan-Newman Algorithm

- **Repeat:**
  - Compute the **betweenness** $\forall e \in E$
  - **Remove** the **edge** with the **highest betweenness**.
- **Output**: **hierarchy of partitions**.

Top-down divisive approach

How to choose **only one** partition?

Modularity

Principles

- Random graphs have no community structure
- Compare edge density in each community w.r.t. randomized version.

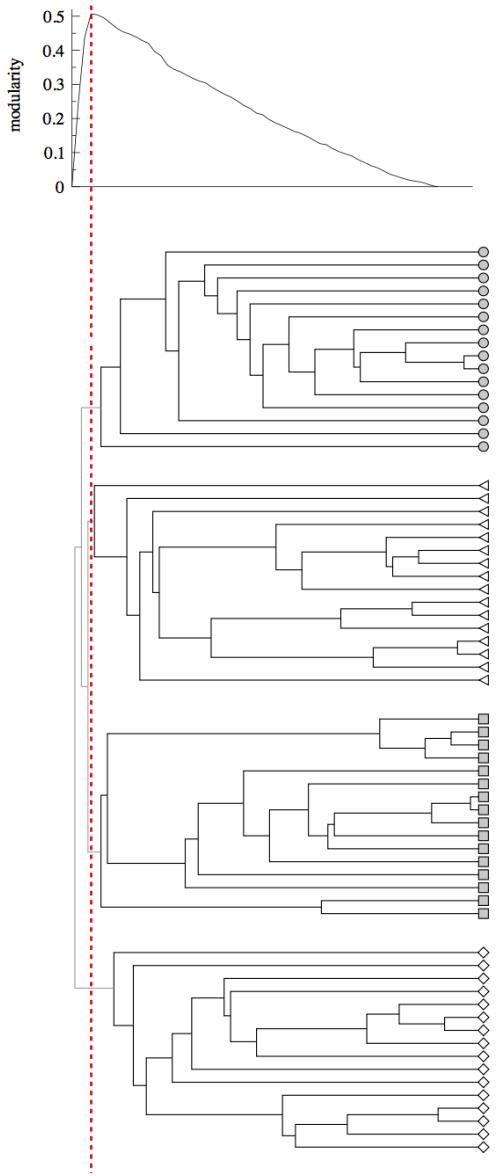
$$Q = \frac{1}{2 \cdot |E|} \sum_{v_i, v_j} \left[A_{ij} - \frac{k_{v_i} \cdot k_{v_j}}{2 \cdot |E|} \right] \delta(v_i, v_j)$$

Where: $k_v = deg(v)$; $\delta(v_i, v_j) = 1$ if $v_i \in C$ and $v_j \in C$ an 0 otherwise.

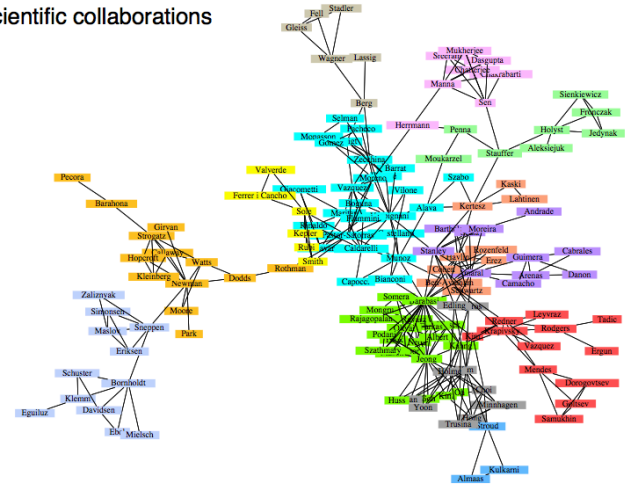
$$Q = \frac{1}{|E|} \sum_{c=1}^{n_c} (|E^c| - \frac{k_c^2}{4 \cdot |E|})$$

Where: Community $C = \langle V^c, E^c \rangle$; n_c nb. of communities; $k_c = \sum_{v \in C} k_v$

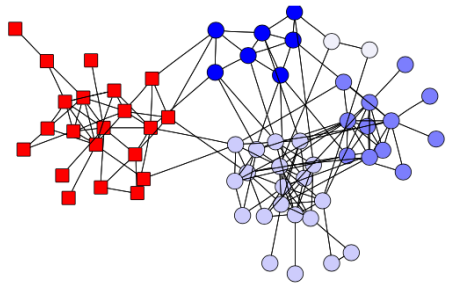
Modularity Based Partition Choice



Scientific collaborations



Dolphins



WWW