CSE 427 Computational Biology

Biological graphs

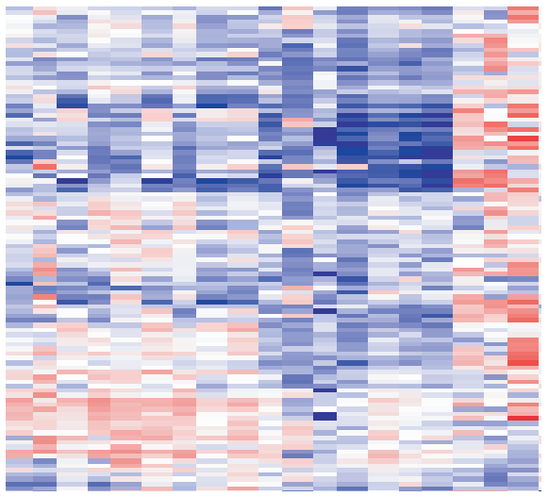
Any dataset can be represented as a graph

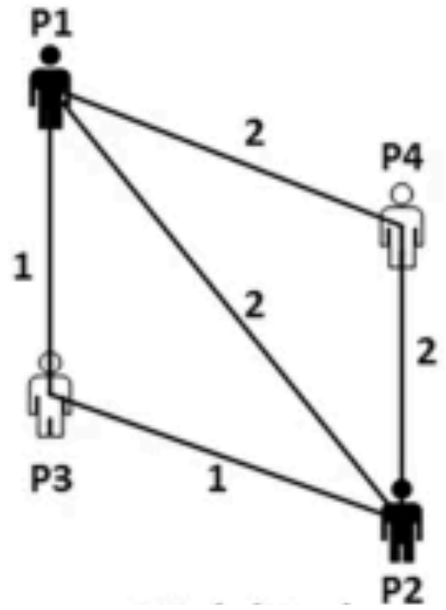
- **Step 1: how to build the graph from a dataset**
  - How to calculate the similarity between two entities?
  - How to define the graph structure based on the similarity
- **Step 2: what graph-based methods should be used for downstream applications**
  - Guilt-by-association
  - Random walk, random walk with restart
  - network embedding, graph neural network
  - Interpretation (which edge or node is important)

Patient graph

20k genes

1000 patients





A patient network of 1000 nodes

- Each node is a patient
- No edge between two patients if the similarity score is too low
- Edge weight is the patient similarity

- How to calculate the similarity? Cosine or Euclidean
  - ML models use Euclidean distance in the loss function

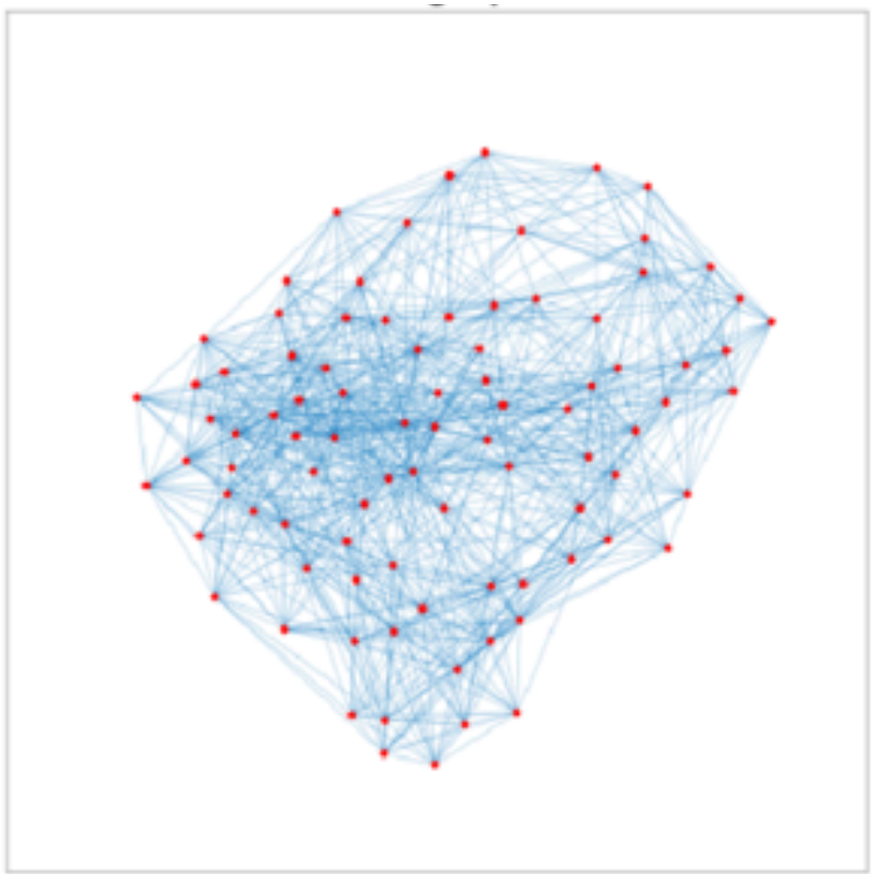
How to define the graph structure: kNN graph or a threshold graph

Use patient graph for drug response prediction

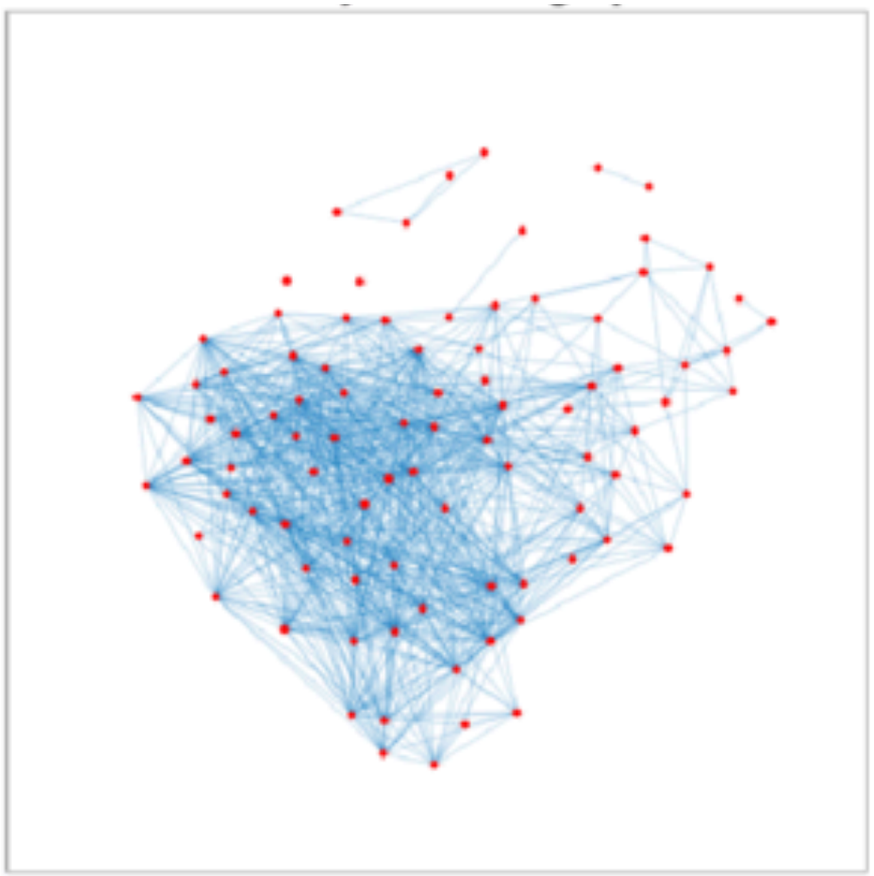
- Nearby patients have similar drug response

kNN graph or Threshold graph?

kNN graph



Threshold graph



kNN graph: each node is connected to k neighbors that have the largest edge weights

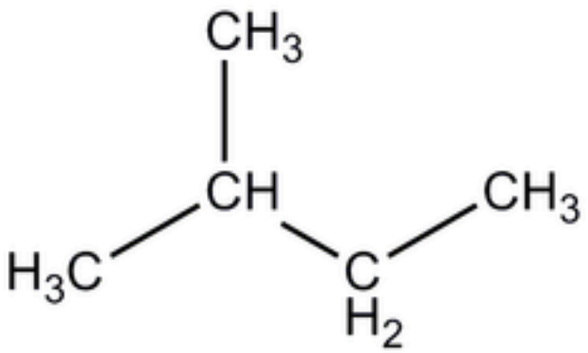
- k is a hyperparameter

Threshold graph: each node is connected to neighbors that have an edge weight larger than x

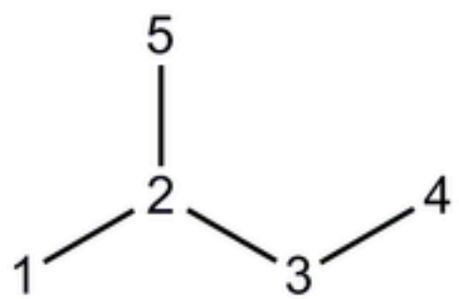
- x is the hyper parameter

Source: Efficient small-world and scale-free functional brain networks at rest using k-nearest neighbors thresholding

Drug graph



Molecule



Graph

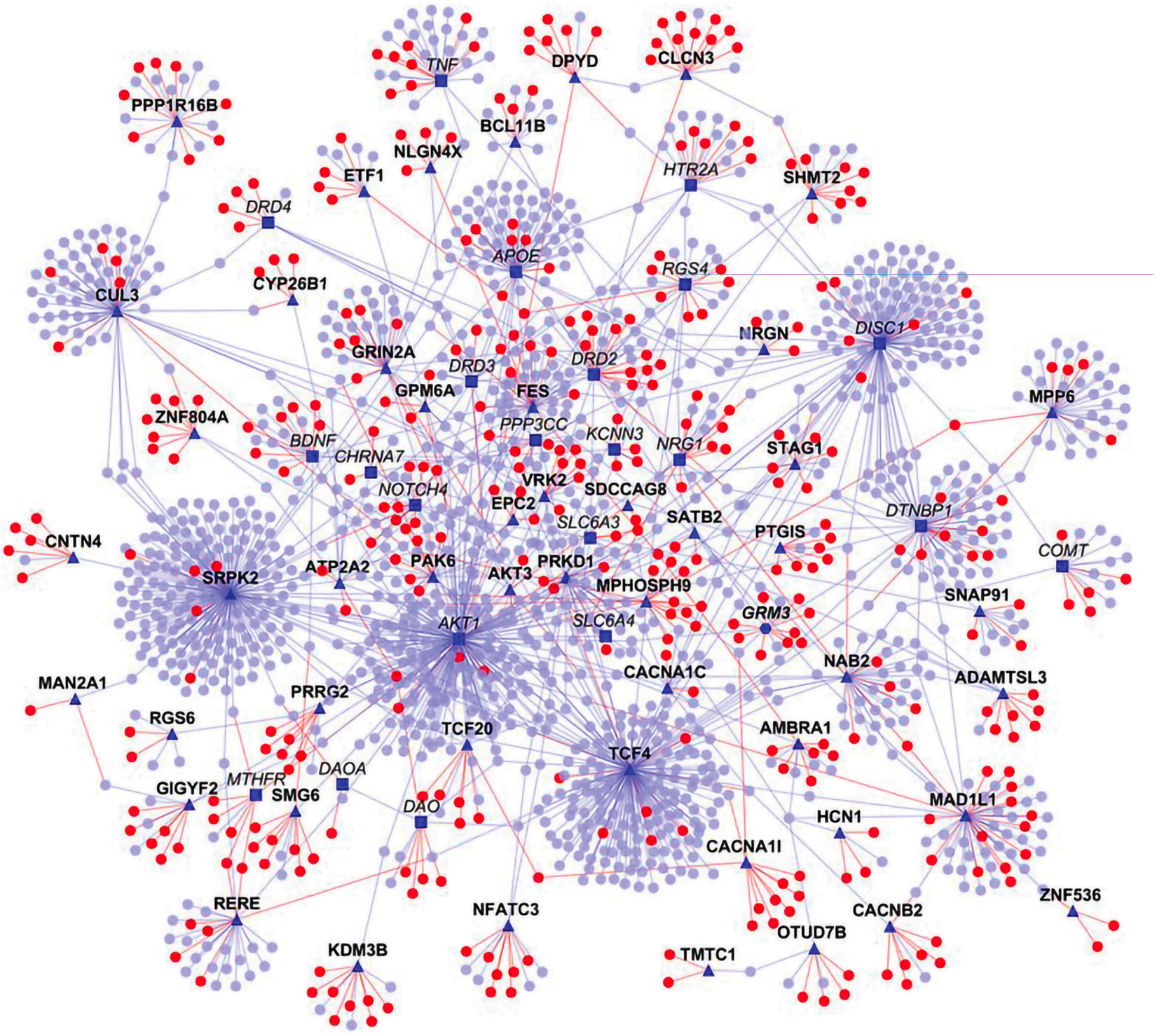
$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad \delta \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 1 \end{bmatrix}$$

Adjacency matrix

Node type

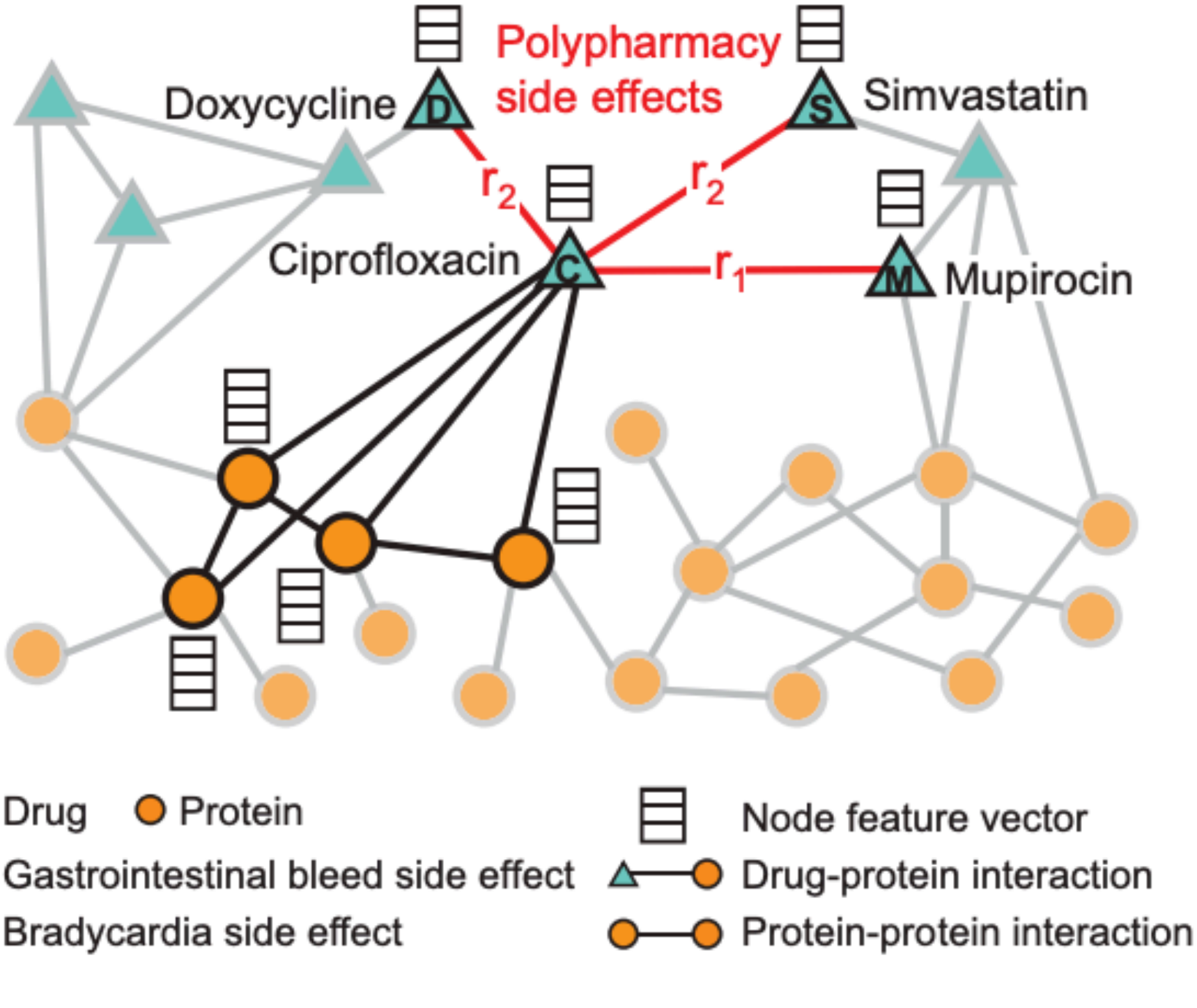
Source: Molecular topology in QSAR and drug design studies

Protein graph



Source: A-Alpha Bio

Use drug and protein graph for side effect prediction



Source:Modeling polypharmacy side effects with graph convolutional networks

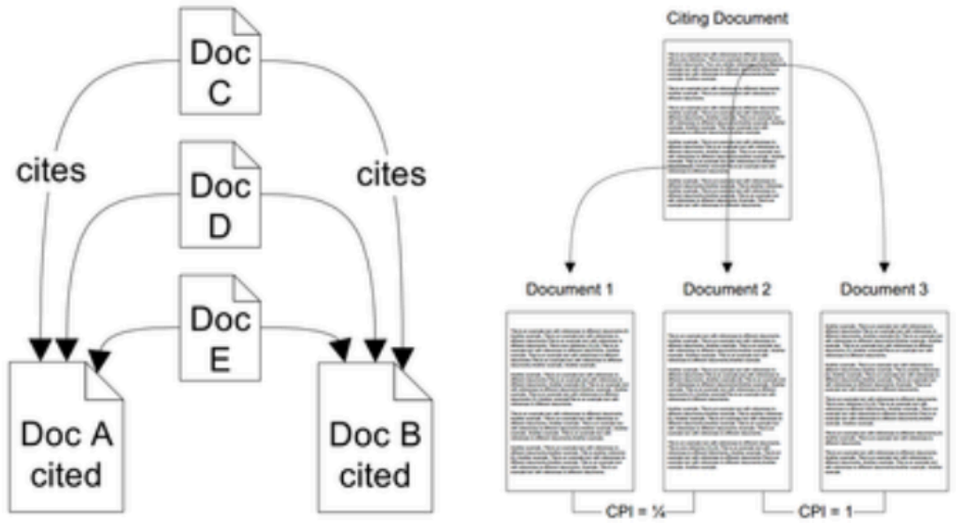
Why do we need to use graph?

- **Use patient by gene matrix or patient by patient network?**
  - Use the original data is like operating on a fully connected network
  - Transformer is like a fully connected graph neural network
- **Advantage of using networks:**
  - Reduce noise in the data
  - Less computationally intensive
  - Visualization and interpretation

Different types of graphs

Graphs from co-citation network

- Co-citation is defined as the frequency with which two documents are cited together by other documents
- Undirected graph.

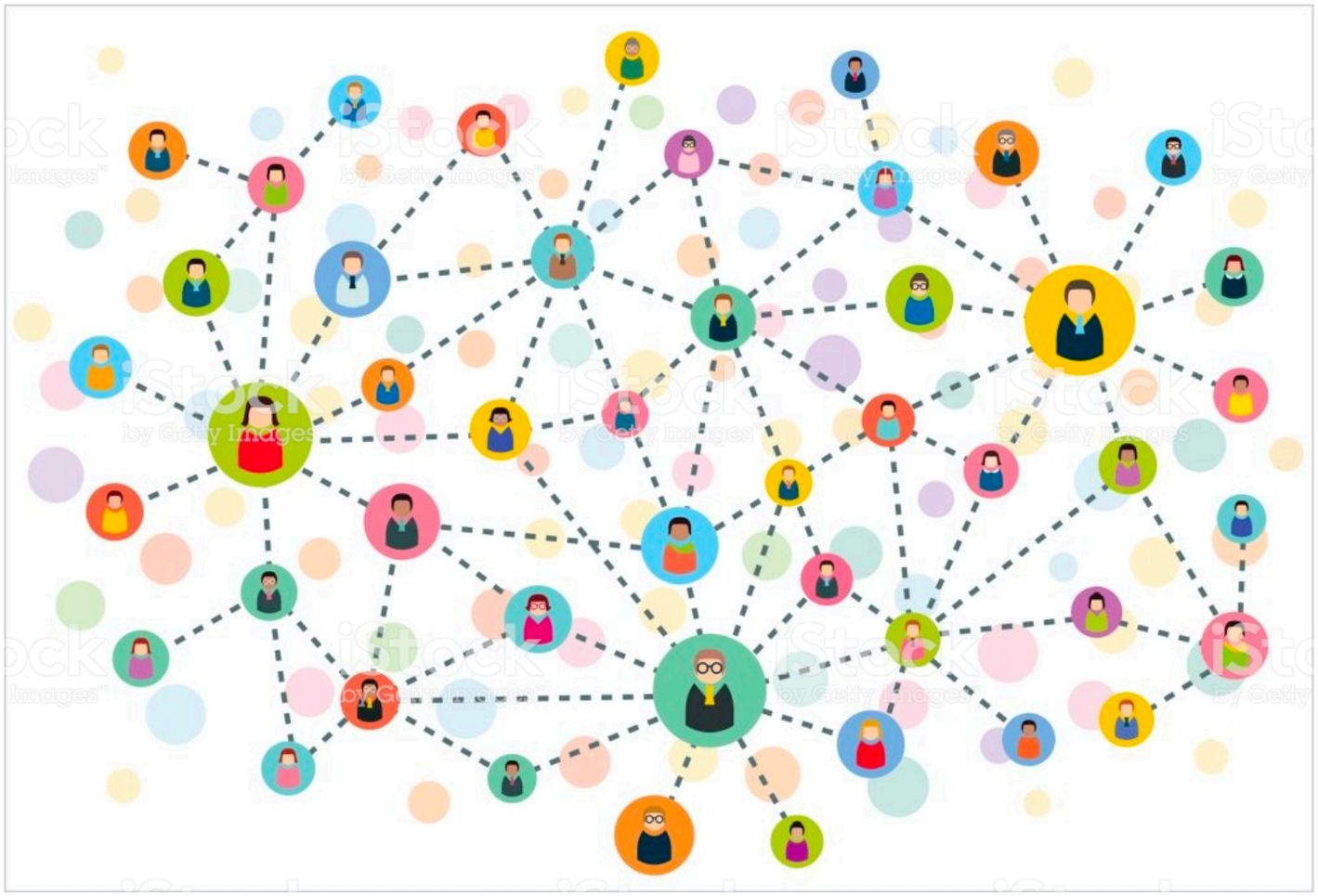


Graphs from citation network

- An edge represents a citation from the current publication to another.
- Directed graph.

Different types of graphs

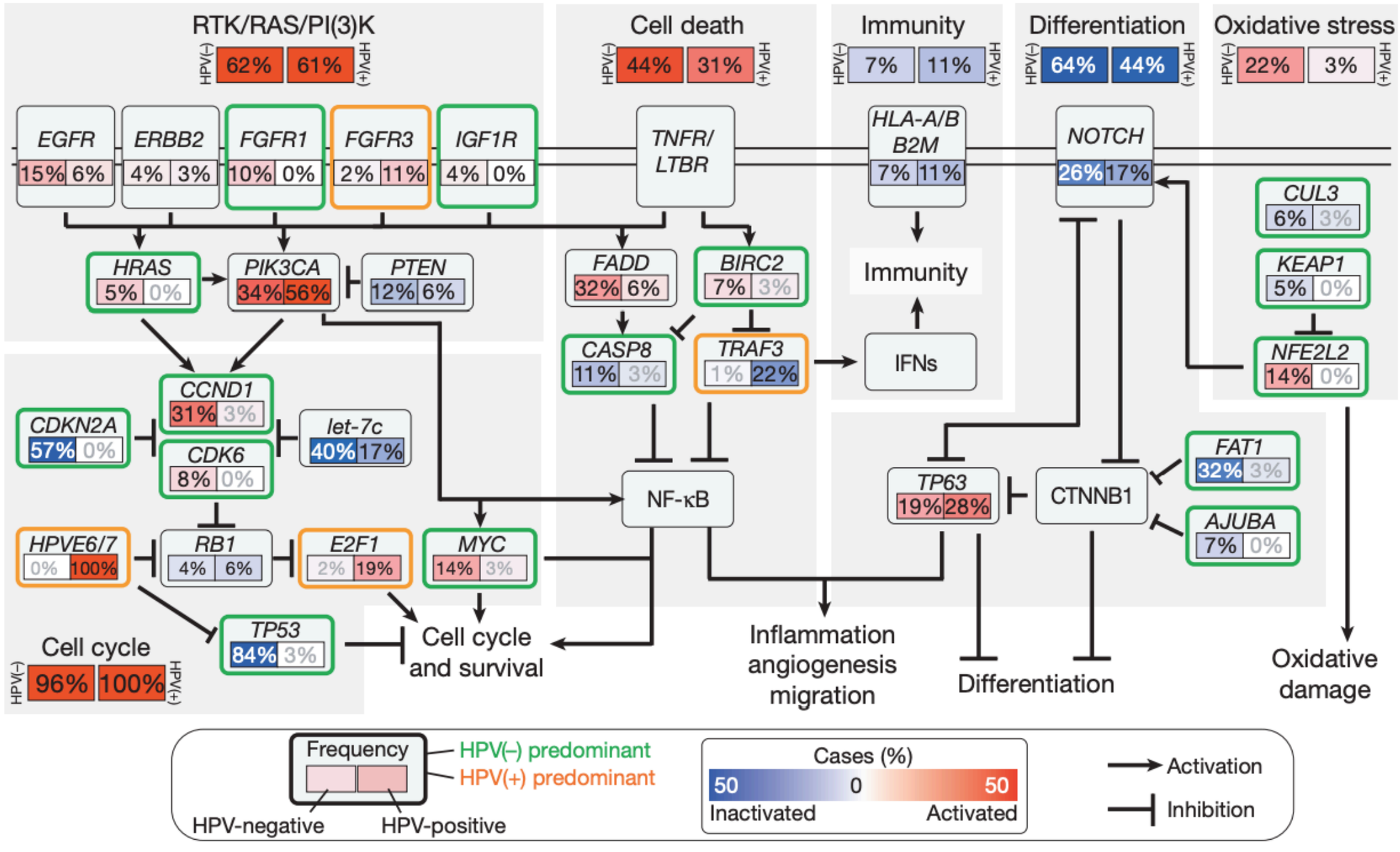
Graphs from social network



- Nodes are users and edges indicates they are friends.

Different types of graphs

Graphs in Cancer



Comprehensive genomic characterization of head and neck squamous cell carcinomas, Nature, 2015

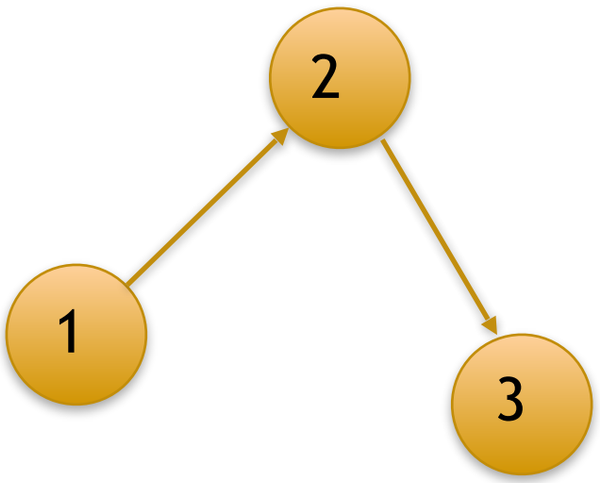
How to build a graph from data ?

- What are nodes and edges?
- In what way nodes interact with each other?
- Directed or Undirected?
- Self-loop?
- Weighted or Unweighted?
- Signed or Unsigned?
- What are node attributes?
- What are edge attributes?

Some examples

- Facebook friendships: undirected, unweighted
- Co-citation networks: undirected, weighted
- Citation networks: directed, unweighted/weighted
- City subway networks: undirected, weighted
- Protein physical networks: undirected, weighted
- Genetic interaction networks: undirected, weighted, signed
- Gene regulatory networks: directed, unweighted, signed

Representing graphs



Adjacent matrix

$$\begin{matrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix}$$

Adjacent list

1: 2
2: 3

Edge set

(1, 2)
(2, 3)

Basic concepts of graph

Degree distribution

- Probability that a randomly chosen node has degree k.

Paths

- A path is a sequence of nodes in which each node is linked to the next one.

Diameter

- The maximum (shortest path) distance between any pair of nodes in a graph.

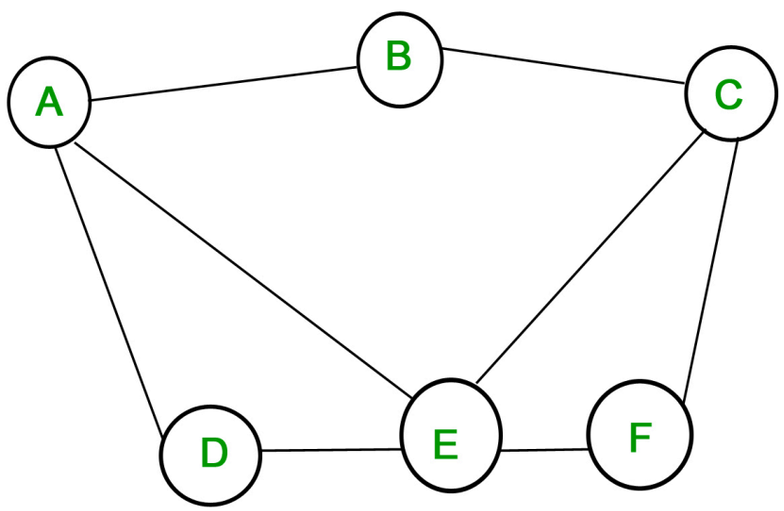
Clustering coefficient

- For each node, what is the portion of its neighbors are connected.

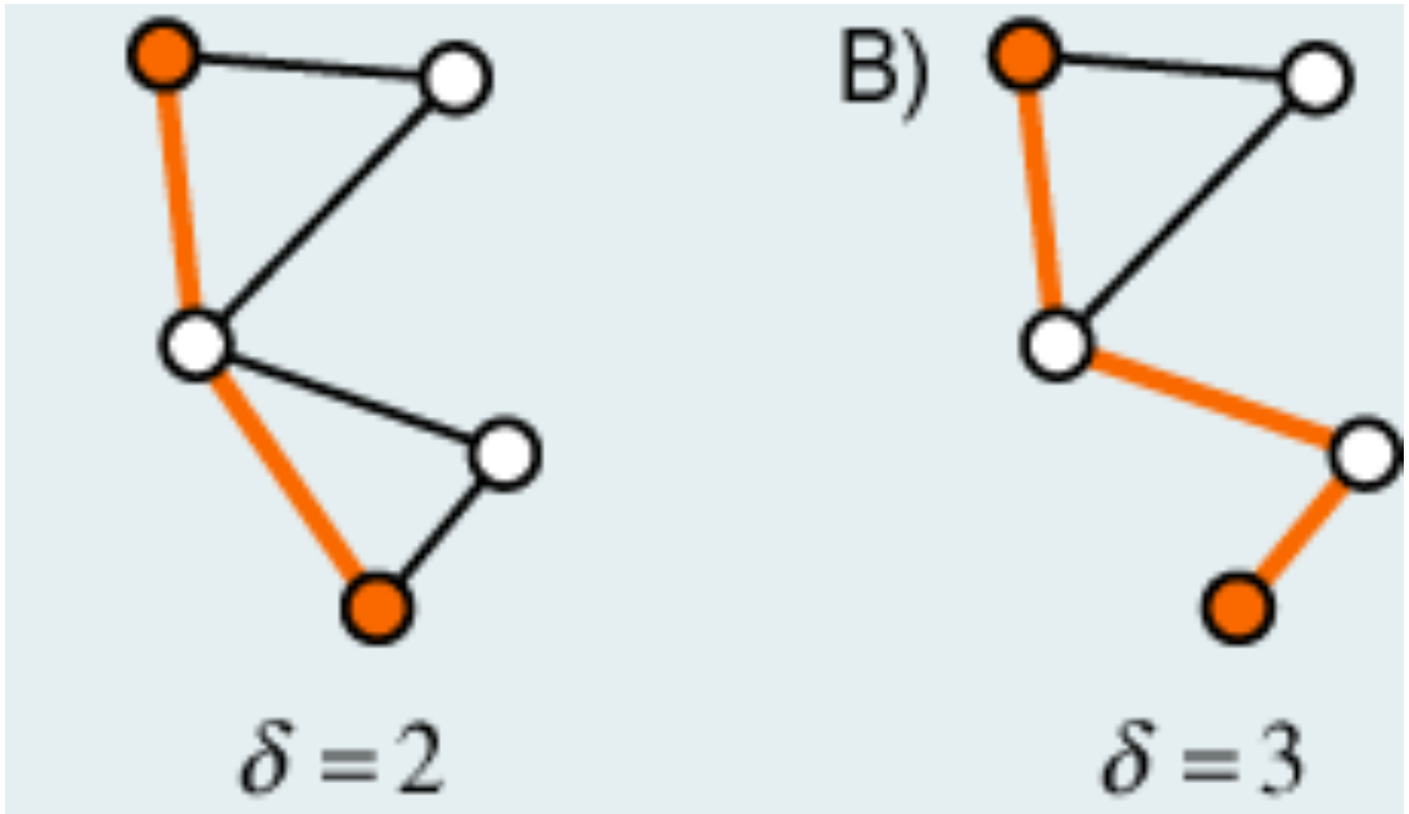
$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

Largest connected component

- Largest node set where any two nodes are connected to each other.



Graph diameter



source: Encyclopedia of Systems Biology

Basic concepts of graph

Degree distribution

- Probability that a randomly chosen node has degree k.

Paths

- A path is a sequence of nodes in which each node is linked to the next one.

Diameter

- The maximum (shortest path) distance between any pair of nodes in a graph.

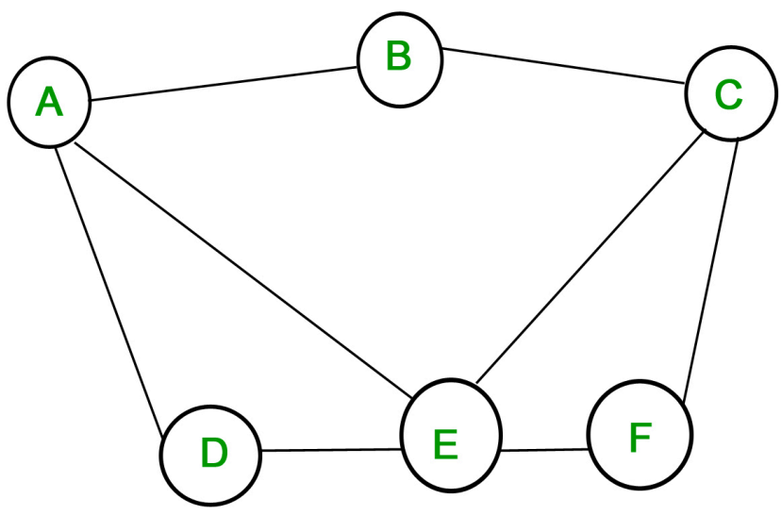
Clustering coefficient

- For each node, what is the portion of its neighbors are connected.

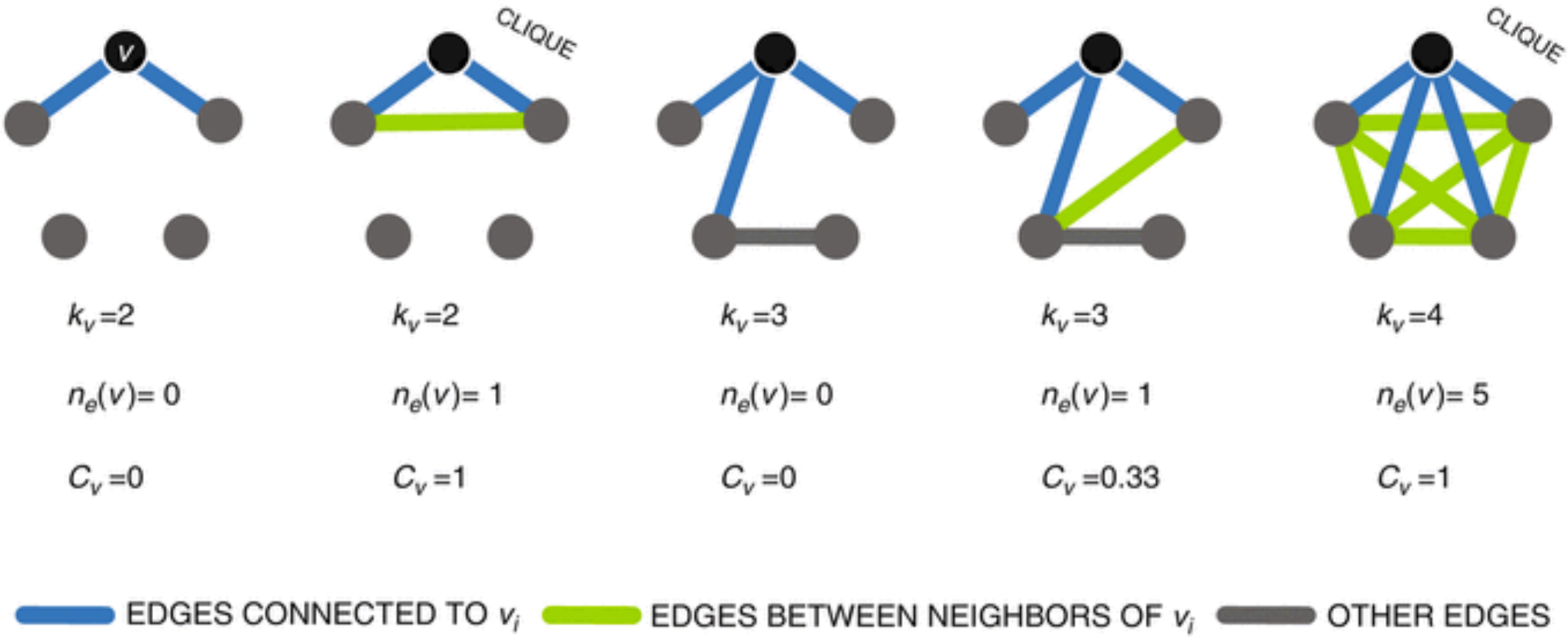
$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

Largest connected component

- Largest node set where any two nodes are connected to each other.



a **Local clustering coefficient**



source: Encyclopedia of Systems Biology

Basic concepts of graph

Degree distribution

- Probability that a randomly chosen node has degree k.

Paths

- A path is a sequence of nodes in which each node is linked to the next one.

Diameter

- The maximum (shortest path) distance between any pair of nodes in a graph.

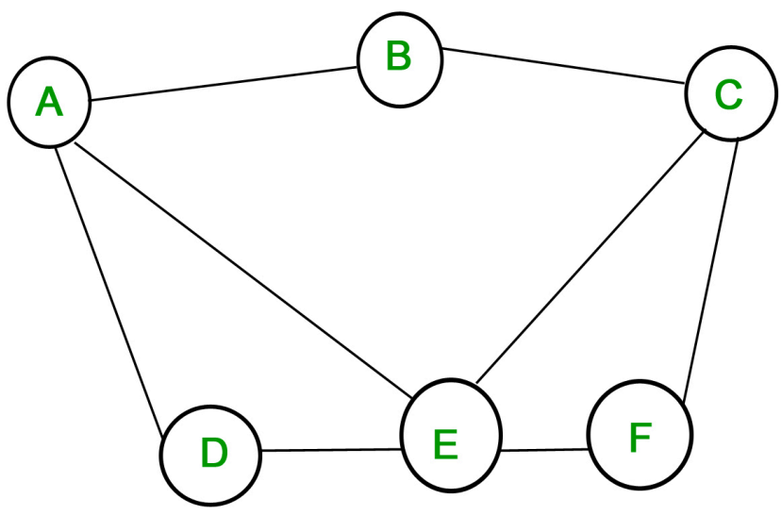
Clustering coefficient

- For each node, what is the portion of its neighbors are connected.

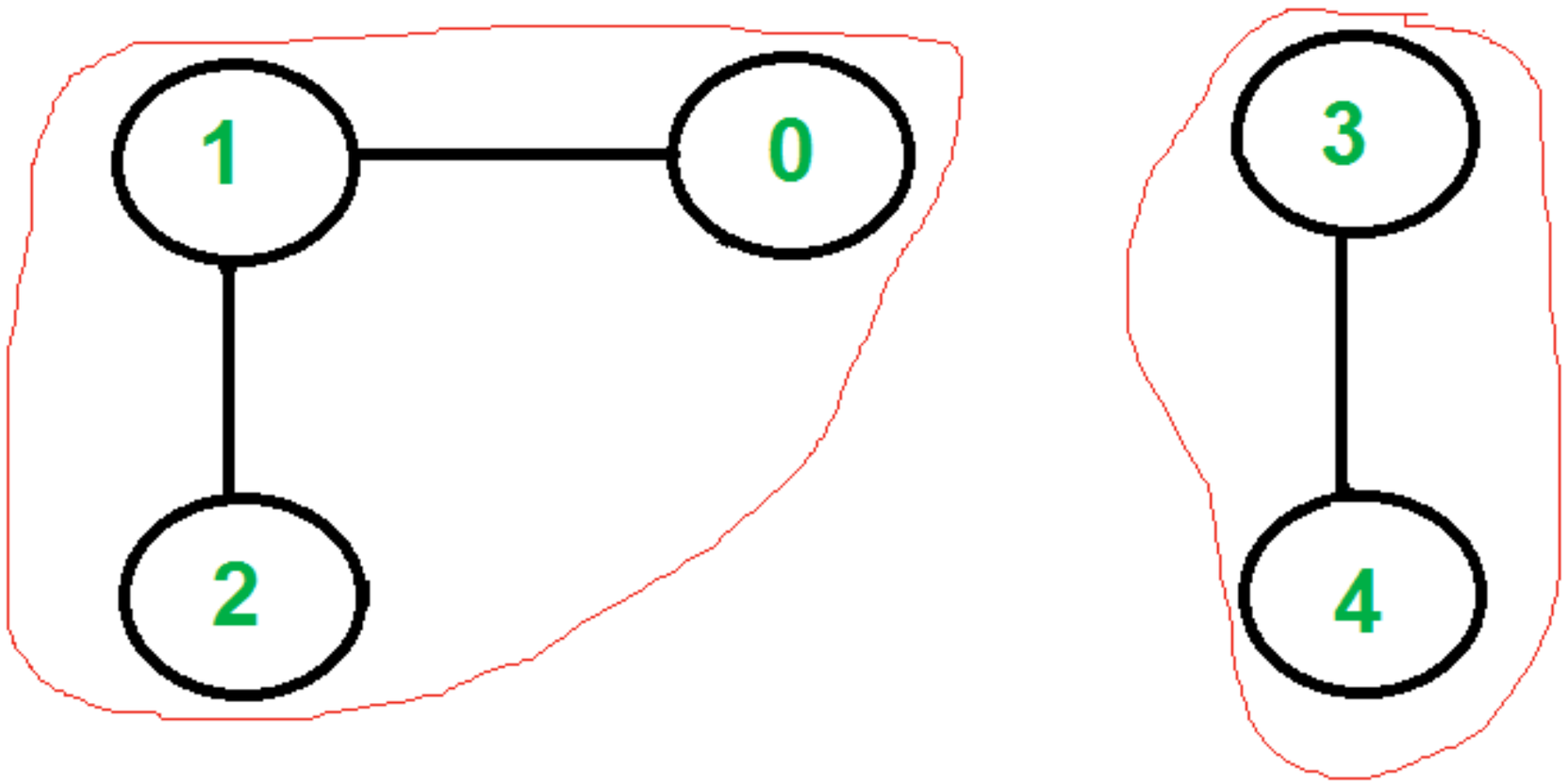
$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

Largest connected component

- Largest node set where any two nodes are connected to each other.



Connected component



There are two connected components in above undirected graph
0 1 2
3 4

source: https://www.geeksforgeeks.org/connected-components-in-an-undirected-graph/

Analysis of networks: Protein-protein interaction network

Download the data

- 5487 nodes, 141,347 edges

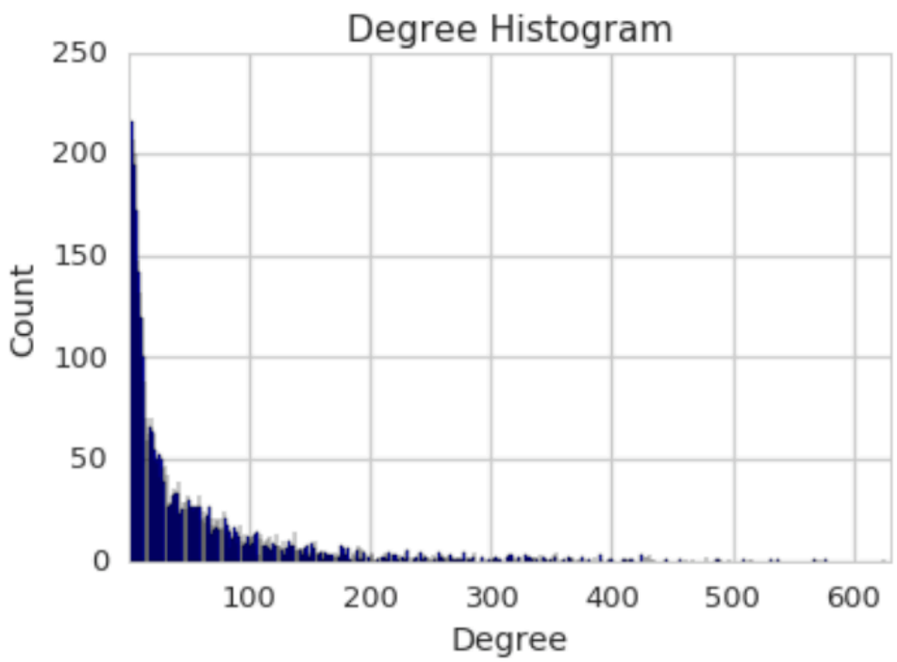
Diameter

- Infinity (disconnected)

Clustering coefficient

- 0.2485

Degree distribution

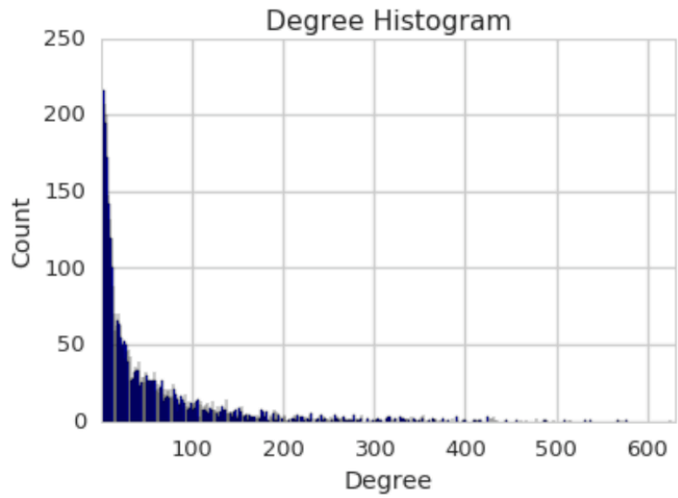


Largest connected component

- 5481
- Diameter: 8

Real networks vs. Random graph ?

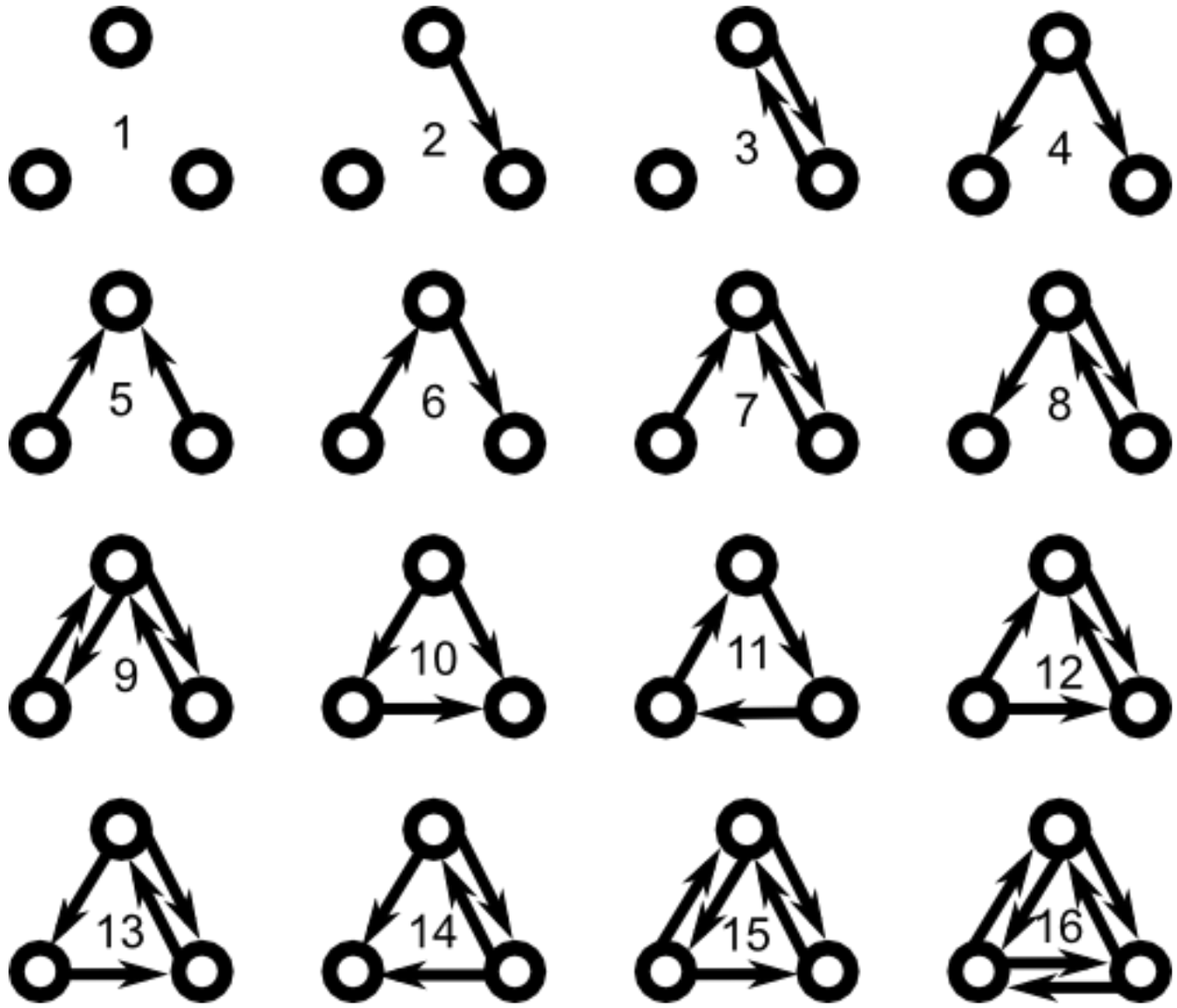
Are real networks like random graphs?



	Random graph	PPI networks
Degree distribution	Binomial	Power law
Clustering coefficient	*P=0.0093*	0.2485
Diameter	Infinity	8

There are local structures

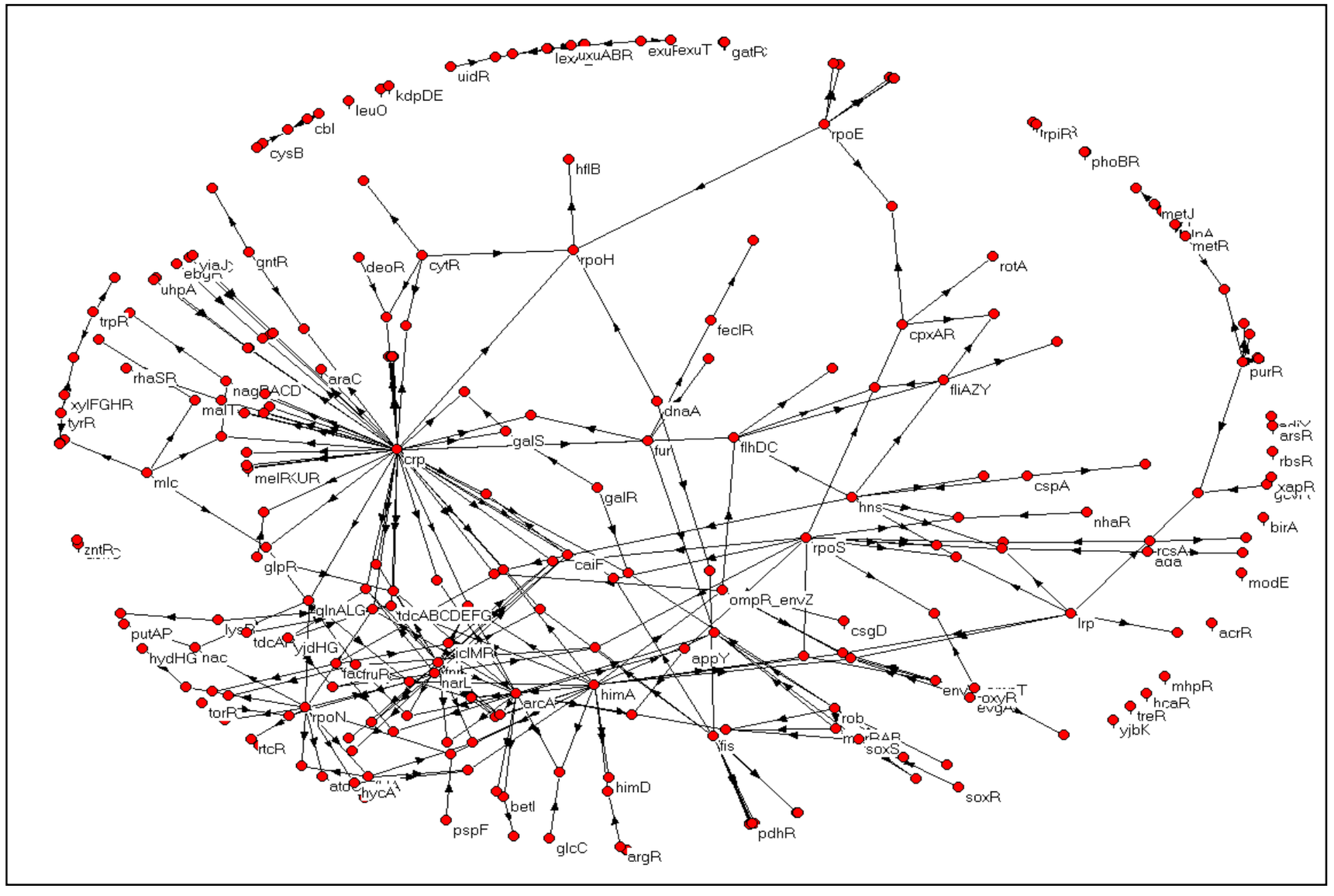
How to understand this biological network?



Network motif

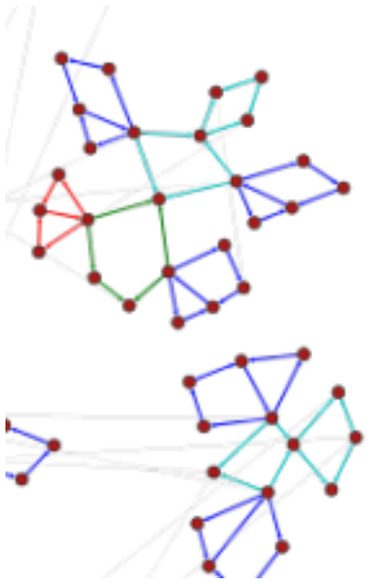
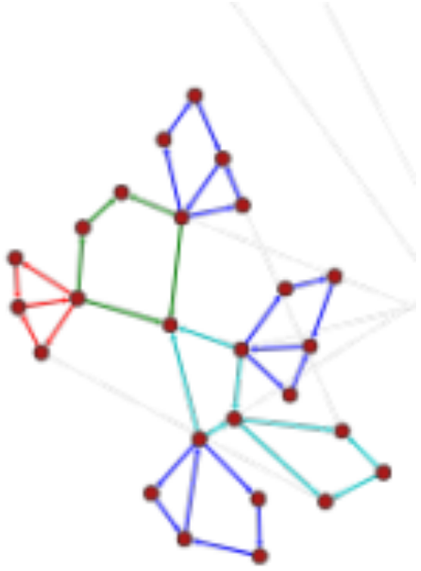
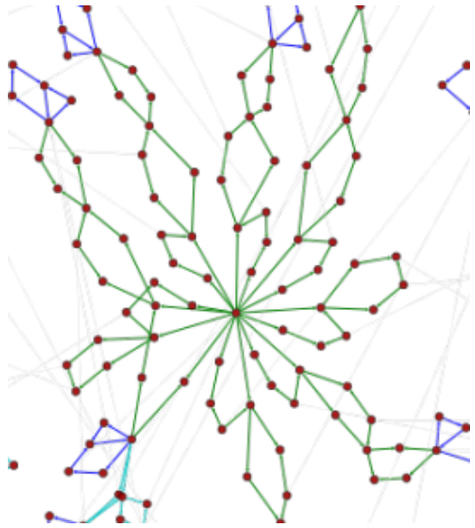
Network motifs

Network motifs were first systematically defined in Escherichia coli (E. coli)

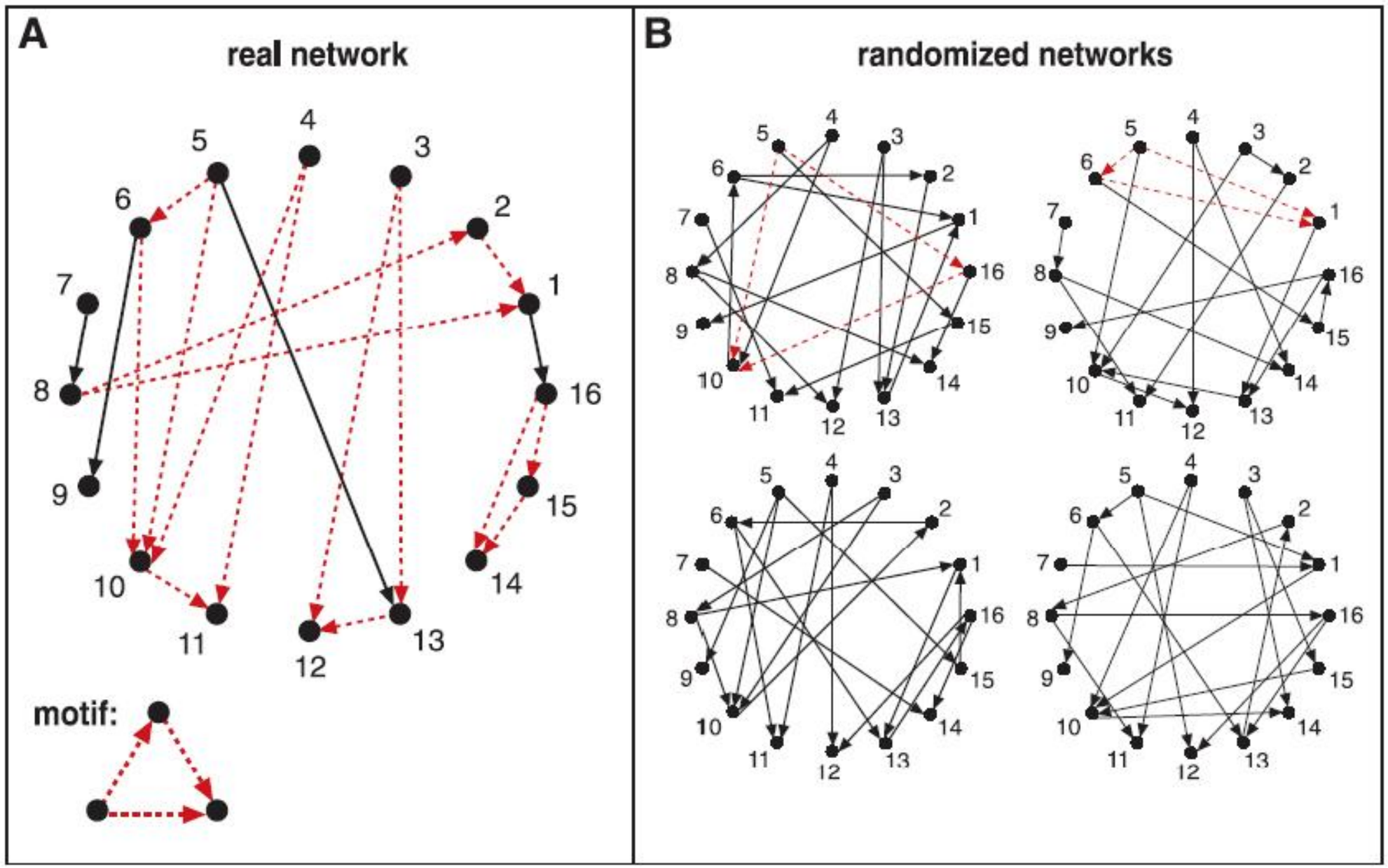


Network motifs

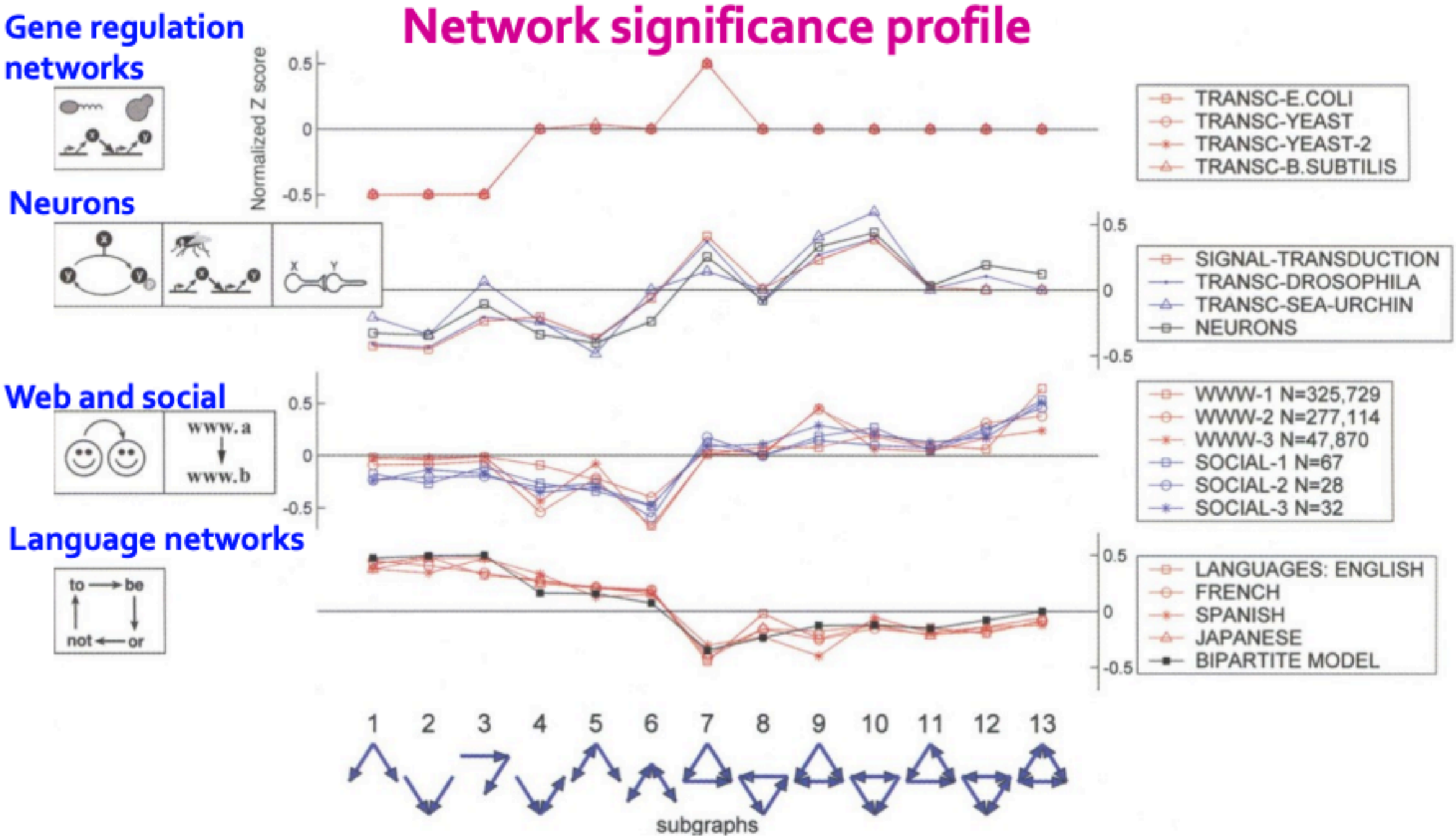
- Network motifs are defined as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks.
- Another definition: Network motifs are statistically overrepresented sub-graph.
- Another definition: Network motifs are the building blocks of networks which can be used to characterize and discriminate networks.
- A network motif has to be a connected subgraph.

Compare with random network



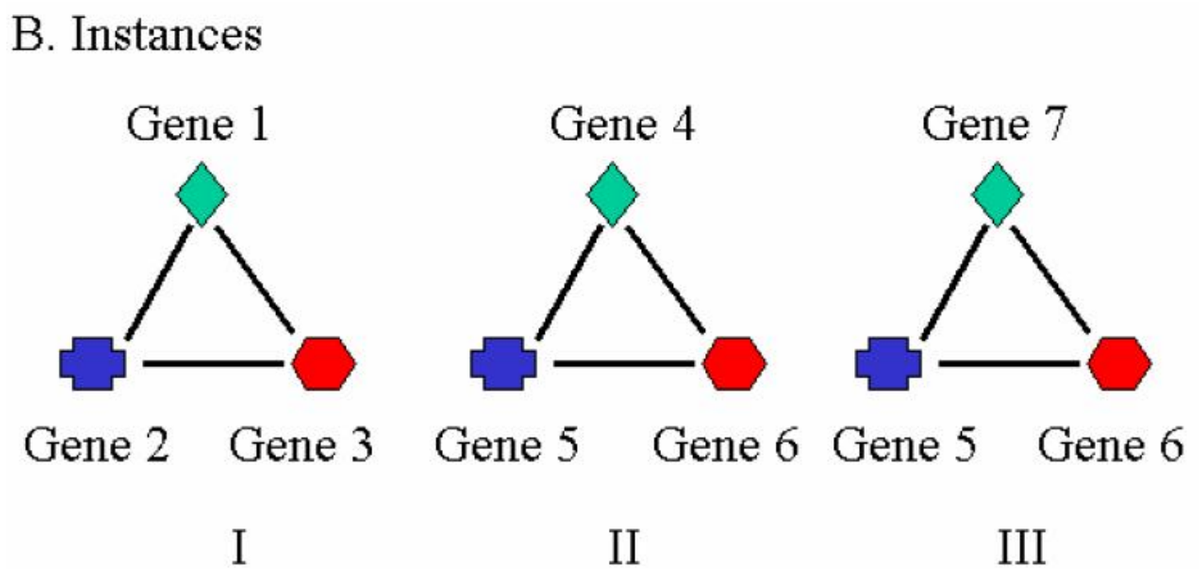
Significance of different sub-graphs



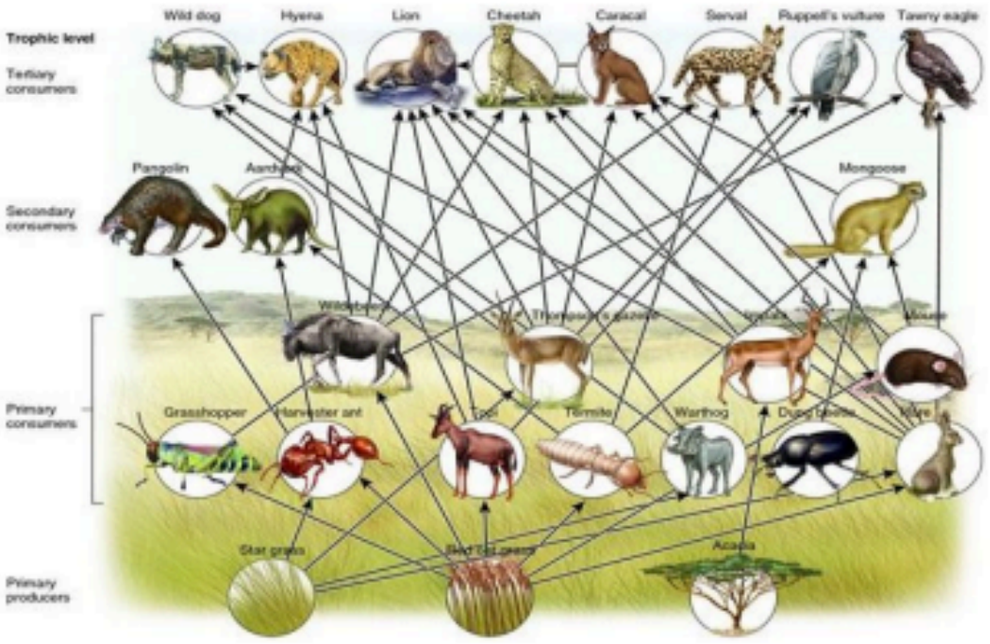
From Jure Leskovec slides

Not only sub-graph frequency

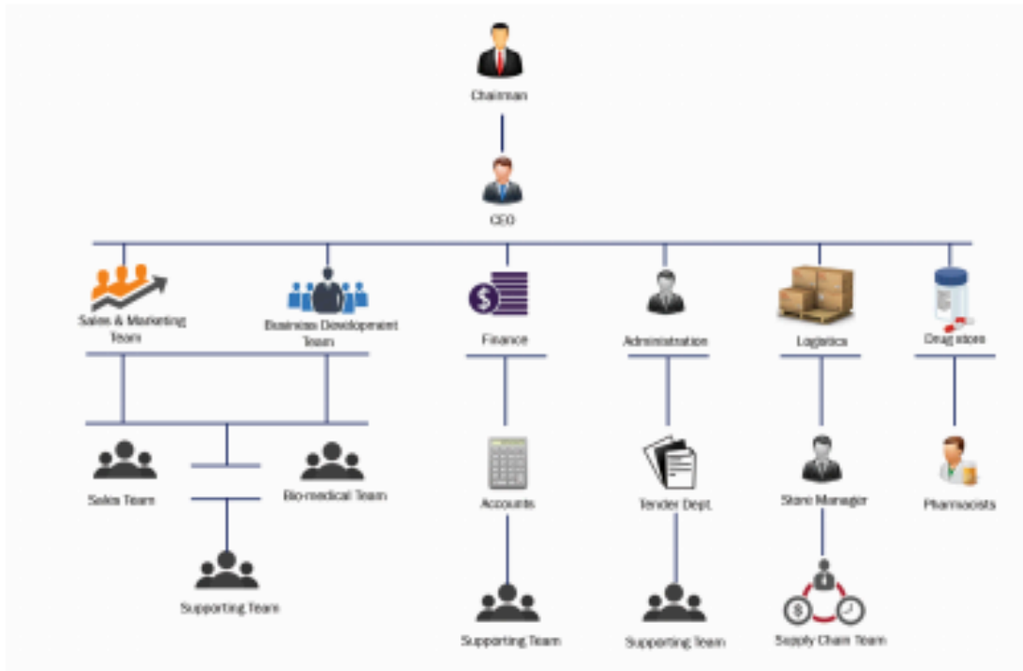
Need to consider node properties



In every life cycle there are predators and plants

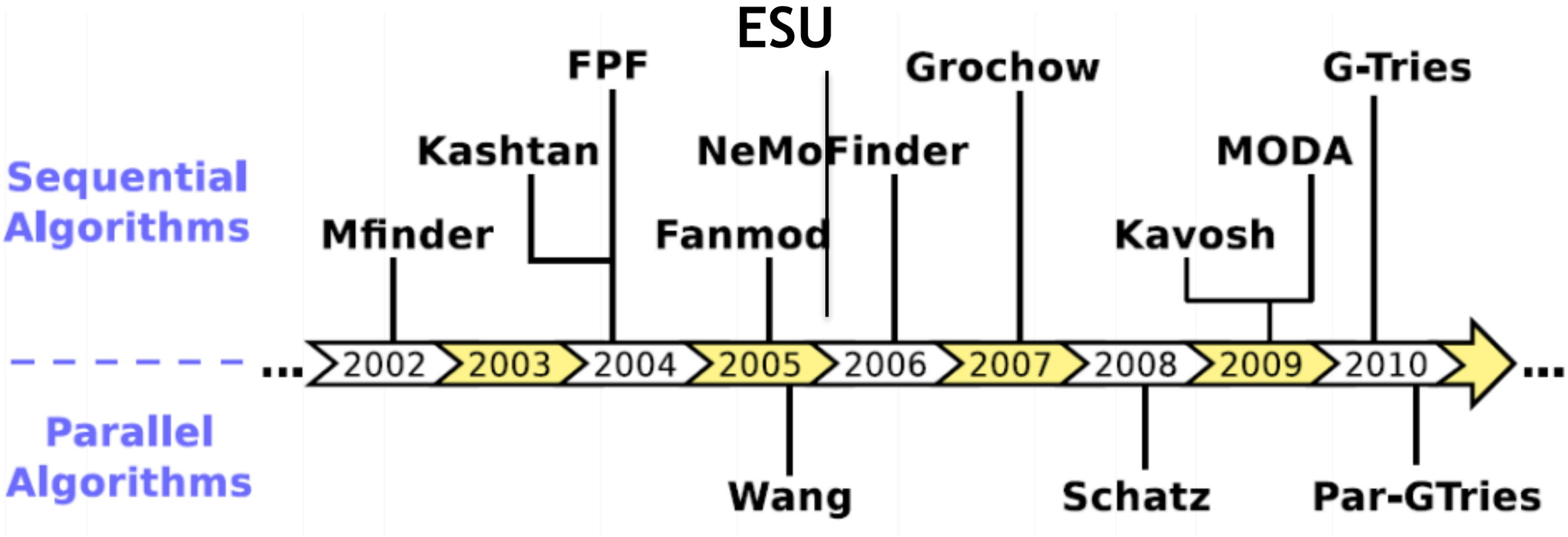


Every company there are managers and employees



Network motif detection algorithm

Timeline



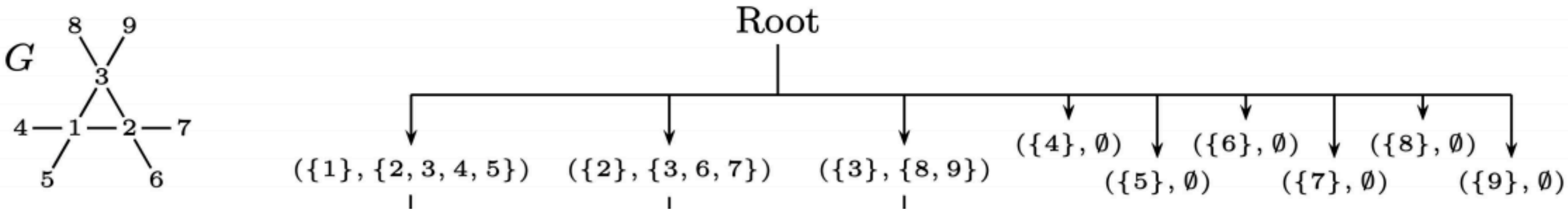
Motif Detection Algorithm

- Finding size-k motifs requires two step:
  - Enumerating all size-k connected subgraphs
  - Counting #(occurrences of each subgraph type)
    - Given two subgraph, we want to tell whether they are isomorphism (NP-complete)

- Exact subgraph enumeration (ESU) [Wernicke 2006]

Exact Subgraph Enumeration

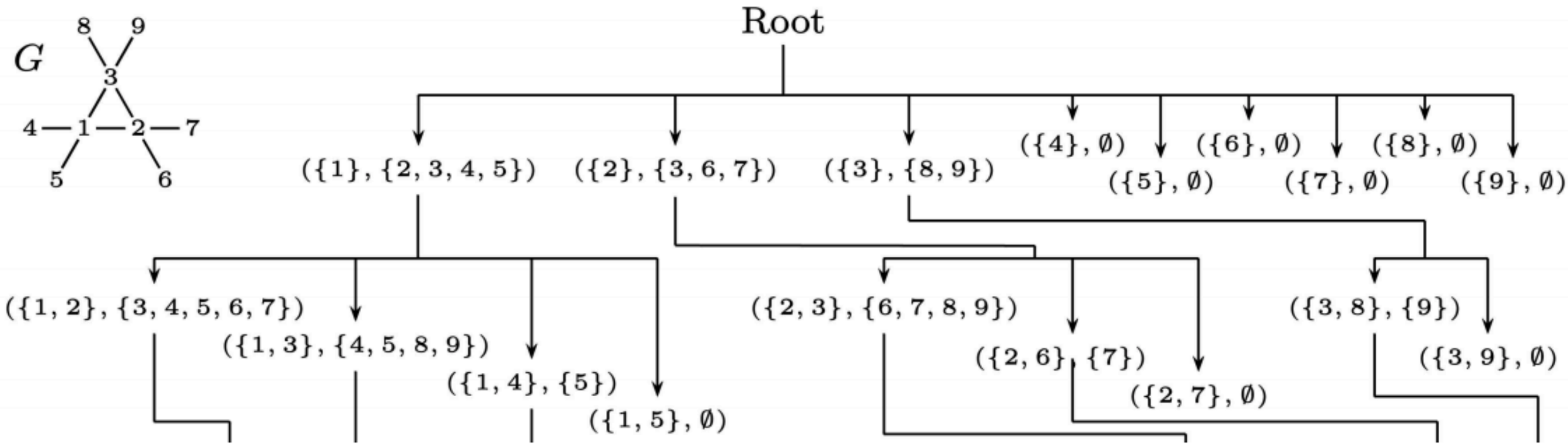
K=3



Next, you want to count all these subgraphs

Exact Subgraph Enumeration

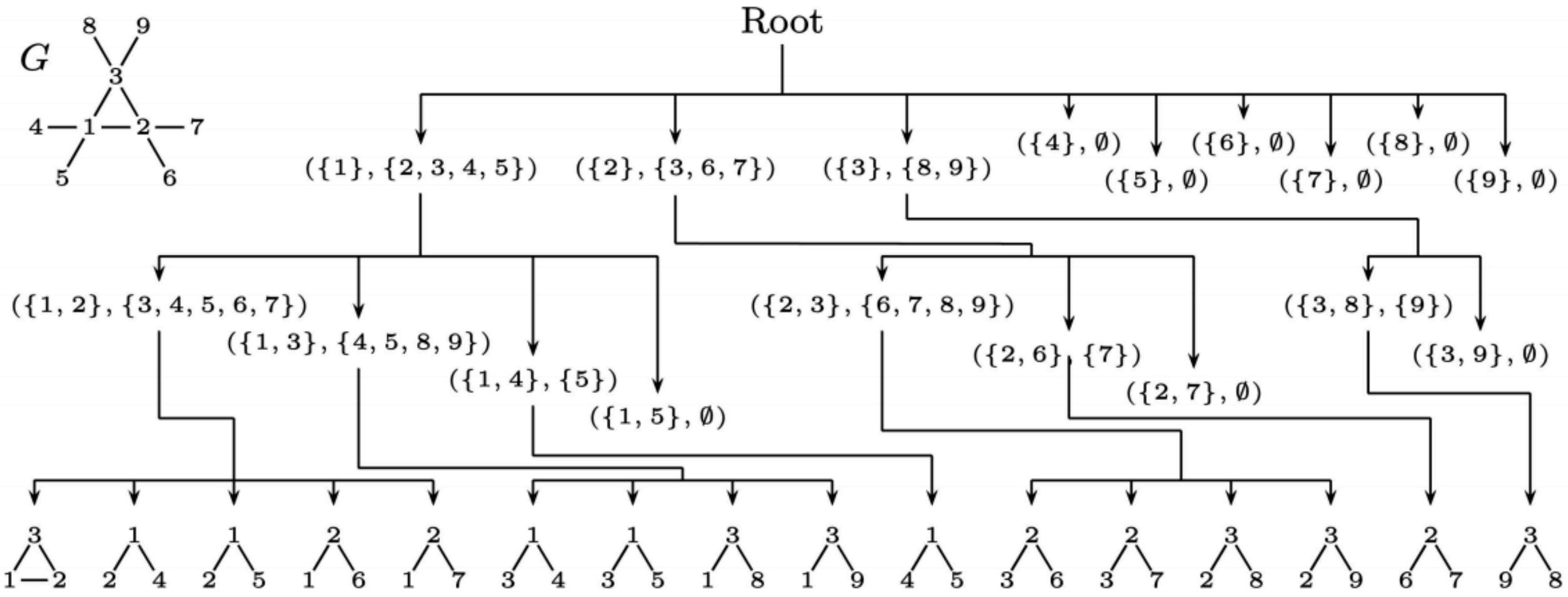
K=3



Next, you want to count all these subgraphs

Exact Subgraph Enumeration

K=3



Next, you want to count all these subgraphs

Enumerate size-k Subgraph Enumeration

- Two sets:
  - $V_{subgraph}$ currently constructed subgraphs.
  - $V_{extension}$ set of candidate nodes to extend the subgraphs.
- Intuition: Starting with a node ***v***, add those nodes ***u*** to $V_{extension}$ with two properties of ***u***:
  - ***u***'s node_id must be larger than that of ***v***
  - ***u*** may only be neighbored to some newly added node w but not of any node already in $V_{subgraph}$
- ESU is implemented as a recursive function:
  - The running of this function can be displayed as a tree-like structure of depth k, called the ESU-Tree

Exact Subgraph Enumeration

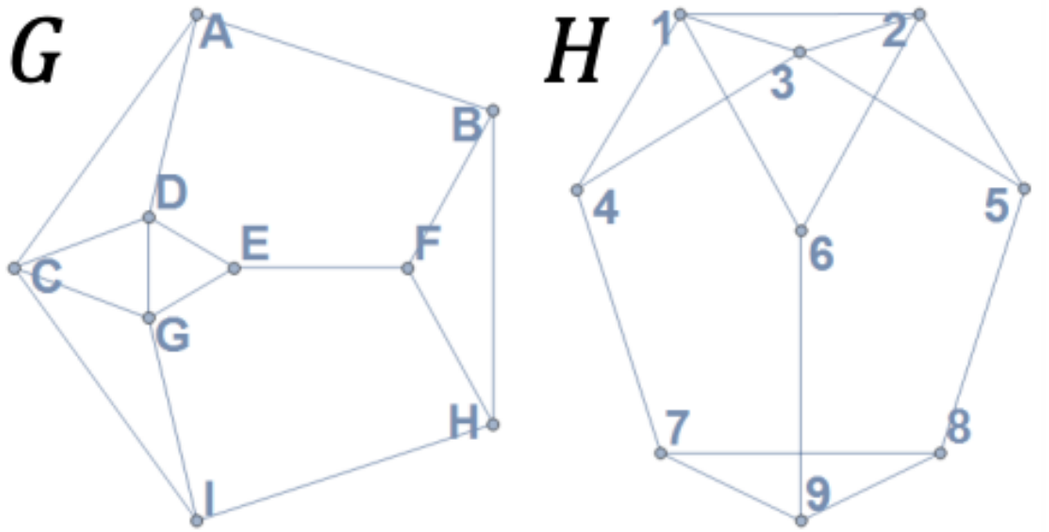
Algorithm: ENUMERATESUBGRAPHS(G, k) (ESU)
Input: A graph $G = (V, E)$ and an integer $1 \leq k \leq |V|$.
Output: All size-k subgraphs in G.

01 **for** each vertex $v \in V$ **do**
02 $V_{Extension} \leftarrow \{u \in N(\{v\}) : u > v\}$
03 **call** EXTENDSUBGRAPH$(\{v\}, V_{Extension}, v)$
04 **return**

EXTENDSUBGRAPH$(V_{Subgraph}, V_{Extension}, v)$
E1 **if** $|V_{Subgraph}| = k$ **then output** $G[V_{Subgraph}]$ and **return**
E2 **while** $V_{Extension} \neq \emptyset$ **do**
E3 Remove an arbitrarily chosen vertex w from $V_{Extension}$
E4 $V'_{Extension} \leftarrow V_{Extension} \cup \{u \in N_{excl}(w, V_{Subgraph}) : u > v\}$
E5 **call** EXTENDSUBGRAPH$(V_{Subgraph} \cup \{w\}, V'_{Extension}, v)$
E6 **return**

Graph Isomorphism

- If you ignore the node types/features and edge types/features, you will find some subgraphs are topologically equivalent.



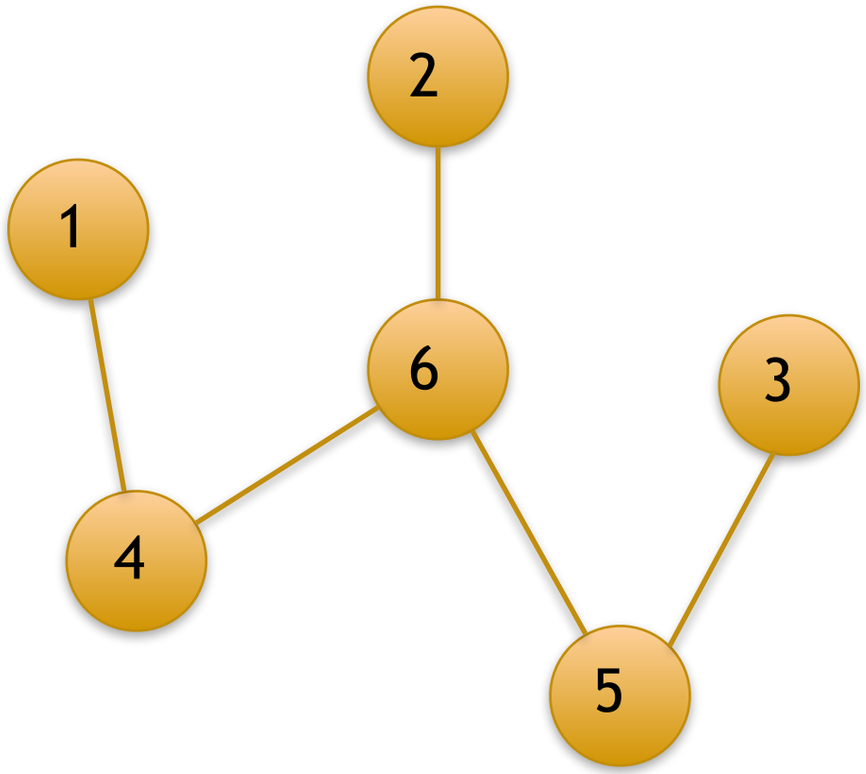
f

A	4
B	7
C	1
D	3
E	5
F	8
G	2
H	9
I	6

- Graphs G and H are isomorphic if there exists a bijection f: V(G) → V(H) such that: Any two nodes u and v of G are adjacent in G iff f(u) and f(v) are adjacent in H.

Graph Isomorphism Detection Algorithm

- **McKay's Canonical Graph Labeling Algorithm:** Nauty, Trace, Bliss all have their own implementations of Mckay's algorithm. [McKay 1981]
- Time complexity exp(O(n2/3))
- Intuition: First hash two graphs as two strings and then compare two strings.
- Label each node according to their degrees first. Iterate over each edge.
- Put a "1" if there is an edge between those two nodes, a "0" if not.



(1,2) (1,3) (1,4) (1,5) (1,6) 00100

(2,3) (2,4) (2,5) (2,6) 0001

(3,4) (3,5) (3,6) 010

(4,5) (4,6) 01

(5,6) 1

001000001010011

N node: N * (N-1)/2 edges
5 node: 15 edges

Graph Isomorphism Detection Algorithm

- But the order of the edge matters in this hash coding. To solve that problem we want to enumerate all the orderings.
- We first sort the all the nodes according to their degrees.
- Within each degree bin, we enumerate all the orderings.

{A,B,C} | {D, E} | {F}

{A,B,C} | {D, E} | {F} {A,C,B} | {D, E} | {F} {B,A,C} | {D, E} | {F} {B,C, A} | {D, E} | {F}

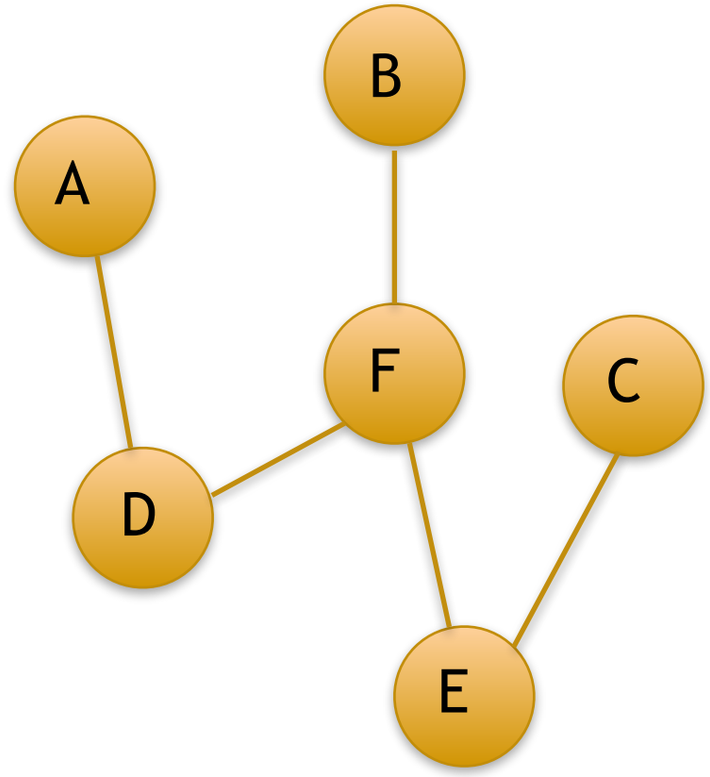
{A,B,C} | {D, E} | {F}

{A,B,C} | {E, D} | {F}

001000001010011

000100101000011

This hash code matches the previous one!



Acknowledgement

- Part of the slides are from
  - Dr. Jianzhu Ma's lecture on graphs in machine learning