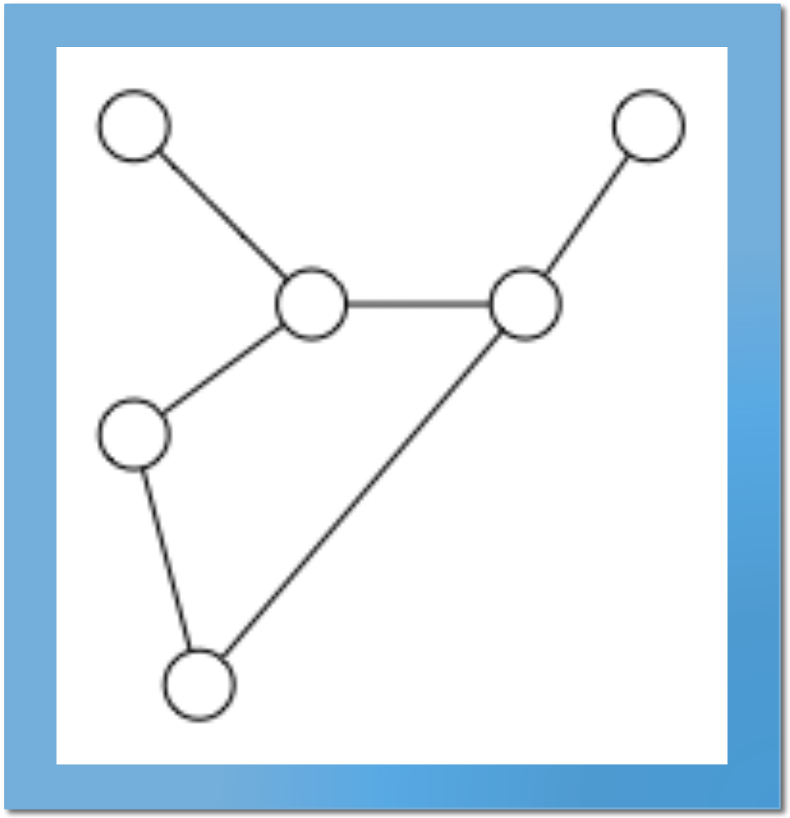
GRAPHS + CLUSTERING

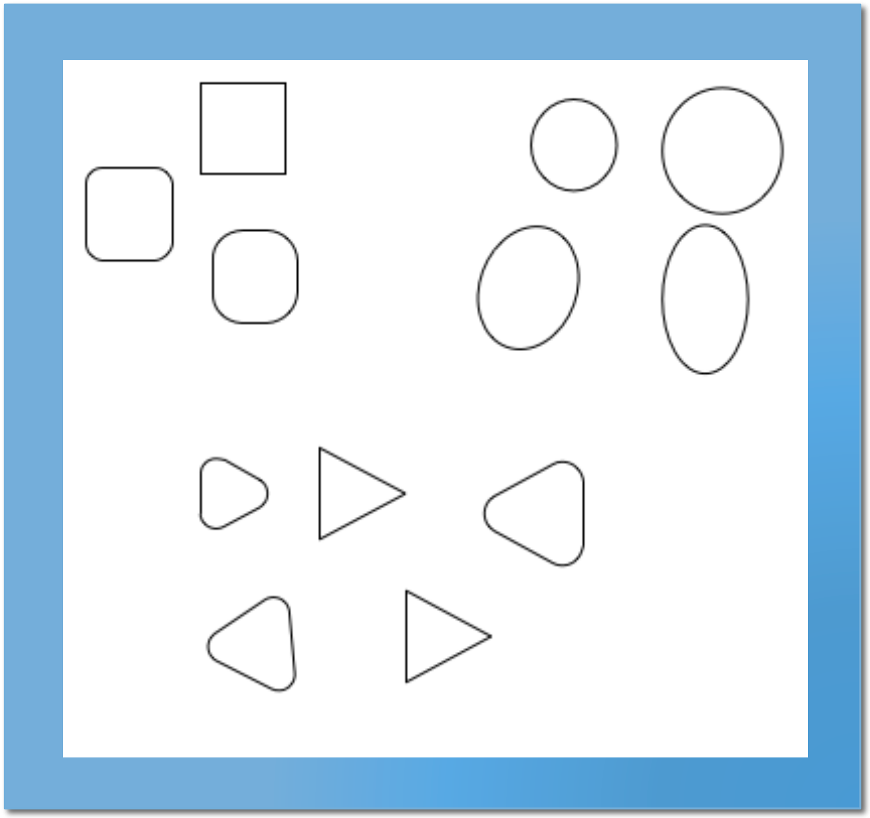
Together At Last!

April 2016

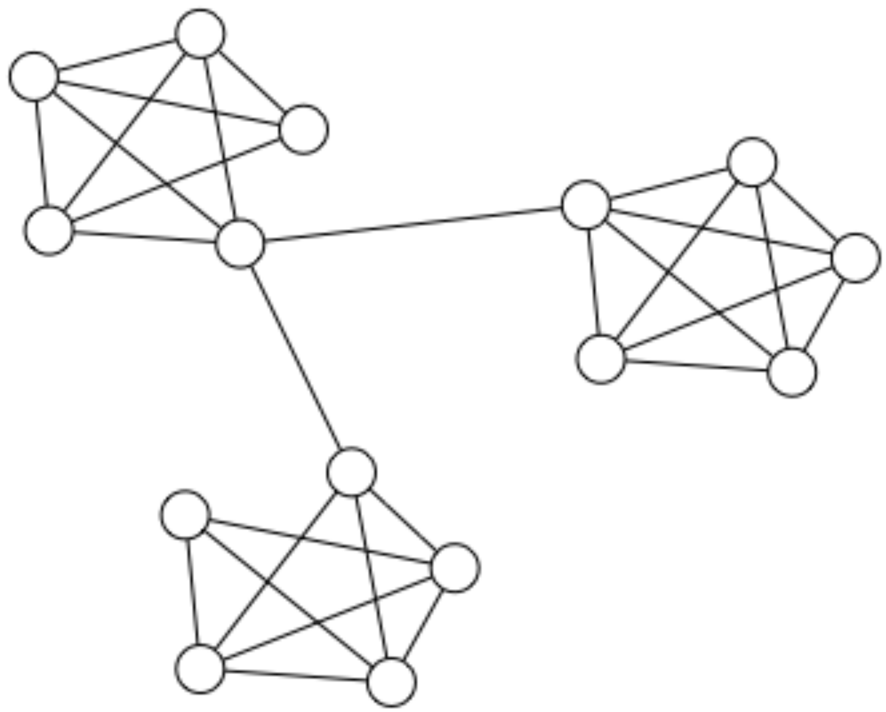
Jen Schellinck



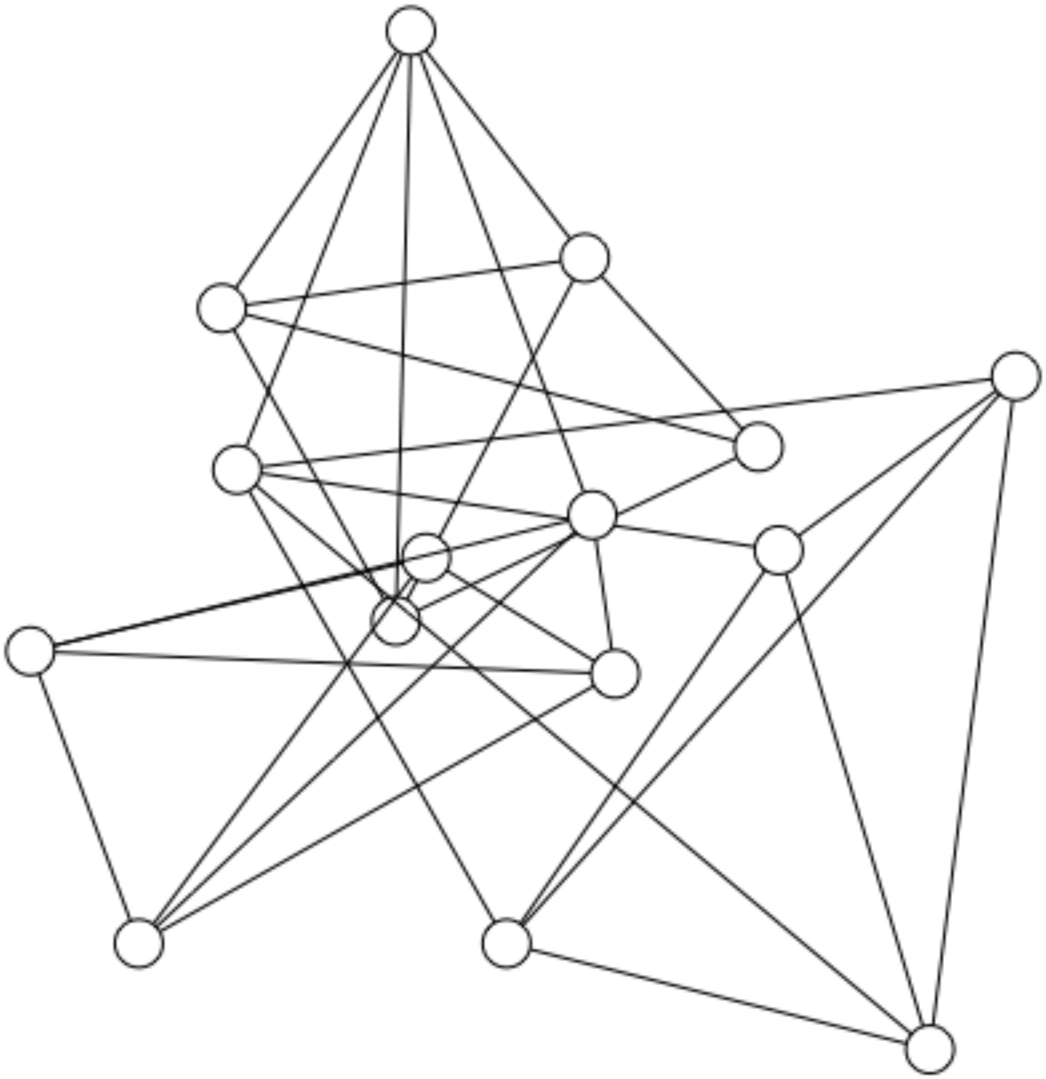
+



graphs + clustering = a lot to talk about!



Where are the clusters? Obvious?

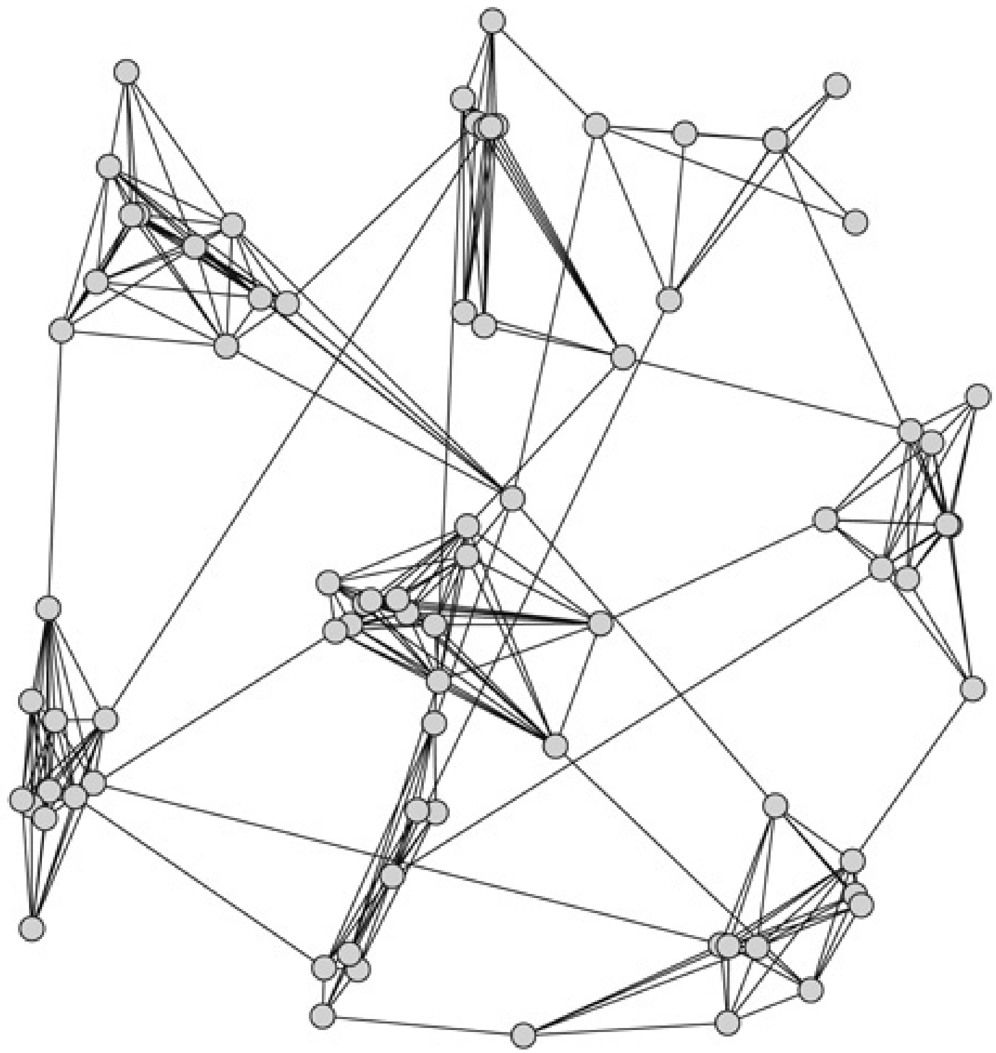


Where are the clusters? Not obvious?

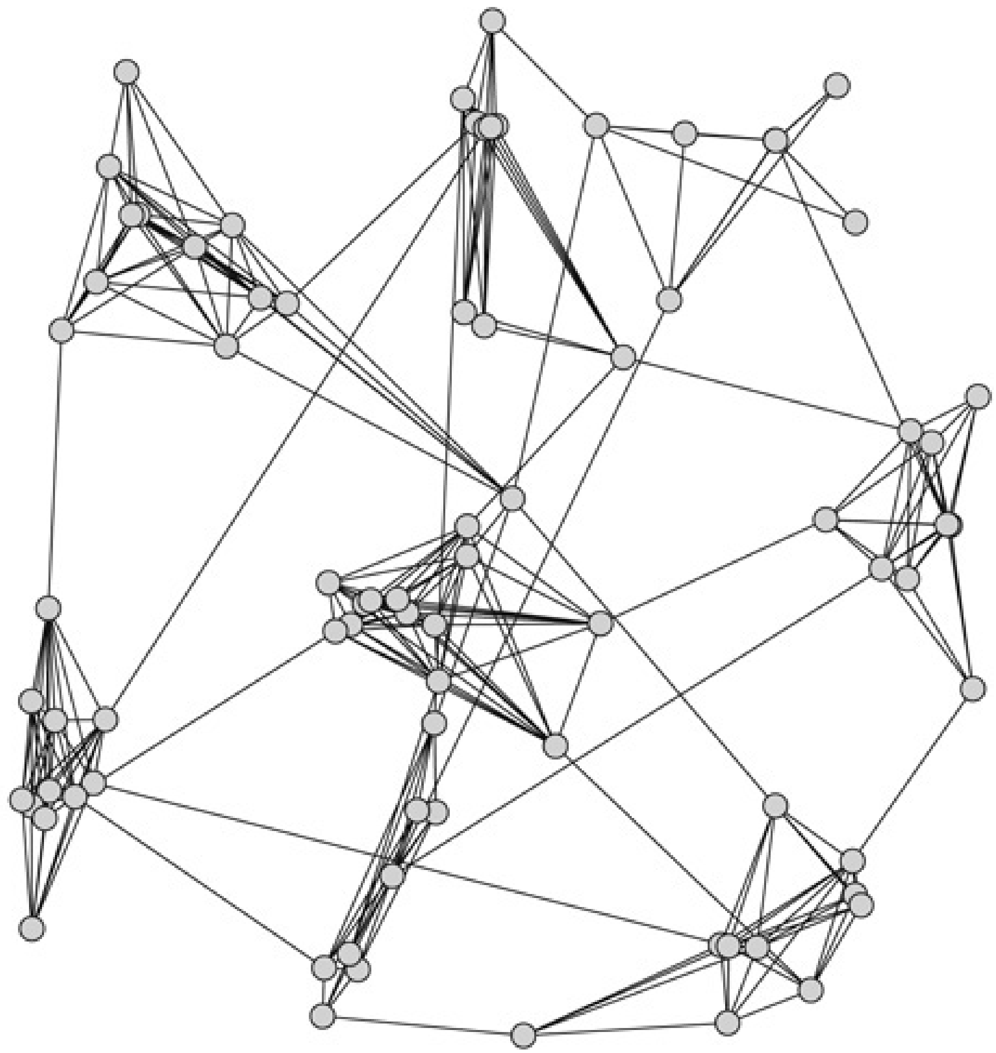
Diagram from: Graph clustering, Satu Elisa Schaeffer COMPUTER SCIENCE REVIEW 1 (2007) 27–64

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

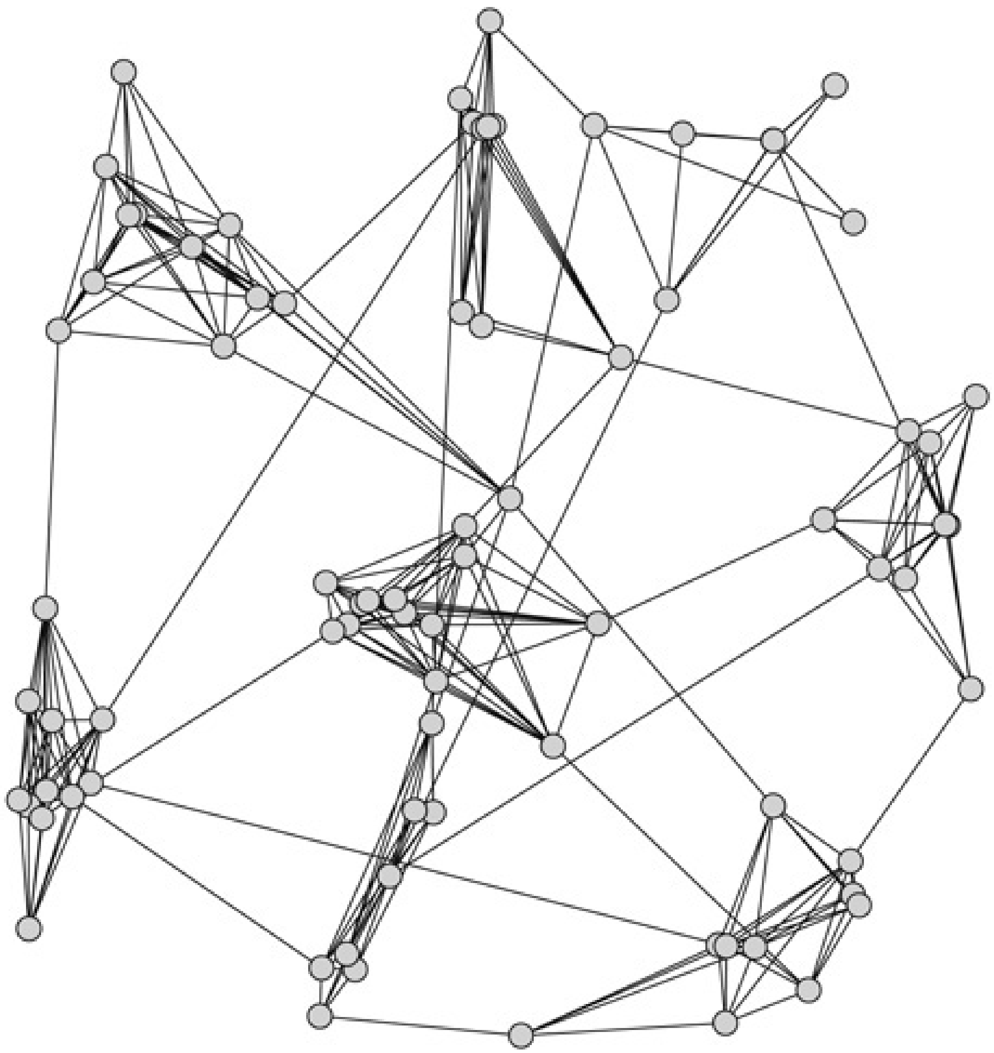
high density separated by low density



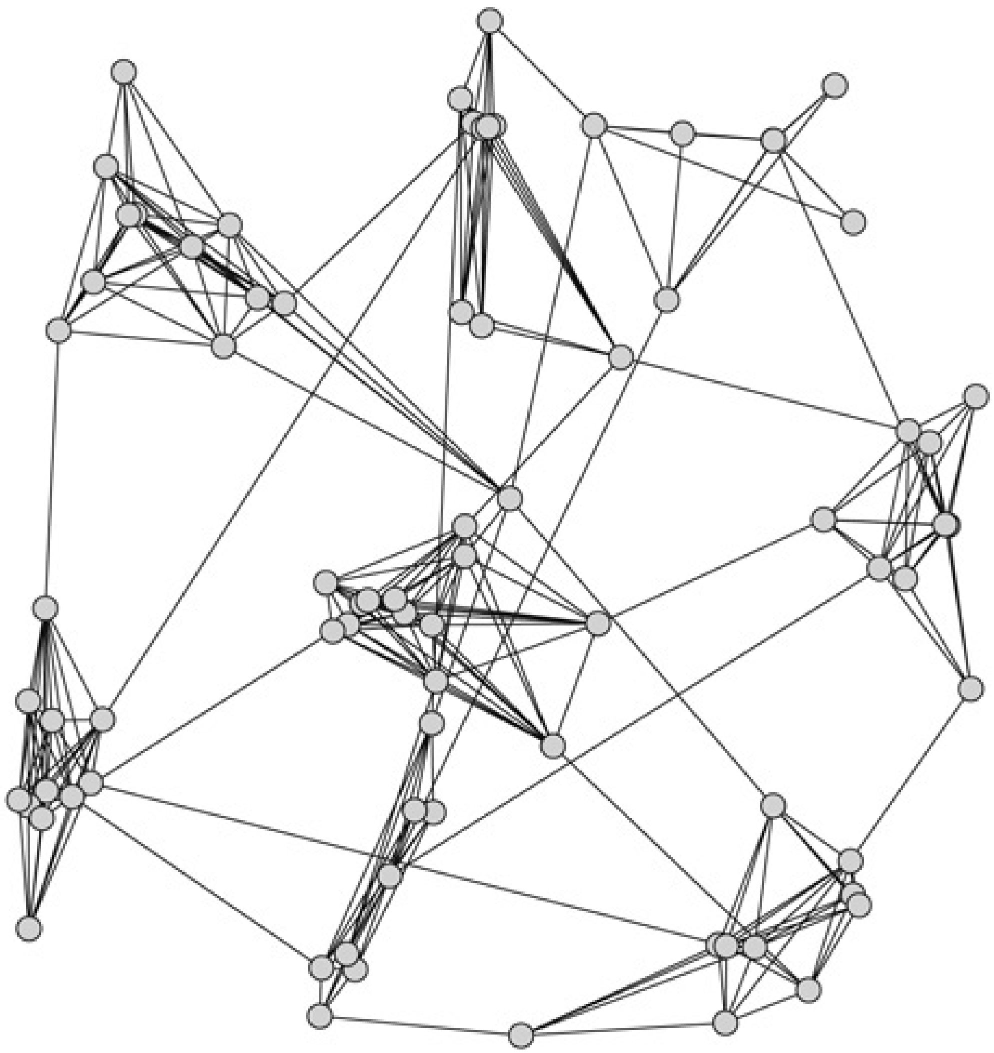
What if this graph represented: a road network?



What if this graph represented: flight paths?



What if this graph represented: neuron connections?

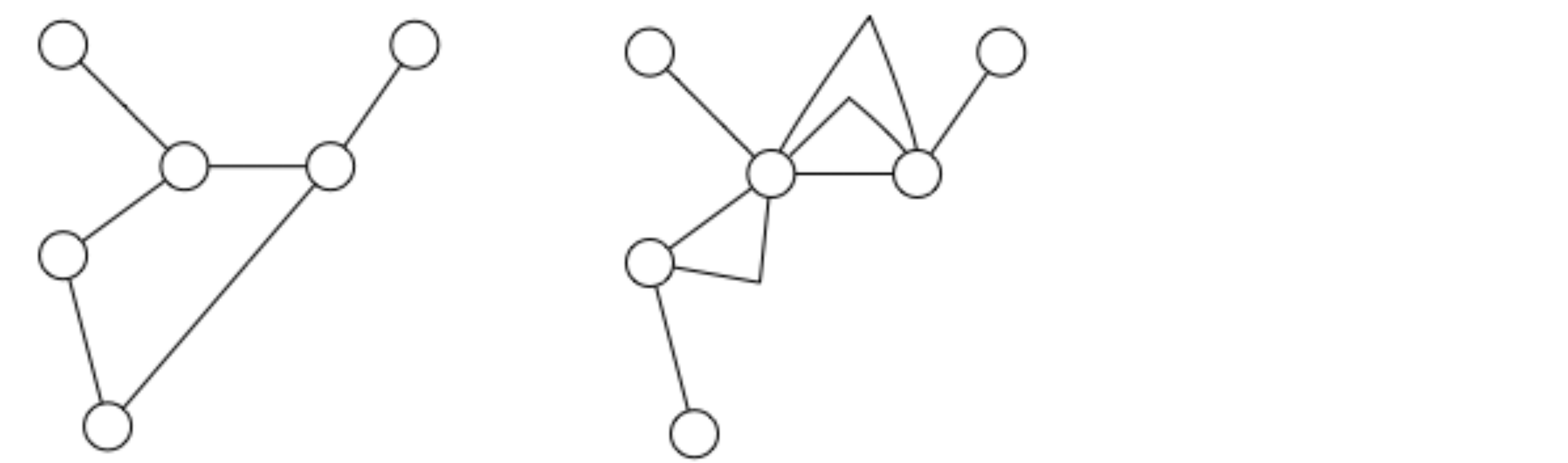


What if this graph represented: people’s friends?

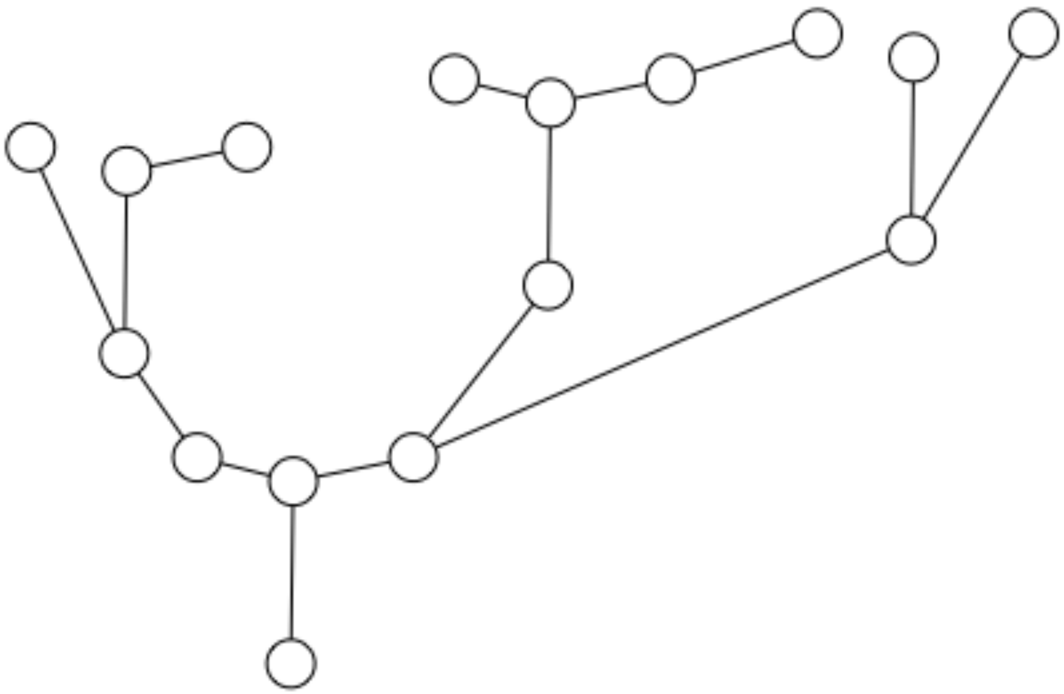
PRESS FOR CLUSTERING

Clustering: The ‘Dark Art’

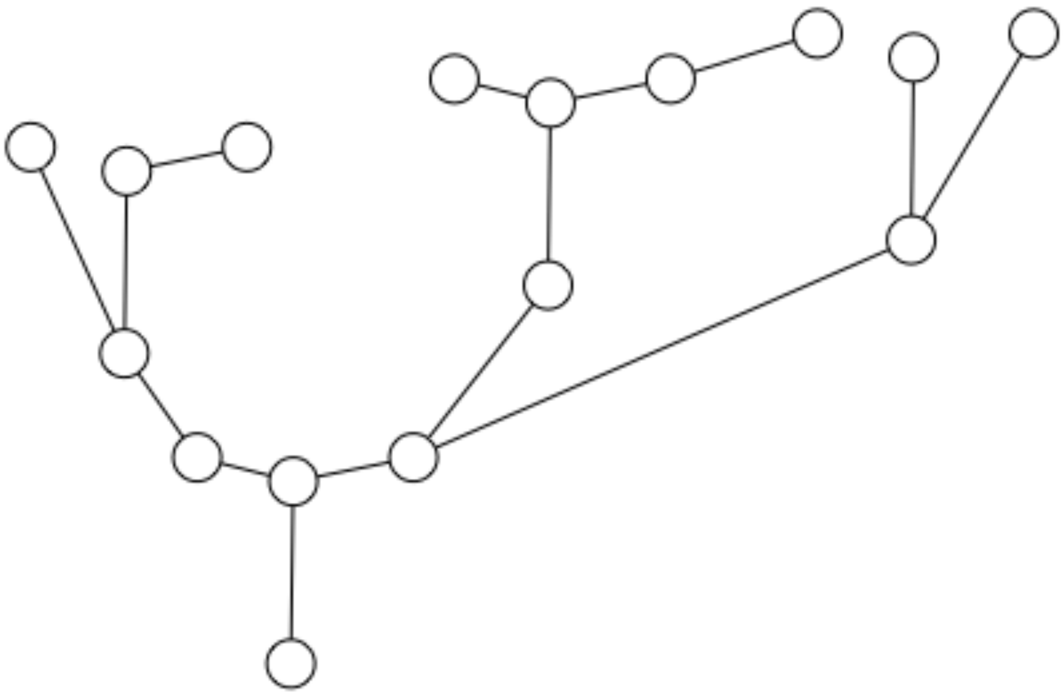
- **No agreed upon definition of a cluster**
- Different algorithms use different definitions
- Algorithms may not be deterministic (due to NP hard problem)



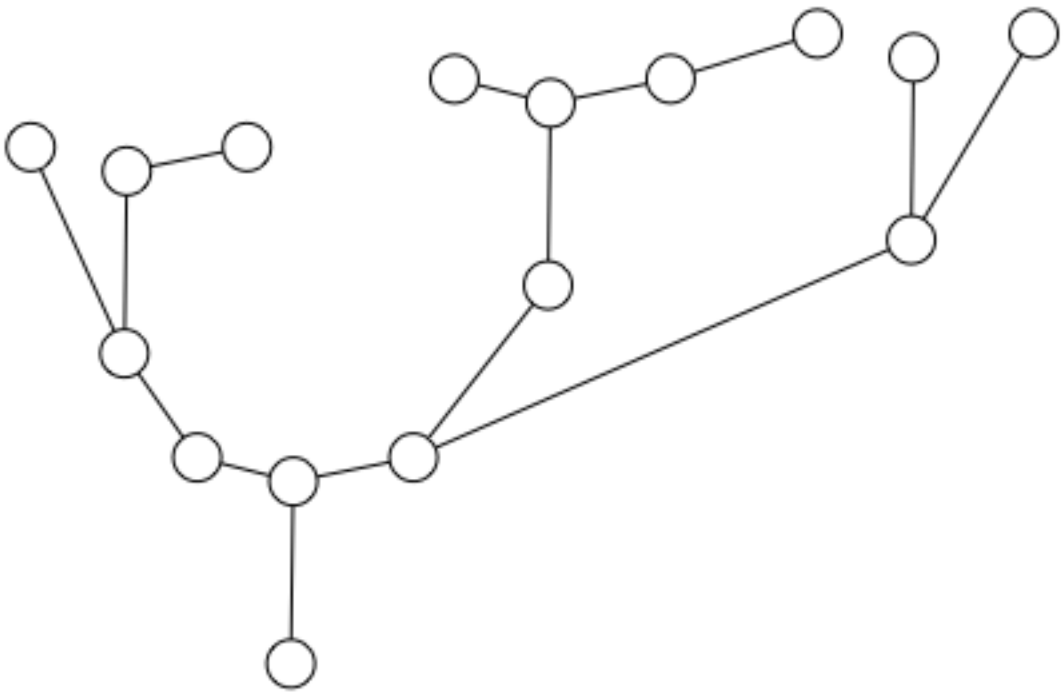
Some basics (not too poetic yet)



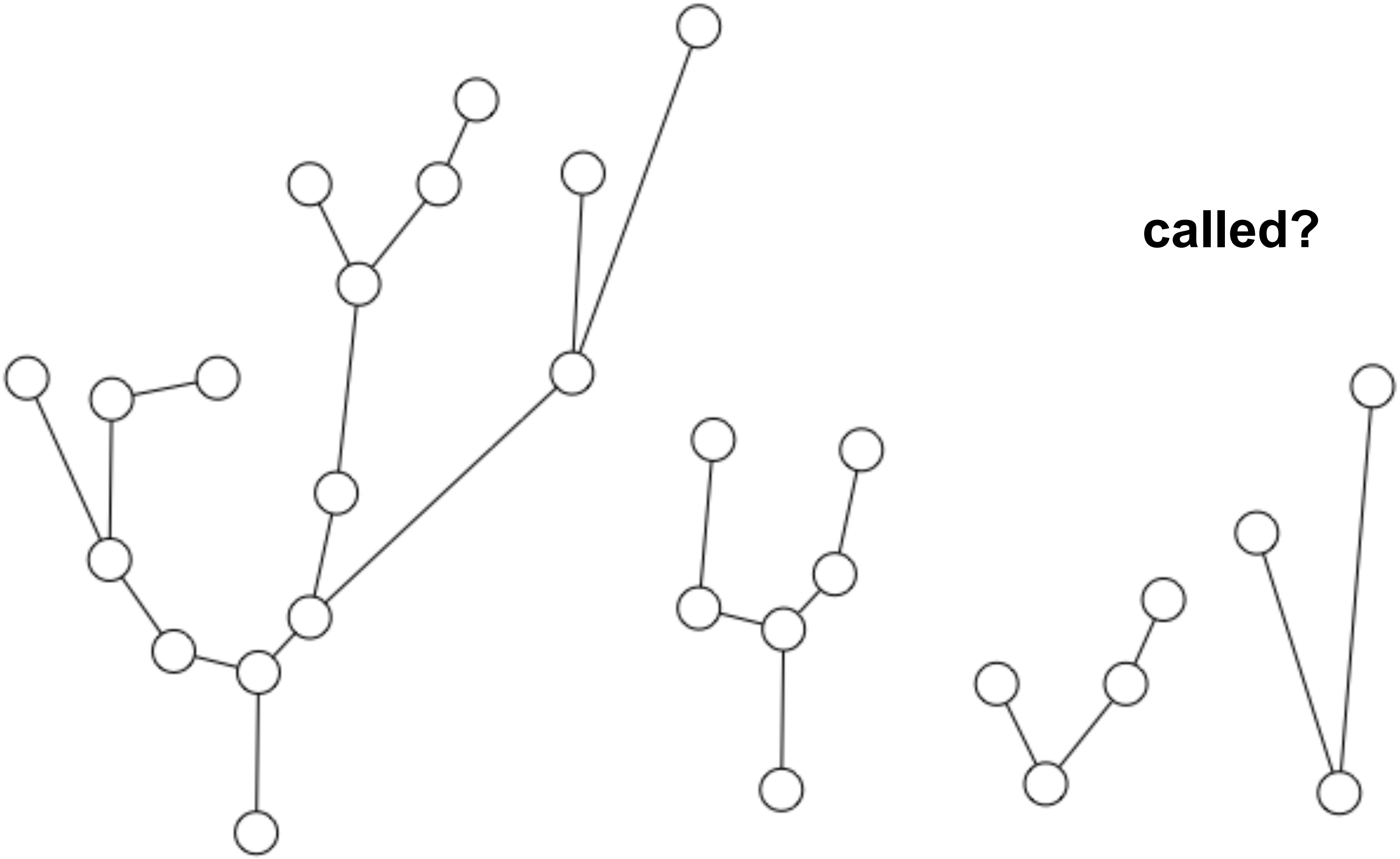
acyclic graph



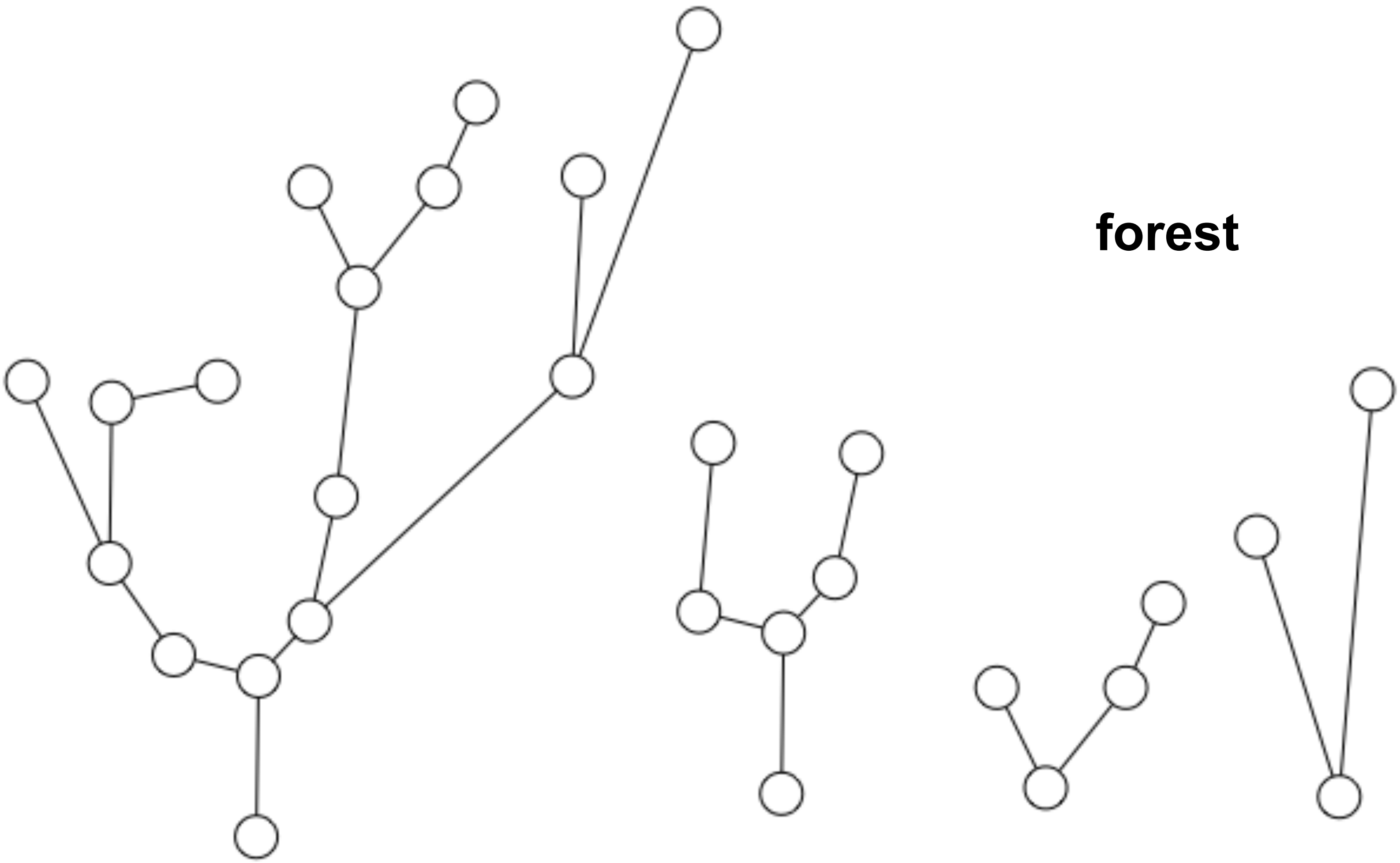
also called?

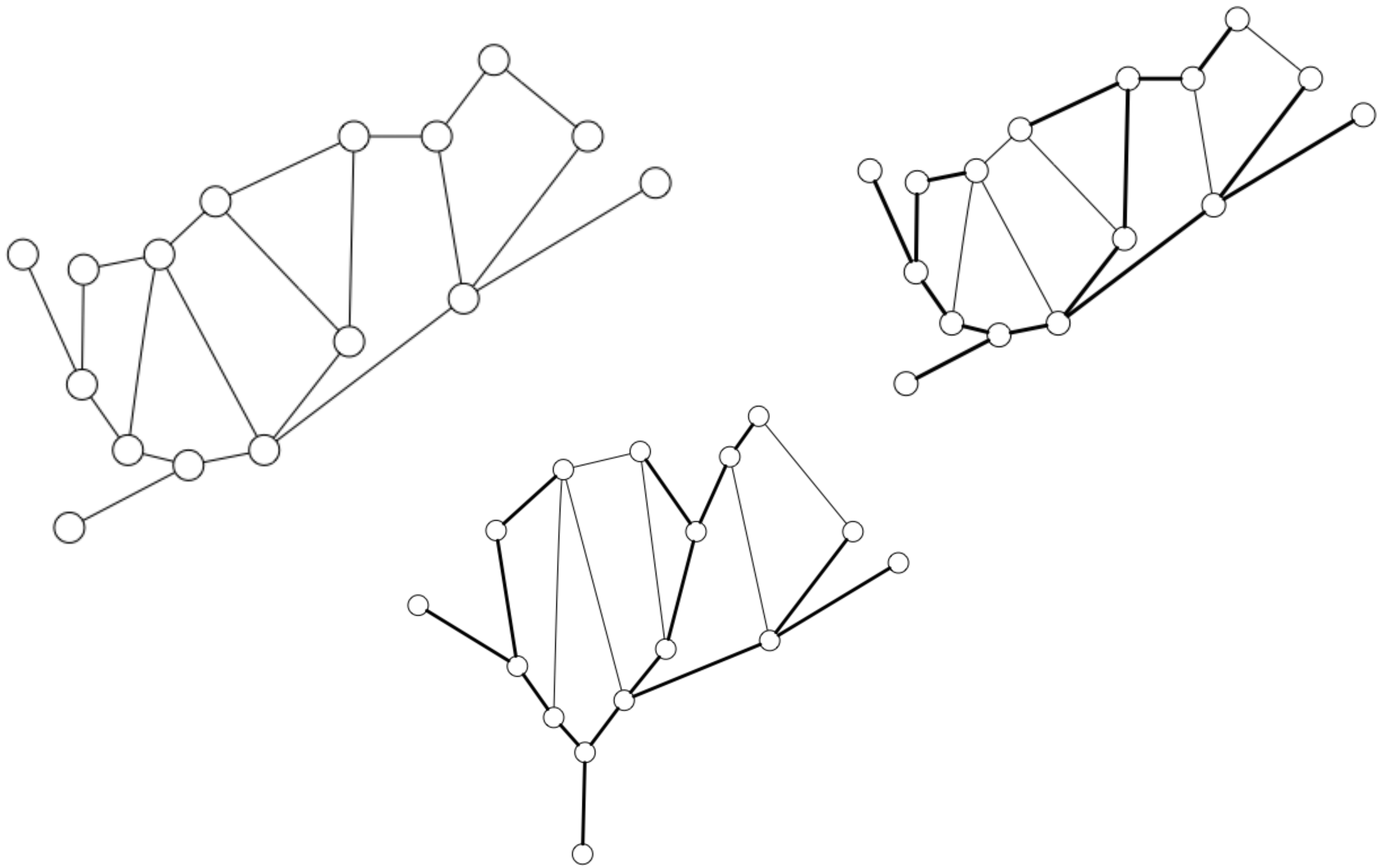


tree

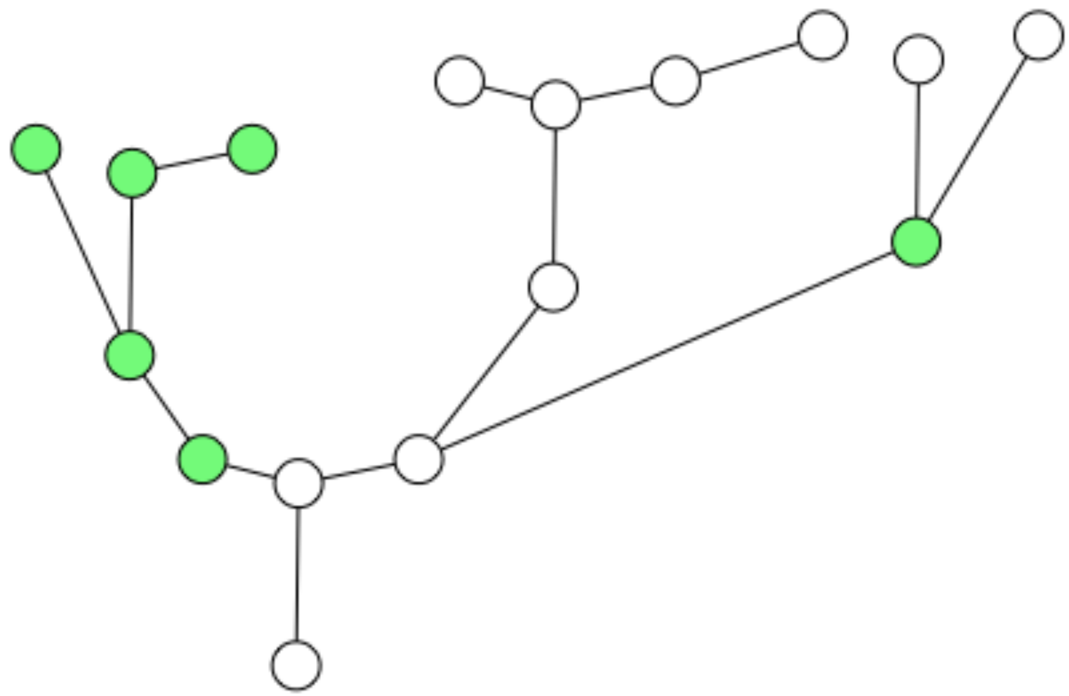


called?

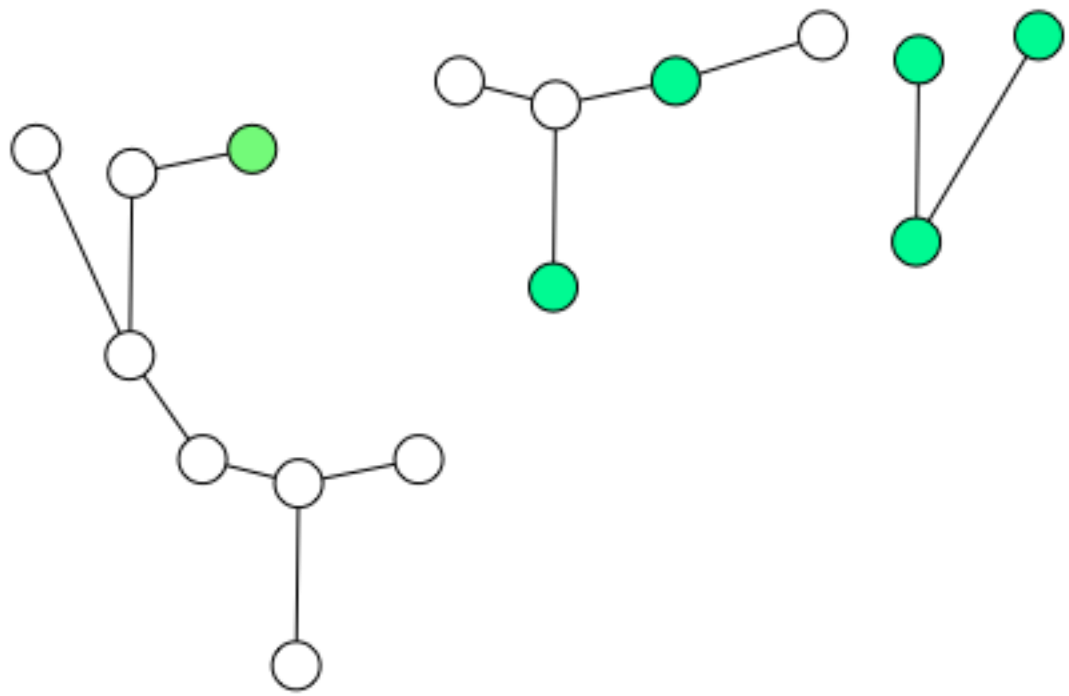

forest



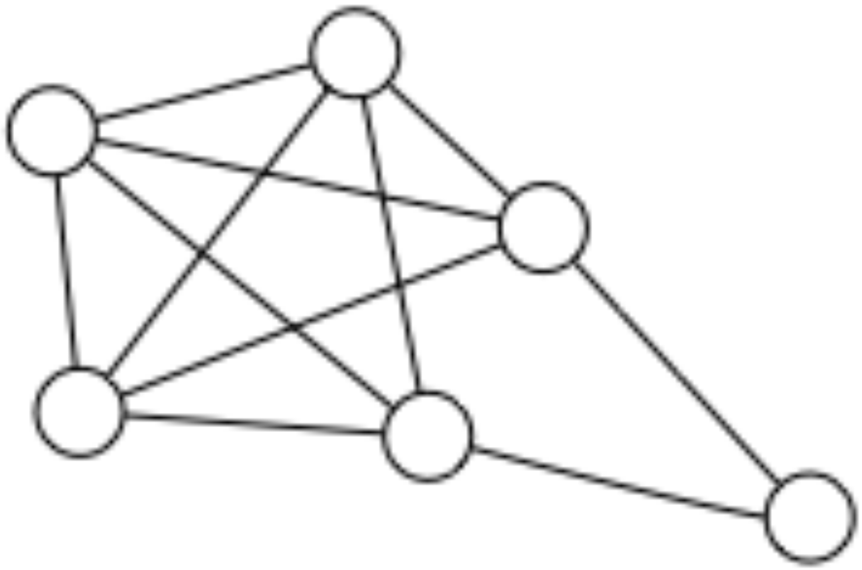
a spanning tree



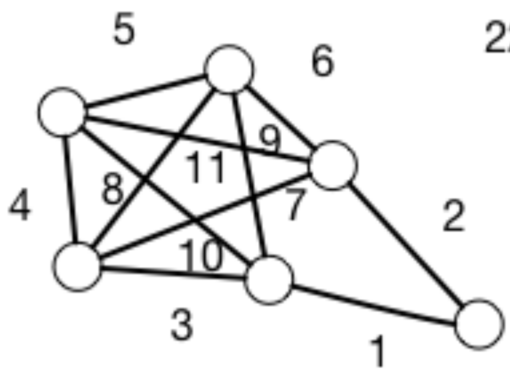
subgraph



also a
subgraph

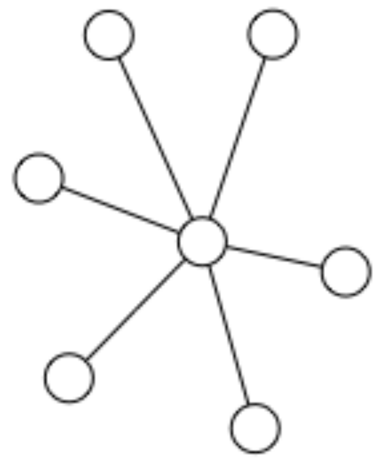


density = number of connections/total number of possible connections

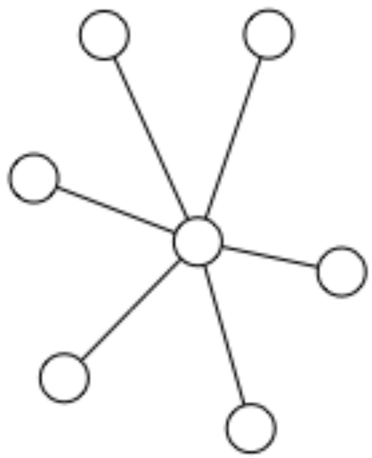


22 connections (both directions)/30 possible connections
= 0.73 density

density = number of connections/total number of possible connections

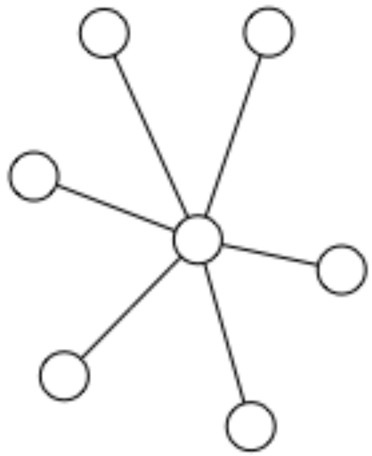


star



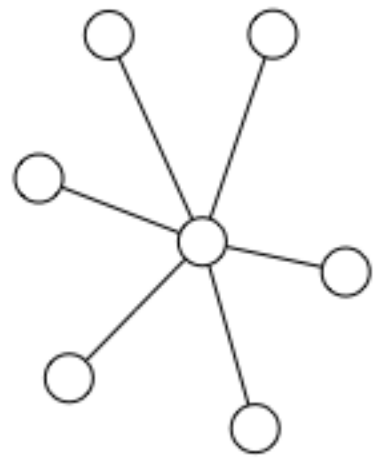
star

is a type of…?

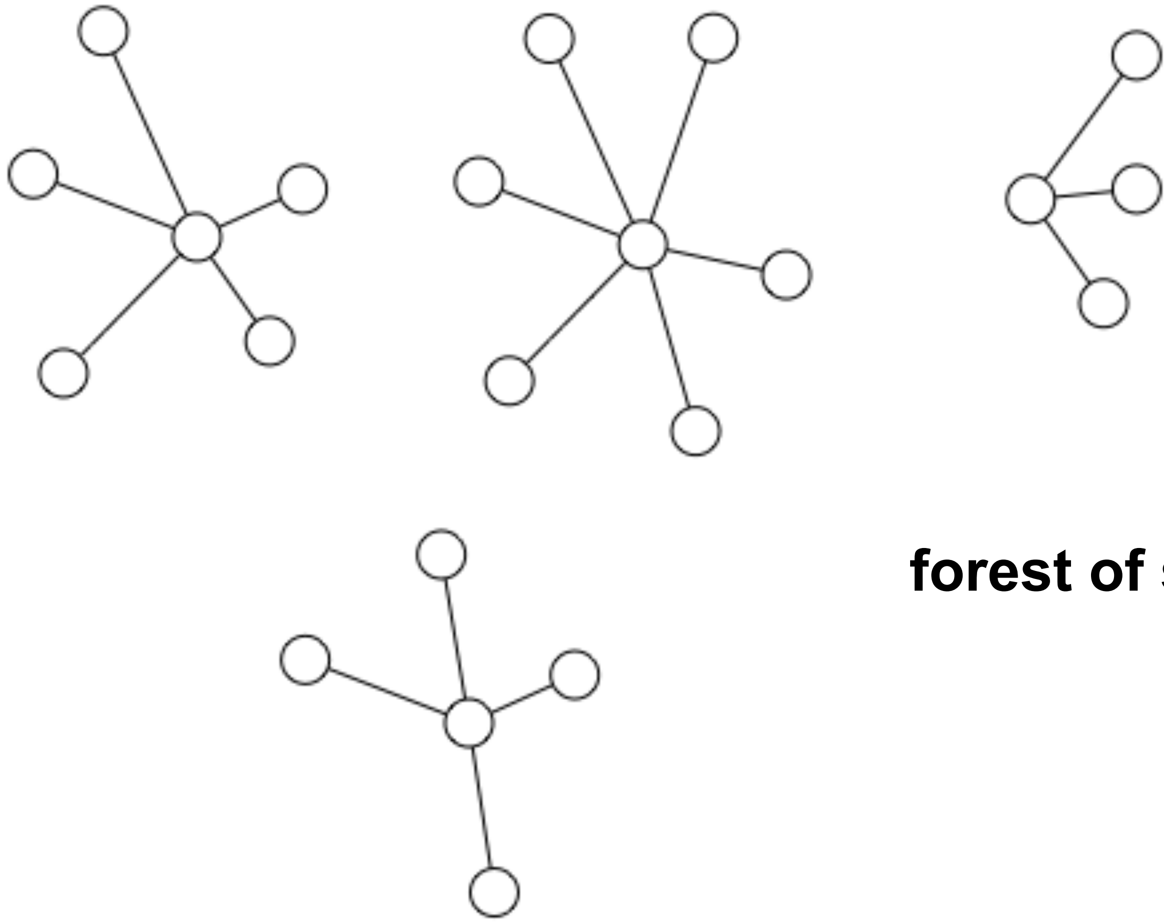


star

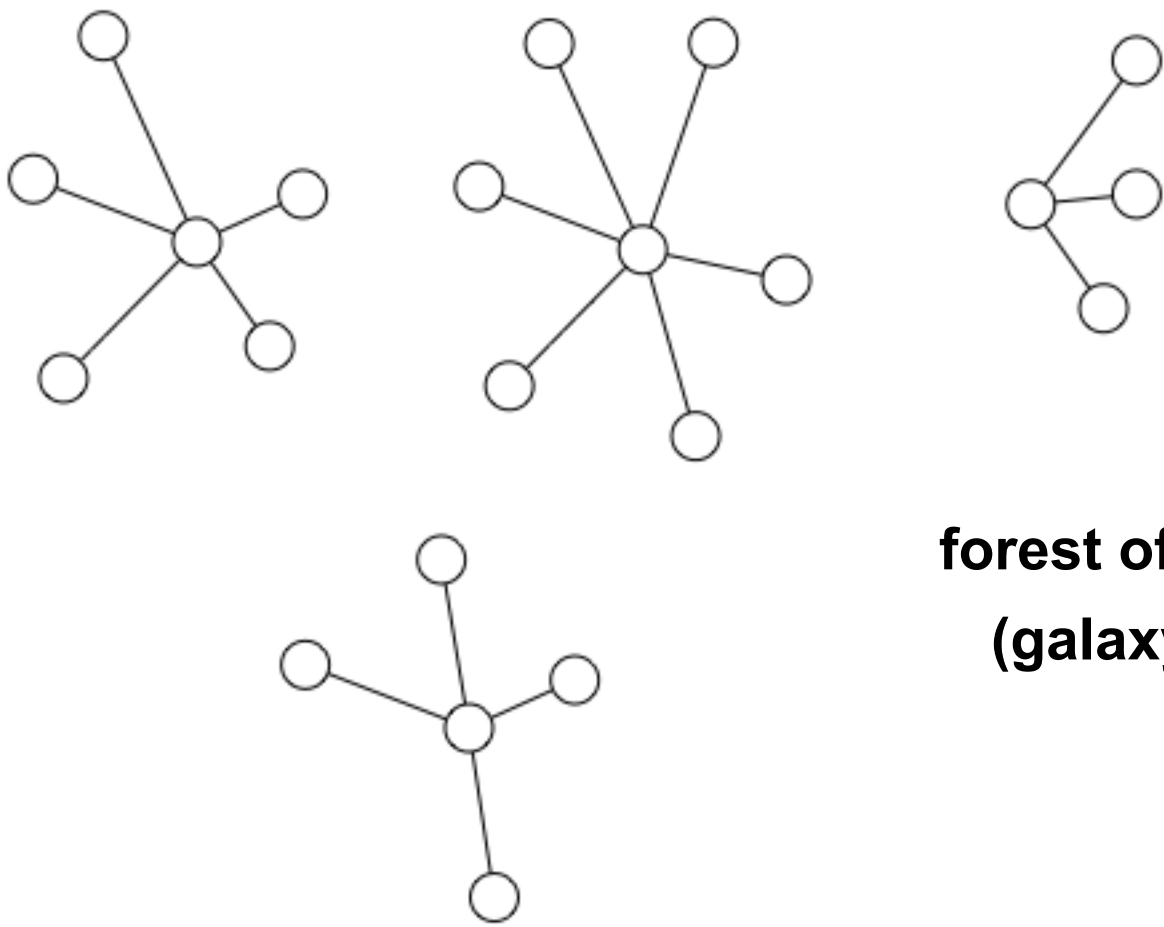
(tree)



edge-graceful star



forest of stars



forest of stars

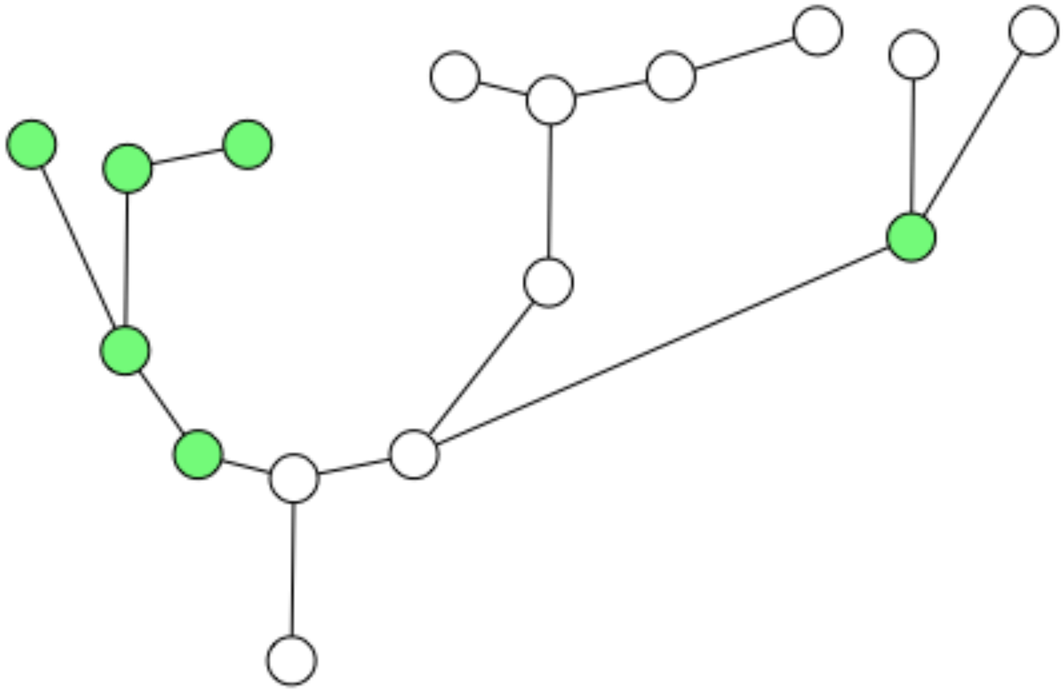
(galaxy)

“Star arboricity is the minimum number of forests that a graph can be partitioned into such that each tree in each forest is a star”

rather poetic…

"Star arboricity is the minimum number of forests that a graph can be **partitioned** into such that each tree in each forest is a star"

rather poetic…



partitioned: divided into sub groups

$$D = \frac{2|E|}{|V|\left(|V| - 1\right)}$$

Είναι όλα ελληνικά για μένα

a note about mathematical notation

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

(Translation: graph density equals the number of graph edges, considered in both directions, divided by the total possible number of edges that the graph could in theory have, which is, combinatorally, the number of vertices times the number of vertices minus 1)

Είναι όλα ελληνικά για μένα

(Translation: It's all greek to me)

a note about mathematical notation

A *graph* G is a pair of sets $G = (V, E)$. V is the set of *vertices* and the number of vertices $n = |V|$ is the *order* of the graph. The set E contains the *edges* of the graph. In an *undirected graph*, each edge is an unordered pair $\{v, w\}$. In a *directed* graph (also called a *digraph* in much literature), edges are ordered pairs. The vertices v and w are called the *endpoints* of the edge. The edge count $|E| = m$ is the *size* of the graph. In a *weighted graph*, a weight function $\omega : E \to \mathbb{R}$ is defined that assigns a weight on each edge. A graph is *planar* if it can be drawn in a plane without any of the edges crossing.

$$\delta(G) = \frac{m}{\binom{n}{2}}.$$

$$a^G_{v,u} = \begin{cases} 1, & \text{if } \{v, u\} \in E, \\ 0, & \text{otherwise.} \end{cases}$$

The number of edges incident on a given vertex v is the *degree* of v and is denoted by $\deg(v)$. A graph is *regular* if all of the vertices have the same degree; if $\forall v \in V$ in $G = (V, E)$ we have $\deg(v) = k$, the graph G is k-regular. The diagonal *degree matrix* of a graph $G = (V, E)$ is

$$D = \begin{pmatrix} \deg(v_1) & 0 & 0 & \cdots & 0 & 0 \\ 0 & \deg(v_2) & 0 & \cdots & 0 & 0 \\ 0 & 0 & \deg(v_3) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \deg(v_{n-1}) & 0 \\ 0 & 0 & 0 & \cdots & 0 & \deg(v_n) \end{pmatrix}. \quad (5)$$

A *path* from v to u in a graph $G = (V, E)$ is a sequence of edges in E starting at vertex $v_0 = v$ and ending at vertex $v_{k+1} = u$;

$$\{v, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_k\}, \{v_k, u\}. \quad (8)$$

The *cut size* is the number of edges that connect vertices in S to vertices in $V \setminus S$:

$$c(S, V \setminus S) = |\{\{v, u\} \in E \mid u \in S, v \in V \setminus S\}|. \quad (6)$$

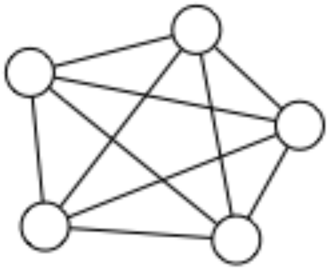
We denote by

$$\deg(S) = \sum_{v \in S} \deg(v) \quad (7)$$

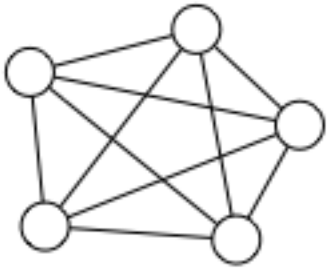
An *induced subgraph* of a graph $G = (V, E)$ is the graph with the vertex set $S \subseteq V$ with an edge set $E(S)$ that includes *all* such edges $\{v, u\}$ in E with both of the vertices v and u included in the set S:

$$E(S) = \{\{v, u\} \mid v \in S, u \in S, \{v, u\} \in E\}. \quad (9)$$

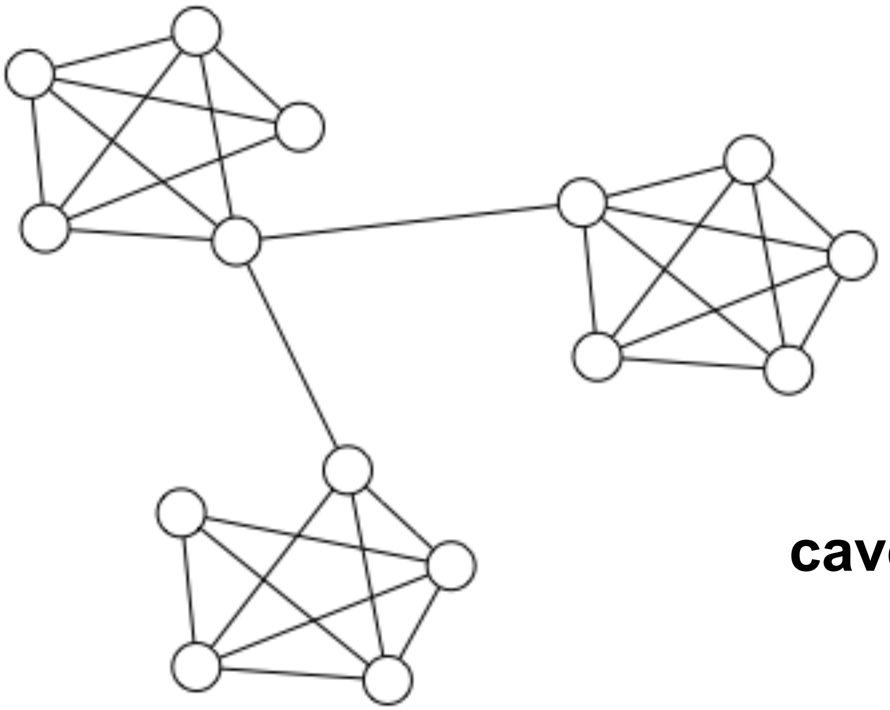
Graph clustering, Satu Elisa Schaeffer COMPUTER SCIENCE REVIEW 1 (2007) 27–64



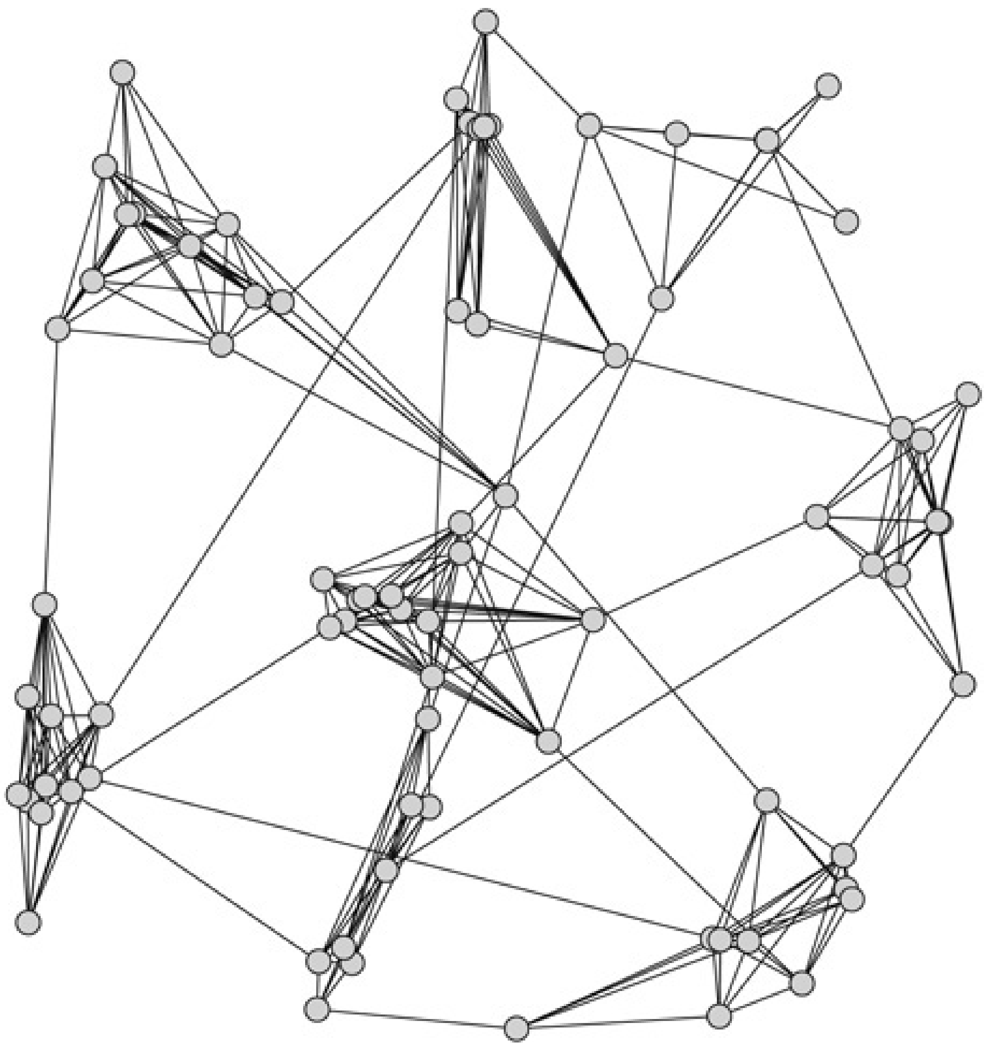
clique



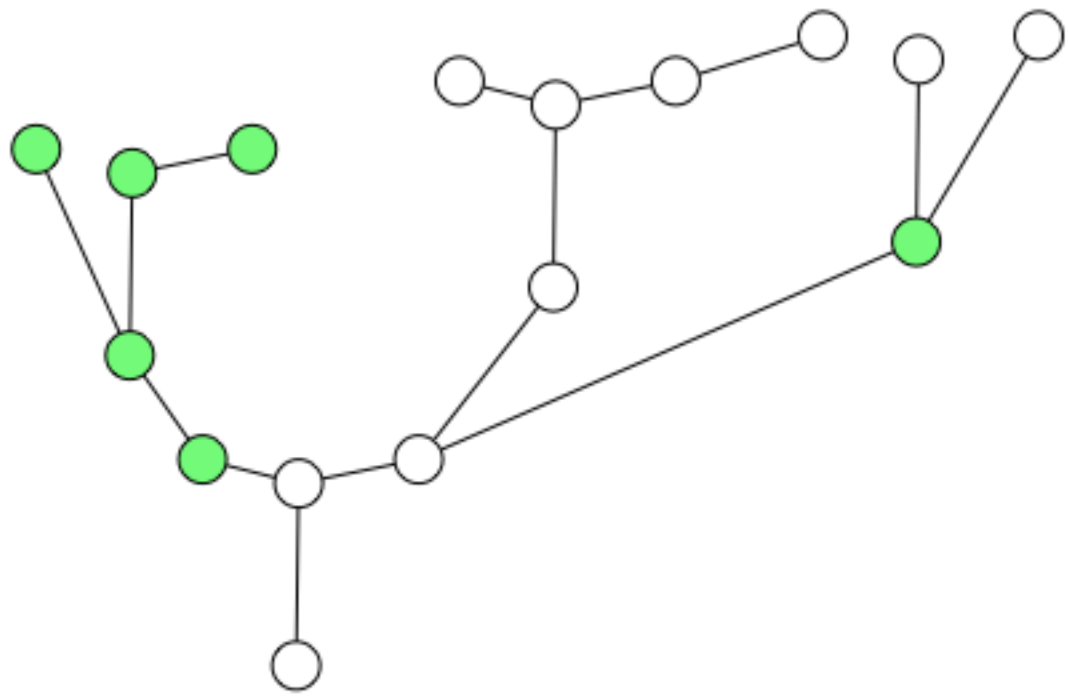
clique
(cave)



caveman graph



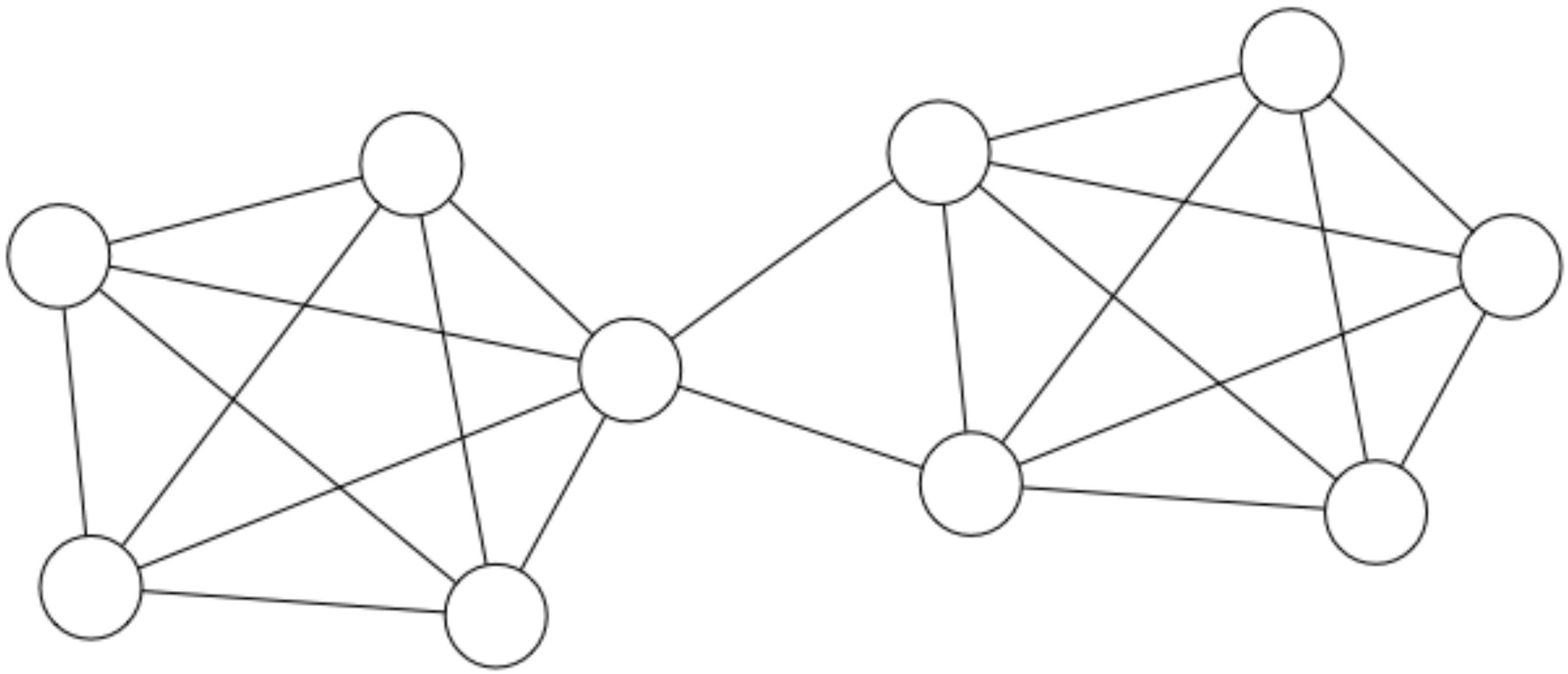
relaxed caveman graph



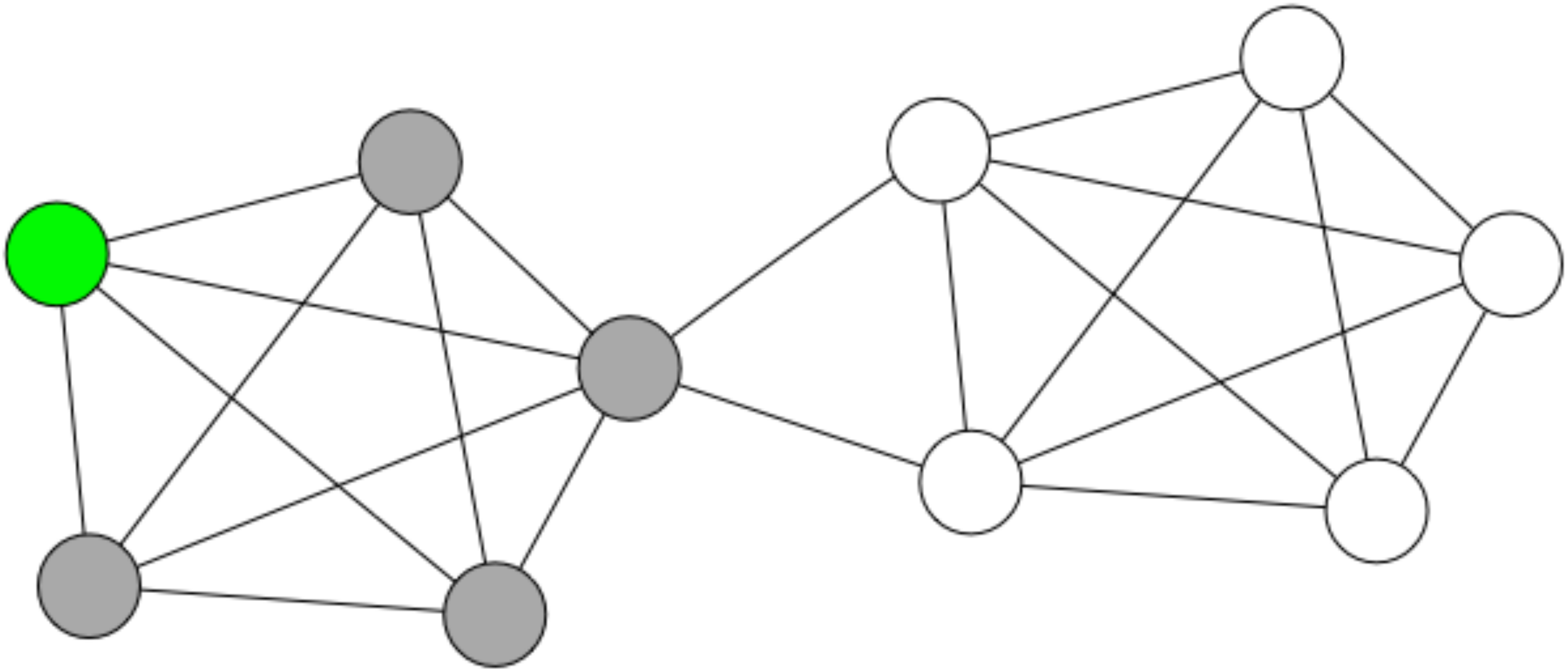
randomly ‘clustered’ (but not really *clustering*, per se)

Defining clusters

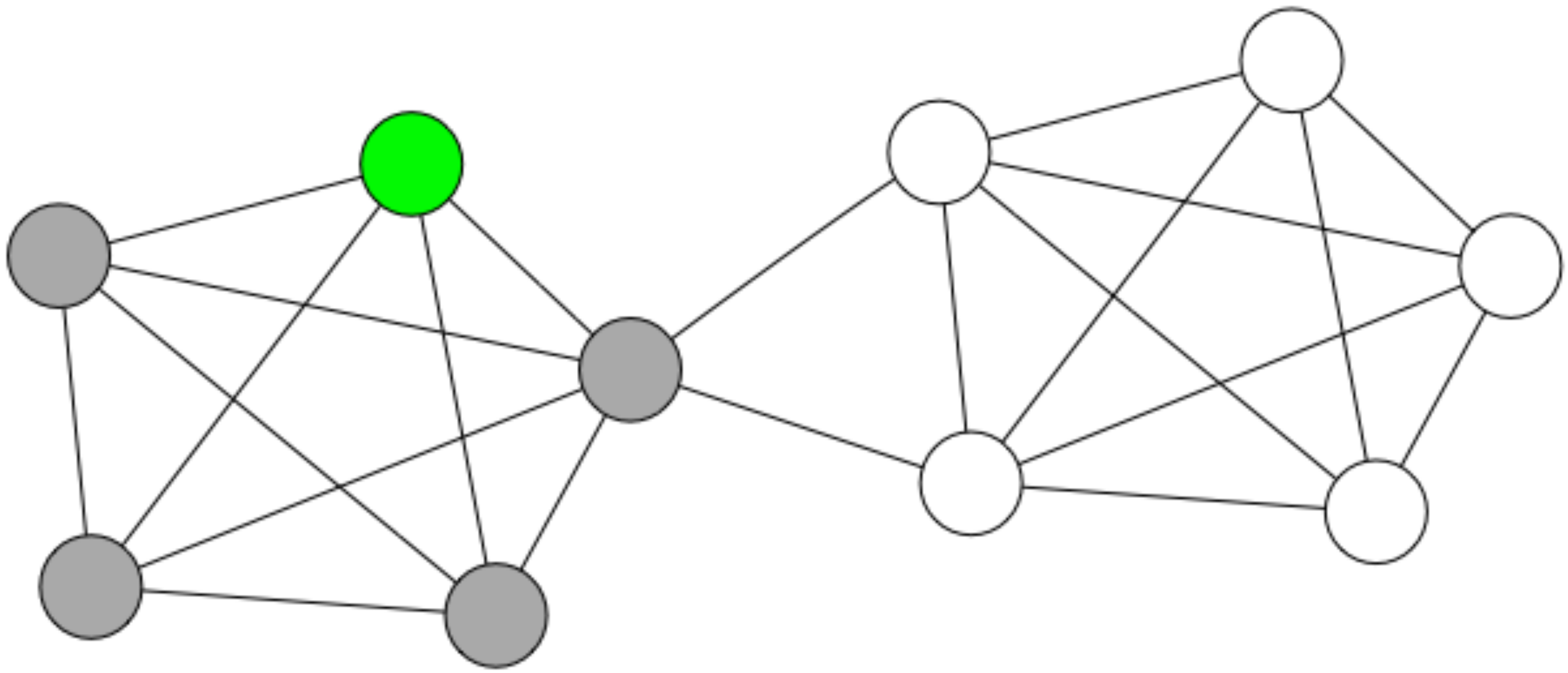
- **Partition based on:**
  - High-Low density
  - Shared neighbours
  - Traversal probability (random walk)
  - Cuts required to separate
  - Other?



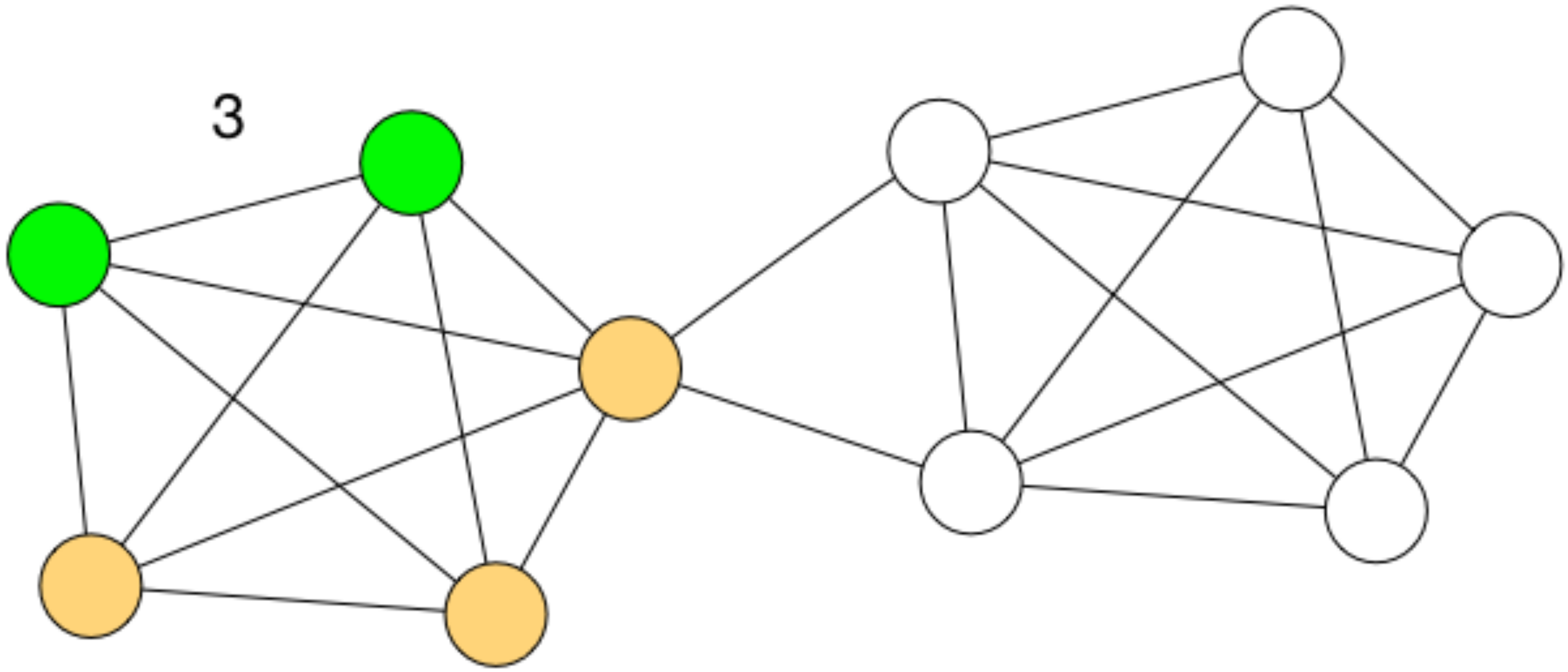
Calculate shared nearest neighbours



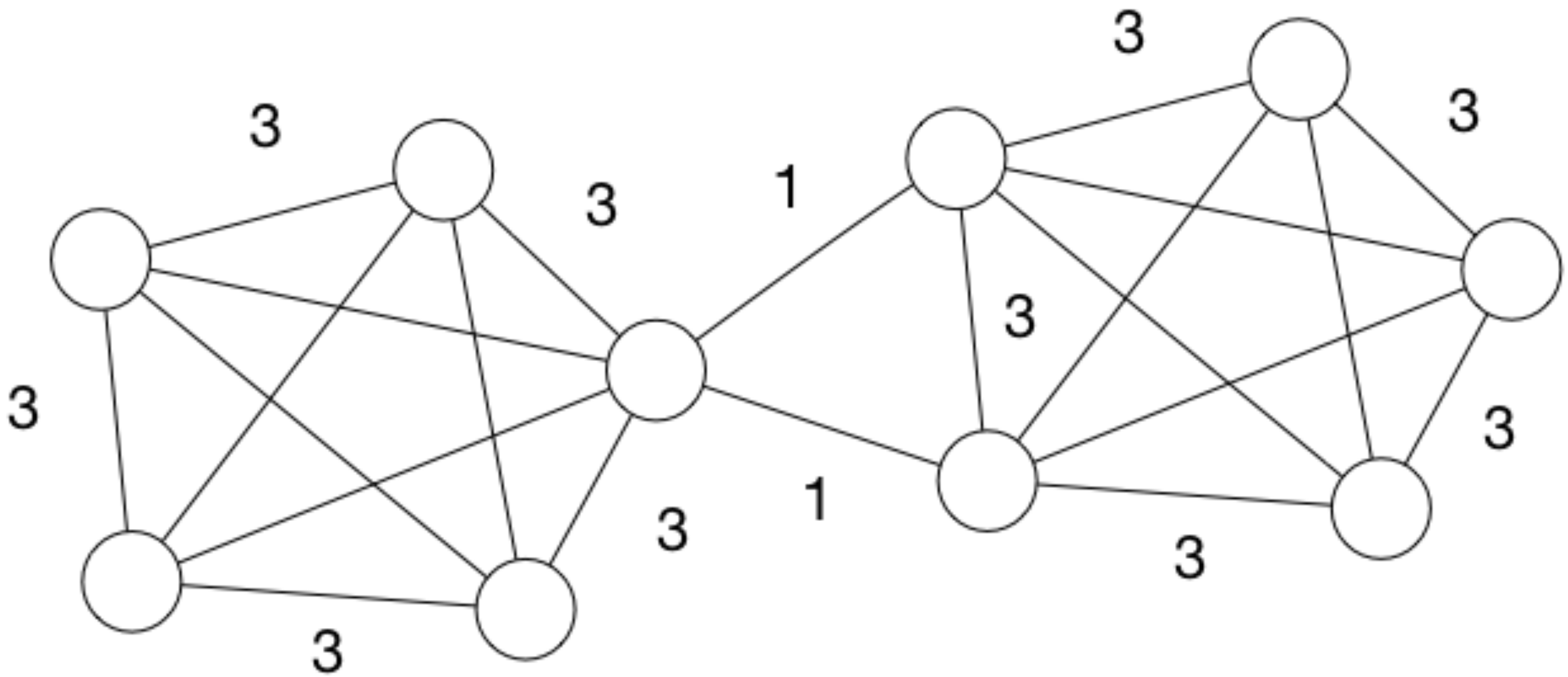
Calculate shared nearest neighbours



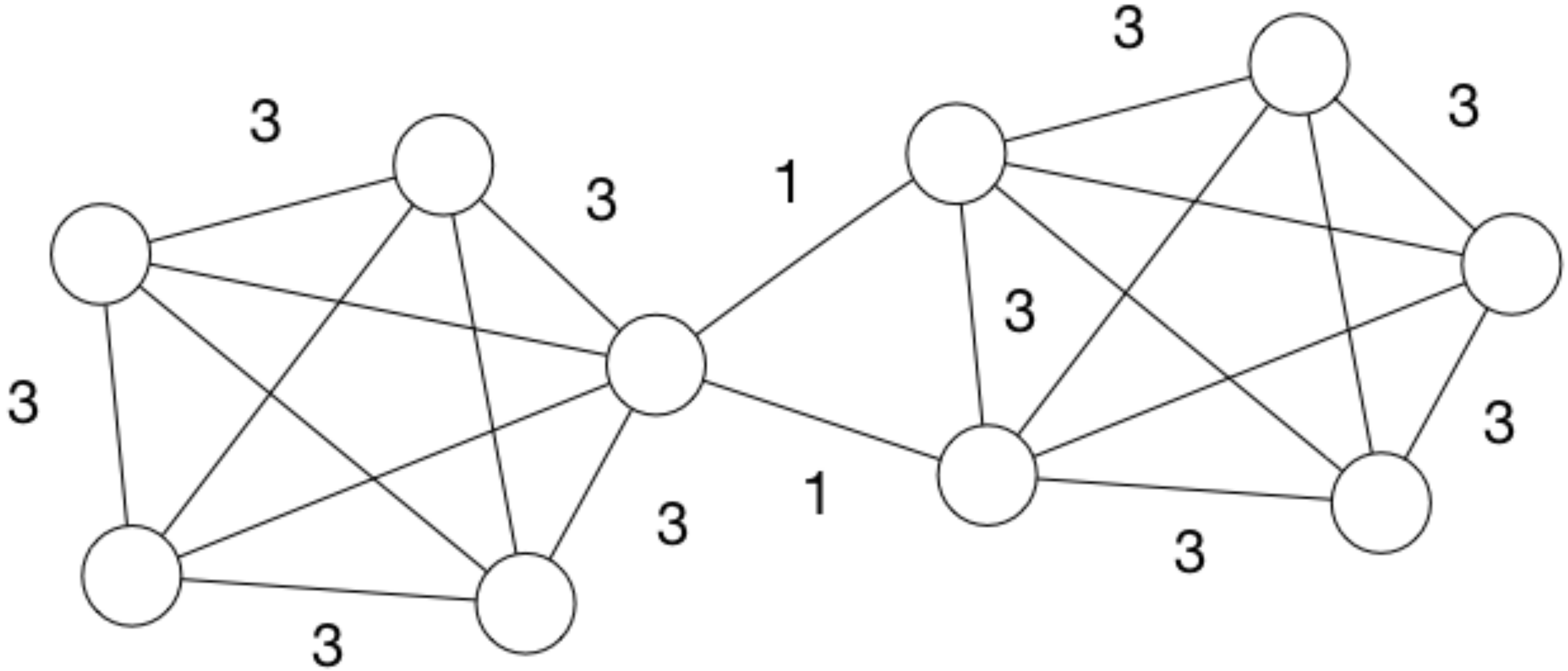
Calculate shared nearest neighbours



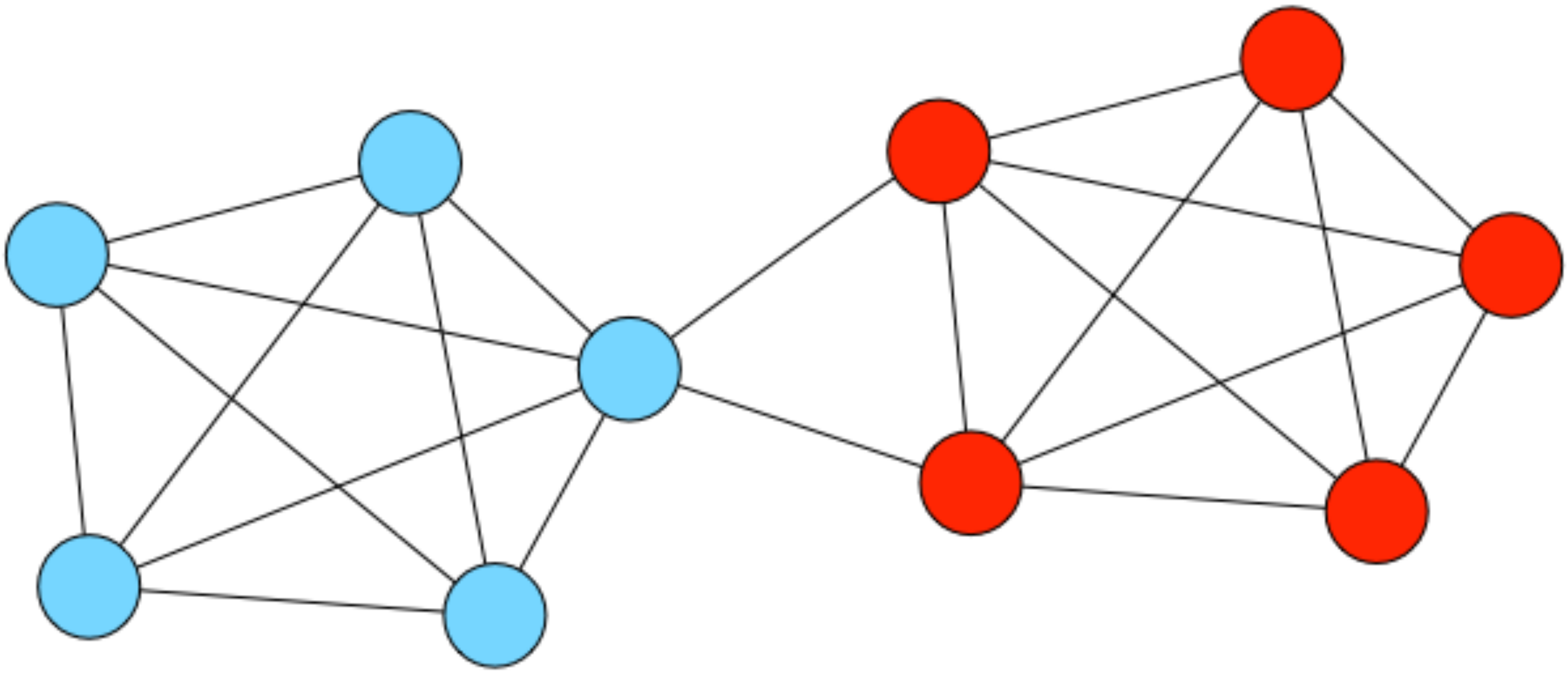
Calculate shared nearest neighbours



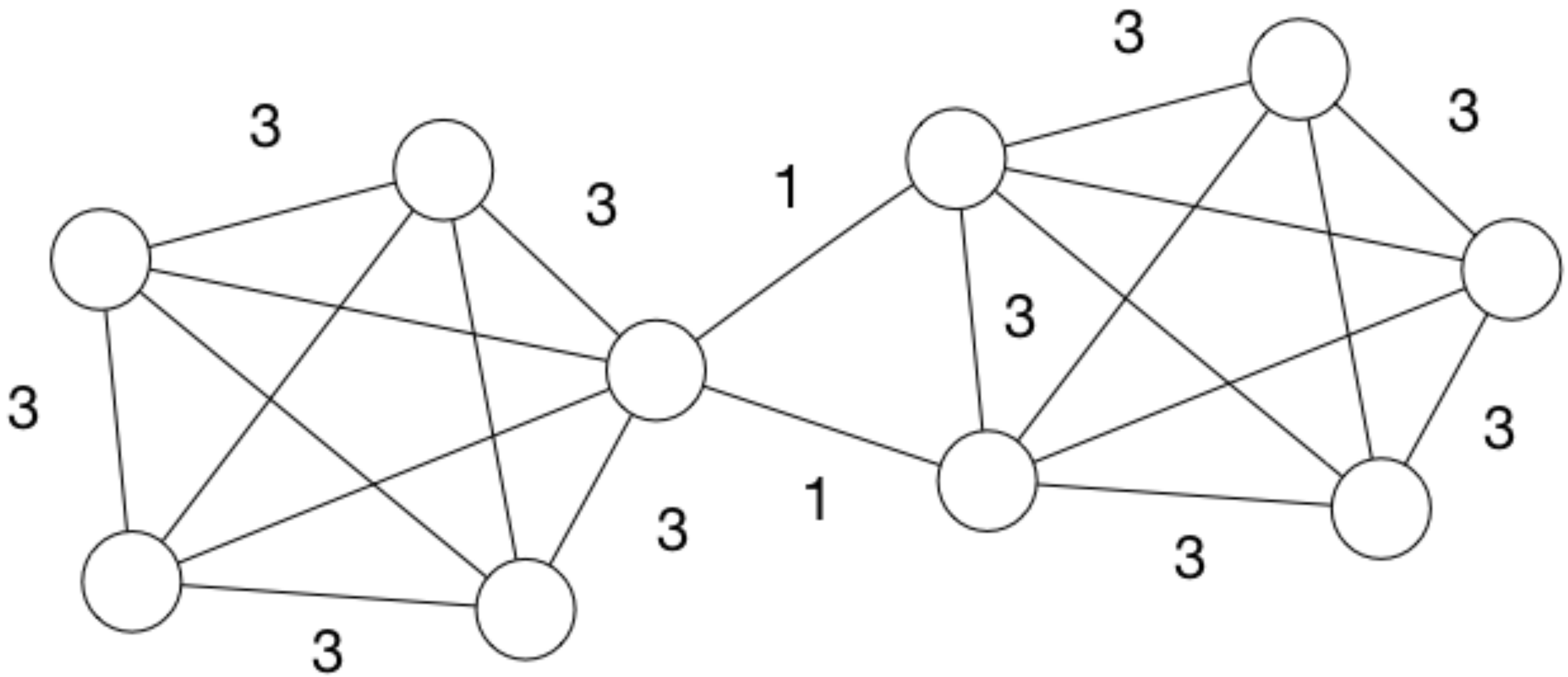
Calculate shared nearest neighbours



pick a threshold



create clusters based on threshold



which threshold to pick?

Some other clustering algorithms

- **k-Spanning Tree**
- **Betweenness Centrality Based**
- **Highly Connected Components**
- **Maximal Clique Enumeration**
- **The Markov Cluster Algorithm**
- **HCS algorithm**

QUESTIONS

?