Social (complex) Networks Analysis

COMPLEX NETWORKS BASICS

Rushed Kanawati

kanawati@sorbonne-paris-nord.fr

https://kanawati.fr

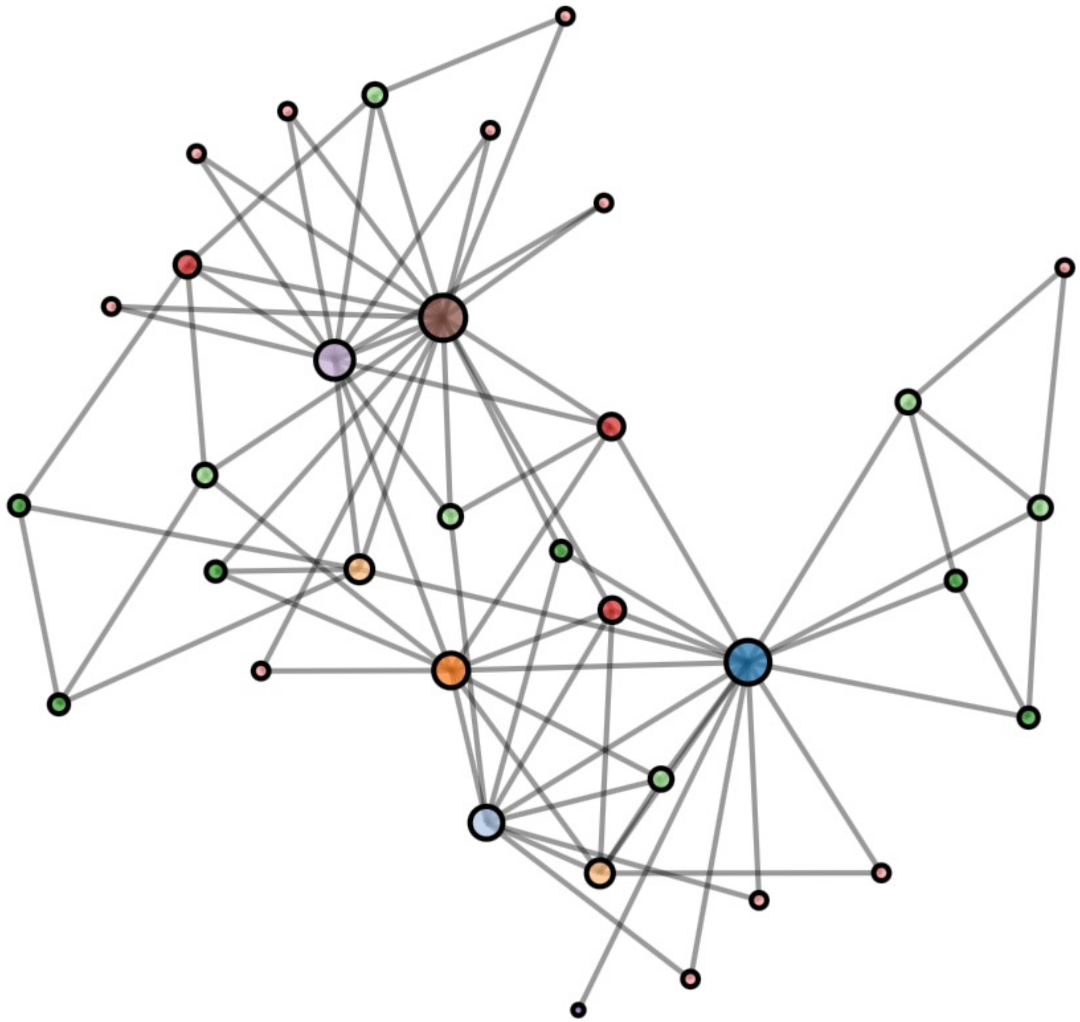
Université Sorbonne Paris Nord

Plan

1. (SOCIAL) INTERACTION NETWORKS
2. BACKGROUND: GRAPH THEORY
3. TOPOLOGICAL FEATURES OF COMPLEX NETWORKS
4. (RANDOM) NETWORKS MODELS
5. COMPLEX NETWORKS ANALYSIS TASKS

1

(Social) Interaction Networks



Interaction networks

Graphs modelling **direct**/**indirect** interactions among actors.

Direct interactions:

- Friendship
- Proximity
- Message exchange
- …

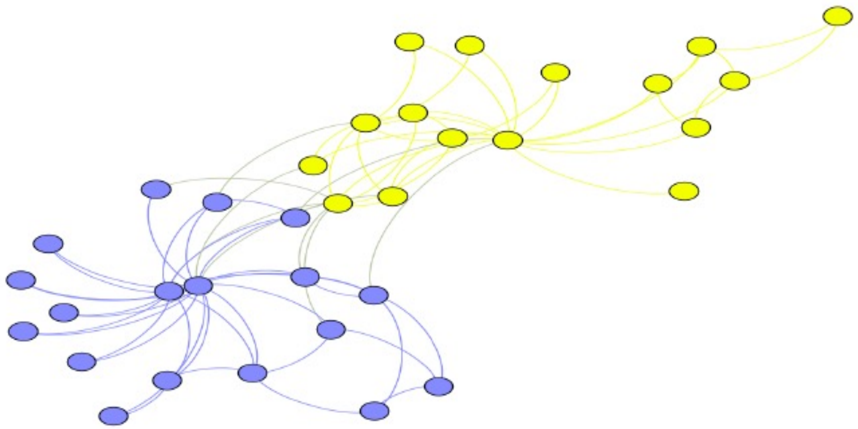
Indirect interactions:

- *Affiliation share*
- Preference share
- **Similarity**
- …

Social networks

Direct interaction

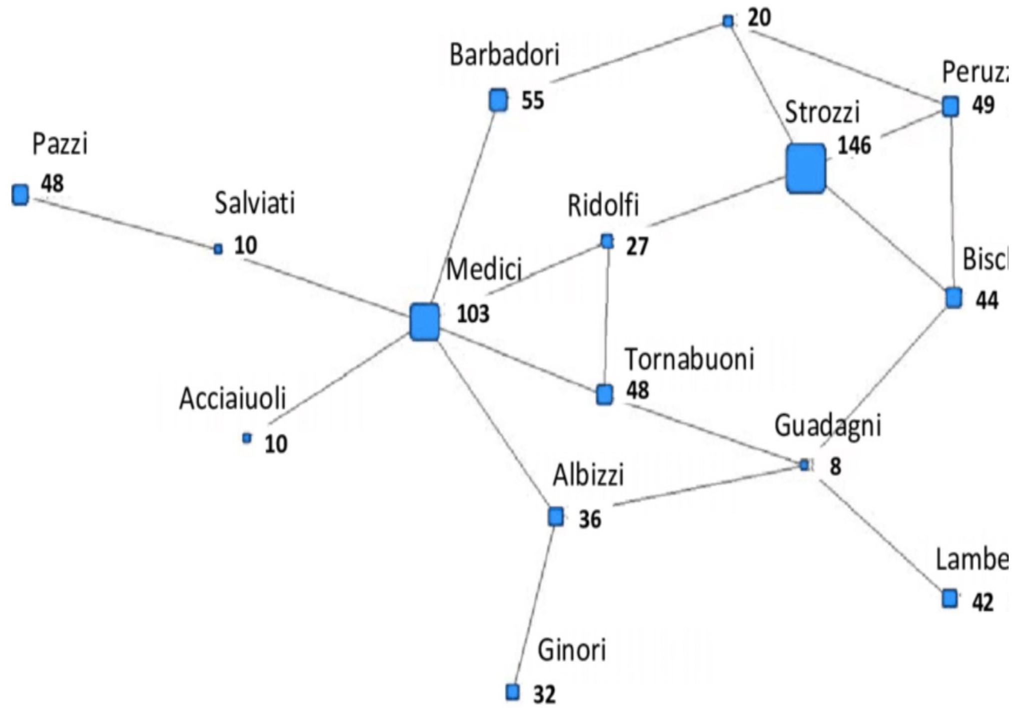
Zachary Karate Club is a friendship-based network observed in the context of a Karate Club that have been split into two clubs after a dispute between the manger and the coach.



Social networks

Direct interaction

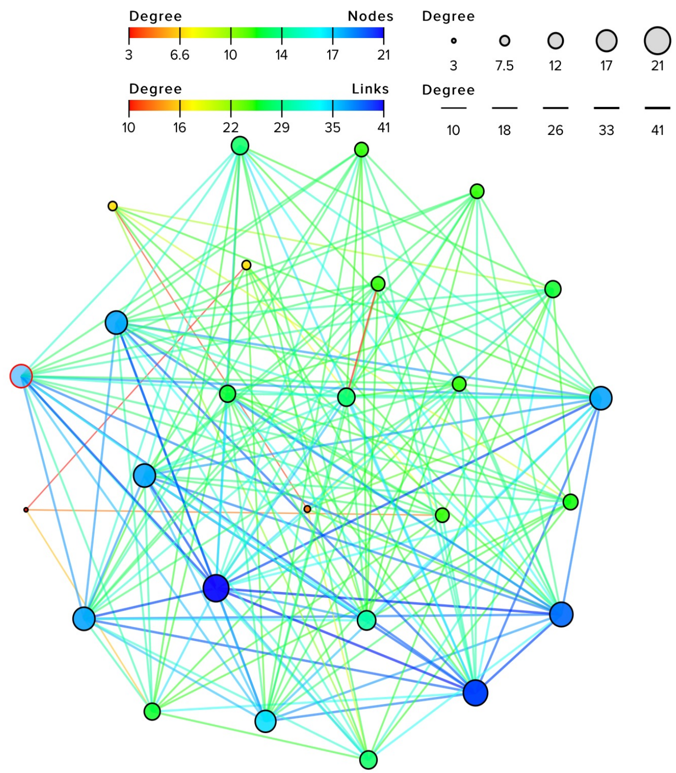
Florentine families network : Marriage network among 16 Florentine families (values reflect wealth).



Social networks

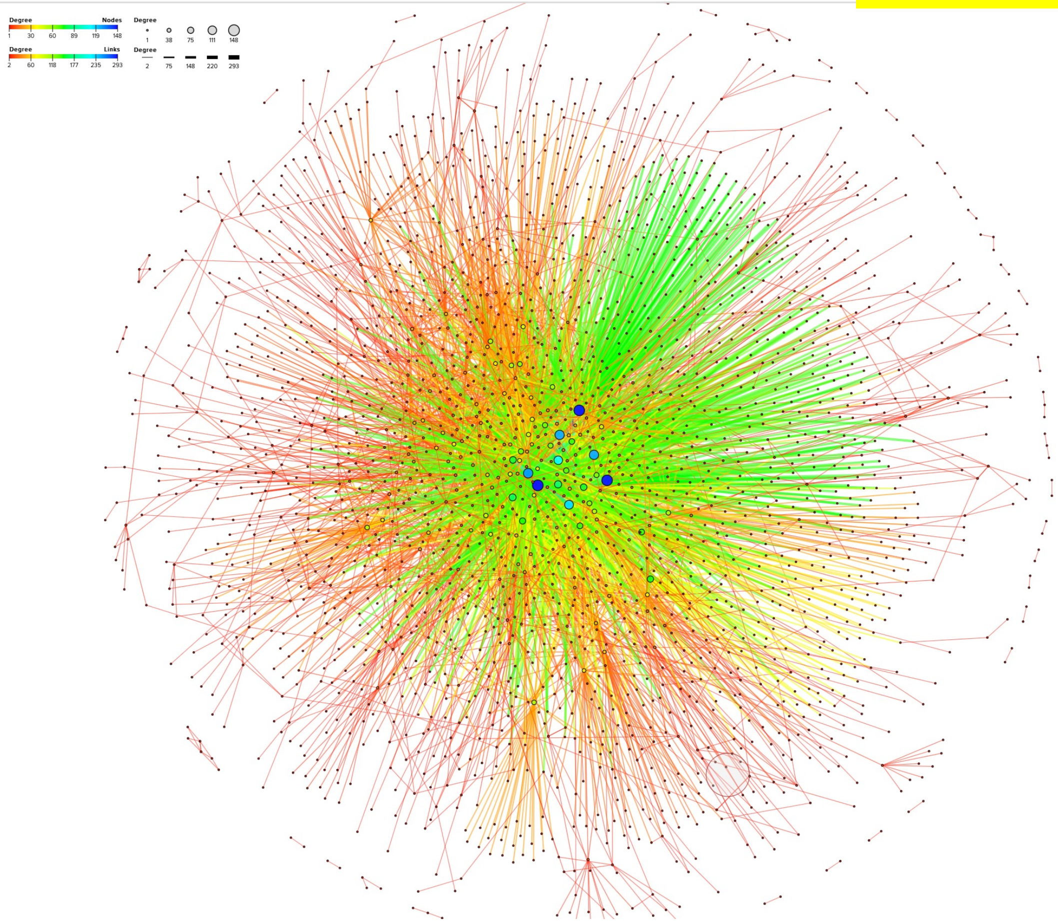
Direct interaction

Social network of **tribes of the Gahuku-Gama** alliance structure of the Eastern Central Highlands of New Guinea, from Kenneth Read (1954). The dataset contains a list of all links, where a link represents signed friendships between tribes.



Social networks

Direct interaction

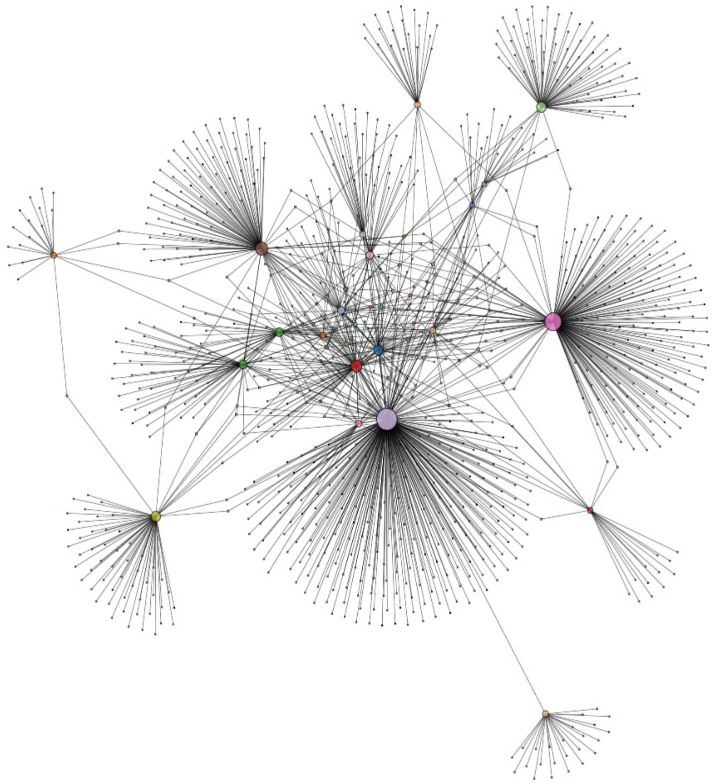


Advogato is a social community platform where users can explicitly express weighted trust relationships among themselves. The dataset contains a list of all of the user-to-user links.

Twitter network

Direct interaction

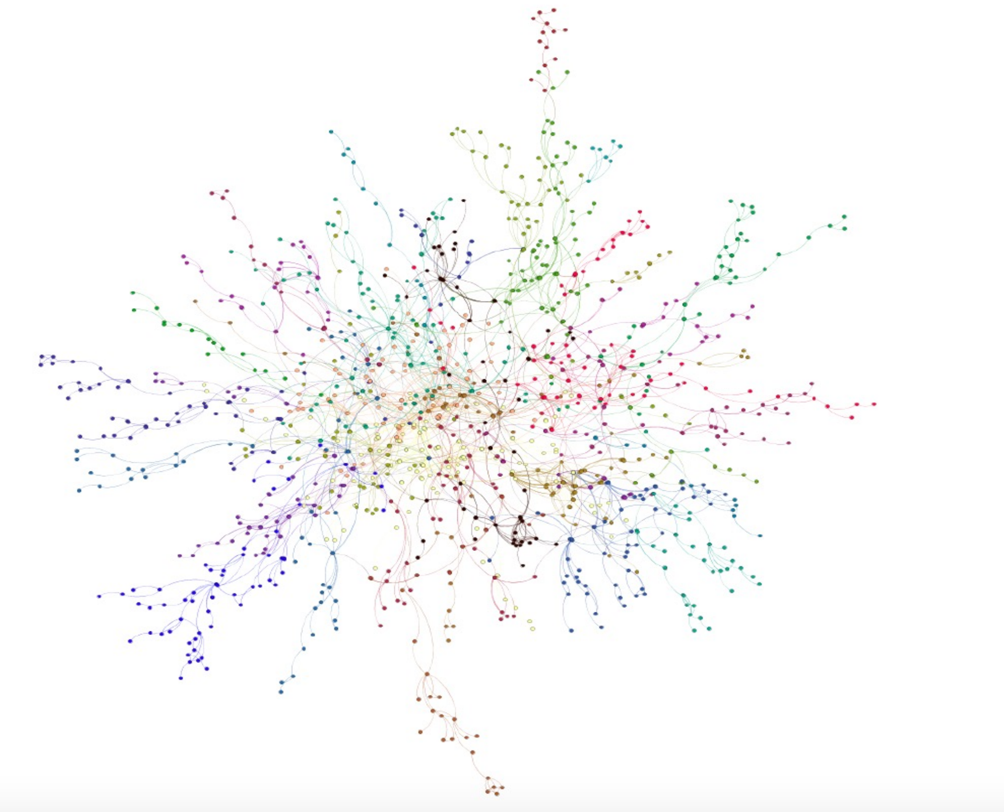
Twitter Higgs Boson This dataset is used to study the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012.



Co-authorship network

indirect interaction

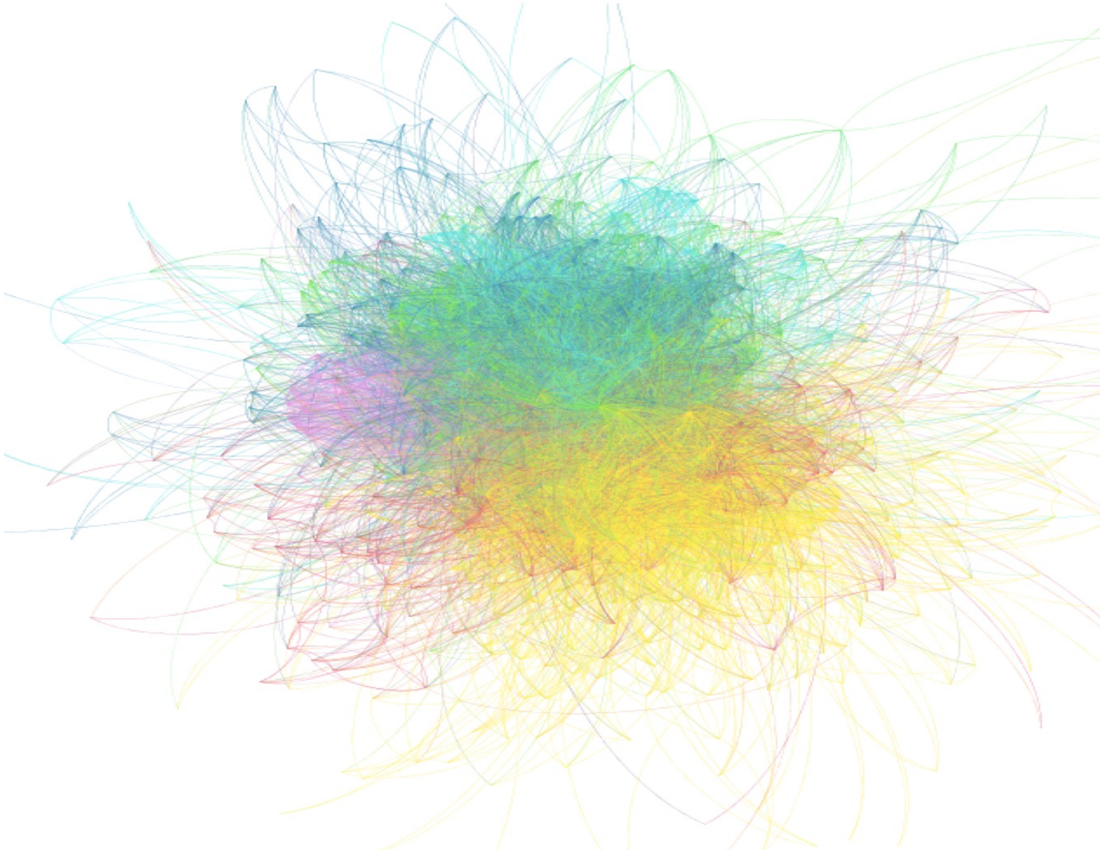
DBLP co-authorship network (1980-1984) - authors are active for more than 10 years.



Movie rating network

indirect interaction

MovieLens co-rating network : users that co-rate by rate 1 at least one movie



Other types of similar networks

http://networkrepository.com

Data & Network Collections. Find and interactively VISUALIZE and EXPLORE hundreds of network data

ANIMAL SOCIAL NETWORKS	816	INTERACTION NETWORKS	29	SCIENTIFIC COMPUTING	11
BIOLOGICAL NETWORKS	37	INFRASTRUCTURE NETWORKS	8	SOCIAL NETWORKS	77
BRAIN NETWORKS	116	LABELED NETWORKS	105	FACEBOOK NETWORKS	114
COLLABORATION NETWORKS	19	MASSIVE NETWORK DATA	21	TECHNOLOGICAL NETWORKS	12
CHEMINFORMATICS	646	MISCELLANEOUS NETWORKS	2668	WEB GRAPHS	36
CITATION NETWORKS	4	POWER NETWORKS	8	DYNAMIC NETWORKS	115
ECOLOGY NETWORKS	6	PROXIMITY NETWORKS	13	TEMPORAL REACHABILITY	38
ECONOMIC NETWORKS	16	GENERATED GRAPHS	221	BHOSLIB	36
EMAIL NETWORKS	6	RECOMMENDATION NETWORKS	36	DIMACS	78
GRAPH 500	8	ROAD NETWORKS	15	DIMACS10	84
HETEROGENEOUS NETWORKS	15	RETWEET NETWORKS	34	NON-RELATIONAL ML DATA	211

Similarity graphs

ε- neighbourhood graph : $\{u, v\}$ are linked if $sim(u, v) \geq \epsilon$

$Complexity\ \mathcal{O}(n^2)$

KNN-graph : Each item is connected to the K most similar items

$Complexity\ \mathcal{O}(n^2)$

Relative neighbourhood graph (RNG) :

$Complexity\ \mathcal{O}(n^3)$

$$\{u, v\}\ are\ linked\ if\ similarity(u, v) \geq \max_x\{sim(u, x), sim(v, x)\}\ \forall x \neq u, v$$

Similarity graphs: illustration

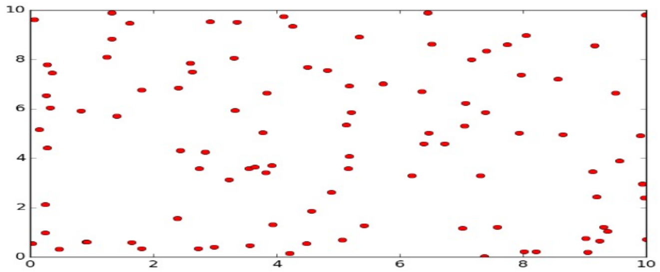


Figure: ϵ-threshold graph

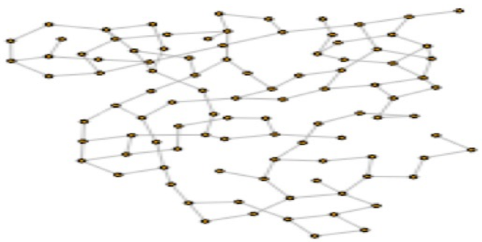


Figure:
RNG graph

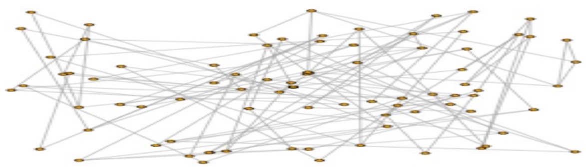


Figure:
Knn graph

RNG graph example

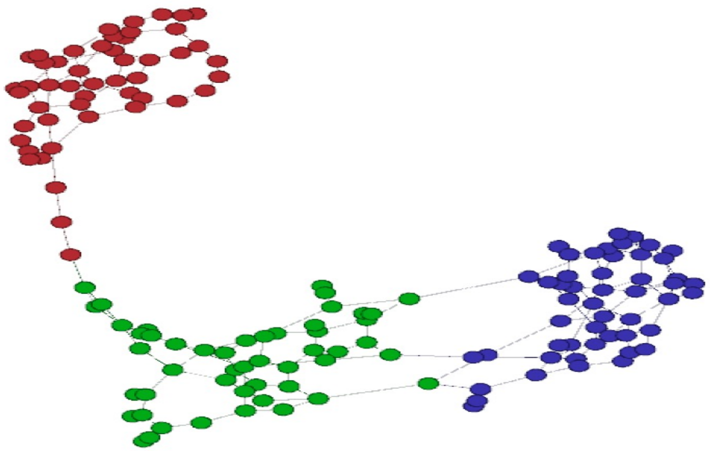

Iris Virginica


Iris Setosa


Iris Versicolor

Iris dataset :

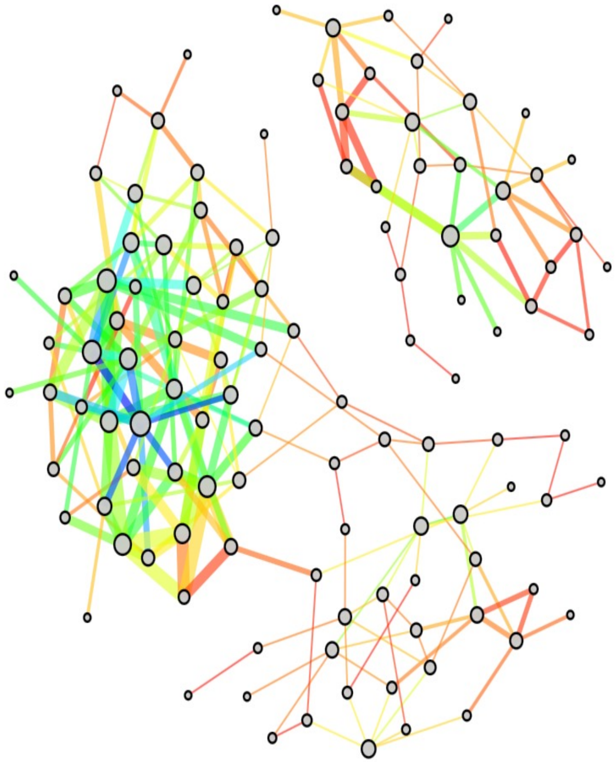
Node's colors code the class



2

Background

Graph Theory



Graphs

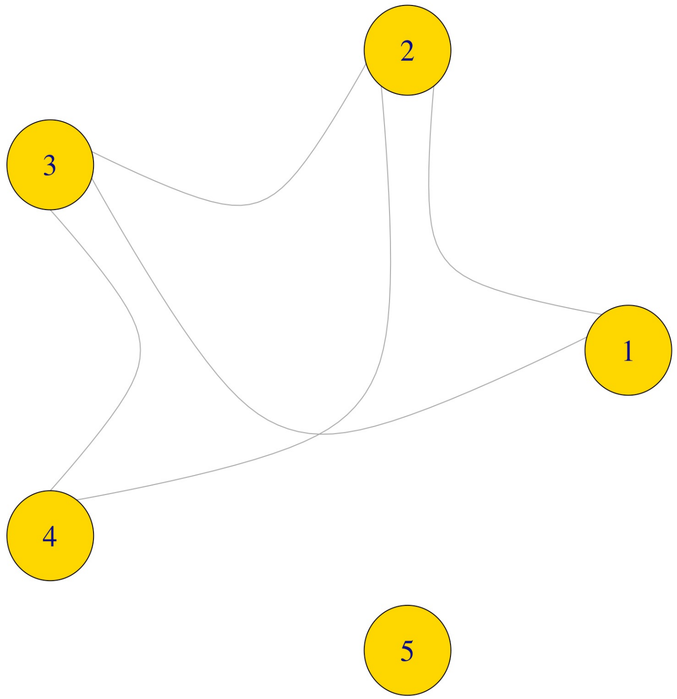
A graph **G** = $< V(G), E(G) >$

V(G) : set of **V**ertices

actors, nodes, sites, …

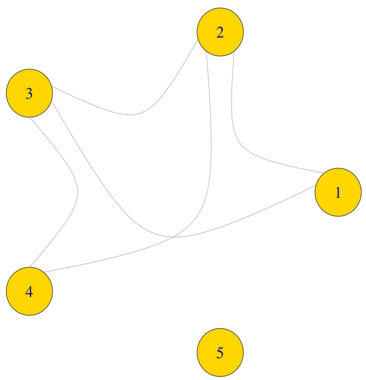
E(G) : set of **E**dges

links, ties, arcs, bonds …

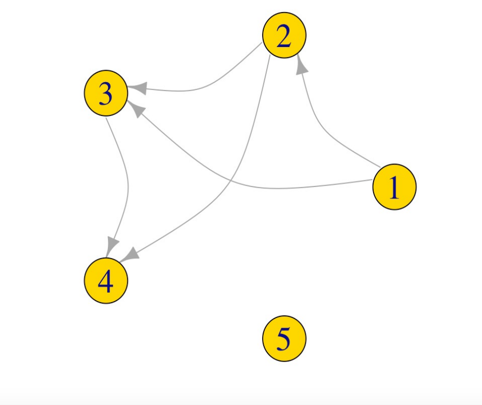


Graph: some types

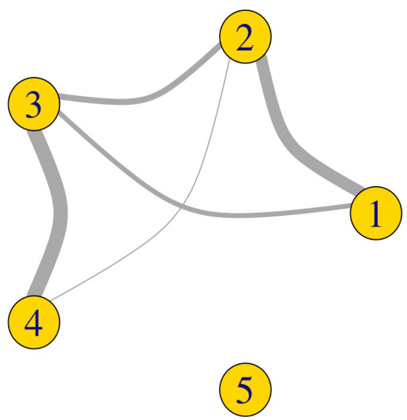
Undirected



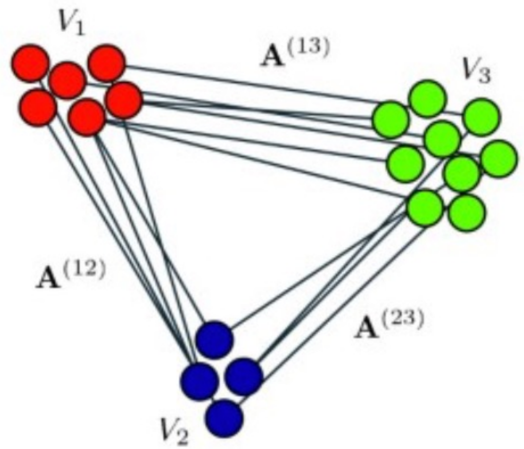
Directed



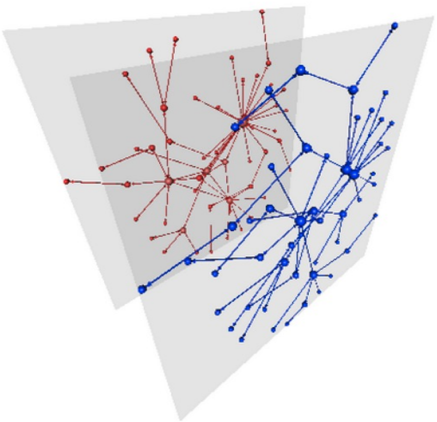
Weighted

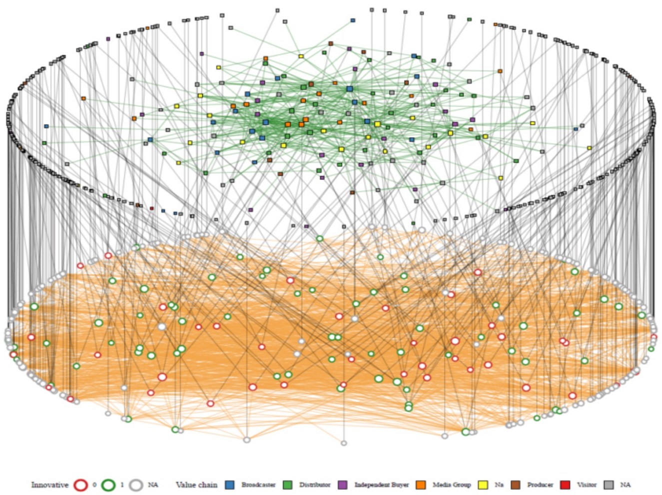


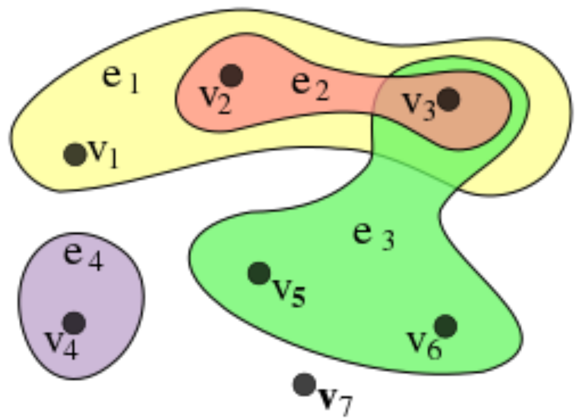
K-partite



Graph: some other types







https://en.wikipedia.org/wiki/Hypergraph

Throughout this course we only consider **undirected binary simple graphs** unless otherwise explicitly cited

Graph: Definitions

G is **simple** if it has no loops, neither multi-edges

E(G) = $\{\{v_i, v_j\}: v_i, v_j \in V(G) \wedge v_i \neq v_j\}$

$\mathbf{n_G}$ = |V(G)| is the **order** of G

$\mathbf{m_G}$ = |E(G)| is the **size** of G

G is **sparse** if $m_G \sim n_G$ for $n_G >> 1$

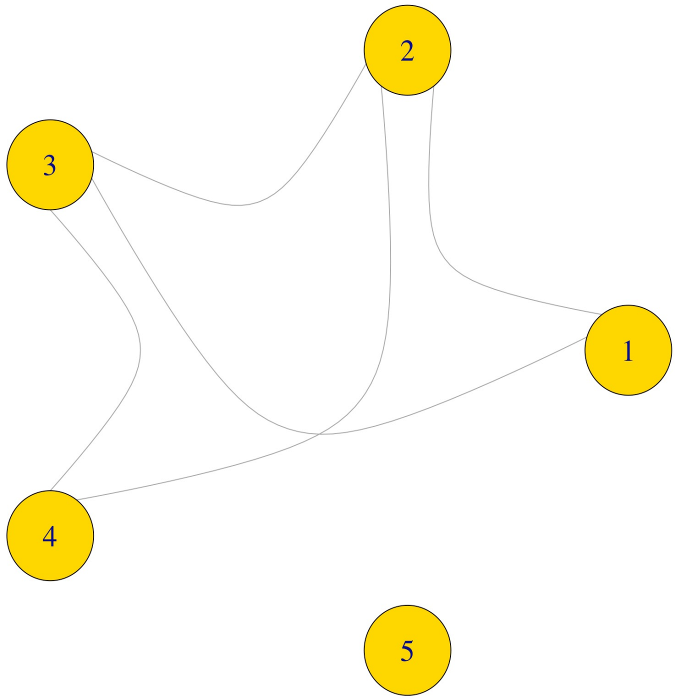
Graph: Illustration

$V(G) = \{1,2,3,4,5\}$

$n_G = 5$

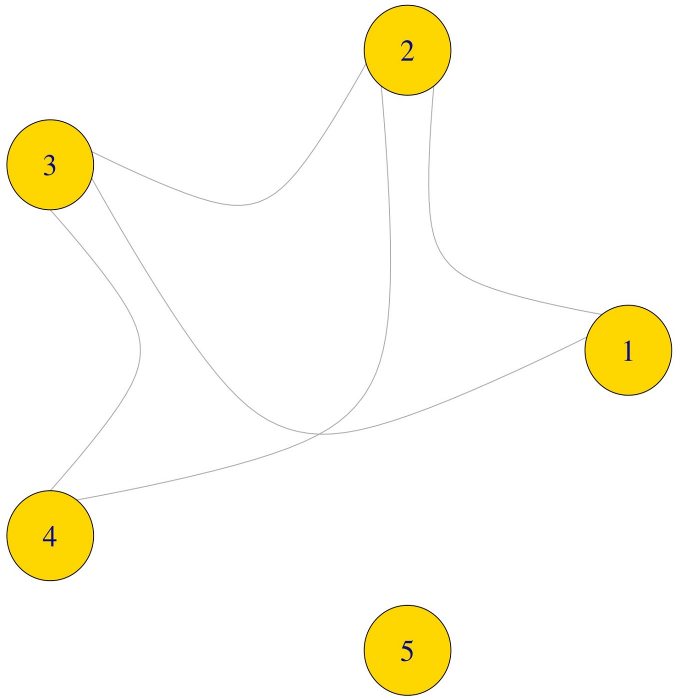
$E(G) = \{\{1,2\},\{1,3\},\{2,3\},\{2,4\},\{3,4\}\}$

$m_G = 5$



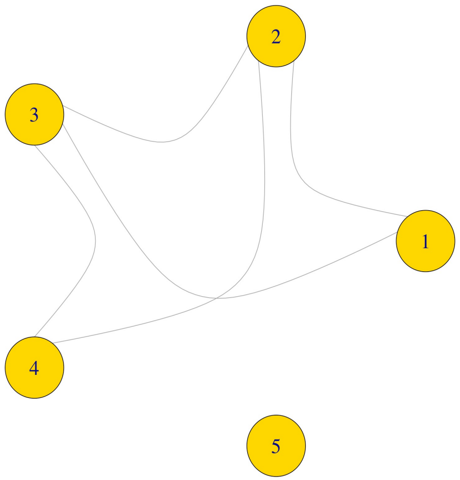
Graph: igraph code

```
>library(igraph)
>g <- graph(edges=c(1,2,2,3,1,3,3,4,2,4), n=5, directed=FALSE)
>vcount(g) ## length(V(g))
[1] 5
>ecount(g) ## length(E(g))
[1] 5
>plot(g, vertex.color="gold", edge.curved=TRUE, layout= layout_in_circle)
```



Graph Density

Density = Probability of having a link between two randomly selected nodes



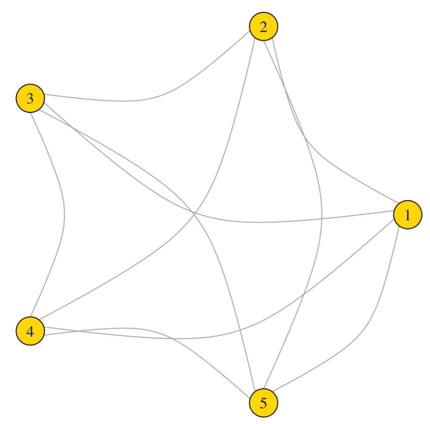
$$\rho(G) = \frac{2\, m_G}{n_G(n_G - 1)} \in [0,1]$$

```
> g <- graph(edges=c(1,2,2,3,1,3,3,4,2,4), n=5, directed=FALSE)
> graph.density(g)
[1] 0.5
```

$\rho(G) = 0.5$

Complete Graph / Cliques

G is **complete** if each node is directly connected to each other node.



K_n : Complete graphs of order n

```
> k <-graph.full(5, directed=False)> graph.density(g)
> graph.density(k)
[1] 1
```

$\rho(G) = 1$

Neighbourhood: definitions

$\Gamma(v) = \{x_i \in V(G): \{v, x_i\} \in E(G)\}$ Open neighbourhood

$\overline{\Gamma}(v) = \Gamma(v) \cup \{v\}$ Closed neighbourhood

$d_v = |\Gamma(v)|$ Degree of vertex v

$\delta_{avg}(G) = \frac{\sum_{v \in V(G)} d_v}{n_G} = \frac{2m_G}{n_G}$ Average degree of G

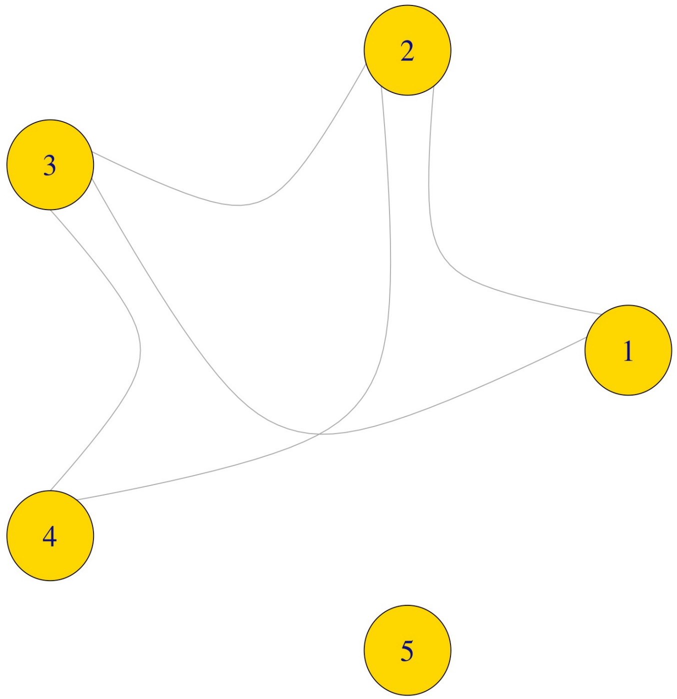
Neighbourhood: Illustration

$\Gamma(2) = \{1,3,4\}$

$\overline{\Gamma}(2) = \{1,3,4\} \cup \{2\} = \{1,2,3,4\}$

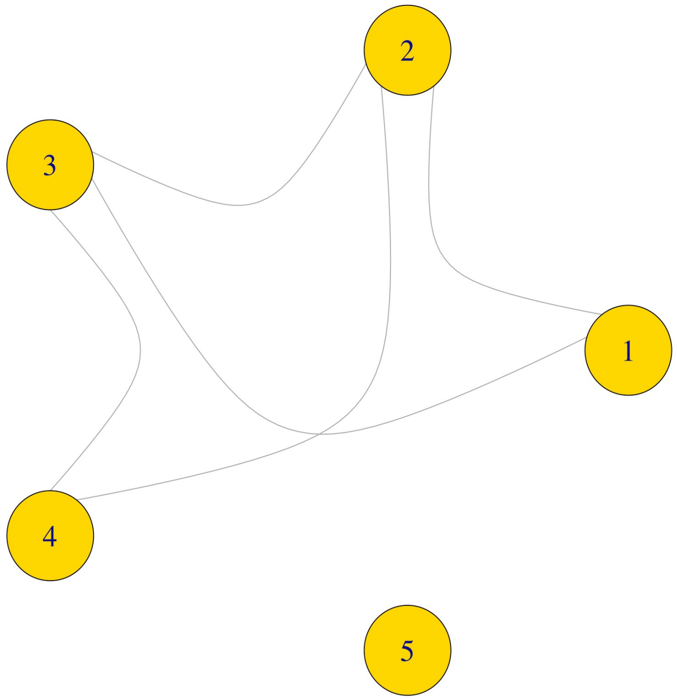
$d_2 = |\Gamma(2)| = 3$

$\delta_{avg}(G) = \frac{\sum_{v \in V(G)} d_v}{n_G} = \frac{2m_G}{n_G} = \frac{2*5}{5} = 2$



Neighbourhood: igraph code

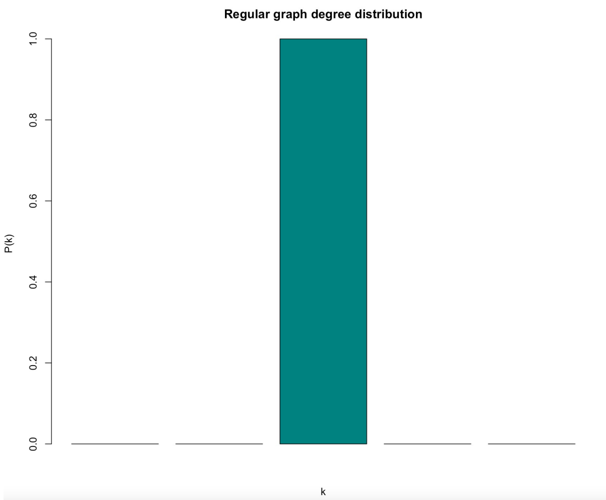
```
> neighbors(g,2)
+ 3/5 vertices, from 18848b1:
[1] 1 3 4
> degree(g,2)
[1] 3
> degree(g)
[1] 2 3 3 2 0
```



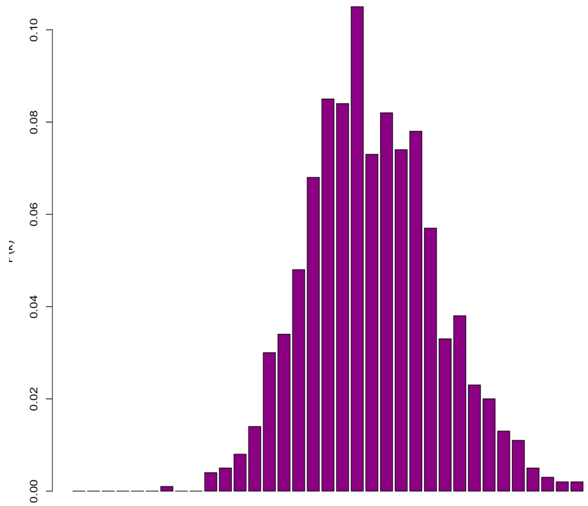
Degree distribution

$$P(k) = \frac{|\{v_i \in V(G): d_{v_i} = k|}{n_G}$$

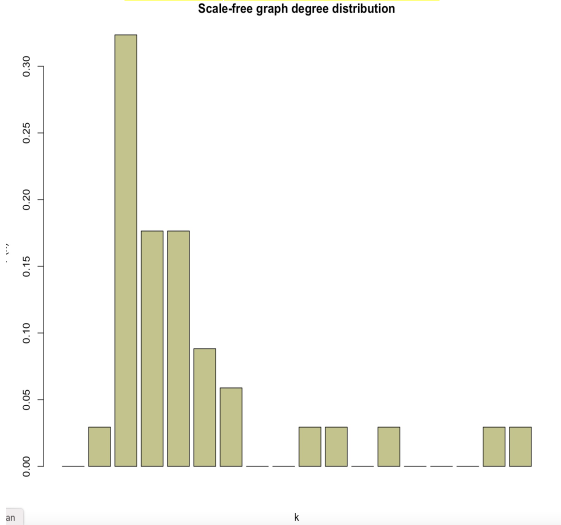
Regular graph



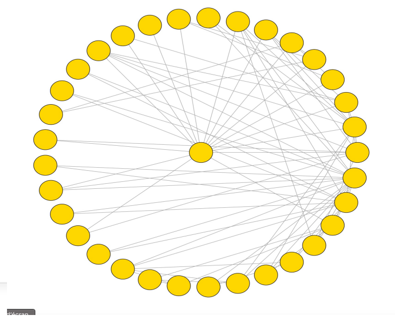
Random graph



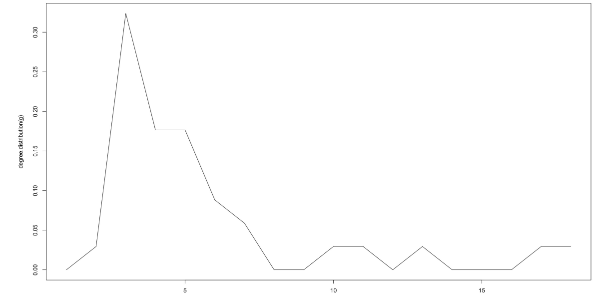
Scale-Free graph

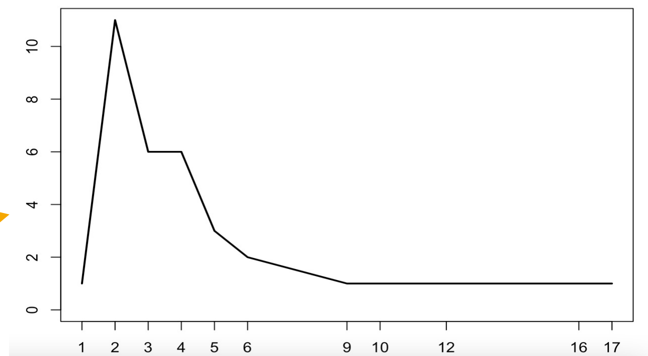


Degree distribution: igraph code



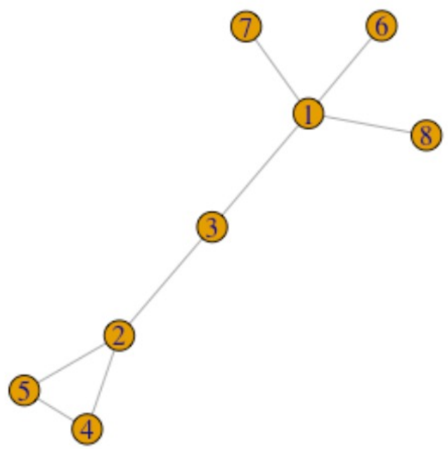
```
> g <- graph.famous("zachary") # Karate club

> plot(g, vertex.label=NA, vertex.color="gold",
layout=layout_as_star)

> degree.distribution(g)

 [1] 0.00000000 0.02941176 0.32352941 0.17647059
0.17647059 0.08823529 0.05882353 0.00000000 0.00000000

[10] 0.02941176 0.02941176 0.00000000 0.02941176
0.00000000 0.00000000 0.00000000 0.02941176 0.02941176

>plot(degree.distribution(g), type="l")

> degree(g) %>% table() %>% plot(type="l")
```

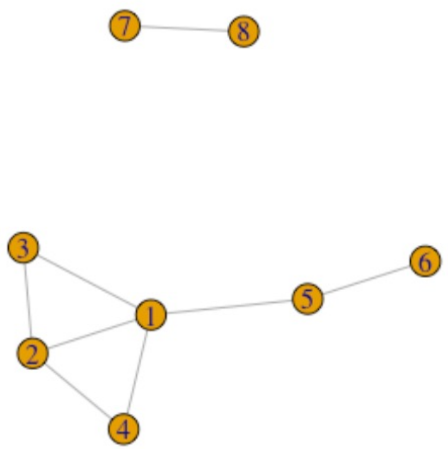




Degree distribution

Different graphs may have the same degree distribution





Paths: definitions

$\sigma^k(u,v) = \{ [x_0, x_1, \dots, x_k]: x_0 = u \wedge x_k = v \wedge$
$\forall i,j\ x_i, x_j \in V(G) \wedge x_i \neq x_j \wedge$
$\{x_i, x_{i+1}\} \in E(G) \}$

Set of elementary paths of length k linking u and v

$SP(u,v) = \min_k \sigma^k(u,v) \neq \phi$

Set of shortest paths linking u and v

$d(u,v) = |\delta \in SP(u,v)|$

Geodesic distance between u and v

$$\delta(G) = \frac{\sum_{u,v \in V(G)} d(u,v)}{\binom{n}{2}}$$

Average distance

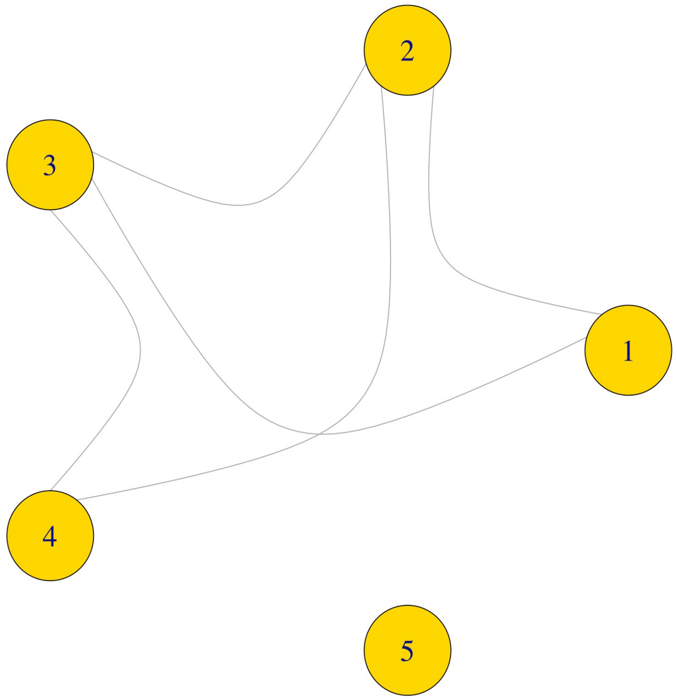
Paths: Illustration

$\sigma^3(1,4) = \{ [1,2,3,4], [1,3,2,4]\}$

$SP(1,4) = \sigma^2(1,4) = \{ [1,3,4], [1,2,4]\}$

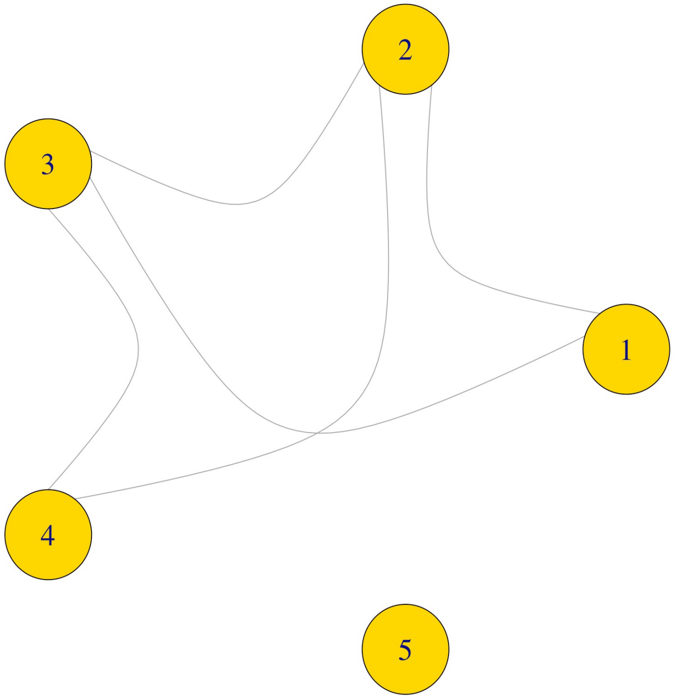
$d(1,4) = 2$

$\delta(G) = \frac{(1+1+2)+(1+1)+1}{6} = \frac{7}{6} = 1.166667$



Paths: igraph code

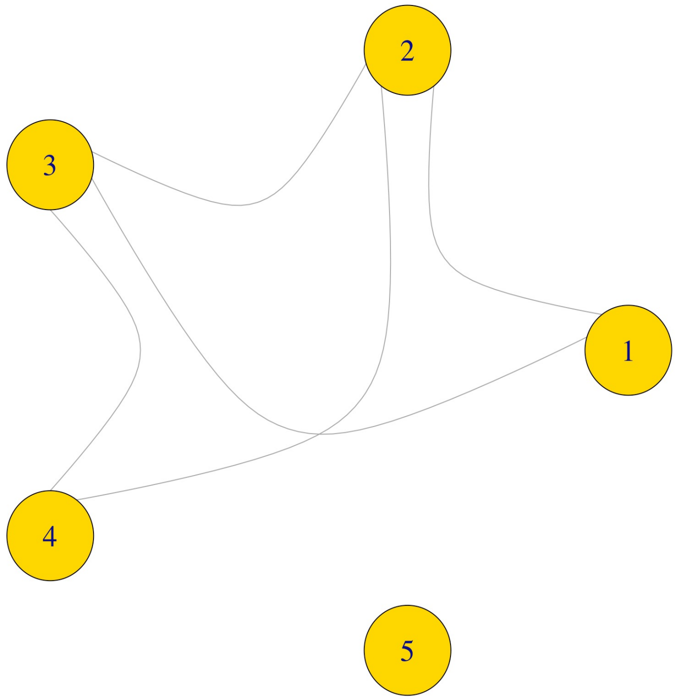
```
>paths<-all_simple_paths(g,1,4)
> paths[lapply(paths,length)==4] #σ³(1,4)
[[1]]
+ 4/5 vertices, from 86325cf:
[1] 1 2 3 4

[[2]]
+ 4/5 vertices, from 86325cf:
[1] 1 3 2 4
```



Exercise : What is the complexity of computing all simple paths ?

Paths: igraph code

```
> all_shortest_paths(g,1,4)
$res
$res[[1]]
+ 3/5 vertices, from 86325cf:
[1] 1 3 4
$res[[2]]
+ 3/5 vertices, from 86325cf:
[1] 1 2 4
$nrgeo #shortest path length between source and each node
[1] 1 1 1 2 0
> > shortest.paths(g,1,4)# Geodesic distance
     [,1]
[1,]    2
> average.path.length(g,directed = FALSE,unconnected = TRUE)
[1] 1.166667
```

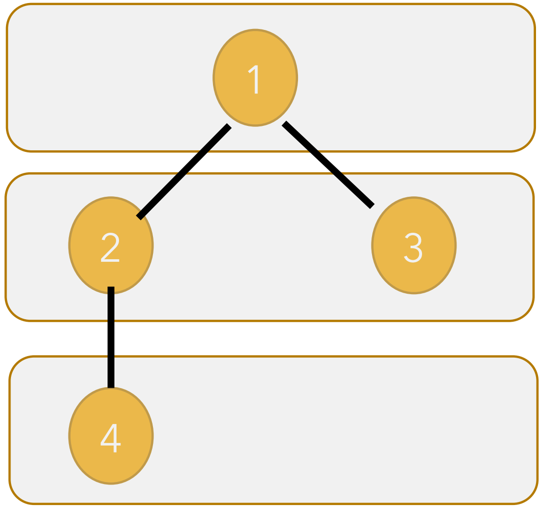


Paths: Algorithms

Single source shortest paths

Graph type	Algorithm	Time complexity
Binary	Breadth First Search	$\mathcal{O}(n_G + m_G)$
Positive weighted undirected	**Dijkstra** (1959)	$\mathcal{O}(n_G^2)$
Positive weighted undirected	Fredman & Tarjan (1987)	$\mathcal{O}(m_G + n_G \log(n_G))$
$\mathbb{N}$ weighted undirected	Thorup (1999)	$\mathcal{O}(m_G)$

Breadth First Search (BFS)



Paths: Algorithms

All pairs shortest paths

Graph type	**Algorithm**	**Time complexity**
Binary	Breadth First Search	$\mathcal{O}(n_G(n_G + m_G))$
Positive weighted undirected	**Floyed-Warshall** (1962)	$\mathcal{O}(n_G^3)$
$\mathbb{N}$ weighted undirected	Thorup (1999)	$\mathcal{O}(n_G m_G)$

Eccentricity, Diameter & Radius

$$ECC(v \in V(G)) = \max_{\{x \in V(G)\}} d(x, v)$$

$$Diam(G) = \max_{\{v \in V(G)\}} ECC(v)$$

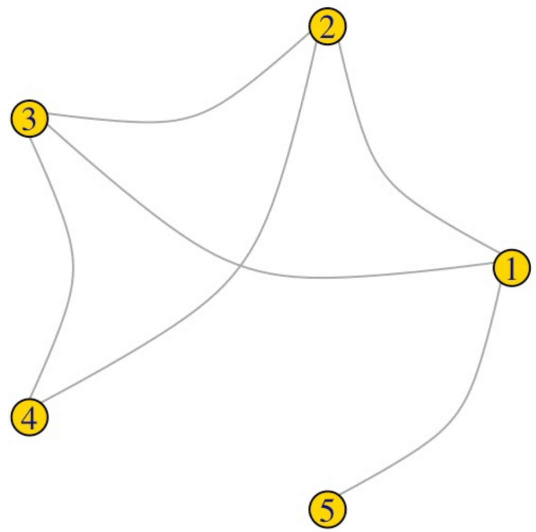
$$Radius(G) = \min_{\{v \in V(G)\}} ECC(v)$$

$$Radius(G) \leq Diam(G) \leq 2\ Radius(G)$$

For connected graphs

Eccentricity, Diameter & Radius: code igraph

```
> eccentricity(g)
[1] 2 2 2 3 3

> diameter(g)
[1] 3

> radius(g)
[1] 2
```



Connectedness

G is **Connected** iif $\forall v_i, v_j \in V(G)\ \exists k > 1;\ \left|\sigma^k(v_i, v_j)\right| \neq \phi$

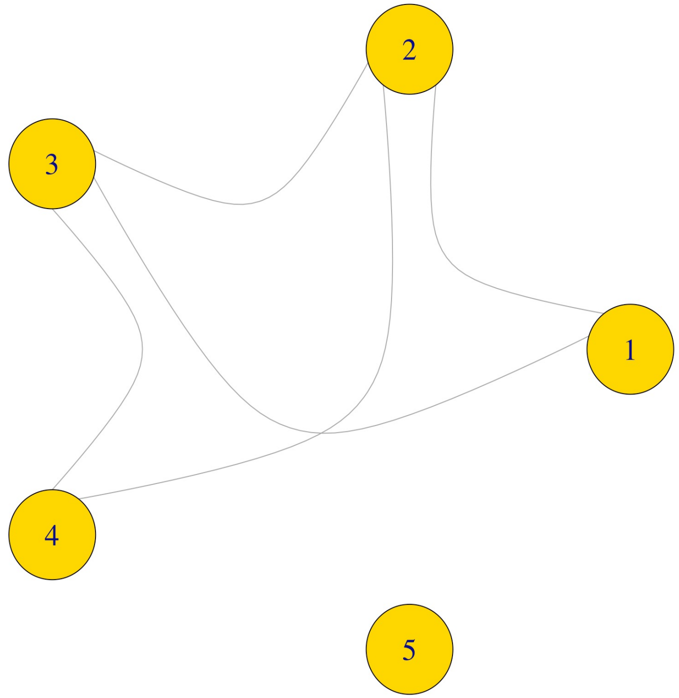
A graph may be composed of several **connected components**.

Giant component : a component with most of the vertices of the network.

Algorithm: **BFS** $\mathcal{O}(n_G + m_G)$

Connectedness: igraph code

```
> is_connected(g)
[1] FALSE
> clusters(g)
$membership
[1] 1 1 1 1 2
$csize
[1] 4 1
$no
[1] 2
```



Clustering coefficient

Probability of having a link between two nodes that share a common neighbour

What is the probability that two friends of a given person are friends themselves ?

Global version

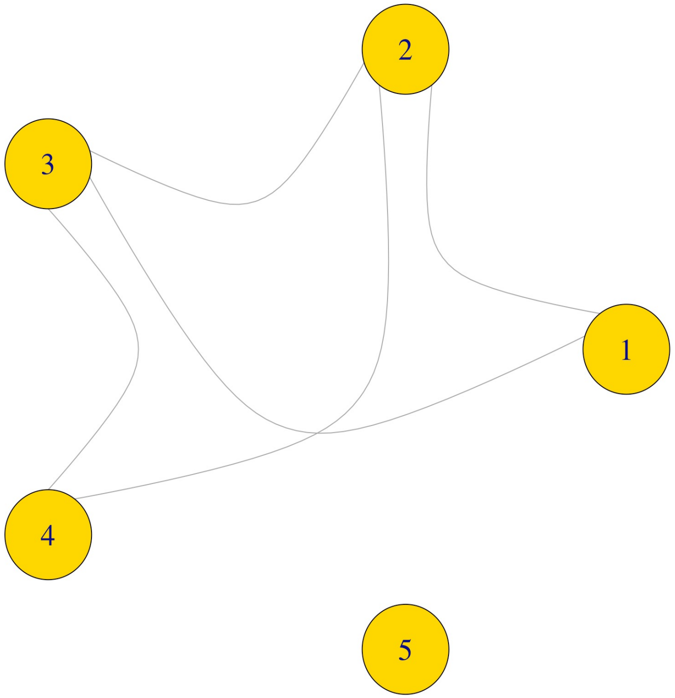
$$CC(G) = \frac{3 \times \# \Delta}{\# \wedge}$$

Local version

$$CC(G, v) = \frac{\# links\ between\ neighbours\ of\ v}{\# potential\ links\ between\ neighbors\ of\ v}$$

Clustering coefficient: illustration

```
> transitivity(g). # global version
[1] 0.75
> transitivity(g,type="local")
[1] 1.0000000 0.6666667 0.6666667 1.0000000
NaN
```

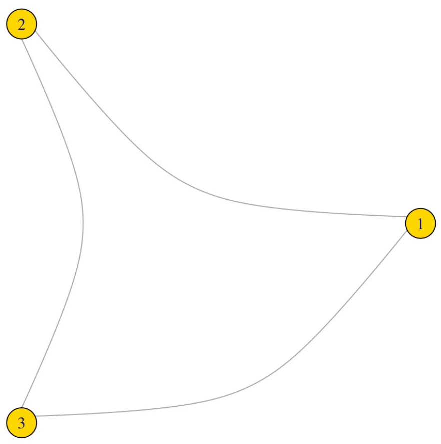


Subgraphs

- G_s is a subgraph of G iif : $V(G_s) \subseteq V(G) \quad E(G_s) \subseteq E(G)$
- **Induced subgraph** : G_s is an induced subgraph of G on S if

$$V(G_S) = S \subseteq V(G) \wedge E(G_s) = \{ \{v_i, v_j\} \in E(G): v_i, v_j \in S\}$$

```
> s <- c(1,2,3)
> gs <- induced.subgraph(g,s)
> plot(gs, vertex.color="gold", edge.curved=TRUE,
layout= layout_in_circle)
```



Special subgraphs : ego-network

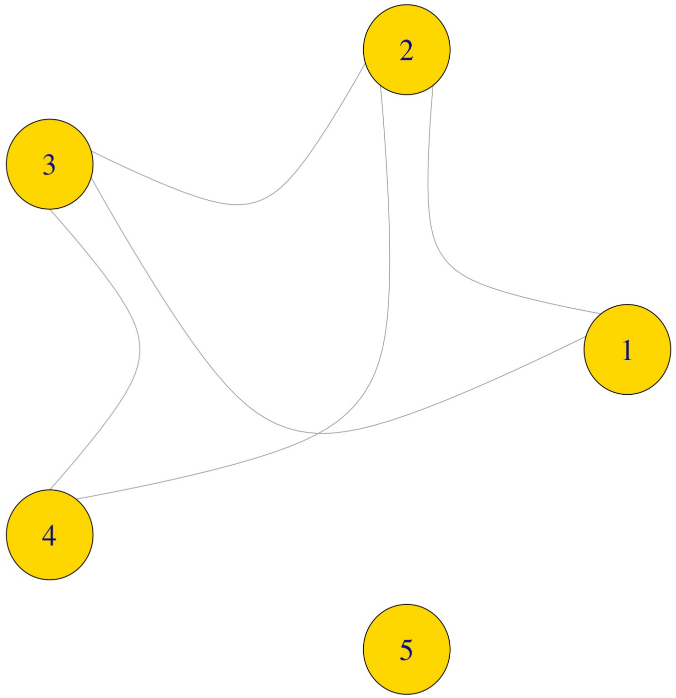
An **ego-network** of a target node $v_t \in V(G)$ is the induced subgraph on the set $\overline{\Gamma}(v_t)$

```
ego_network <- function(v,g){
      return(induced_subgraph(g,c(v,neighbors(g,v))))
}
```

Special subgraphs : cliques & quasi cliques

- **K-Clique** : $G_c = \big(V(G_c) \subseteq V(G), E(G_c) \subseteq E(G)\big) : n_{G_c} = k \wedge \rho(G_c)$=1
- **K-plex** : $G_c = \big(V(G_c) \subseteq V(G), E(G_c) \subseteq E(G)\big) : \min_{\{v \in V(G_c)\}} d_{\{G_c\}}(v) \geq n_{G_c} - k$
- **K-club :** $G_c = \big(V(G_c) \subseteq V(G), E(G_c) \subseteq E(G)\big) : Diam(G_c) \leq k$
- **K-core :** $G_c = \big(V(G_c) \subseteq V(G), E(G_c) \subseteq E(G)\big)$; $\forall\, v \in V(G_c)$: $d_v^{\{G_c\}} \geq k$

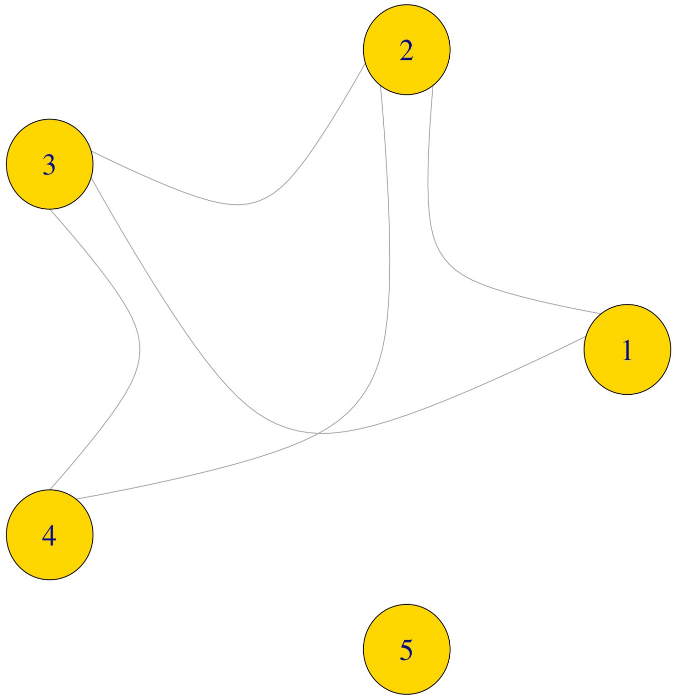
Graph representation: Adjacency Matrix



$$a_{ij} = 1 \; if \; \{v_i, v_j\} \in E(G); 0 \; otherwise$$

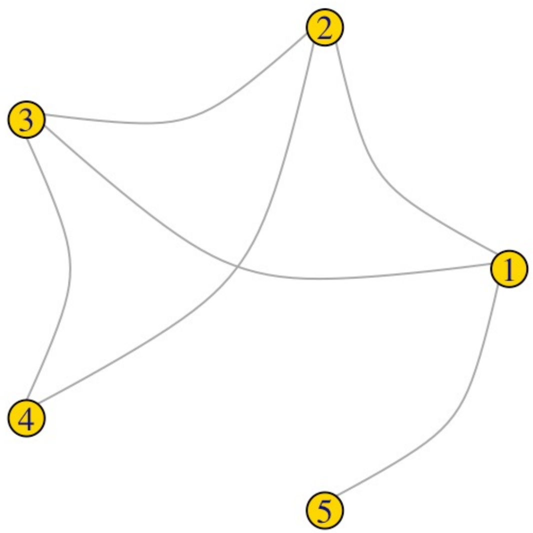
$$A_G = \begin{pmatrix} 0\,1\,1\,0\,0 \\ 1\,0\,1\,1\,0 \\ 1\,1\,0\,1\,0 \\ 0\,1\,1\,0\,0 \\ 0\,0\,0\,0\,0 \end{pmatrix}$$

Adjacency Matrix: igraph code




```
>library(igraph)
>g <- graph(edges=c(1,2,2,3,1,3,3,4,2,4), n=5, directed=FALSE)
>a <- as_adjacency_matrix(g)
> a
5 x 5 sparse Matrix of class "dgCMatrix"

[1,] . 1 1 . .
[2,] 1 . 1 1 .
[3,] 1 1 . 1 .
[4,] . 1 1 . .
[5,] . . . . .
```

Adjacency Matrix: igraph code



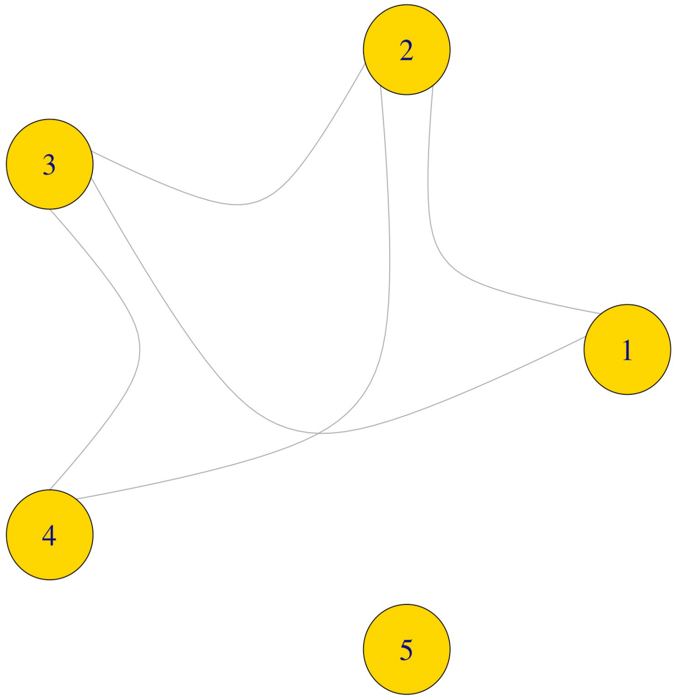

```
> a[1,5]<-1
> g <- graph_from_adjacency_matrix(a,
weighted=FALSE)
> plot(g, vertex.color="gold", edge.curved=TRUE,
layout= layout_in_circle)
```

Adjacency Matrix: some proprieties

- A_G is **square symmetric** matrix ($A_G = A_G^t$)
- $D_G =< d_1, \dots, d_{n_G} >= A_G 1_{n_G}$

  where 1_{n_G} is one vector of dimension n_G
- A_G has at most n **real eigen values**
- $\exists\, \boldsymbol{P}: \boldsymbol{A_G} = \boldsymbol{P^{-1}} \boldsymbol{\Delta} \mathbf{P}$
- $A_G^k[i,j]$: number of de walks of length k between v_i, v_j

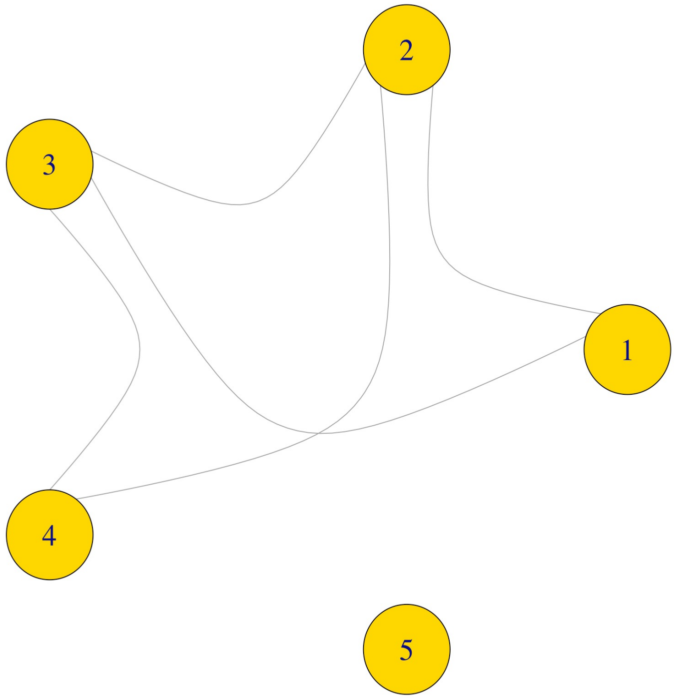
Graph representation: Incidence Matrix



$$b_{ij} = 1 \; if \; v_i is \; incident \; to \; edge \; k;$$
$$0 \; otherwise$$

$$B_G = \begin{pmatrix} 1\,1\,0\,0\,0 \\ 1\,0\,1\,1\,0 \\ 0\,1\,1\,0\,0 \\ 0\,0\,0\,1\,1 \\ 0\,0\,0\,0\,0 \end{pmatrix}$$

Graph representation: Edge List



The IJ-form of A_G

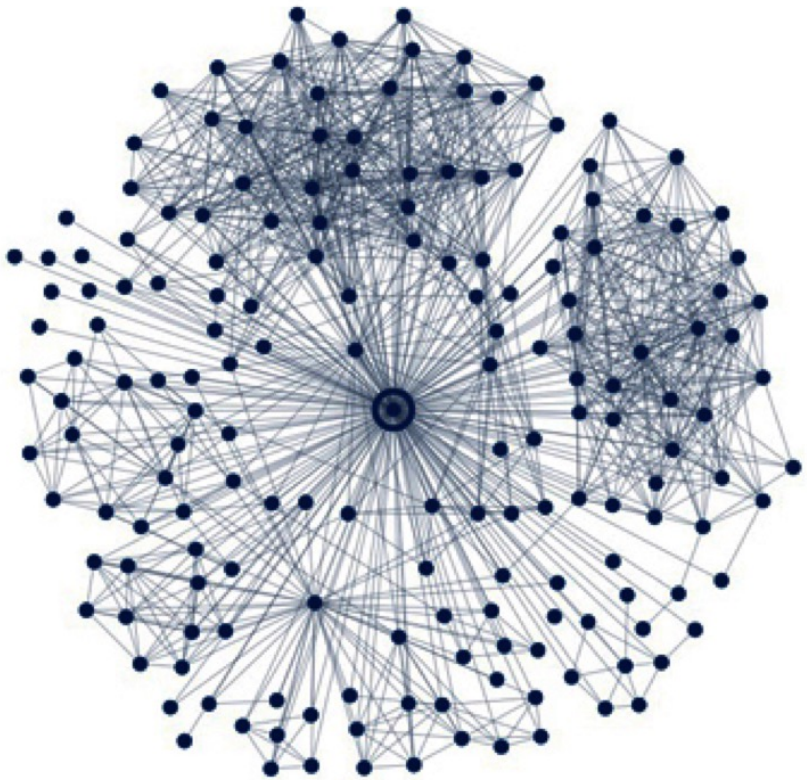
1,2
1,3
2,3
2,4
3,4

Suitable for sparse graphs

3

Complex Networks

Topological Features



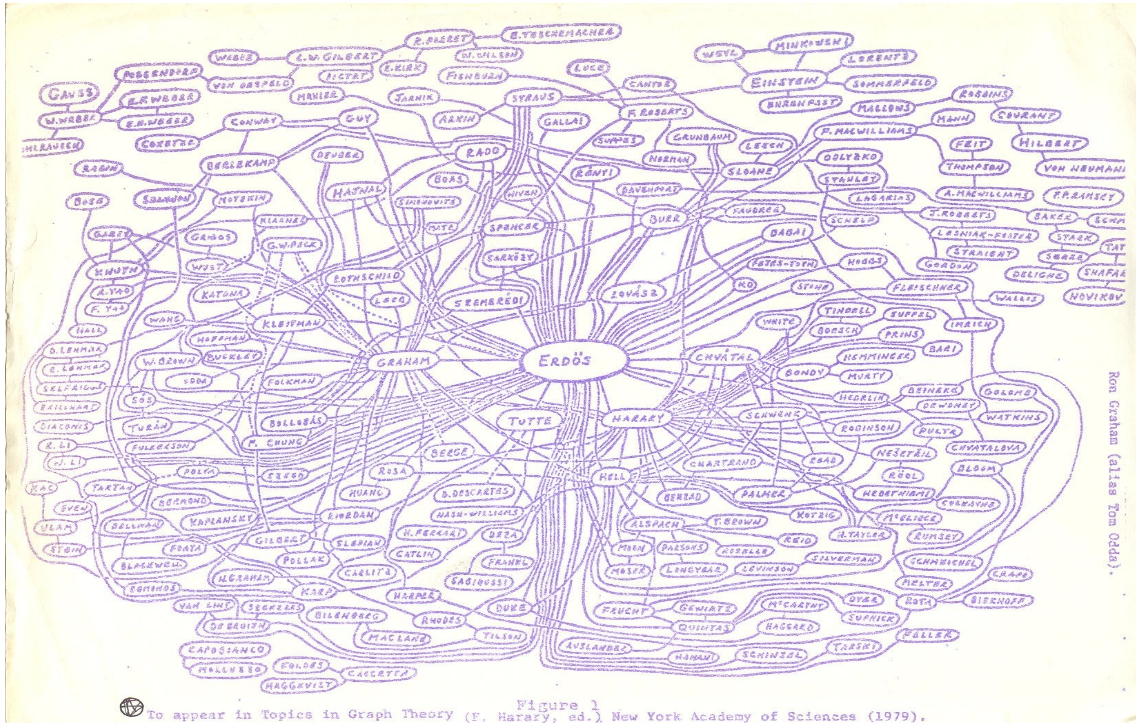
Small-World experiment (1967)

Satnley Milgram
(1933-1984)

- Goal: delivering letters to a Broker at Boston
- Source : Random selected persons from Boston area and also from Nebraska
- **217** letters have been sent : **64** has been received.
- Average path length : **5.2**

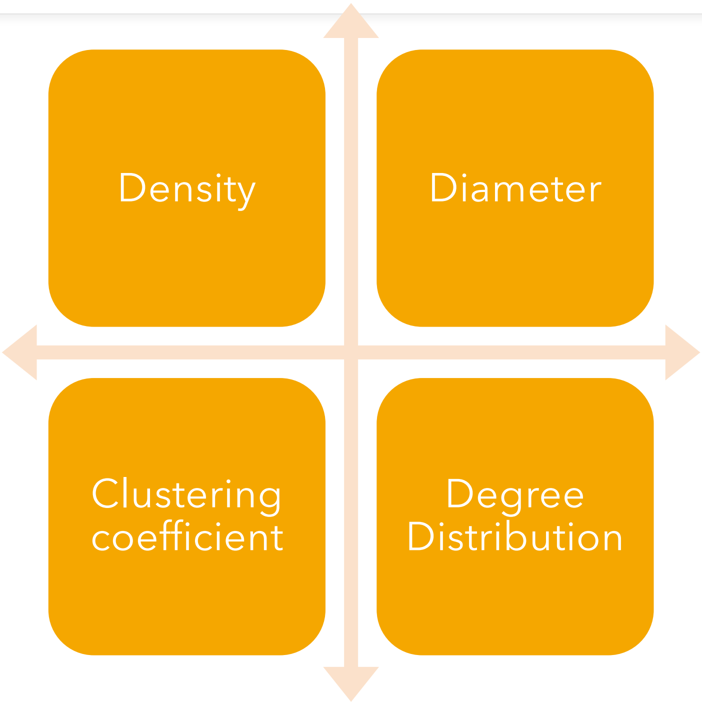
Paul Erdös collaboration network

Paul Erdös
(1913–1996)



- A very productive Hungarian Mathematician who has published over 1500 scientific paper
- **Erdös Number** : distance from Erdös in co-authorship network
- https://www.oakland.edu/enp/
- Diameter of 2-Erdös is 5

Basic topological features



Social networks

Advogato

ρ_G

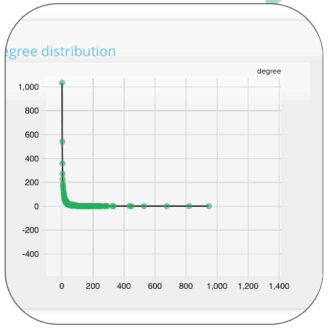
$2\ 10^{-3}$

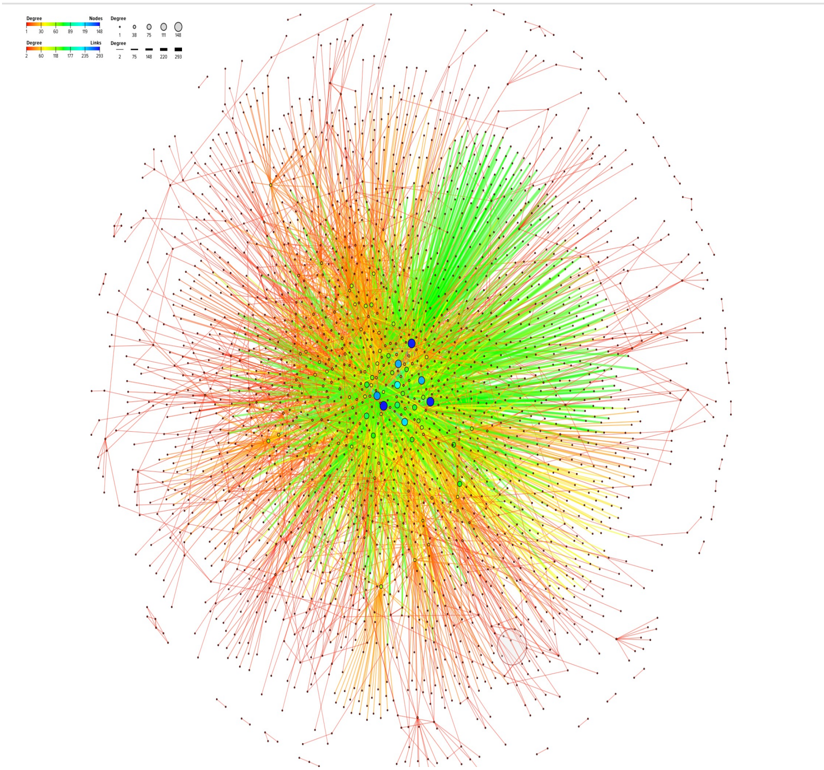
$Diam(G)$

11

CC

$9\ 10^{-2}$





Social networks

Direct interaction

Twitter Higgs Boson

ρ_G

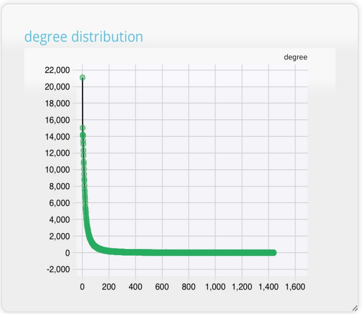
10^{-3}

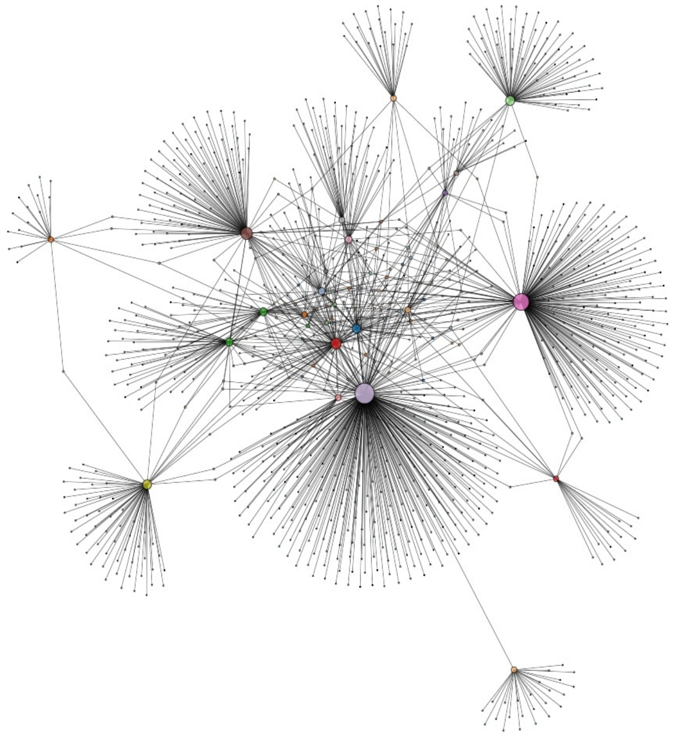
$Diam(G)$

7

CC

$3\ 10^{-2}$





Social networks

indirect interaction

DBLP co-authorship network (1980-1984)

ρ_G

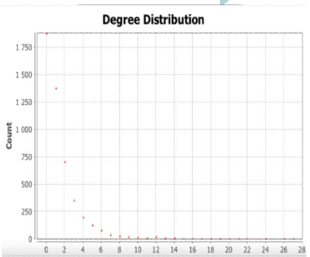
10^{-4}

$Diam(G)$

24

CC

0.67



Social networks

Direct interaction

Tribes of the Gahuku-Gama

$\rho(G)$

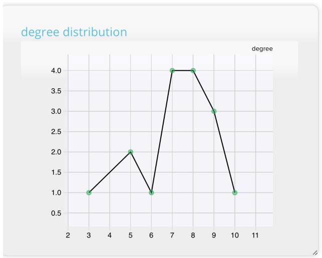
0.54

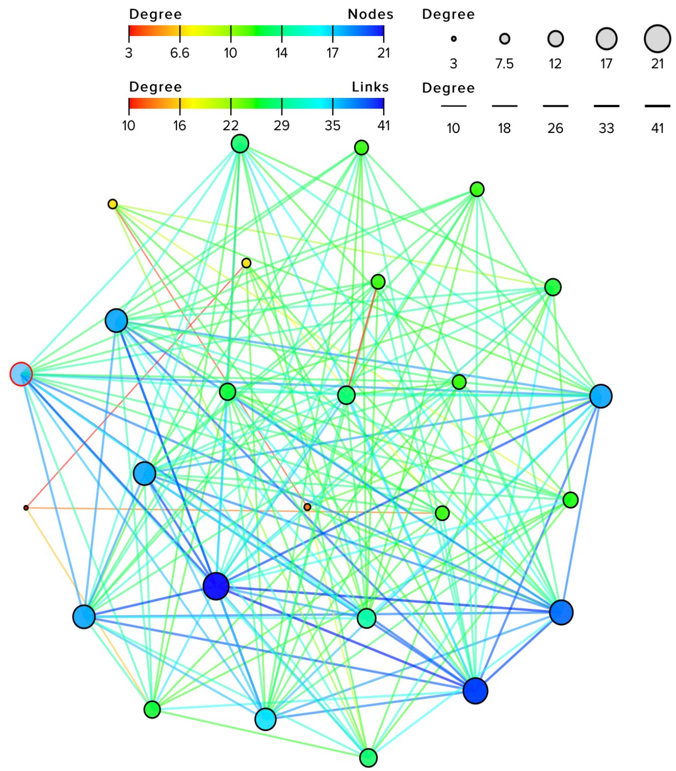
$Diam(G)$

3

CC

0.65





Wikipedia network

Direct interaction

$\rho(G)$

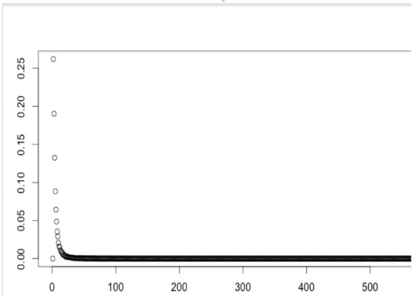
2^{-3}

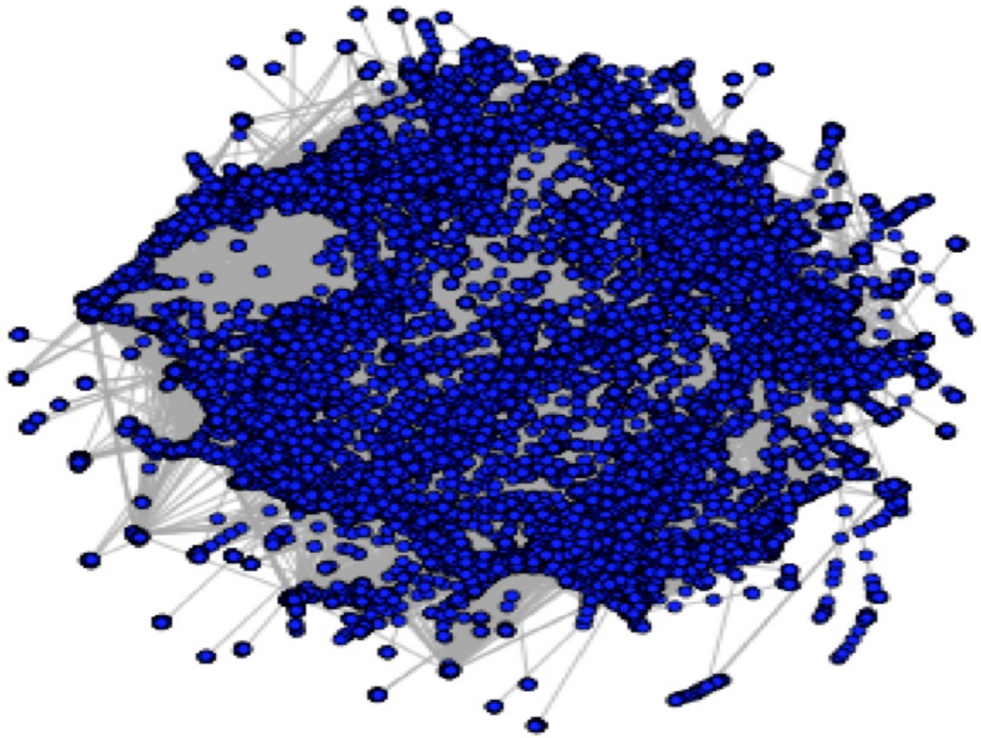
$Diam(G)$

28

CC

0.47





Similarity network

$\rho(G)$

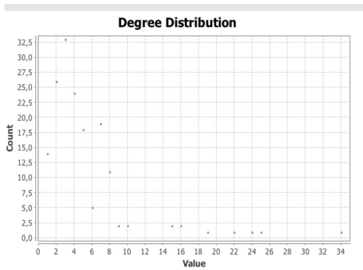
0.052

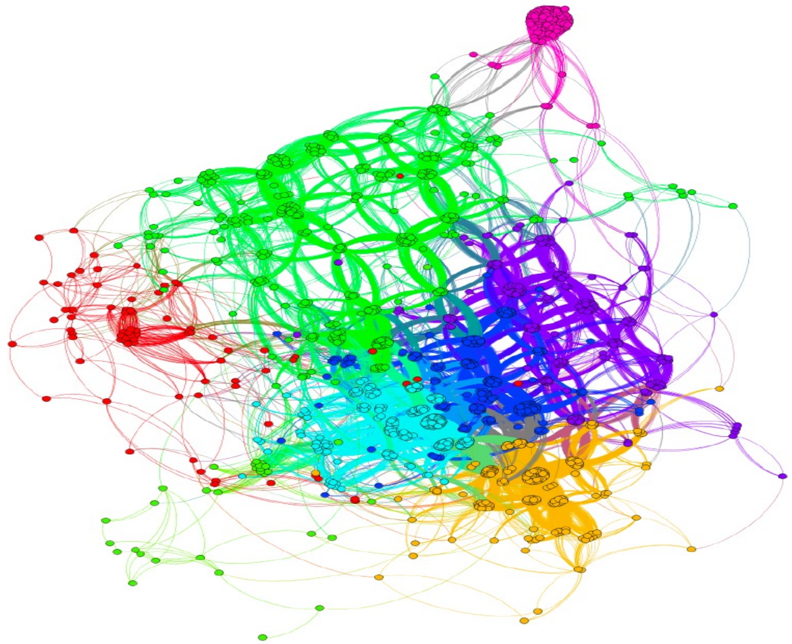
$Diam(G)$

7

CC

0.97

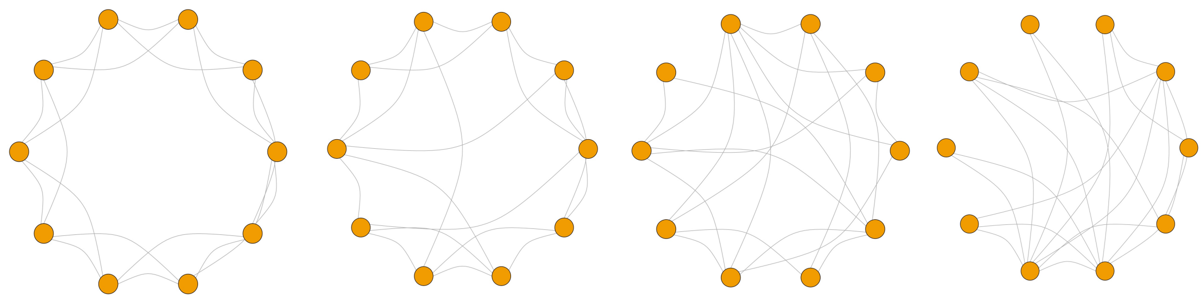




Public sites accessibility in Bourget District, Paris Area

4

(Random) Networks Models



Erdös-Rényi model

Paul Erdös
(1913-1996)

Alfréd Rényi
(1921-1970)

$G_{\{n,p\}}$

n : number of nodes
p : *probability of an edge*

Complexity : $\mathcal{O}(n^2)$

$G_{\{n,m\}}$

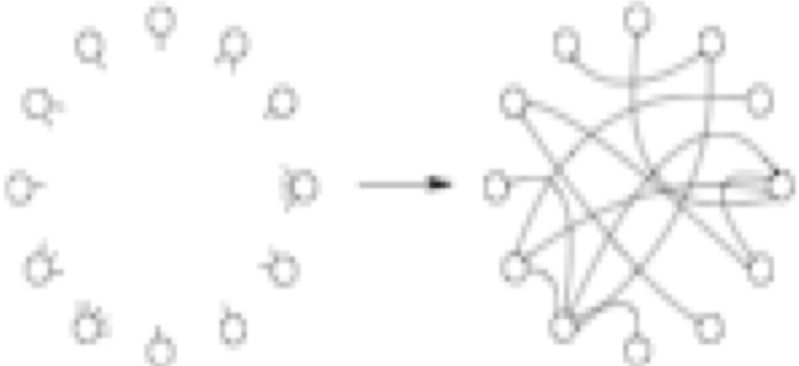
n : number of nodes
m *randomly chosen edges*

Complexity : $\mathcal{O}(m)$

Generated graphs properties (see lab 1)

- **Low diameter**
- **Sparse**
- **Homogeneous degree distribution**
- **Low clustering coefficient**

Molloy & Reed Model



Force a heterogenous degree distribution, with random linking

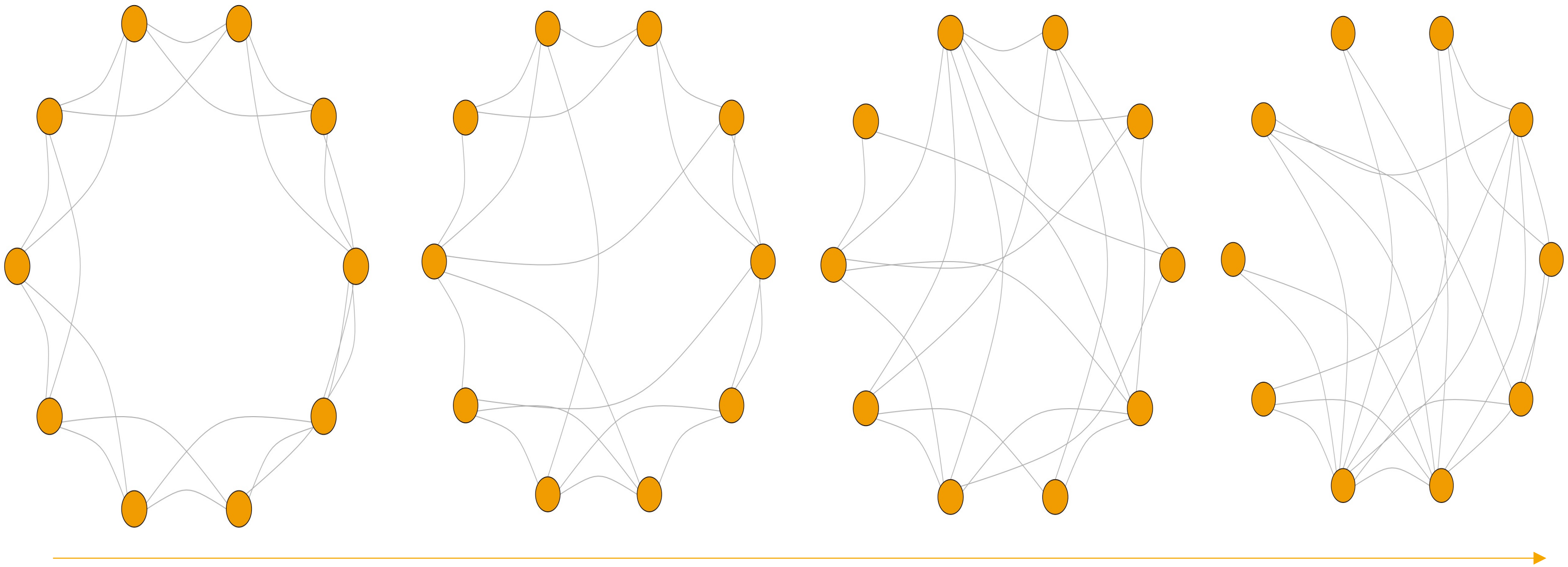
Generated graphs properties (see lab 1)

- **Low diameter**
- **Sparse**
- **Heterogenous degree distribution**
- **Low clustering coefficient**

Strogatz-Watts model

http://www.stevenstrogatz.com

https://en.wikipedia.org/wiki/Duncan_J._Watts



0 p 1

Strogatz-Watts model

Generated graphs properties (see lab 1)

- **Low diameter**
- **Sparse**
- **Homogeneous degree distribution**
- **High clustering coefficient**

Preferential Attachment model

Albert

Barbasi

Preferential attachment law

Nodes are joining the network one by one.
The probability of node v_i to connect to a new comer is proportional to $d(v_i)$

Algorithm 2: BA generation (n nodes, $(n - n_0)\alpha + m_0$ edges)

Parameters: n , G, α (degree arriving node)
with G connected graph with n_0 nodes and m_0 edges,
for i *from* $(n_0 + 1)$ *to* n **do**
  add node i to G
  $num = 0$ // num: number of links of i
  while $num < \alpha$ **do**
    draw $j \in [\![0; i-1]\!]$ with probability $\mathcal{P}(j) = \frac{k_j}{\sum_{q=0}^{i-1} k_q}$
    if $(i, j) \notin G$ **then**
      add edge (i, j) in G
      num++
    end
  end
end

Generated graphs properties (see lab 1)

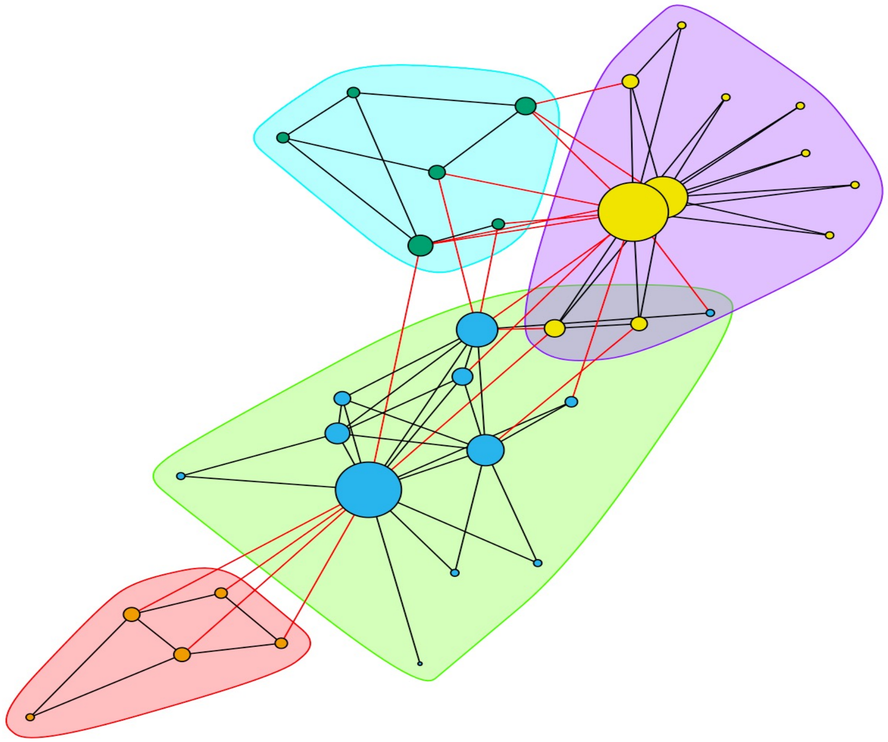
- **Low diameter**
- **Sparse**
- **Heterogenous degree distribution**
- **Very low clustering coefficient**

Random Graph Models

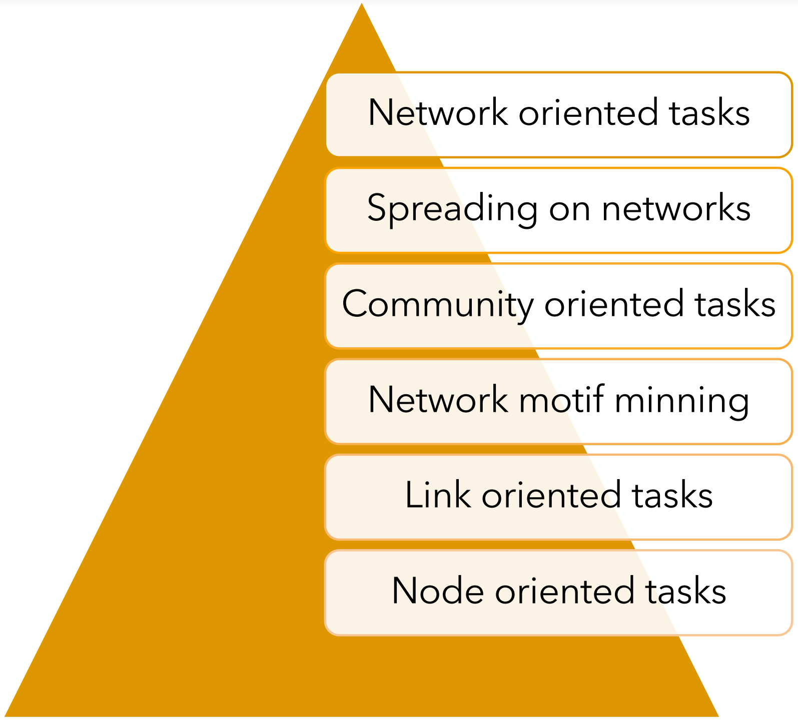
	Sparsity	Low diameter	Hetergenous degree distribution	Hich clustering coefficint
Erdös-Renyei	✅	✅	❌	❌
Molloy & Reed	✅	✅	✅	❌
Strogatz & Watts	✅	✅	❌	✅
Albert & Barbasi	✅	✅	✅	❌

5

Complex Networks Analysis Tasks



Complex Networks Analysis



Node oriented tasks

- **Compute node's importance**

  influence, vulnerability, Hub, Authority, ..

- **Applications**

  Viral marketing

  Influential nodes in social networks

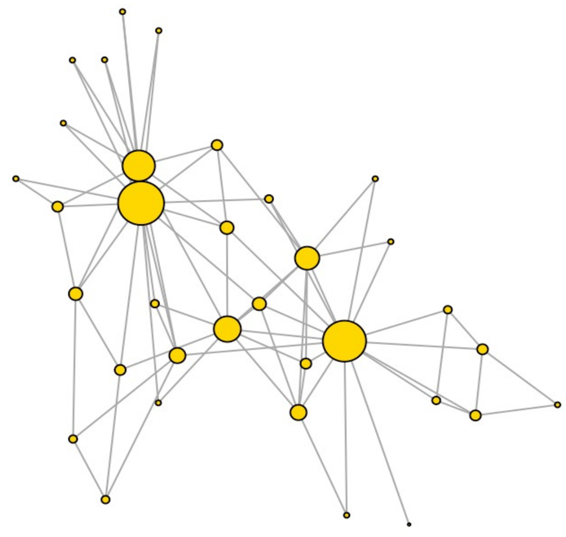
  Ranking web pages returned by a search engine

  Important sites in the web

  Where to vaccinate in priority to stop pandemic

  Control nodes in human contact networks

  …



…

Take home notes

Complex networks

- are everywhere
- are not random
- are not captured by a simple mathematical model
- share a set of untrivial topological features :
  - sparsity
  - low diameter
  - heterogeneous degree distribution
  - high clustering coefficient
  - community structure
- analysis can allow gaining insights in a wide range of domains.