Descriptive Network Analysis A

Dr Daniele Rotolo

SPRU (Science Policy Research Unit)
Business School
University of Sussex

Network Analysis
& Infographics

Week 4

Learning Outcomes

	Learning outcome	Assessment mode
1	Explain the concept of network and list the main network indicators	ESS
2	Describe and apply the major techniques for the collection of network data and their statistical analysis	ESS, GPN + GWS
3	Identify the main characteristics of networks by means of network measures	ESS, GPN + GWS
4	Employ network analysis techniques to produce network data-based infographics	GPN + GWS

Note: ESS: Essay; GPN: Group Presentation; GWS: Group Written Submission

Overview

1. Approaches to the analysis of networks [recap]

2. Network-level measures
   - Diameter
   - Average Path Length (APL)
   - Density
   - Components
   - Cutpoints and bridges
   - Connectivity
   - Cliques
   - Inclusiveness
   - Reachable pairs
   - Transitivity

Approaches to the analysis of networks [recap]

Approaches to the analysis of networks [recap]

Descriptive network analysis

- An observed network is analysed by means of measures
- Network-level measures
- Node-level measures

Modelling and inference of networks

- Mathematical models
  Based on ‘simple’ probabilist rules to capture specific mechanisms (e.g. Erdós-Rényi networks, ‘the rich get richer’)
- Statistical models
  The observed network is considered as one of the possible realisation of a process – a model that aims to fit to the observed data is specified (e.g. explanatory power of certain variables)

Approaches to the analysis of networks [recap]

Network Analysis is not a theory *per se*, but it a methodological tool to support the development of theories [Borgatti and Halgin, 2011]

- Network theory: mechanisms and processes that interact with network structures to produce certain outcomes for individuals, groups, and organisations (e.g. firms' performance, individuals' creativity)

- Theory of networks: mechanisms and processes that explain why certain networks have certain structures (i.e. antecedents of network properties)

	Dependent variable	
Independent variable	Nonnetwork variable as outcome	Network variable as outcome
Nonnetwork variable as antecedent	(Nonnetwork theory)	Theory of networks
Network variable as antecedent	Network theory	Network theory of networks

Source: [Borgatti and Halgin, 2011]

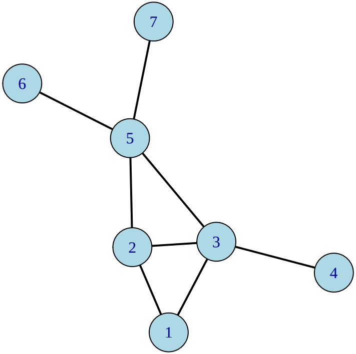
Network-level measures

Network-level measures

1. Diameter
2. Average Path Length (APL)
3. Density
4. Components
5. Cutpoints and bridges
6. Point/Line connectivity
7. Cliques
8. Inclusiveness
9. Reachable pairs
10. Transitivity

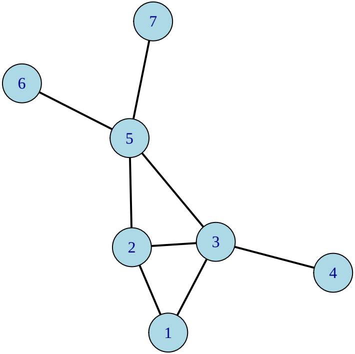
Network-level measures

1. Diameter
2. Average Path Length (APL)
3. Density
4. Components
5. Cutpoints and bridges
6. Point/Line connectivity
7. Cliques
8. Inclusiveness
9. Reachable pairs
10. Transitivity



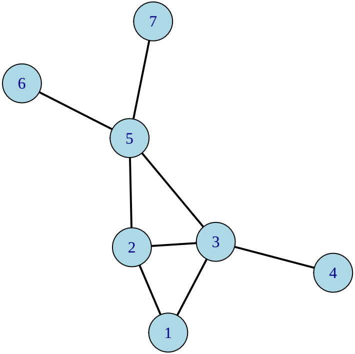
Note: We will mostly focus on undirected and unweighted networks

- To define the diameter of a network, we need first to recall the definition of geodesic distance



$d(n_i, n_j)$

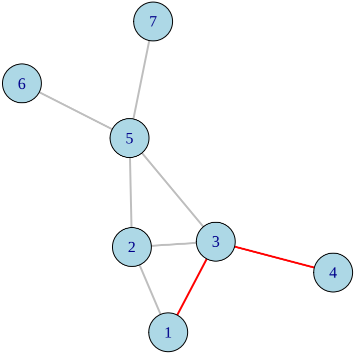
- To define the diameter of a network, we need first to recall the definition of geodesic distance
- The geodesic distance between two nodes n_i and n_j is the shortest path* between these nodes



$d(n_i, n_j)$

*A path is a sequence of nodes and lines (i.e. a walk) in which all nodes and links are distinct

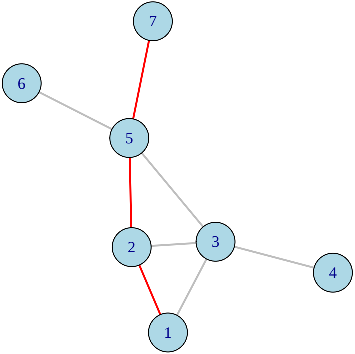
- To define the diameter of a network, we need first to recall the definition of geodesic distance
- The geodesic distance between two nodes n_i and n_j is the shortest path* between these nodes



$d(1,4) = 2$

*A path is a sequence of nodes and lines (i.e. a walk) in which all nodes and links are distinct

- To define the diameter of a network, we need first to recall the definition of geodesic distance
- The geodesic distance between two nodes n_i and n_j is the shortest path* between these nodes



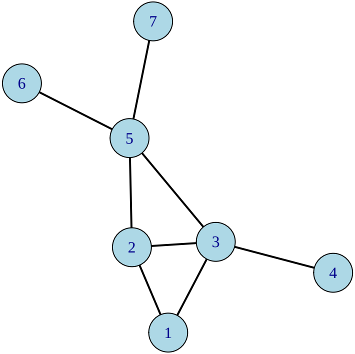
$d(1, 7) = 3$

*A path is a sequence of nodes and lines (i.e. a walk) in which all nodes and links are distinct

- The largest geodesic distance between any pair of nodes in a network is called diameter

$$D = max_i max_j d(n_i, n_j)$$

- The diameter of a network can range from 1 to $N - 1$

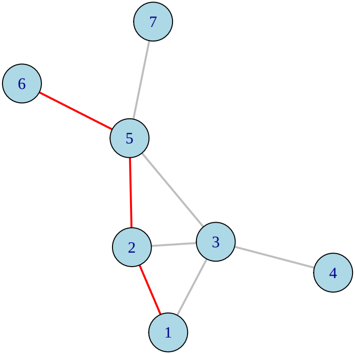


$$D = max_i max_j d(n_i, n_j)$$

- The largest geodesic distance between any pair of nodes in a network is called diameter

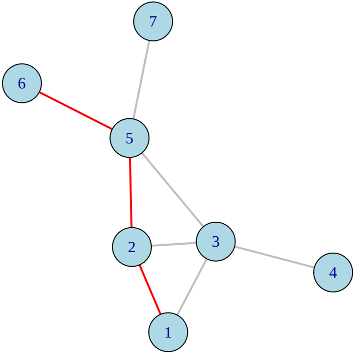
$$D = max_i max_j d(n_i, n_j)$$

- The diameter of a network can range from 1 to $N - 1$



$D = 3$

Nodes	Geodesic distance
1-2	1
1-3	1
1-4	2
1-5	2
1-6	3
1-7	3
2-3	1
2-4	2
2-5	1
2-6	2
2-7	2
3-4	1
3-5	1
3-6	2
3-6	2
4-5	2
4-6	3
4-7	3
5-6	1
5-7	1
6-7	2

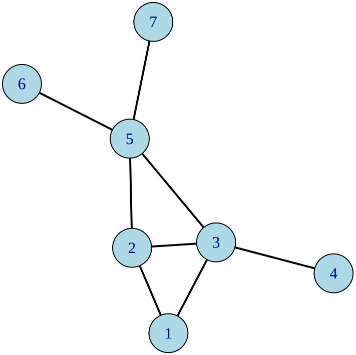


$$D = max_i max_j d(n_i, n_j) = 3$$

- Average Path Length (APL) of a network is defined as

$$APL = \frac{\sum_{n_i \neq n_j} d(n_i, n_j)}{\frac{N(N-1)}{2}}$$

- APL cannot be larger than the diameter of the network

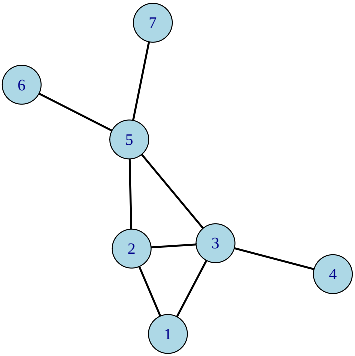


$APL = 38/21 = 1.81$

- The density of a network is defined as number of edges in the network out the number of possible edges

$$\Delta = \frac{E}{\frac{N(N-1)}{2}}$$

- N nodes, E edges
- The density of a network ranges from 0 (no edges between nodes) to 1 (fully-connected network)

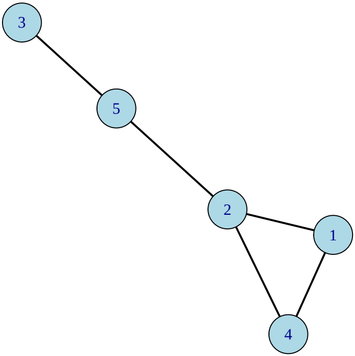


$$\Delta = \frac{8}{\frac{7(7-1)}{2}} = 0.38$$

- The density of a network is defined as number of edges in the network out the number of possible edges

$$\Delta = \frac{E}{\frac{N(N-1)}{2}}$$

- N nodes, E edges
- The density of a network ranges from 0 (no edges between nodes) to 1 (fully-connected network)

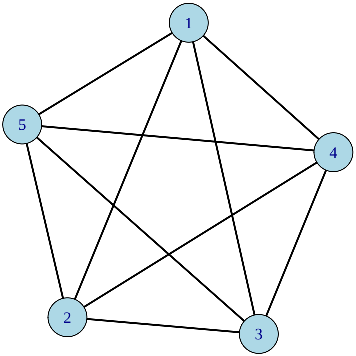


$$\Delta = \frac{5}{\frac{5(5-1)}{2}} = 0.5$$

- The density of a network is defined as number of edges in the network out the number of possible edges

$$\Delta = \frac{E}{\frac{N(N-1)}{2}}$$

- N nodes, E edges
- The density of a network ranges from 0 (no edges between nodes) to 1 (fully-connected network)



$$\Delta = \frac{10}{\frac{5(5-1)}{2}} = 1.0$$

Warnings when using the density measure for comparison purposes

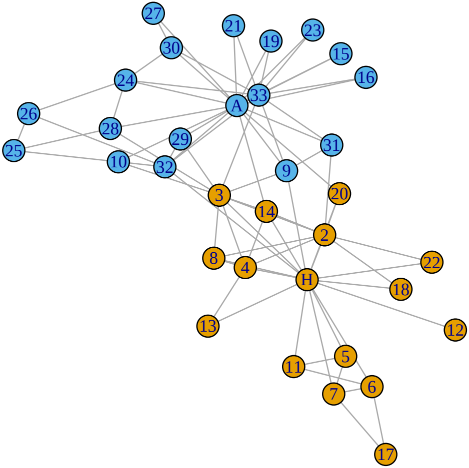
- The density measure is dependent on the size of the network (larger networks are likely to be less dense, i.e. more sparse)
- Comparison between different types of networks (e.g. who knows whom vs. who has a love affair with whom in an academic department)

Network-level measures
Example: Diameter, APL, density

Karate data including the social network between members of a university karate club

- $N = 34$
- $E = 78$
- $Diameter = 13$
- $APL = 2.41$
- $\Delta = 0.14$

```r
library(igraph)
library(igraphdata)
data(karate)
diameter(karate)
mean_distance(karate)
edge_density(karate)
setwd("YOUR WORKING DIRECTORY")
V(karate)$size <- 12
pdf(file = "karate.pdf",
    width = 4, height = 4)
plot(karate)
dev.off()
```

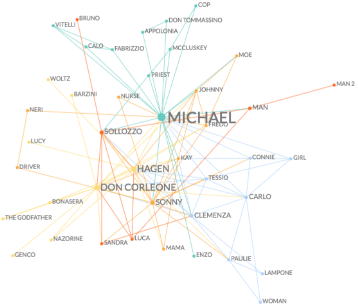


Source: [Zachary, 1977]

Network-level measures

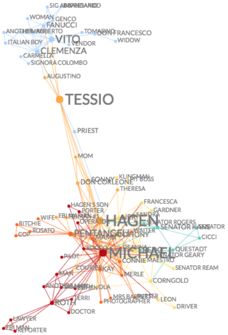
Example: Diameter, APL, density

The Godfather (1972)



Source: http://moviegalaxies.com

The Godfather (1974)



Source: http://moviegalaxies.com

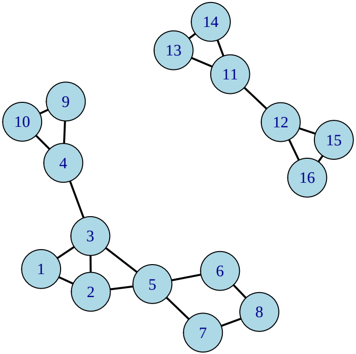
- $N = 42$; $E = 104$
- $Diameter = 4$; $APL = 2.26$; $\Delta = 0.12$

```
library(igraph)
g1 <- read_graph("gf1.gml", format = "gml")
diameter(g1)
mean_distance(g1)
edge_density(g1)
```

- $N = 78$; $E = 219$
- $Diameter = 7$; $APL = 3.06$; $\Delta = 0.07$

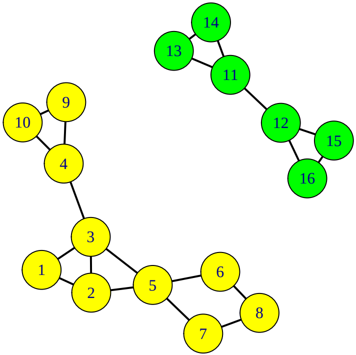
```
library(igraph)
g2 <- read_graph("gf2.gml", format = "gml")
diameter(g2)
mean_distance(g2)
edge_density(g2)
```

- A component is a connected subgraph of a *disconnected network*, i.e. a path between all pairs of nodes in the subgraph exists
- The number of components provides some indication about network connectivity
- The component with the largest number of nodes is called giant or largest component
- Network measures that are based on distances between nodes (e.g. APL) are assessed on the largest component of an disconnected network

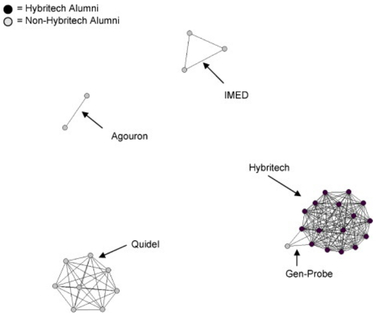


Network-level measures
Components

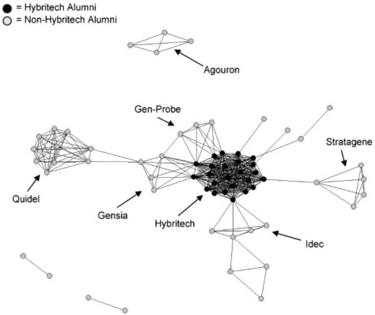
- A component is a connected subgraph of a *disconnected network*, i.e. a path between all pairs of nodes in the subgraph exists
- The number of components provides some indication about network connectivity
- The component with the largest number of nodes is called giant or largest component
- Network measures that are based on distances between nodes (e.g. APL) are assessed on the largest component of an disconnected network



Network-level measures
Components

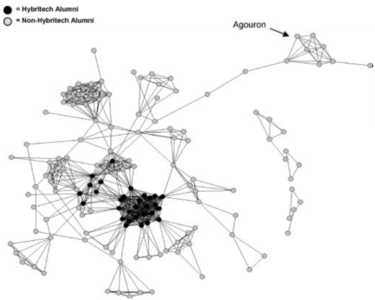


Source: Career affiliation network, San Diego 1984 [Casper, 2007]

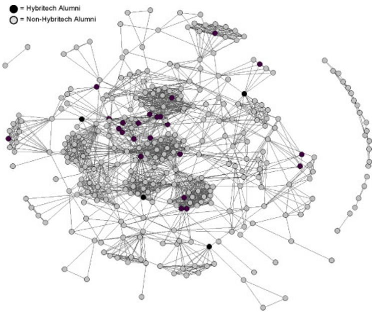


Source: Career affiliation network, San Diego 1987 [Casper, 2007]

Network-level measures
Components



Source: Career affiliation network, San Diego 1990 [Casper, 2007]



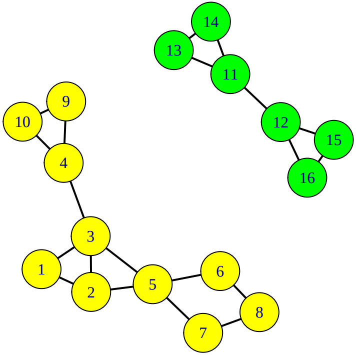
Source: Career affiliation network, San Diego 1995 [Casper, 2007]

Table 2
Descriptive statistics, San Diego career affiliation networks, 1978–2005

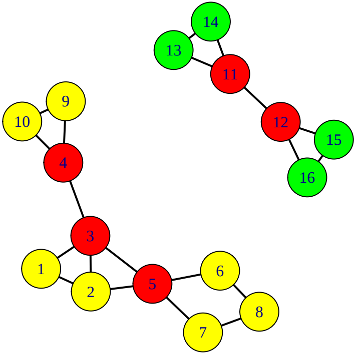
Year	Total individuals	Avg. people per firm	Size of main component	Percent in main component	Average path length	Network density
1978	2	2.0	2	1	1.0	1.00
1979	4	4.0	3	75%	1.0	0.50
1980	7	7.0	4	57%	1.0	0.33
1981	9	4.5	4	44%	1.0	0.19
1982	19	3.8	9	47%	1.0	0.25
1983	27	3.9	15	56%	1.0	0.33
1984	39	3.5	17	44%	1.1	0.21
1985	47	4.3	24	51%	1.2	0.25
1986	59	3.9	35	59%	1.8	0.19
1987	78	3.5	57	73%	2.8	0.13
1988	107	3.6	81	76%	3.1	0.10
1989	132	3.5	103	78%	3.3	0.07
1990	165	4.3	135	82%	3.9	0.06
1991	188	4.1	151	80%	3.7	0.05
1992	232	4.2	204	88%	3.8	0.05
1993	273	4.5	243	89%	4.1	0.04
1994	317	5.0	290	92%	4.0	0.04
1995	342	5.0	300	88%	3.6	0.04
1996	397	5.2	347	87%	3.5	0.04
1997	452	4.8	409	91%	3.6	0.03
1998	503	4.9	466	93%	3.6	0.03
1999	547	5.1	498	91%	3.6	0.03
2000	624	5.0	559	90%	3.8	0.02
2001	702	5.5	648	92%	3.8	0.02
2002	771	5.9	719	93%	3.8	0.02
2003	817	6.4	760	93%	3.8	0.02
2004	852	6.8	806	95%	3.9	0.02
2005	867	7.2	824	95%	4.2	0.02

Source: [Casper, 2007]

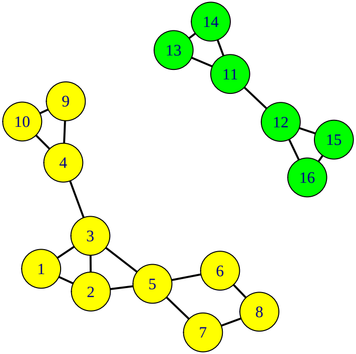
- A cutpoint is a node the removal of which increases the number of components



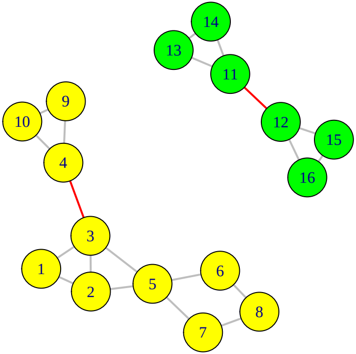
- A cutpoint is a node the removal of which increases the number of components



- A cutpoint is a node the removal of which increases the number of components
- A bridge is a link the removal of which increases the number of components

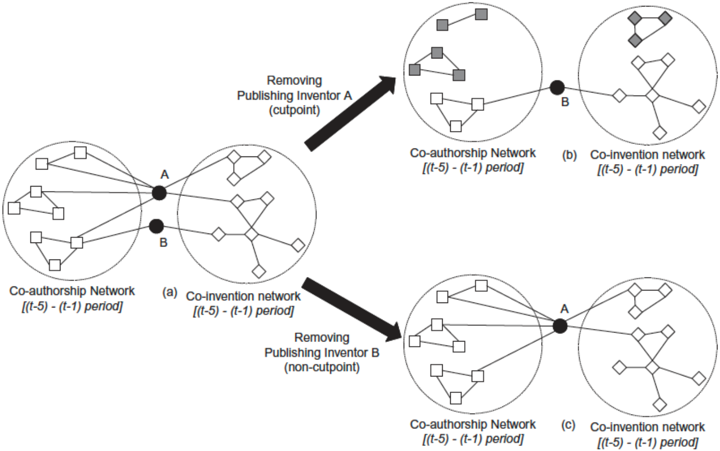


- A cutpoint is a node the removal of which increases the number of components
- A bridge is a link the removal of which increases the number of components



Network-level measures

Cutpoints and bridges

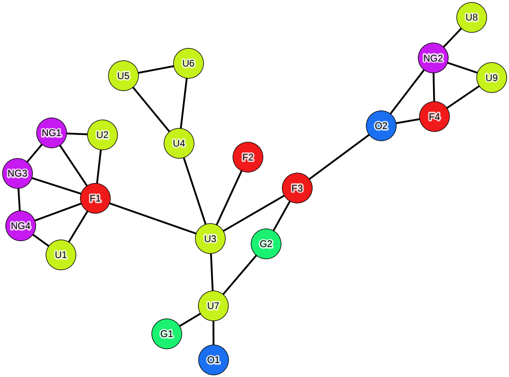


Source: [Cattani and Rotolo, 2013]

Network-level measures
Cutpoints and bridges

R&D project	List of partners
Proj01	U2, F1, NG1
Proj02	U1, NG4, F1
Proj03	NG3, NG1, F1
Proj04	NG3, NG4, F1
Proj05	U3, F1
Proj06	U3, F2
Proj07	U3, F3
Proj08	U3, U4
Proj09	F1
Proj10	U5
Proj11	U4, U5, U6
Proj12	U3, U7
Proj13	U7, G1
Proj14	U7, O1
Proj15	U7, G2
Proj16	G2, F3
Proj17	F3, O2
Proj18	O2, F4, NG2
Proj19	F4, U9, NG2
Proj20	NG2, U8

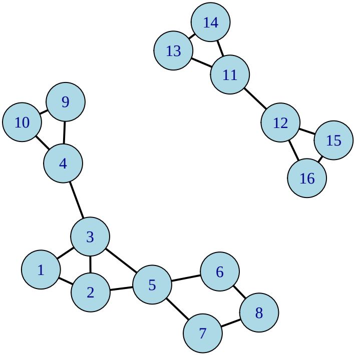
Firm (F); University (U); Gov. org. (G); Non-Gov. org. (NG); Other (O)



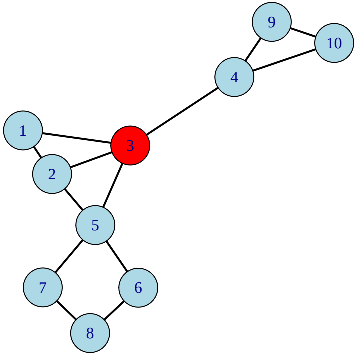
- Which nodes are cutpoints?
- How many bridges exist?

- The point-connectivity of a network is the minimum number of nodes we need to remove to make the network disconnected

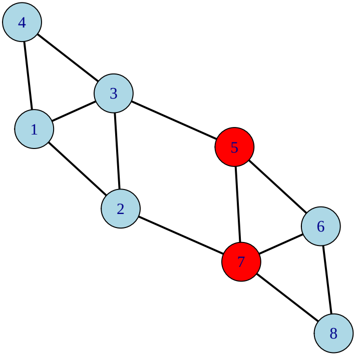
- The point-connectivity of a network is the minimum number of nodes we need to remove to make the network disconnected
- If the network is disconnected: $k = 0$



- The point-connectivity of a network is the minimum number of nodes we need to remove to make the network disconnected
- If the network is disconnected: $k = 0$
- If the network includes at least one cutpoint: $k = 1$

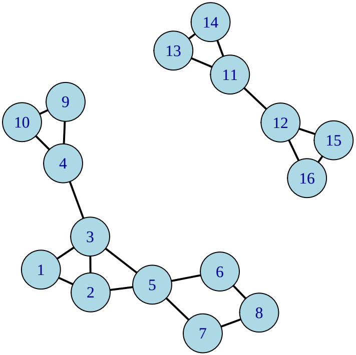


- The point-connectivity of a network is the minimum number of nodes we need to remove to make the network disconnected
- If the network is disconnected: $k = 0$
- If the network includes at least one cutpoint: $k = 1$
- If we need to remove at least two nodes to disconnect the network: $k = 2$

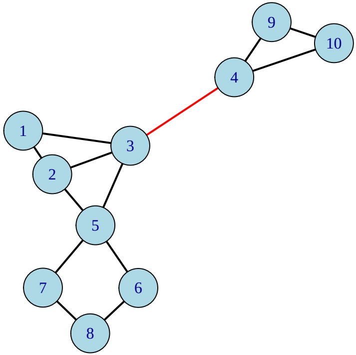


- The line-connectivity of a network is the minimum number of lines/edges we need to remove to disconnect the network

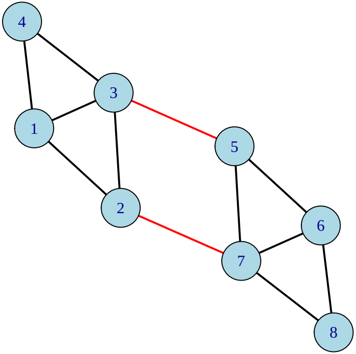
- The line-connectivity of a network is the minimum number of lines/edges we need to remove to disconnect the network
- If the network is disconnected: $l = 0$



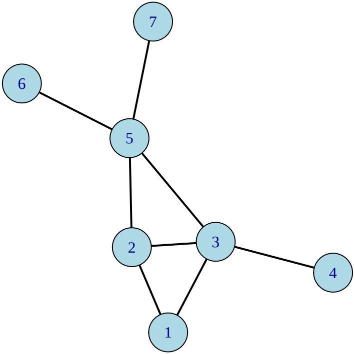
- The line-connectivity of a network is the minimum number of lines/edges we need to remove to disconnect the network
- If the network is disconnected: $l = 0$
- If the network includes one bridge: $l = 1$



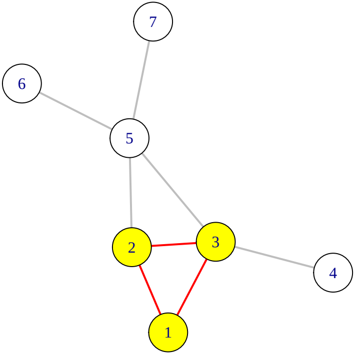
- The line-connectivity of a network is the minimum number of lines/edges we need to remove to disconnect the network
- If the network is disconnected: $l = 0$
- If the network includes one bridge: $l = 1$
- If we need to remove at least two lines to disconnect the network: $l = 2$



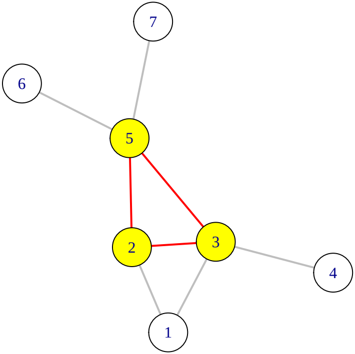
- A clique is a subgraph of three or more nodes where ties exist between every pair of nodes (maximal complete subgraph)



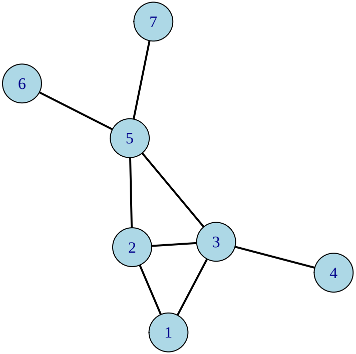
- A clique is a subgraph of three or more nodes where ties exist between every pair of nodes (maximal complete subgraph)



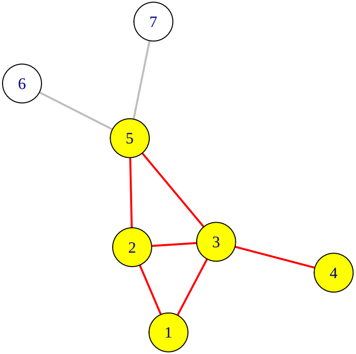
- A clique is a subgraph of three or more nodes where ties exist between every pair of nodes (maximal complete subgraph)



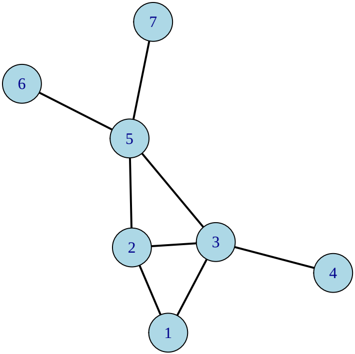
- A clique is a subgraph of three or more nodes where ties exist between every pair of nodes (maximal complete subgraph)
- An n-clique is a subgraph with largest geodesic distance between any pair of nodes not larger than n



- A clique is a subgraph of three or more nodes where ties exist between every pair of nodes (maximal complete subgraph)
- An n-clique is a subgraph with largest geodesic distance between any pair of nodes not larger than n

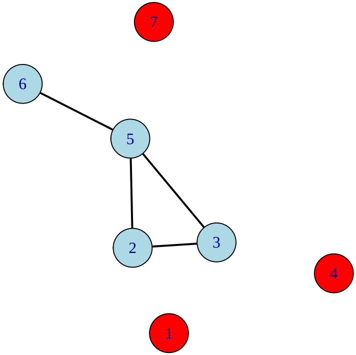


- Inclusiveness is a defined as the number of connected nodes out the total number of nodes in a network



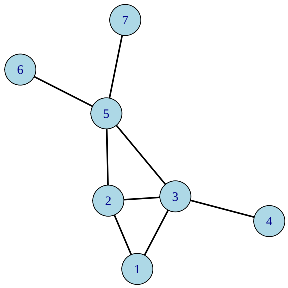
$inclusiveness = 7/7 = 1.00$

- Inclusiveness is a defined as the number of connected nodes out the total number of nodes in a network
- Nodes that have no ties are called isolates



$$\textit{inclusiveness} = 4/7 = 0.57$$

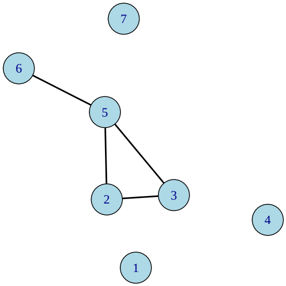
- Two nodes are reachable if a path between them exists (this property is called reachability)
- The number of reachable node pairs out the total number of node pairs would provide an indication of network connectivity
- Geodesic distance matrix



$$\mathbf{D} = \begin{pmatrix} - & . & . & . & . & . & . \\ 1 & - & . & . & . & . & . \\ 1 & 1 & - & . & . & . & . \\ 2 & 2 & 1 & - & . & . & . \\ 2 & 1 & 1 & 2 & - & . & . \\ 3 & 2 & 2 & 3 & 1 & - & . \\ 3 & 2 & 2 & 3 & 1 & 2 & - \end{pmatrix}$$

$$21/21 = 1.00$$

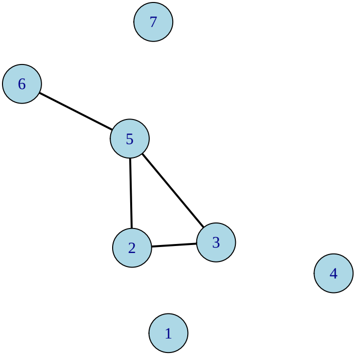
- Two nodes are reachable if a path between them exists (this property is called reachability)
- The number of reachable node pairs out the total number of node pairs would provide an indication of network connectivity
- Geodesic distance matrix



$$\mathbf{D} = \begin{pmatrix} - & . & . & . & . & . & . \\ \infty & - & . & . & . & . & . \\ \infty & 1 & - & . & . & . & . \\ \infty & \infty & \infty & - & . & . & . \\ \infty & 1 & 1 & \infty & - & . & . \\ \infty & 2 & 2 & \infty & 1 & - & . \\ \infty & \infty & \infty & \infty & \infty & \infty & - \end{pmatrix}$$

$$6/21 = 0.28$$

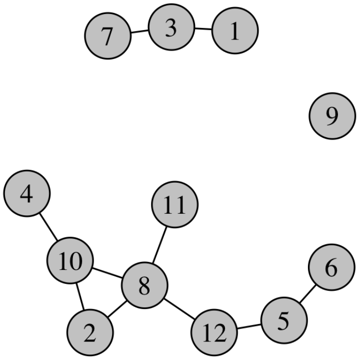
- Transitivity is defined as the number of *closed triads* out the number of *closed* and *open triads*
- Closed triad:
  - $n_i \leftrightarrow n_j$
  - $n_j \leftrightarrow n_k$
  - $n_i \leftrightarrow n_k$
- Open triad :
  - $n_i \leftrightarrow n_j$
  - $n_j \leftrightarrow n_k$
  - no tie between n_i and n_k



$$Transitivity = 3/5 = 0.60$$

Characterise the network in terms of

- Diameter
- APL
- Components
- Cutpoints
- Transivity



Network-level measures

Summary

Measure	Interpretation
Diameter	Maximum time/resources for communication, transfer, ...
APL	Average time/resources for communication, transfer, ...
Density	Connectivity of a network
Components	Presence of unconnected groups, bridging opportunities, ...
Cutpoints and bridges	Vulnerability/resilience of a network
Point/Line connectivity	Vulnerability/resilience of a network
Cliques	Highly connected sub-groups, exclusion, ...
Inclusiveness	Presence of unconnected nodes, exclusion, ...
Reachable pairs	Unconnected nodes or groups, bridging opportunities, ...
Transitivity	Social interactions, 'friends of my friends are my friends', ...

Questions

Next time ...

Next time ...

- **Seminar: Descriptive network analysis A**
  - Assessment of network-level measures in igraph
- **Lecture: Descriptive network analysis B**
  - Node-level measures (centrality measures)

References I

Borgatti, S. P. and Halgin, D. S. (2011).
On Network Theory.
Organization Science, 2(1):71–87.

Casper, S. (2007).
How do technology clusters emerge and become sustainable? Social network formation and inter-firm mobilitywithin the San Diego biotechnology cluster.
Research Policy, 36(4):438–455.

Cattani, G. and Rotolo, D. (2013).
An exploratory study of the role of publishing inventors in nanotechnology.
In Aharonson, B. S., Stettner, U., Amburgey, T. L., Ellis, S., and Drori, I., editors, *Understanding the Relationship Between Networks and Technology, Creativity and Innovation (Technology, Innovation, Entrepreneurship and Competitive Strategy)*, volume 13, chapter 4, pages 97–122. Emerald Group Publishing Limited, Bingley, UK.

Zachary, W. W. (1977).
An Information Flow Model for Conflict and Fission in Small Groups.
Journal of Anthropological Research, 33(4):452–473.