Social Networks

LECTURE 4

Sibylle Mohr
Sibylle.Mohr@glasgow.ac.uk

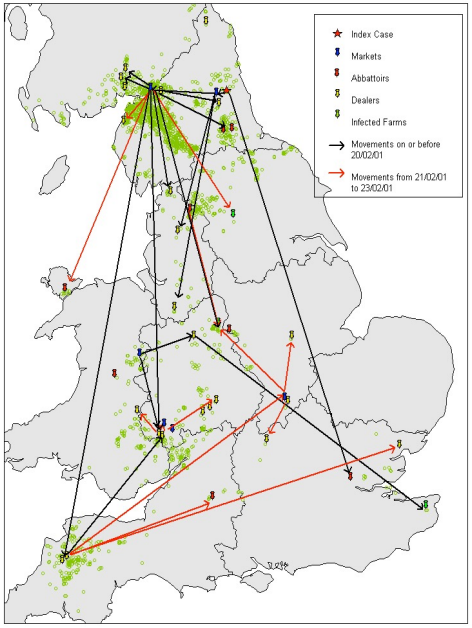
Lecture plan

- Network concepts and terminology
- Network centrality measures (social network analysis)
- Network properties and their effect on epidemic outbreaks
  - Node and edge properties
  - Larger-scale network structure (path length, communities)

- Network Models
  - Random networks
  - Small-world networks
  - Scale-free networks

Network Science: Crossing Disciplines

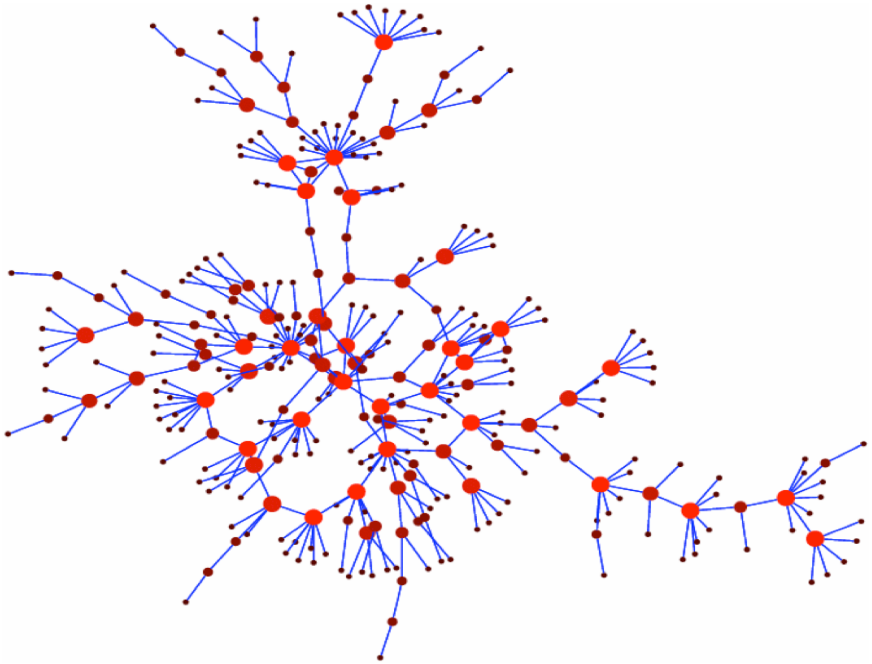
- Study of complex systems through their network representations (economy, metabolism, brain, society, Web, . . .)
- Universal language for describing complex systems and data
- Striking similarities in networks across science, nature, technology
- Understand complex systems ⇔ Understand networks behind them

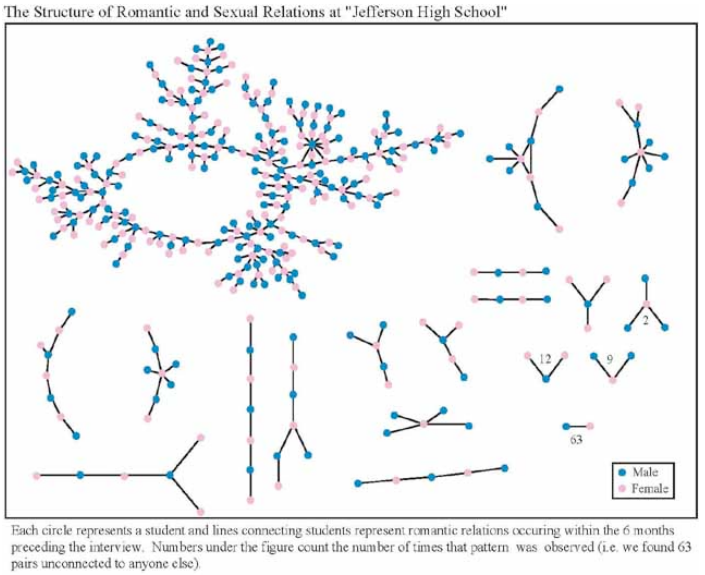


Initial spread of FMD (2001, UK) through livestock movement

(a) Gibbens, J.C. *et al.* (2001) Descriptive epidemiology of the 2001 foot-and-mouth
disease epidemic in Great Britain: the first five months. *Vet. Rec.* 149, 729–743.
(b) Rowland R. Kao (2002) The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. Trends in Microbiology, 279-286.

Why networks?





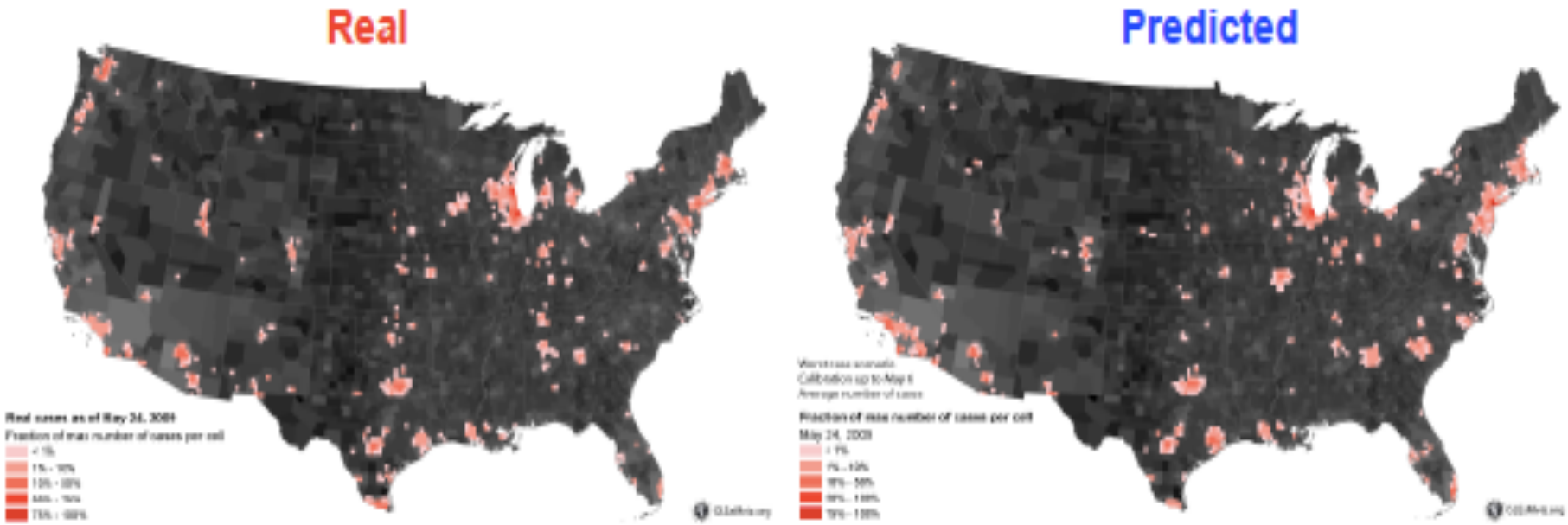
- Diseases are spread through social networks.
- This is especially relevant to sexually transmitted diseases such as AIDS.
- "*Contact tracing*" is an important part of any strategy to combat outbreaks of diseases (e.g. smallpox, FMD, …)
- Is one of these networks more dangerous for disease spread?
- Are there important nodes for policy to target?

Network Models

- Provide insight into why we see certain phenomena:
  - Why do social networks have short average path lengths?
- Allow for comparative statics:
  - How does component structure change with density? (important in contagion, diffusion...)
- Predict out of sample:
  - What will happen with a new policy (vaccine, movement restrictions, ...)?
- Allow for statistical estimation:
  - Is there significant clustering on a local level or did it appear at random?

Healthcare Impact

- Prediction of epidemics , e.g. the 2009 H1N1 pandemic



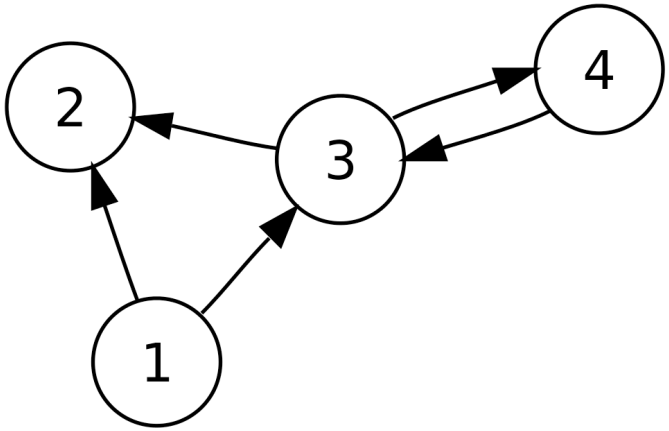
What do We Know?

- Easy and intuitive representation of contacts / interactions within complex populations
- Networks play role in many settings
- Network position and structure matters
  - Better estimate of “at risk” individuals
  - Identify “key player” e.g. “super spreaders”
  - Develop targeted surveillance / control strategies
  - Identify source of disease (More realistic models for disease spread?)
- ``Social’’ Networks have special characteristics
  - small worlds, degree distributions...

Analyse / Visualize Graphs

Lots of available software:

- Packages in R (igraph , sna, network, statnet….)
- Gephi
- Pajek
- Python packages: networkx



File formats:
- Many!
- csv, DOT files are read by most

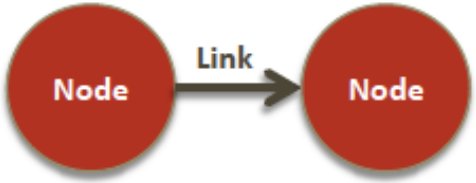
```
digraph untitled {
  node [shape=circle, width="0.03"];
  1 -> 2;
  1 -> 3;
  3 -> 2;
  3 -> 4;
  4 -> 3;
}
```

Network concepts and terminology

Terminology

- **Node**: N={1,…,n}, the entity of interest
  - vertices, actors, players…

- **Edge**: the relationship of interest
  - also called a tie, link
  - "weighted" or "unweighted"
  - "undirected'' or "directed''



- **Network**: a collection of nodes and links
  - Also called a graph

Types of nodes

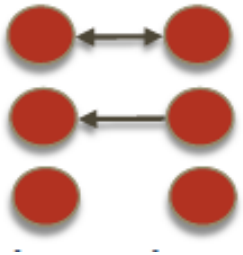
- Individual units
  - Humans
  - Animals
  - Airports
  - Computers
  - Genes
- Collectivities
  - Countries,cities
  - Families
  - Species
  - Organs

Types of links

- Social
- Affective (like/dislike, trust/do not trust)
- Kinship / social role (mother of, brother of, boss of)
- Exchange (advice seeking, sexual intercourse, trade)
- Cognitive (knows/does not know)
- Affiliation (belongs to, is a member of)
- …

Edge properties

- Directed (e.g., “likes”)
  - Mutual
  - Asymmetric
  - Null



Nodes are now classified as senders and receivers

- A directed graph is also called a di-graph
- A directed edge is also called an arc

- Undirected (e.g., “dances with”)

Edge attributes

- Examples
  - weight (e.g. frequency of communication, number of animals moved)
  - ranking (best friend, second best friend…)
  - type (friend, relative, co-worker)
  - properties depending on the structure of the rest of the graph: e.g. betweenness

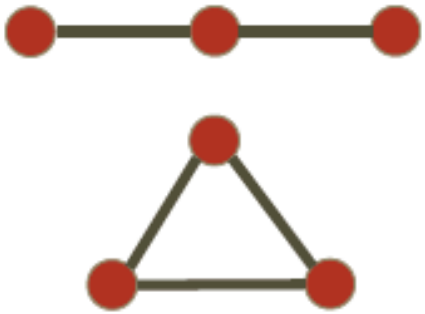
Configurations

Any collection of nodes and links can be defined as a configuration

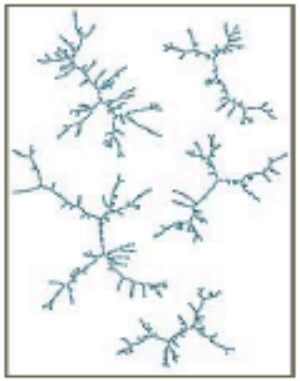
Dyads

Stars

Triples & Triangles

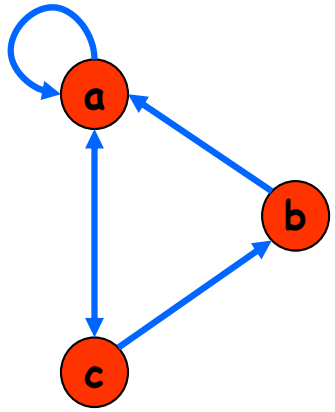


Components



Representing networks mathematically

G = (V, E)



Loop

$$G = \{a, b, c\}$$

$$E(G) = \{aa, ac, ba, ca, cb\}$$

Edge List

a	a
a	c
b	a
c	a
c	b

Adjacency Matrix

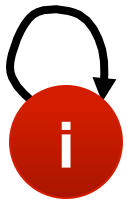
$$A(G) = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

Adjacency Lists

Adjacency matrices

- Representing edges (who is adjacent to whom) as a matrix
  - A_{ij} = 1 if node *i* has an edge to node *j*
    = 0 if node *i* does not have an edge to j

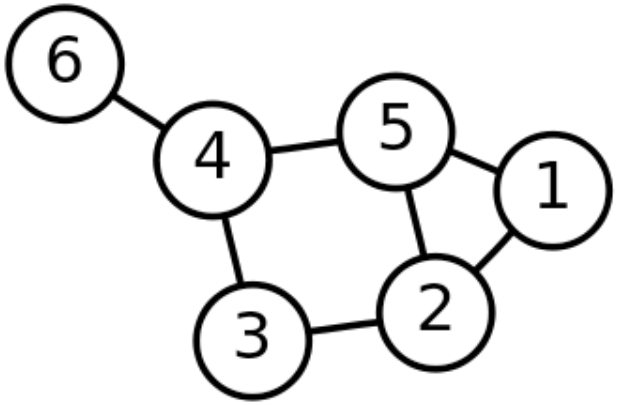
  - A_{ii} = 0 unless the network has self-loops



  - $A_{ij} = A_{ji}$ if the network is undirected, or if *i* and *j* share a reciprocated edge

Example Adjacency Matrix: Undirected Graph

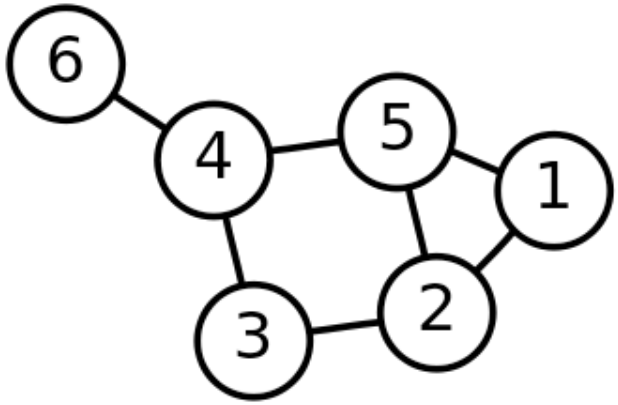
Adjacency matrix:



$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Example Adjacency Matrix: Undirected Graph

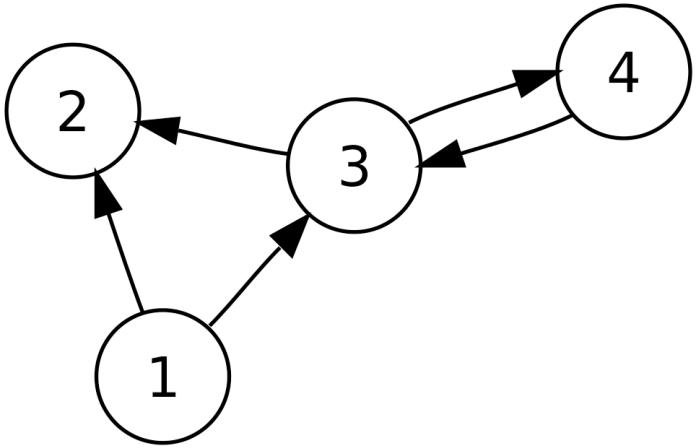
Adjacency matrix:



A =

	1	2	3	4	5	6
1	1	0	0	1	0	
2		1	0	1	0	
3			1	0	0	
4				1	1	
5					0	
6						

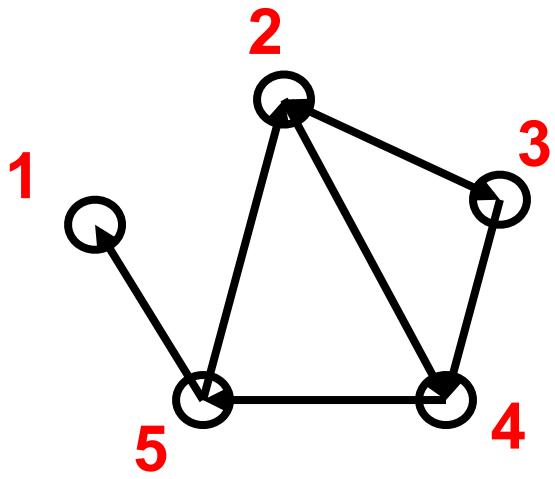
Example Adjacency Matrix: Directed Graph



$$A = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 1 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 1 \\ 4 & 0 & 0 & 1 & 0 \end{pmatrix}$$

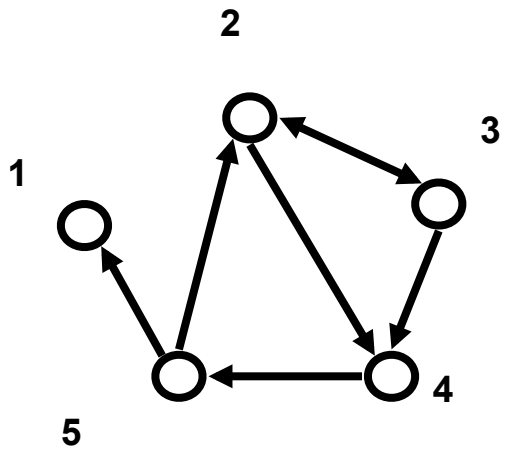
Edge list

- Edge list
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1



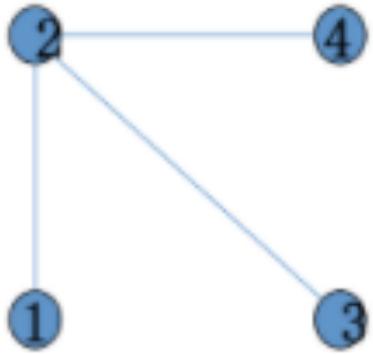
Adjacency lists

- Adjacency list
  - is easier to work with if network is
    - large
    - sparse
  - quickly retrieve all neighbors for a node
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2

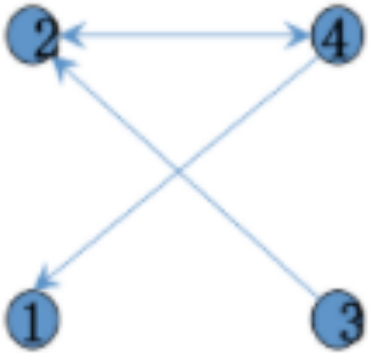


Exercise

Which of the following represent(s) **undirected** network(s)?



a)



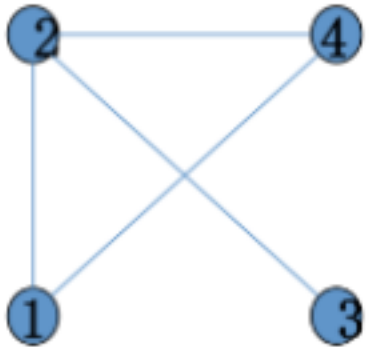
b)

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

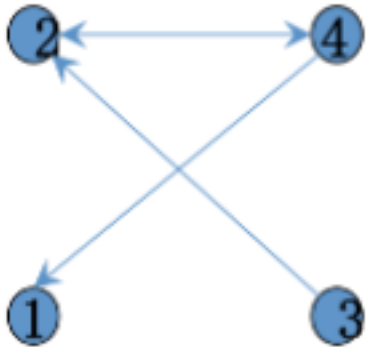
c)

Exercise

Which of the following represent(s) **directed** network(s)?



a)



b)

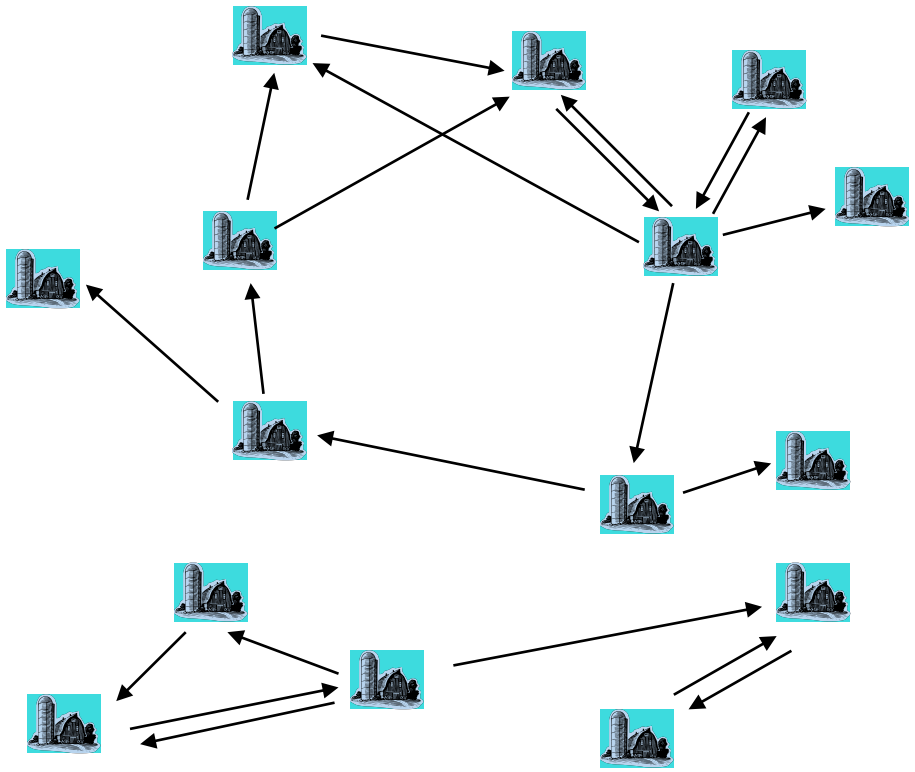
$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

c)

{12, 23, 34, 32, 14}

d)

How do we measure and make sense of networks?



Network of livestock movement between Farms (Markets and Dealers)

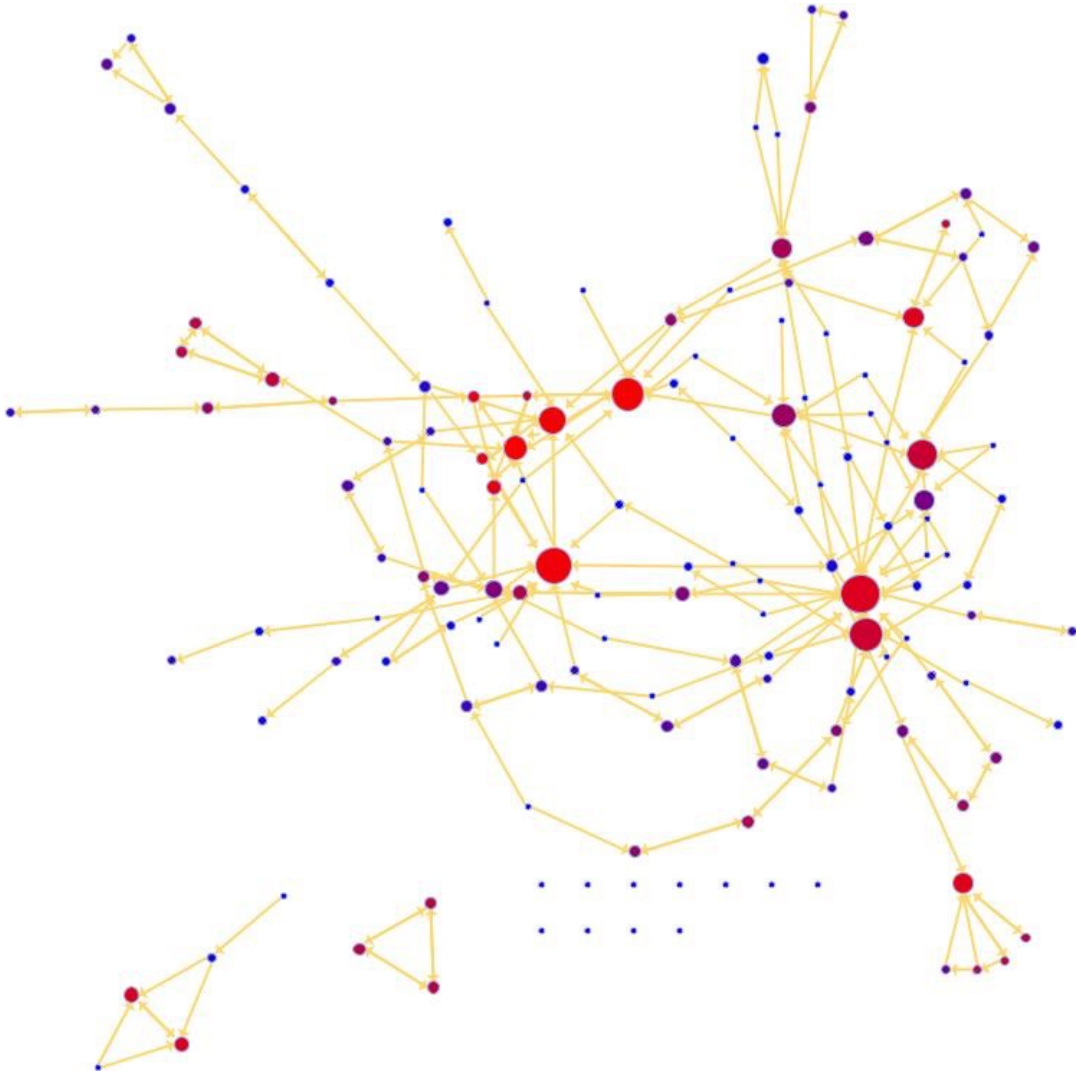
- Can we say whether this network is a 'bad' network?
  - **How fast and how far will the disease spread?**
- Can we identify vertices and/or links that are particularly dangerous?

How do we go about analysing networks?

Network measures and metrics (SNA)

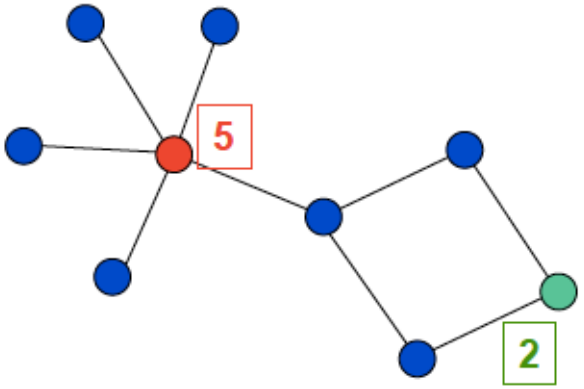
- Node properties
  - Node degree (number of connections)
    - Distribution of degree
- Edge properties
  - Direction
  - Timing
  - Weighting
- Mixing properties (who is connected to whom, is it at random?)
  - Clustering
  - Proportionate/ assortative/ disassortative mixing
- Large-scale properties
  - Path lengths
  - Components & communities

Degree: which node has the most edges?

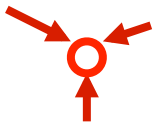


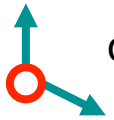
Node Degree

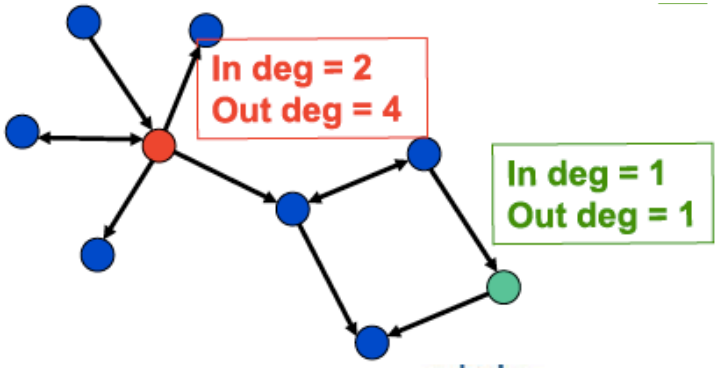
- Node degree: number of contacts made by a node to all other nodes (number of neighbours)

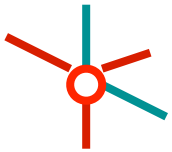


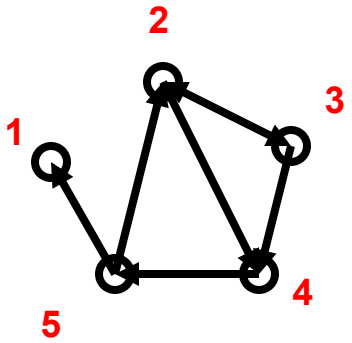
- Directed graphs:
  - In-degree: the number of edges directed towards the node
  - Out-degree: the number of edges leaving a node


indegree=3


outdegree=2




degree=5



Node degree from matrix values

- Outdegree = $\sum_{j=1}^{n} A_{ij}$

example: outdegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd *row* $\sum_{j=1}^{n} A_{3j}$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

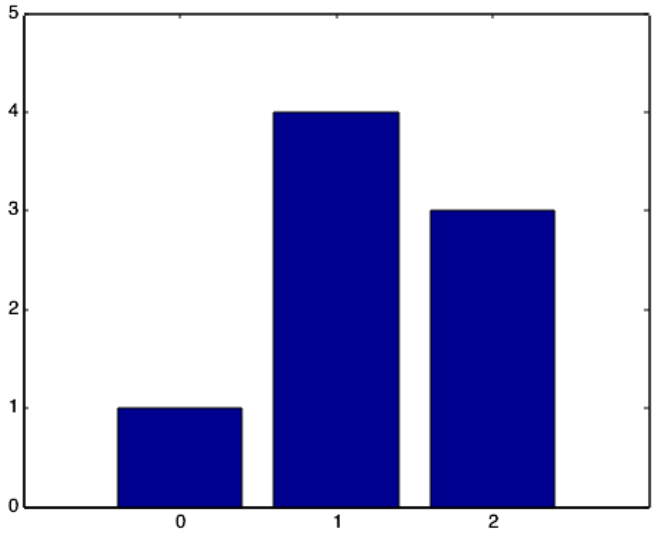
- Indegree = $\sum_{i=1}^{n} A_{ij}$

example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd *column* $\sum_{i=1}^{n} A_{i3}$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Network metrics: degree distribution

- Degree distribution: A frequency count of the occurrence of each degree
  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]



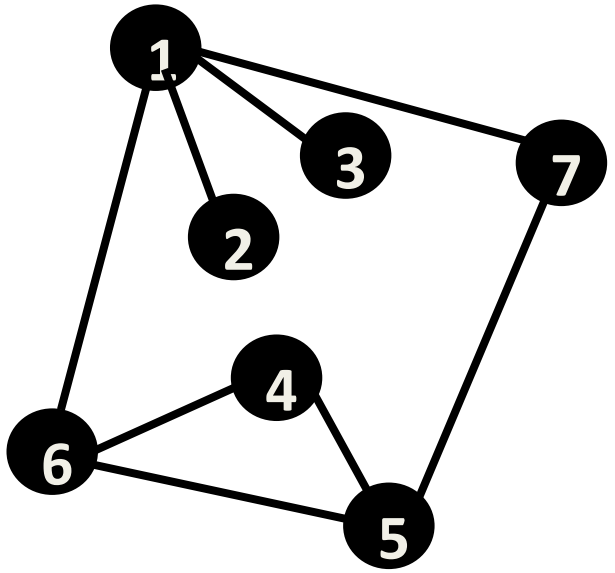
- **Undirected network**
  - higher vertex degree ~ more likely to become infected and likely to infect many
- **Directed networks**
  - in-degree ~ if high, more likely to become infected
  - out-degree ~ if low, less likely to infect many

Degree distribution *P(k)* = how many nodes or what proportion of nodes have degree *k (most often proportion is used)*

Average Degree

- Average number of edges/ links per vertex/node, that is averaged over all nodes in the network
- More edges mean more routes for disease to transmit

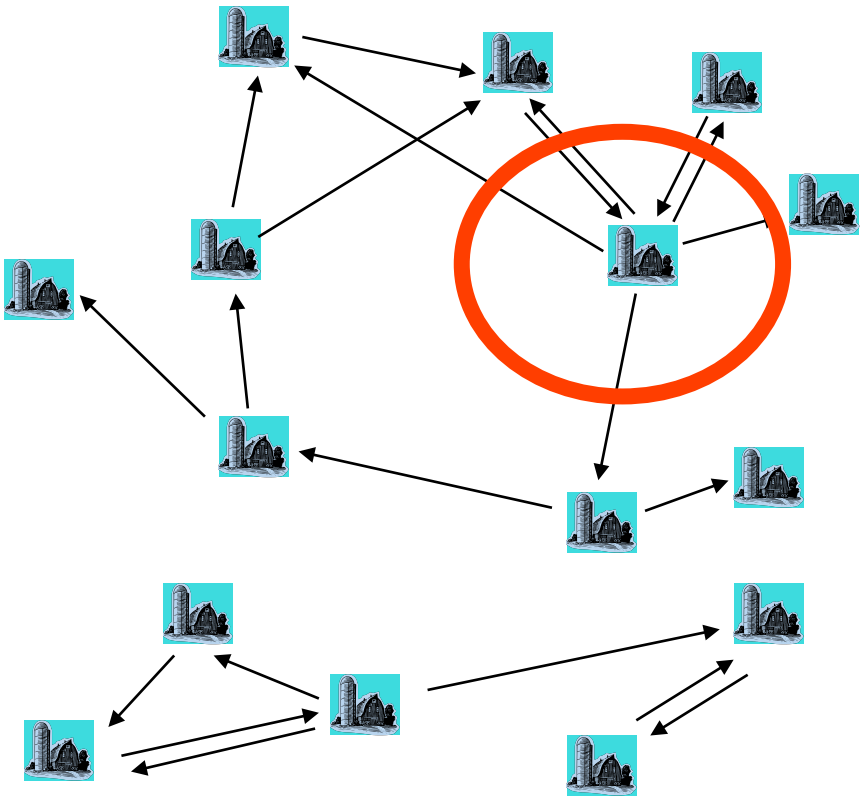
Degree(1)=4,Degree(2)=1,Degree(3)=1,Degree(4)=2,Degree(5)=3,Degree(6)=3,Degree(7)=2



$$\langle k \rangle = \sum_{i=k_{min}}^{k_{max}} i \times P(i)$$

$$\langle k \rangle = \frac{4+1+1+2+3+3+2}{7}$$

Degree heterogeneity



- Can we say whether this network is a ‘bad’ network?
- Can we identify vertices and/or links that are particularly dangerous?

High chance of becoming infected and transmitting infection (many in and out links)

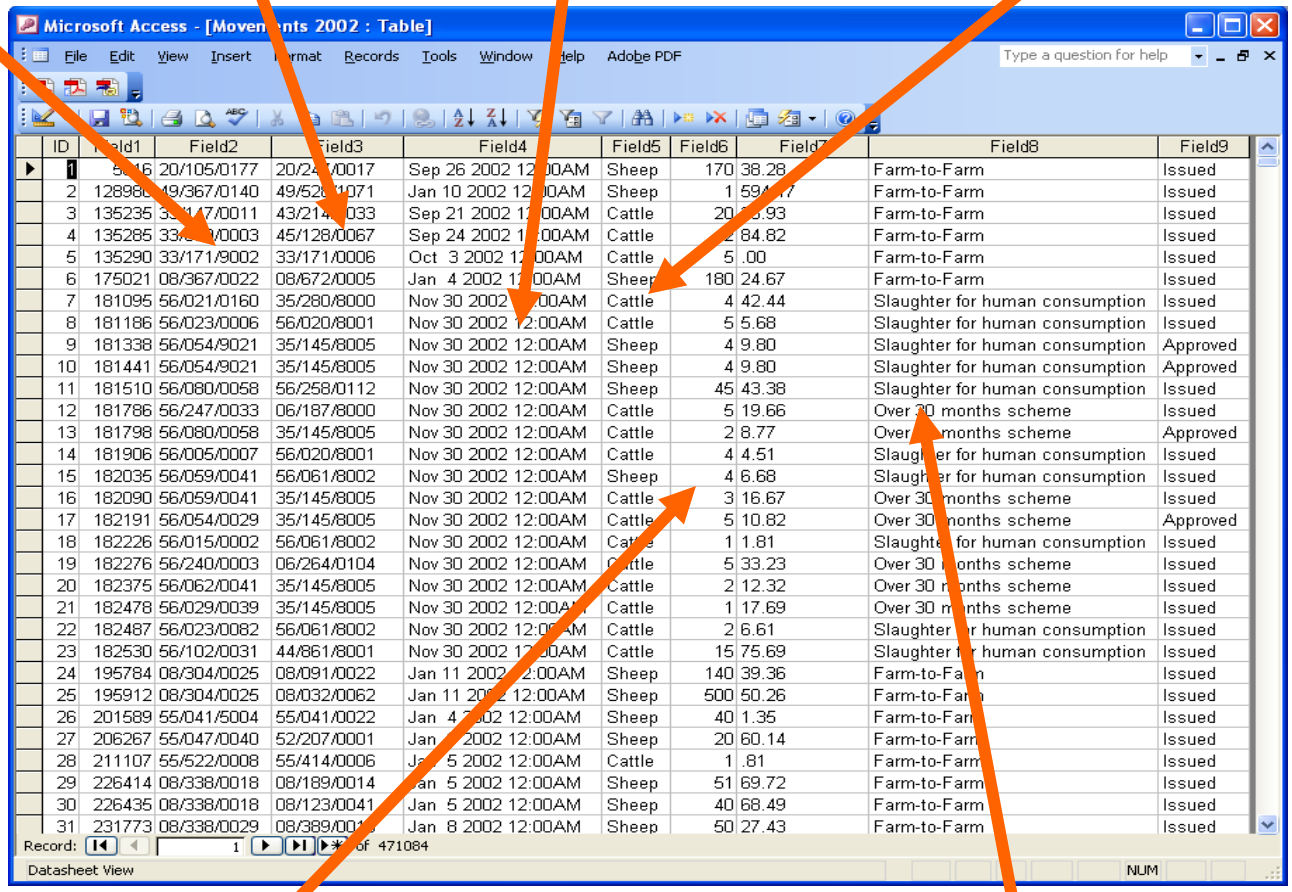
Livestock Movement Database

Movement From

Movement To

Date of Movement

Type of Livestock



ID	Field1	Field2	Field3	Field4	Field5	Field6	Field7	Field8	Field9
1	5016	20/105/0177	20/24[illegible]/0017	Sep 26 2002 12:00AM	Sheep	170	38.28	Farm-to-Farm	Issued
2	128980	49/367/0140	49/52[illegible]/071	Jan 10 2002 12:00AM	Sheep	1	594.17	Farm-to-Farm	Issued
3	135235	33/147/0011	43/214/0033	Sep 21 2002 12:00AM	Cattle	20	3.93	Farm-to-Farm	Issued
4	135285	33/[illegible]/0003	45/128/0067	Sep 24 2002 12:00AM	Cattle	2	84.82	Farm-to-Farm	Issued
5	135290	33/171/9002	33/171/0006	Oct 3 2002 12:00AM	Cattle	5	.00	Farm-to-Farm	Issued
6	175021	08/367/0022	08/672/0005	Jan 4 2002 12:00AM	Sheep	180	24.67	Farm-to-Farm	Issued
7	181095	56/021/0160	35/280/8000	Nov 30 2002 12:00AM	Cattle	4	42.44	Slaughter for human consumption	Issued
8	181186	56/023/0006	56/020/8001	Nov 30 2002 12:00AM	Cattle	5	5.68	Slaughter for human consumption	Issued
9	181338	56/054/9021	35/145/8005	Nov 30 2002 12:00AM	Sheep	4	9.80	Slaughter for human consumption	Approved
10	181441	56/054/9021	35/145/8005	Nov 30 2002 12:00AM	Sheep	4	9.80	Slaughter for human consumption	Approved
11	181510	56/080/0058	56/258/0112	Nov 30 2002 12:00AM	Sheep	45	43.38	Slaughter for human consumption	Issued
12	181786	56/247/0033	06/187/8000	Nov 30 2002 12:00AM	Cattle	5	19.66	Over 30 months scheme	Issued
13	181798	56/080/0058	35/145/8005	Nov 30 2002 12:00AM	Cattle	2	8.77	Over [illegible] months scheme	Approved
14	181906	56/005/0007	56/020/8001	Nov 30 2002 12:00AM	Cattle	4	4.51	Slaughter for human consumption	Issued
15	182035	56/059/0041	56/061/8002	Nov 30 2002 12:00AM	Sheep	4	6.68	Slaughter for human consumption	Issued
16	182090	56/059/0041	35/145/8005	Nov 30 2002 12:00AM	Cattle	3	16.67	Over 30 months scheme	Issued
17	182191	56/054/0029	35/145/8005	Nov 30 2002 12:00AM	Cattle	5	10.82	Over 30 months scheme	Approved
18	182226	56/015/0002	56/061/8002	Nov 30 2002 12:00AM	Cattle	1	1.81	Slaughter for human consumption	Issued
19	182276	56/240/0003	06/264/0104	Nov 30 2002 12:00AM	Cattle	5	33.23	Over 30 months scheme	Issued
20	182375	56/062/0041	35/145/8005	Nov 30 2002 12:00AM	Cattle	2	12.32	Over 30 months scheme	Issued
21	182478	56/029/0039	35/145/8005	Nov 30 2002 12:00AM	Cattle	1	17.69	Over 30 months scheme	Issued
22	182487	56/023/0082	56/061/8002	Nov 30 2002 12:00AM	Cattle	2	6.61	Slaughter for human consumption	Issued
23	182530	56/102/0031	44/861/8001	Nov 30 2002 12:00AM	Cattle	15	75.69	Slaughter for human consumption	Issued
24	195784	08/304/0025	08/091/0022	Jan 11 2002 12:00AM	Sheep	140	39.36	Farm-to-Farm	Issued
25	195912	08/304/0025	08/032/0062	Jan 11 2002 12:00AM	Sheep	500	50.26	Farm-to-Farm	Issued
26	201589	55/041/5004	55/041/0022	Jan 4 2002 12:00AM	Sheep	40	1.35	Farm-to-Farm	Issued
27	206267	55/047/0040	52/207/0001	Jan 2002 12:00AM	Sheep	20	60.14	Farm-to-Farm	Issued
28	211107	55/522/0008	55/414/0006	Jan 5 2002 12:00AM	Cattle	1	.81	Farm-to-Farm	Issued
29	226414	08/338/0018	08/189/0014	Jan 5 2002 12:00AM	Sheep	51	69.72	Farm-to-Farm	Issued
30	226435	08/338/0018	08/123/0041	Jan 5 2002 12:00AM	Sheep	40	68.49	Farm-to-Farm	Issued
31	231773	08/338/0029	08/389/001[illegible]	Jan 8 2002 12:00AM	Sheep	50	27.43	Farm-to-Farm	Issued

Record: 1 of 471084

Datasheet View NUM

Number of Livestock Moved

Type of Premises (i.e. Farm, Market, etc)

Highly heterogeneous Network

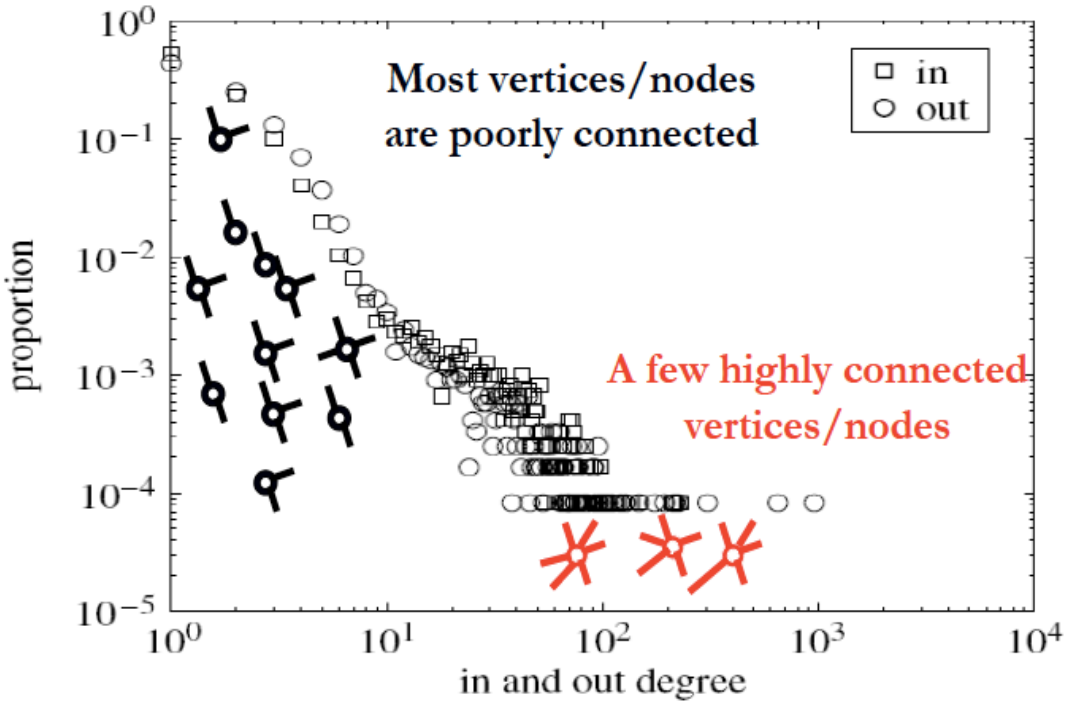


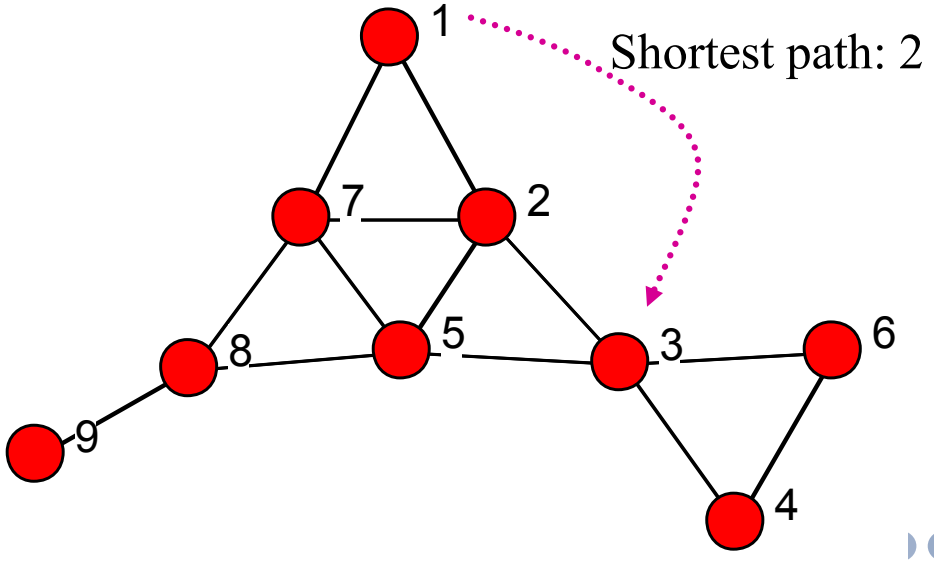
Figure 2. The in and out degree distribution of the sheep movement network starting on 8 September 2004.

- Vaccination/ Contact tracing
- Highly connected nodes can be preferentially targeted for control
- Where degree distribution is homogeneous it is difficult to use targeted control

I. Z. Kiss, D. M. Green and R. R. Kao (2006) The network of sheep movements within Great Britain: network properties and their implication for infectious disease spread. J. R. Soc. Interface 3, 669 - 677

Network Measures: Path length

- Path: A trail that no node is visited more than once
- Shortest path length: shortest directed path linking any given pair of nodes

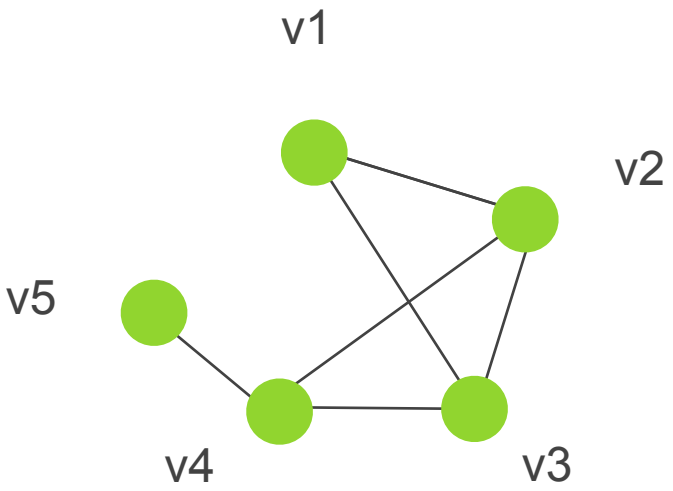


The higher the path lengths the more difficult it is to go from one vertex to another
= long chains of transmission, many generations of infection

--> more opportunities to control spread

Geodesic Path

- Shortest path ⇔ geodesic path
- Length of shortest path often called geodesic distance / shortest distance

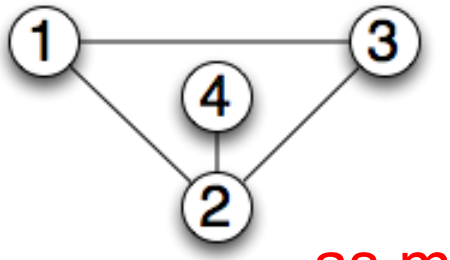


g(1,2) = 1
g(1,3) = 1
g(1,4) = 2
g(1,5) = 3
g(2,3) = 1
g(2,4) = 1
g(2,5) = 2
g(3,4) = 1
g(3,5) = 2
g(4,5) = 1

Exercise

- Create the adjacency matrix, adjacency list, and edge list for this graph (in Excel or text editor; -> save as .csv -> read into R).

  Or do it in R only.



as.matrix()
graph.adjacency
?graph_from_

- How would you get a vertex's degree directly from the matrix?

- Calculate degree per node in R-igraph and plot degree distribution

- Calculate average degree

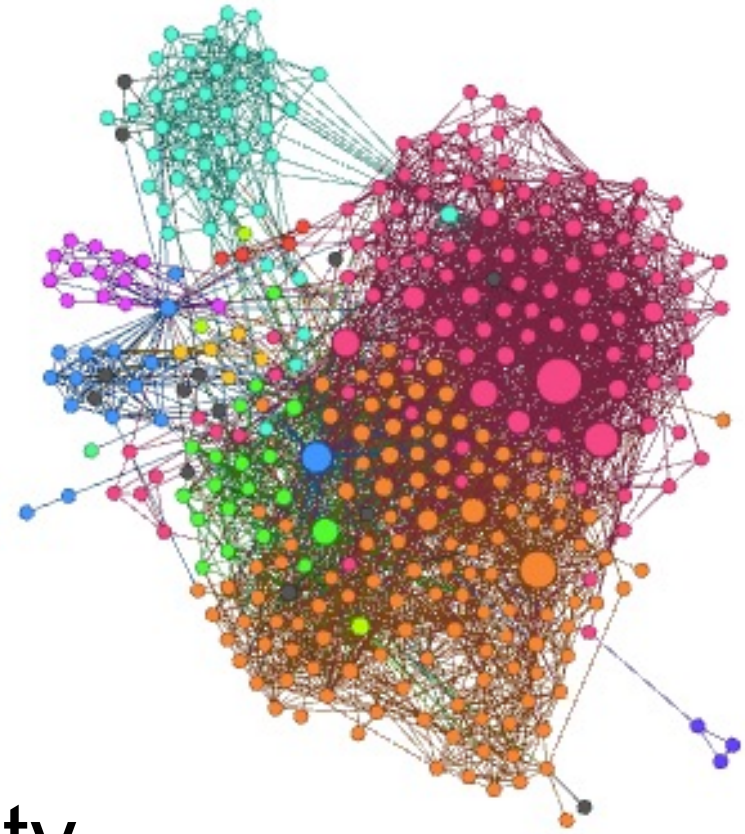
- Plot it in R

	1	2	3	4
1	0	1	0	1
2	1	0	1	0
3	0	1	0	1
4	1	0	1	0

Possible plotting solution

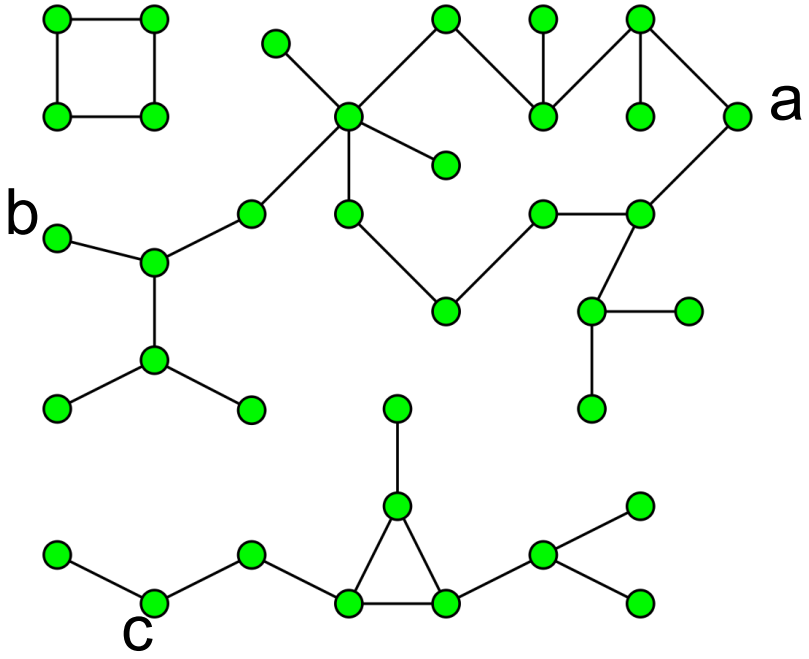
```
library(igraph)
dat=read.csv("adj.csv",header=TRUE,row.names=1,check.names=FALSE)
#write.csv(dat, "adj.csv")
m=as.matrix(dat)
g=graph.adjacency(m,mode="undirected",weighted=NULL,diag=FALSE)
plot.igraph(g)

deg <- degree(g)
hist(deg,xlab="Degree (k)", ylab="Frequency",
 main=NULL,col="red3",border=8,breaks=seq(0.5,3.5),xlim=c(0,4),plot=T)

plot.igraph(g,vertex.label=V(g$name,vertex.size=30,,vertex.label.color="yellow",
vertex.label.font=2,vertex.color="darkblue",edge.color="black")


# try this with sparse adjacency list
dat2=read.csv("adjlist.csv",header=TRUE,row.names=1,check.names=FALSE)
m2=as.matrix(dat2)
g2=graph.adjacency(m,mode="undirected",weighted=NULL,diag=FALSE)
plot.igraph(g2)
```

Connectivity
Is everything connected?



Cycles
Paths
Connectivity
Components
Giant Component

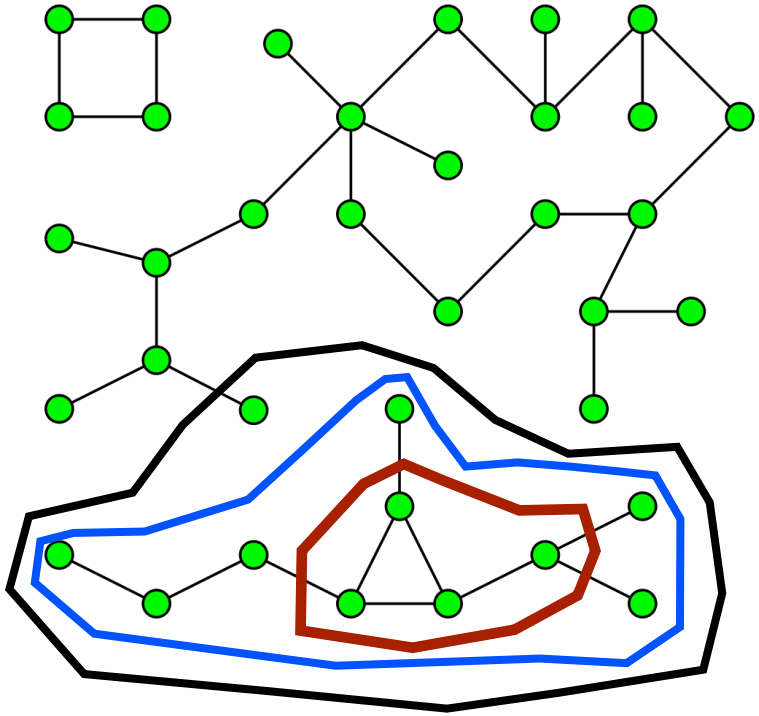
Connectivity



Two vertices are connected if there is a path between them.

A graph is connected if every two vertices are connected.

Connectivity

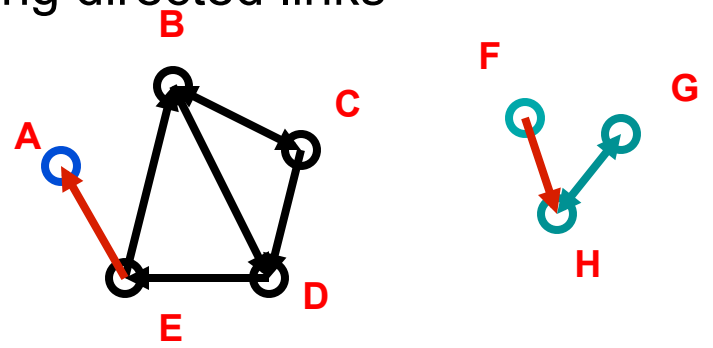


A connected component of a graph is a maximal set of vertices such that every pair of them are connected.

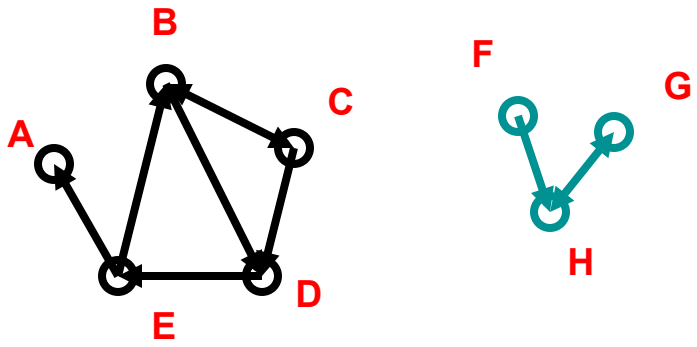
Connected components

- Strongly connected components
  - Each node within the component can be reached from every other node in the component by following directed links

  - Strongly connected components
    - B C D E
    - A
    - G H
    - F



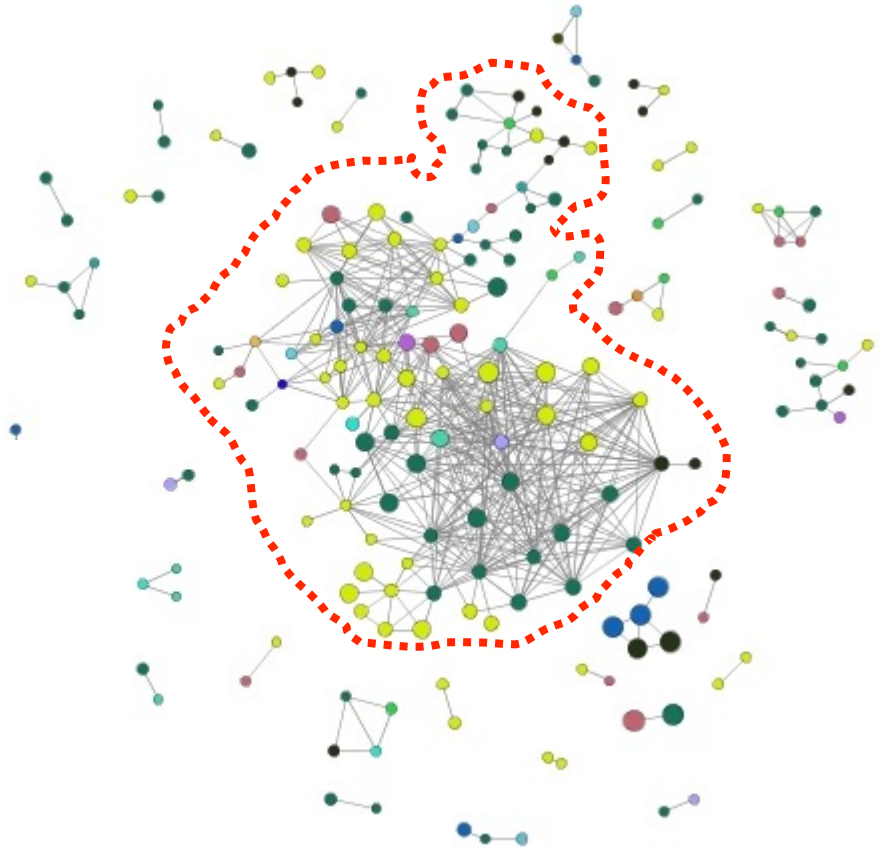
- Weakly connected components: every node can be reached from every other node by following links in either direction
  - Weakly connected components
    - A B C D E
    - G H F

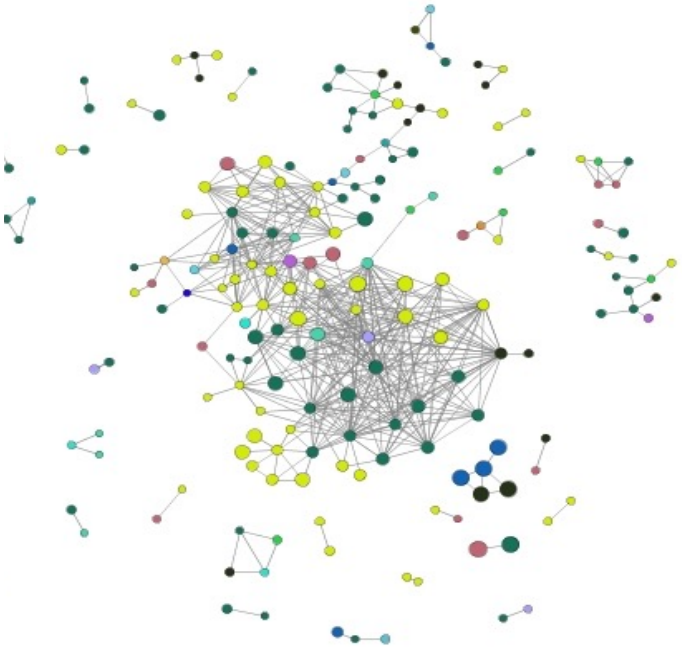


- In undirected networks one talks simply about ‘connected components’

Giant component

- if the largest component encompasses a significant fraction of the graph, it is called the **giant component;** fills most of the network;





Centrality measures

Simplifying the Complexity

- Global patterns of networks
  - degree distributions, path lengths…
- Local Patterns
  – Clustering, Transitivity, Support…
- Positions in networks
  – Neighborhoods, Centrality, Influence...

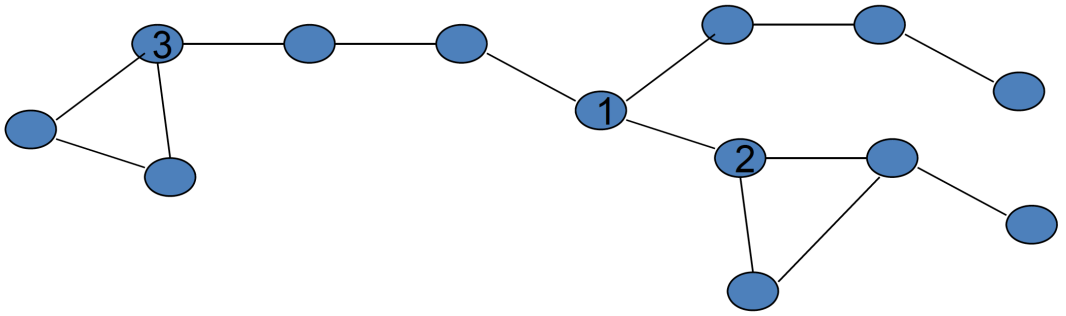
Position in Network

How to describe individual characteristics?

- Degree
- Betweenness
- Eigenvector Centrality
- Closeness
- PageRank
- …

Degree Centrality

- How ``connected'' is a node?
  - degree captures connectedness
  - normalize by n-1



  - Node 3 is considered as ``central'' as 1 and 2

Centrality, different things to measure

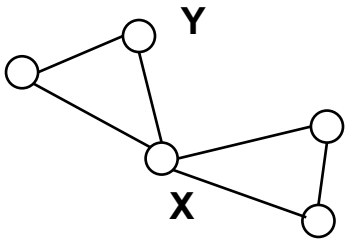
- Degree – connectedness
- Betweenness – role as an intermediary, connector
- Eigenvectors –

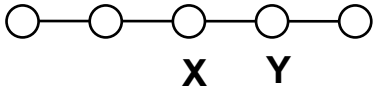
``not what you know, but who you know..''

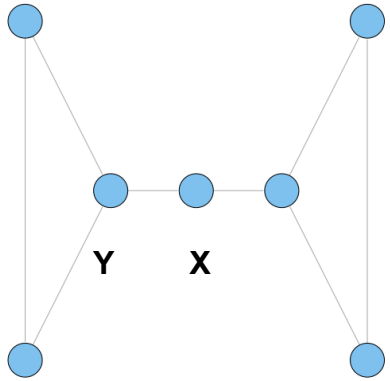
Another approach…. Betweenness

- Intuition
  How many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?

- Who has higher betweenness, X or Y?







Betweenness Centrality

betweenness(v) = $\frac{\text{number of shortest paths from s to t through v}}{\text{number of shortest paths from s to t}}$ summed over all s, t pairs

and then normalized by total number of pairs of vertices

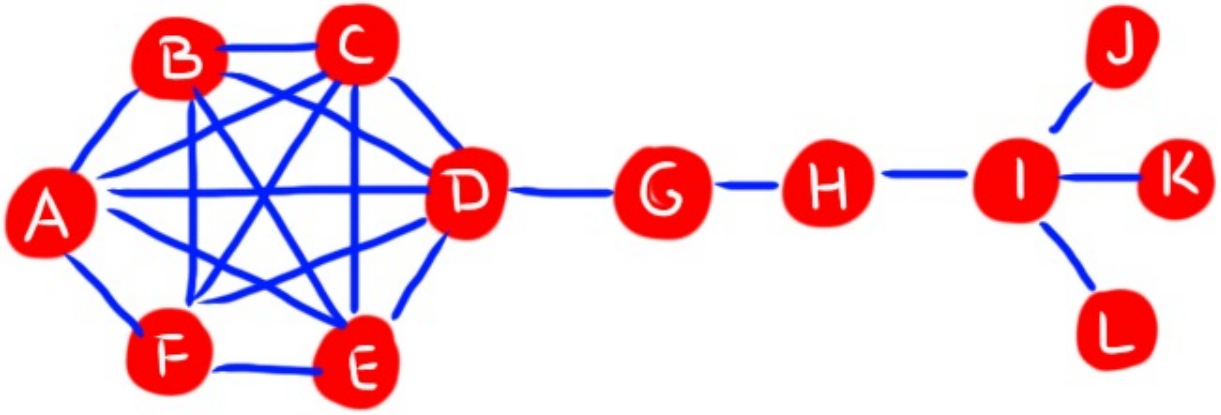
- Undirected and directed network (normalized)

$$C_B(i) = \frac{\sum_{j<k} g_{jk}(i)/g_{jk}}{[(n-1)(n-2)/2]}$$

where
g_{jk} is the number of geodesic paths between two vertices j and k and
$g_{jk}(i)$ is the number of geodesic paths of the two vertices that contain i

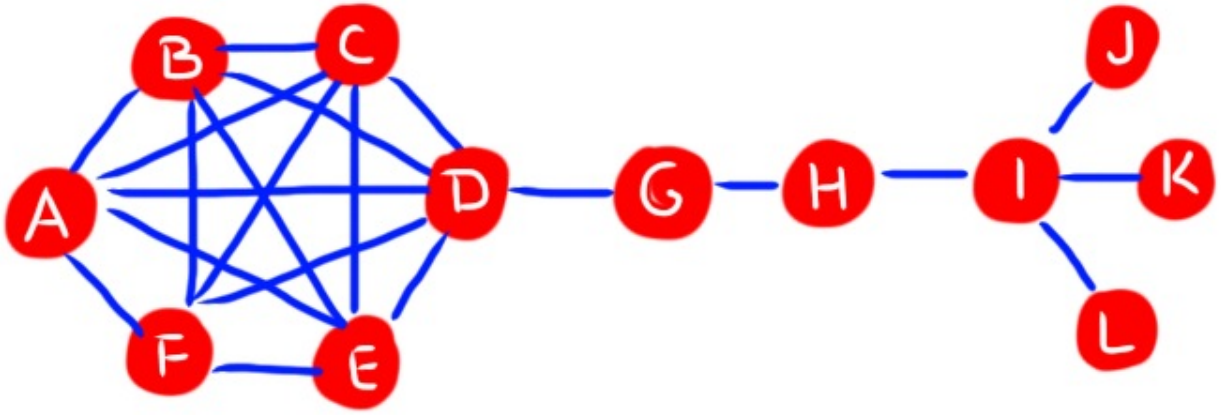
Exercise!

- Find a node that has high betweenness but low degree



Exercise!

- Find a node that has low betweenness but high degree

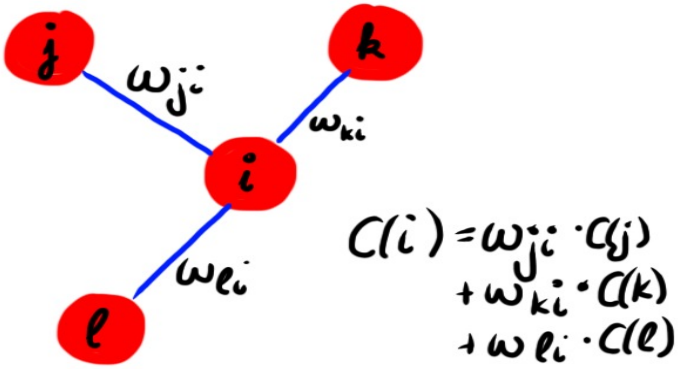


Eigenvector Centrality

- Simple measures of centrality ignore global structure of network.
- In general, connections to people who are themselves influential will give a person more influence than connections to less influential people.
- Eigenvector centrality takes into account centrality of neighbours in identifying the key players.

Eigenvector Centrality

- How central you are depends on how central your neighbors are
- Now distinguishes more ``influential'' nodes



Centrality is proportional to the sum of neighbors' centralities

C_i proportional to $\Sigma_{j:\ \text{friend of i}}$ Cj

$C_i = a\Sigma j\ g_{ij}\ Cj$

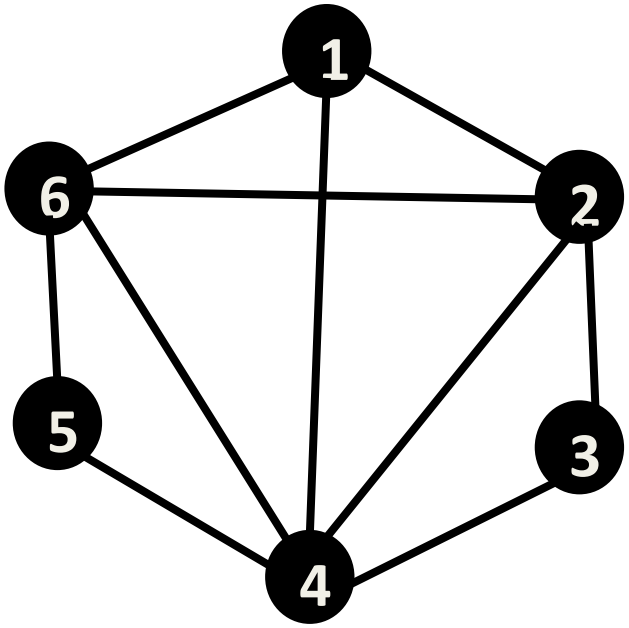
Example: Google Page rank

Clustering

- Transitivity - “neighbours of a node are neighbours of each other”
- Common property of many networks
- Due to spatial effects (farming premises in close spatial proximity)
- or otherwise (i.e. with well connected small communities where friends of an individual are also friends of each other)
- How do we quantify clustering and what are its implications?

Clustering

Clustering = (1) probability that two of your friends are also friends
(2) ratio of triangles to triples



- A Triple is formed by three nodes/vertices ABC where A is connected to B and B is connected to C
- A Triangle is formed of three nodes where each node is connected to all the other
- A triangle is a triple but the opposite is not true
- **When counting triangles and triples always start with the node in the middle of the triple/triangle**

5,4,3 – triple but not a triangle

6,4,2 – triple and triangle

Clustering

- Local clustering coefficient

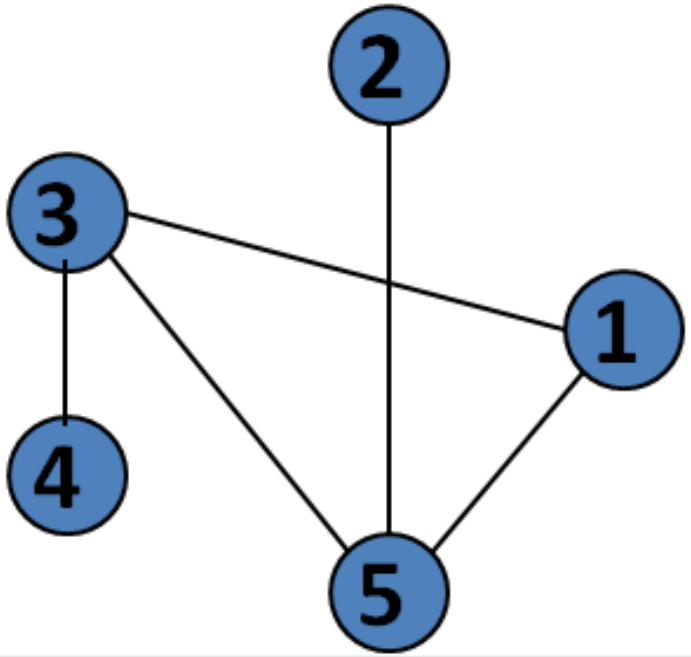
$$cc(v) = \frac{\text{pairs of neighbors of } v \text{ connected by edges}}{\text{total pairs of } v}$$

- Global clustering coefficient

$$C = \frac{(\text{number of triangles})}{(\text{number of connected triples})}$$

Exercise

What is the clustering of node 5, $Cl_5(g)$?

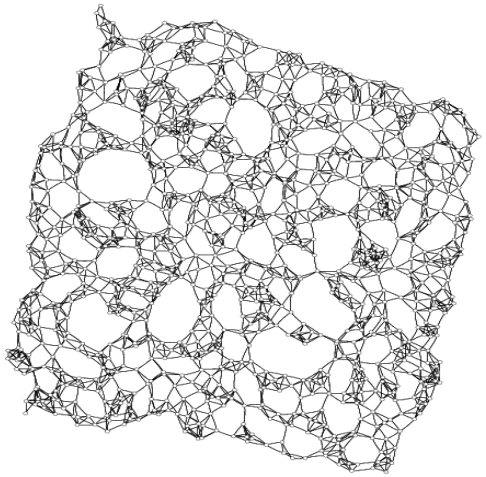


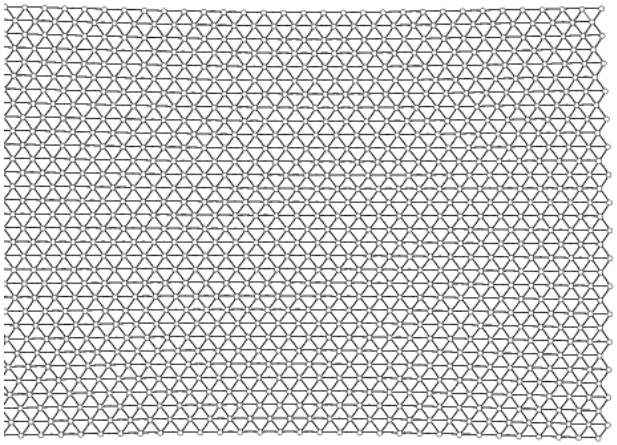
- 1
- 1/2
- 4
- 1/3

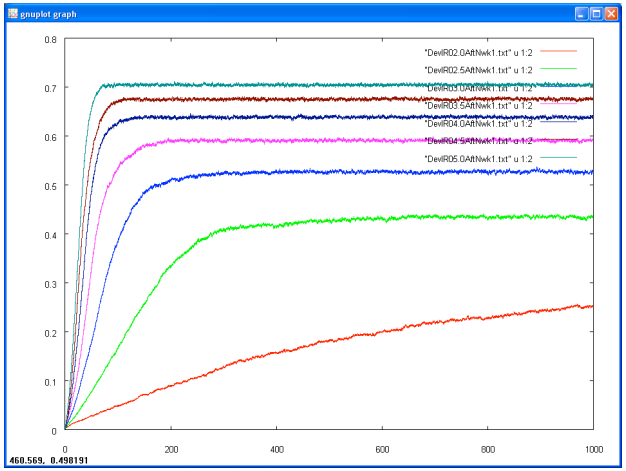
Implications of Clustering

Very different graphs can have similar clustering!

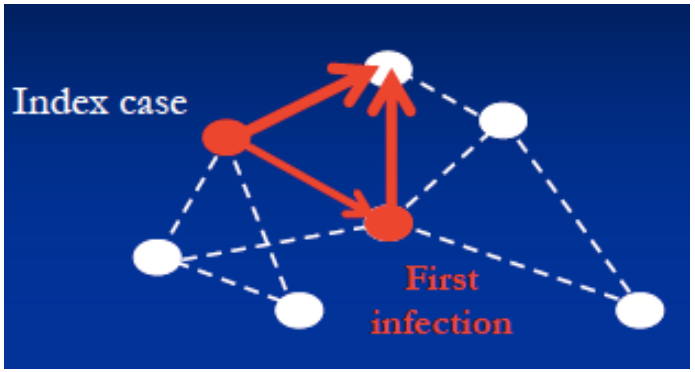
Can show very different disease spread.





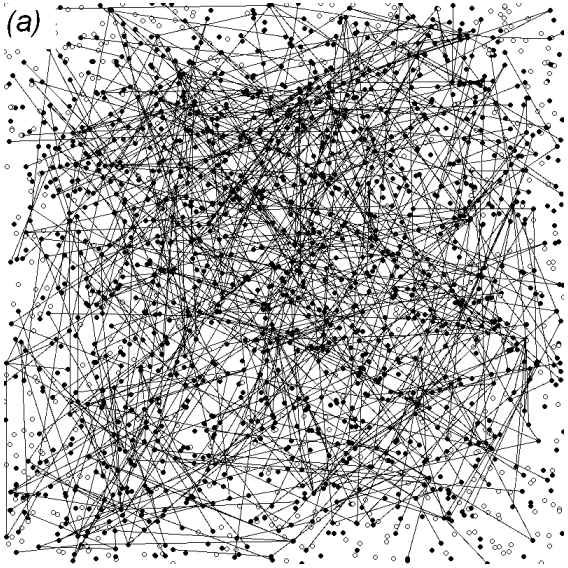


Implications of Clustering

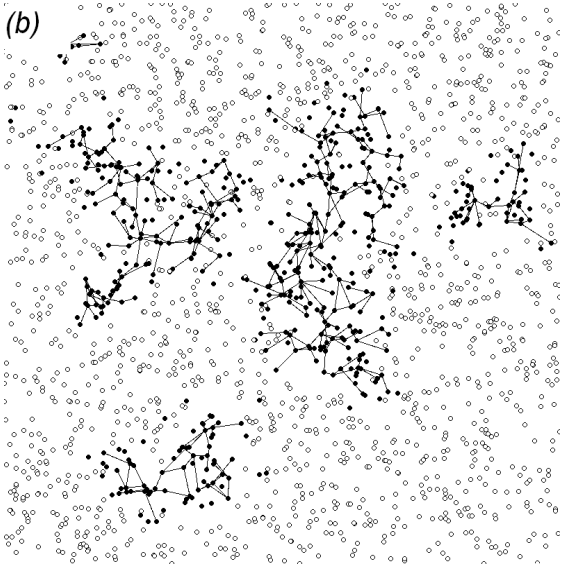


1. Competition for susceptible neighbours
2. Transmission enhanced locally
3. Limited further spread

(a)



(b)



Spatial epidemic spread for random (a) and clustered (b) networks

Disease Persistence in Networks

- A better way of looking at (disease) population persistence in spatial and network models is as a percolation phenomenon

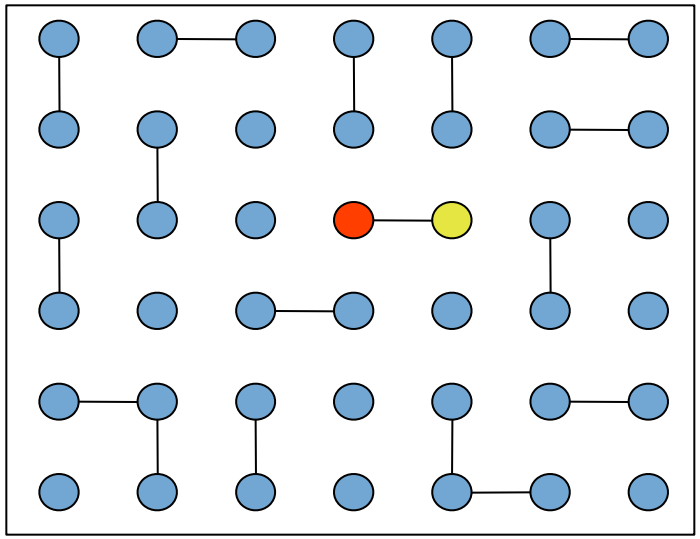
- Percolation
  - We say the network percolates when a giant component forms.

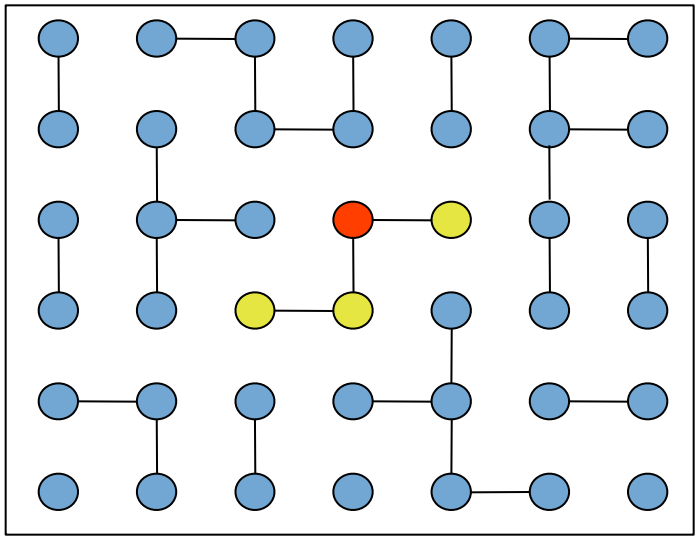
Percolation threshold

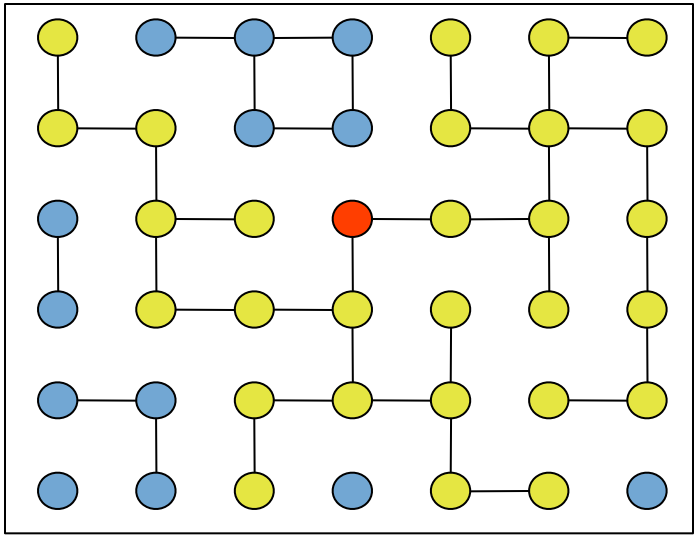
- Percolation threshold
  - refers to simplified lattice models of networks
  - above it, a connected component* exists; below it, it does not
- This transition point is important in determining if there will be epidemic on a network

* connected component
  - a subgraph in which any two nodes are connected to each other by paths

Percolation

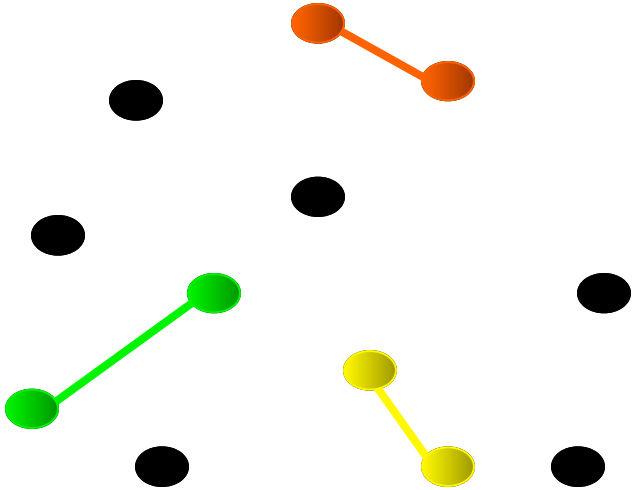




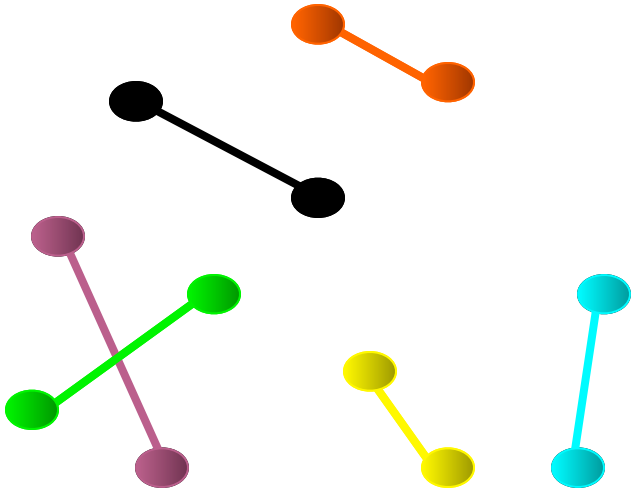


- A network with few links has connected components of small size
- As more links are added, a threshold is found, above which the largest component is comparable to the size of the whole network
- A network with many small components can only support very small epidemics

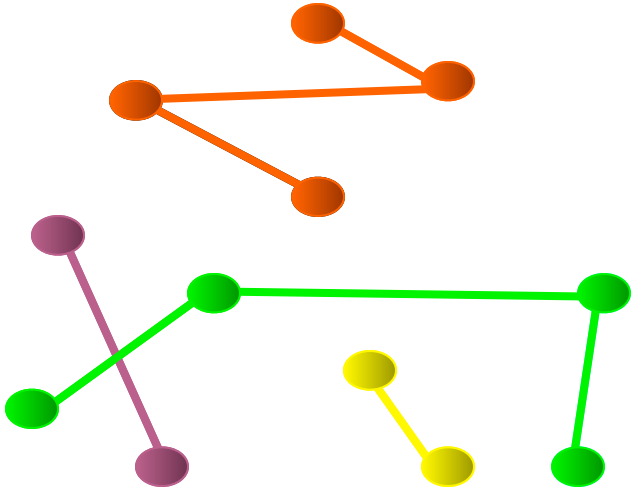
Connected components and the size of the GCC



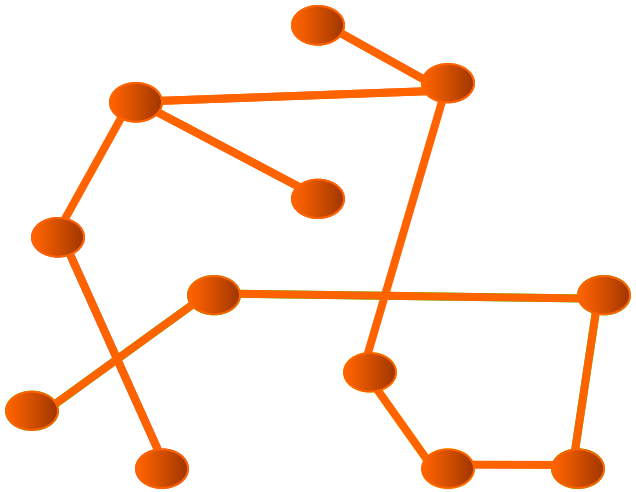
 # Comp of size (1)=6, # Comp of size (2)=3



 # Comp of size (2)=6



 # Comp of size (2)=2, # Comp of size (4)=2



 # Comp of size (12)=1

R0 vs. percolation

- For simple population structures they are the same (but dynamic vs. structural)

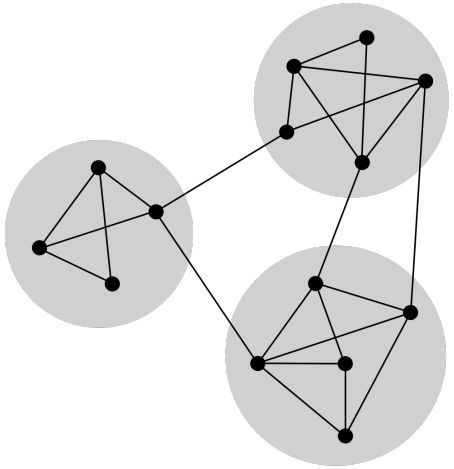
- For more complex structures with clustering and community structures the concept of R_0 breaks down

- Percolation a much more robust concept (though difficult to calculate)

Community Structure

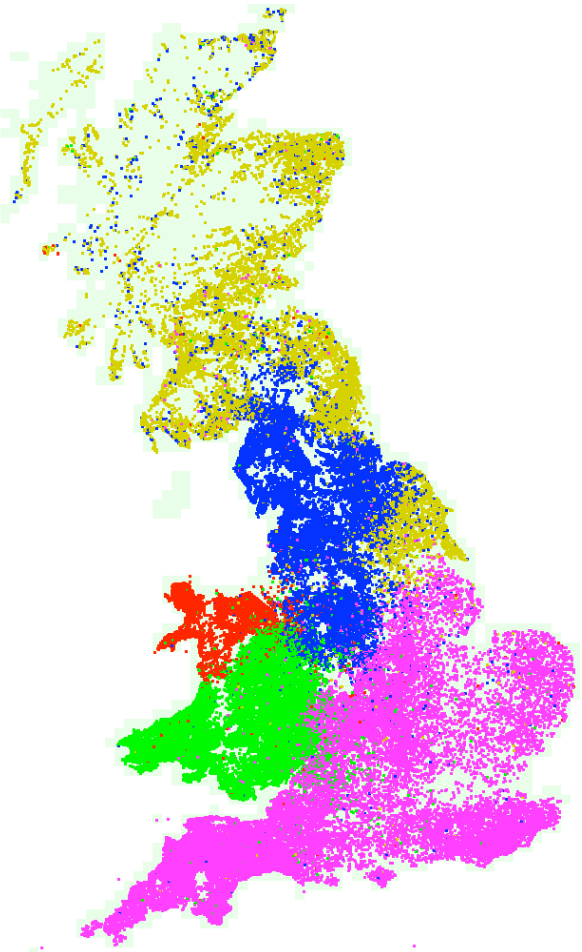
There are groups of nodes that are more densely connected within than between.

- Mutually exclusive communities
- Overlapping communities



Communities in networks

- **Higher-level mixing properties**
  - **Components**
    - Subsets of nodes such that any two can be connected
  - **Communities**
    - Subsets of nodes that share more links within the set than outside of it
      - Maximum modularity index
      - $Q=(1/M)\sum_{ij}(A_{ij} - k_i k_j/M)$ $[c_i=c_j]$
  - **Tend to limit epidemics and reduce R_0**



- GB sheep network communities (2003).
- Largely, but not entirely, geographically based.

Exercise

Say we have a disease spreading on a network.
Which would make the epidemic more likely/bigger?

(a) Large GCC?

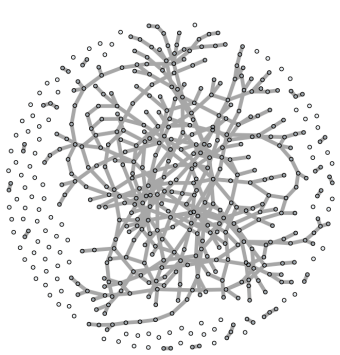
(b) Network has lots of community structure?

(c) Network has short shortest-path lengths?

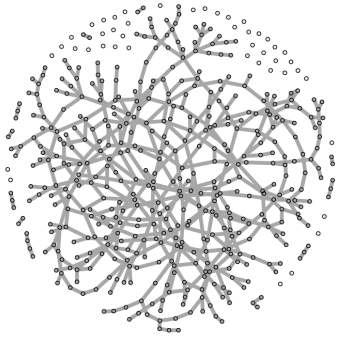
Network models:
Random, small-world, scale-free

Random Graphs

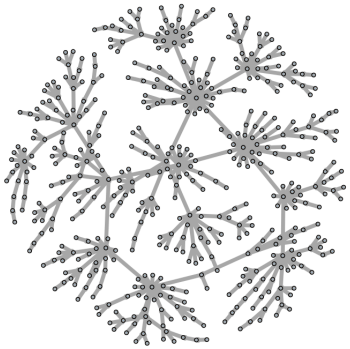
- Erdös-Renyi
- BA - scale-free networks
- Watts-Strogatz - small-world networks



Random Network

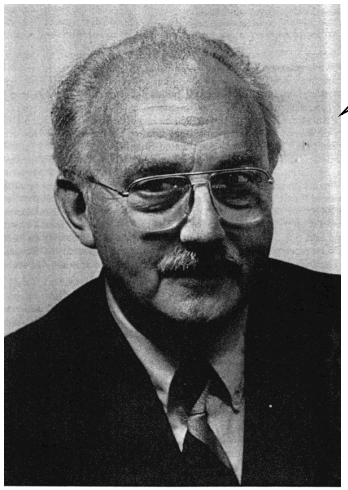


Small World Network



Scale-free Network

“All models are wrong, but some are useful.”



George Edward Pelham Box

- A theory that is consistent with the data
- Coarse-grained abstraction of the system
- Need to be cognitively and mathematically understood
- Represent the structure and/or dynamics of the real system
- Need to be predictive

Königsberg Bridge Problem

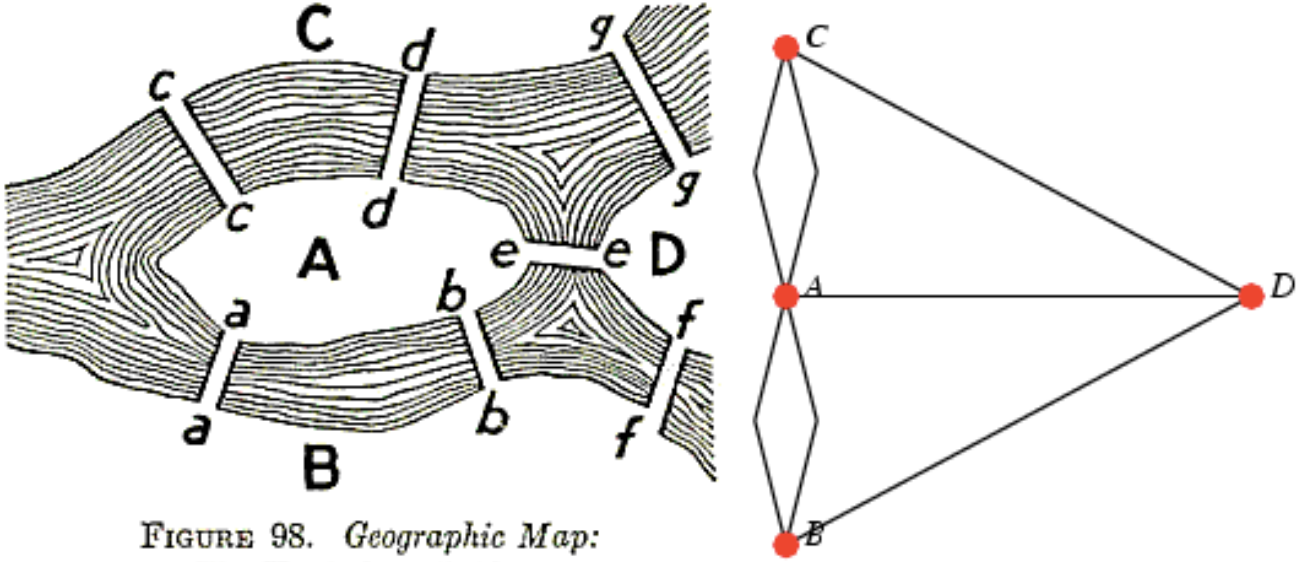
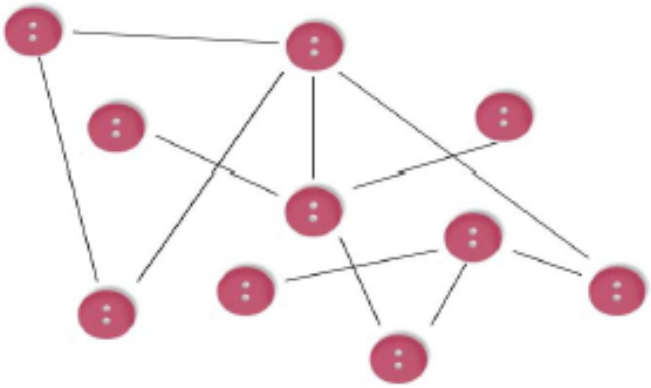


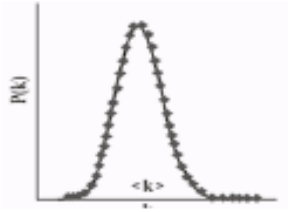
FIGURE 98. *Geographic Map: The Königsberg Bridges.*

The Königsberg bridge problem asks if the seven bridges of the city of Königsberg (left figure; Kraitchik 1942), formerly in Germany but now known as Kaliningrad and part of Russia, over the river Preger can all be traversed in a single trip without doubling back, with the additional requirement that the trip ends in the same place it began. This is equivalent to asking if the multigraph on four nodes and seven edges (right figure) has an Eulerian circuit. This problem was answered in the negative by Euler (1736), and represented the beginning of graph theory.

http://mathworld.wolfram.com/KoenigsbergBridgeProblem.html

Erdös-Renyi Random Networks





P. Erdos A. Rényi. Publ. Math. (Debrecen) 6, 290-297 (1959)

Poisson distribution

- In the 1960's Paul Erdös and Alfred Renyi studied the properties of random graphs.
- What are the mathematical consequences of throwing on the floor a random number of buttons and randomly connecting them with a random number of links?
- Reference point for network analysis / modeling

“Real” Networks are “Small World”

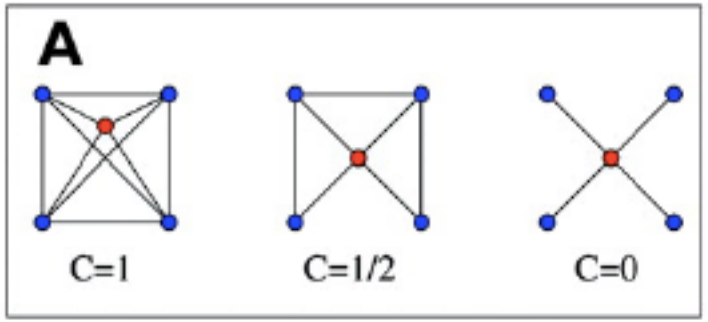
Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Characteristic path length L and clustering coefficient C for three real networks, compared to random graphs with the same number of vertices (n) and average number of edges per vertex (k). (Actors: $n = 225{,}226$, $k = 61$. Power grid: $n = 4{,}941$, $k = 2.67$. *C. elegans*: $n = 282$, $k = 14$.) The graphs are defined as follows. Two actors are joined by an edge if they have acted in a film together. We restrict attention to the giant connected component[16] of this graph, which includes ~90% of all actors listed in the Internet Movie Database (available at http://us.imdb.com), as of April 1997. For the power grid, vertices represent generators, transformers and substations, and edges represent high-voltage transmission lines between them. For *C. elegans*, an edge joins two neurons if they are connected by either a synapse or a gap junction. We treat all edges as undirected and unweighted, and all vertices as identical, recognizing that these are crude approximations. All three networks show the small-world phenomenon: $L \geq L_{random}$ but $C \gg C_{random}$.

Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998 Jun 4;393(6684):440-2.

High Clustering Coefficient



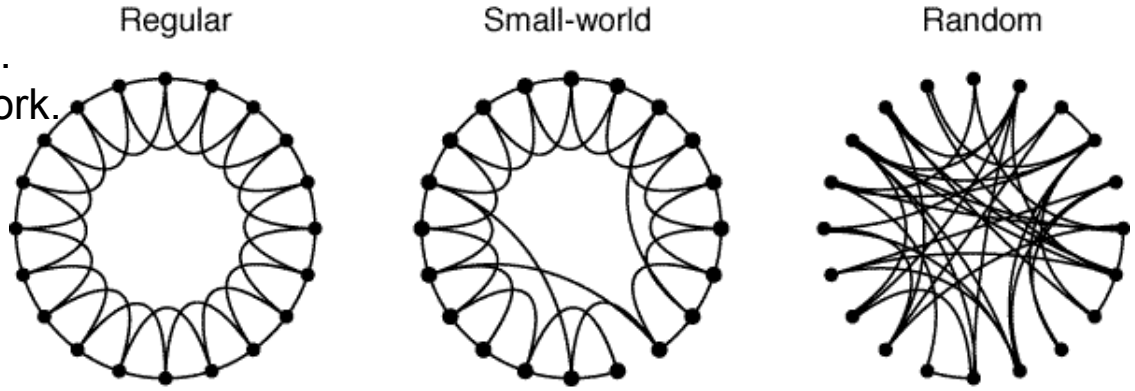
Ravasz et al. *Science* **297**, 1551 (2002)

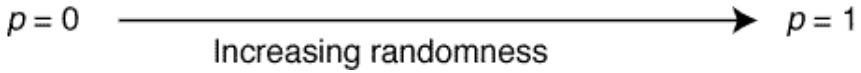
Characteristic Path Length

Small average shortest path between all possible pairs of nodes

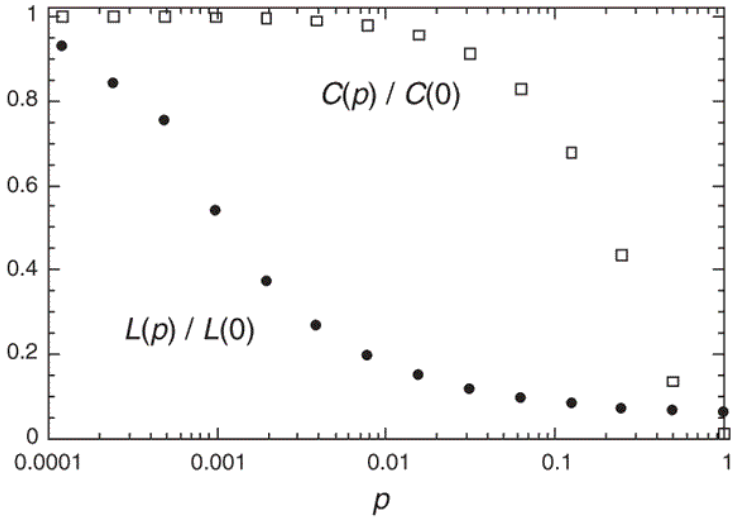
Creating Small-World Networks

- p=0: regular lattice.
- p=1: random network.





- For intermediate values of rewiring, we observe networks with high clustering but short paths: small-world network.

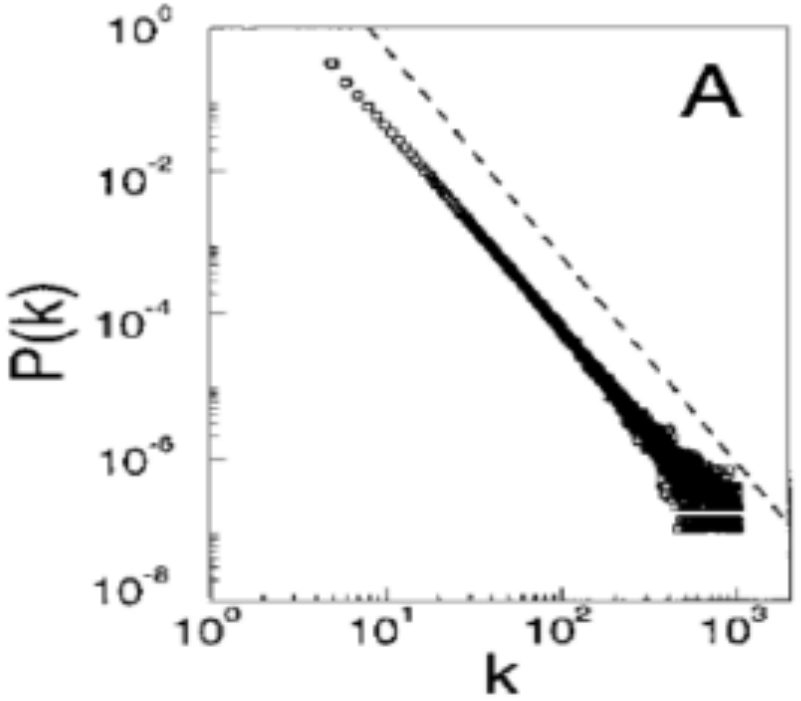


- Small average shortest path length (grows at log N)
- Large clustering coefficient

Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998 Jun 4;393(6684):440-2.

Rich-Get-Richer: Creating Scale-Free Networks

We next show that a model based on these two ingredients naturally leads to the observed scale-invariant distribution. To incorporate the growing character of the network, starting with a small number (m_0) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that link the new vertex to m different vertices already present in the system. To incorporate preferential attachment, we assume that the probability Π that a new vertex will be connected to vertex i depends on the connectivity k_i of that vertex, so that $\Pi(k_i) = k_i/\Sigma_j k_j$. After t time steps, the model leads to a random network with $t + m_0$ vertices and mt edges. This network evolves into a scale-invariant state with the probability that a vertex has k edges, following a power law with an exponent $\gamma_{model} = 2.9 \pm 0.1$ (Fig. 2A). Because the power law observed for real networks describes systems of rather different sizes at different stages of their development, it is expected that a correct model should provide a distribution whose main features are independent of time. Indeed, as Fig. 2A demonstrates, $P(k)$ is independent of time (and subsequently independent of the system size $m_0 + t$), indicating that despite its continuous growth, the system organizes itself into a scale-free stationary state.

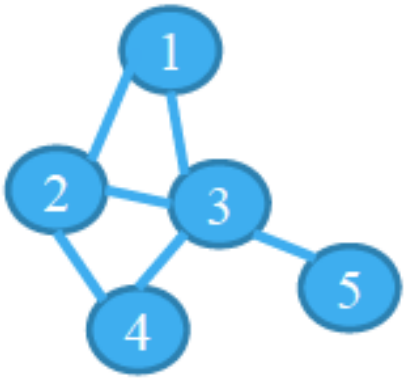


Barabasi and Albert. *Science* **286**, 509 (1999)

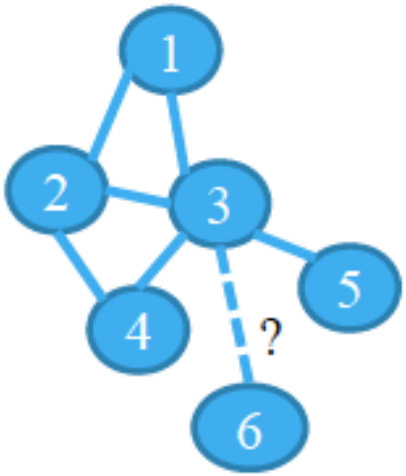
Rich-Get-Richer

- High fitness agents increase their connections and fitness faster than the less connected

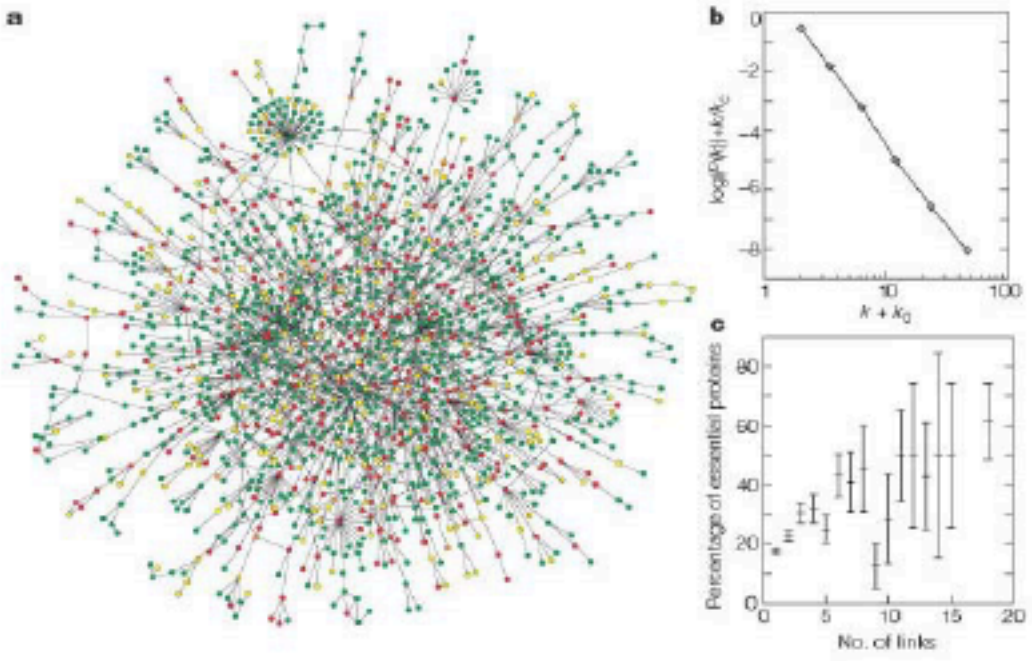
The Rich-Get-Richer Network Growth Model



Node ID	Connection	p
1	2	0.17
2	3	0.25
3	4	0.33
4	2	0.17
5	1	0.08
sum	12	



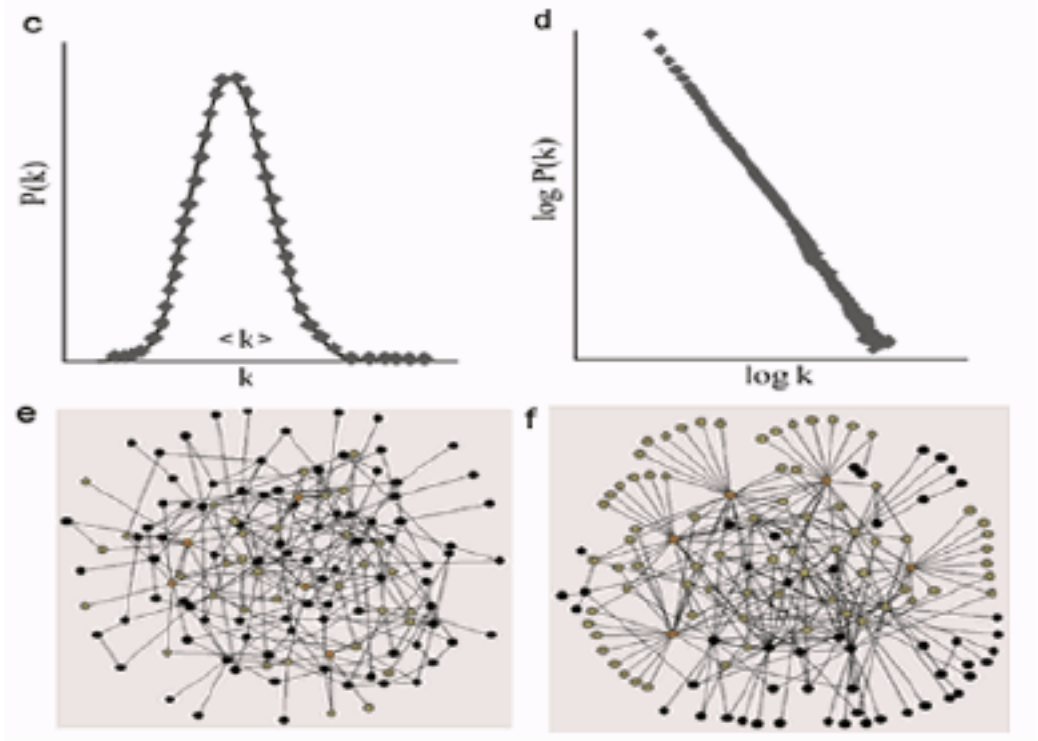
The Importance of Hubs



H. Jeong, S. P. Mason, A.-L. Barabási and Z. N. Oltvai. **Lethality and centrality in protein networks.** *Nature* 411, 41-42 (2001)

Albert R, Jeong H, Barabasi A-L: **Error and attack tolerance of complex networks.** *Nature* 2000, 406(6794):378-382.

Erdös-Renyi random networks
vs.
Barabasi-Albert scale-free networks

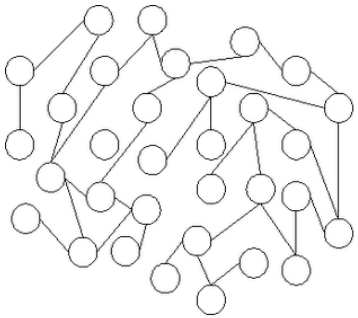


Barabasi, *Physics World*, July 2001

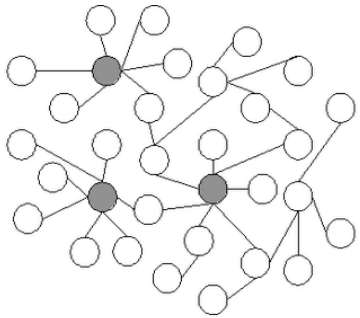
- Scale-free more robust against random removal
- Scale-free less robust against targeted attacks
- Resistant to random vaccination, but targeted interventions will work well.

Scale-free Networks

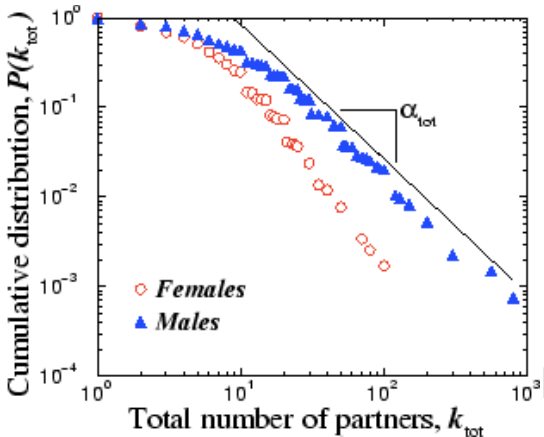
- Clustering coefficient decreases as degree increases: Power-law degree distribution
- "Scale-free" degree distributions have P(degree=k) ~ $k^{-\alpha}$
- "Rich-get-richer" model: preferential attachment
- Many real networks thought to be scale-free
- We expect to find super-spreading hubs
- Path length increases with log log N



(a) Random network



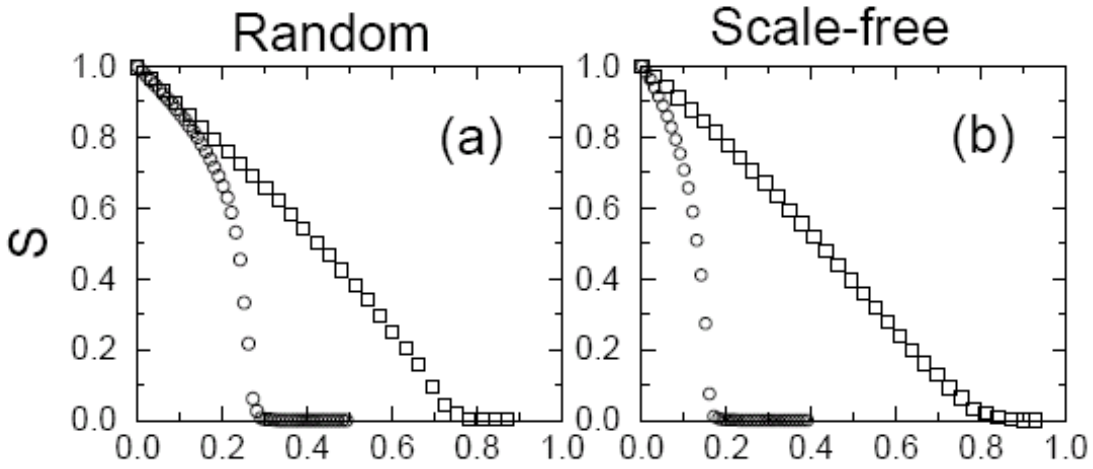
(b) Scale-free network



Liljeros et al, 2001

Network resilience to targeted attacks

- Scale-free graphs are resilient to random attacks, but sensitive to targeted attacks.
- For random networks there is smaller difference between the two

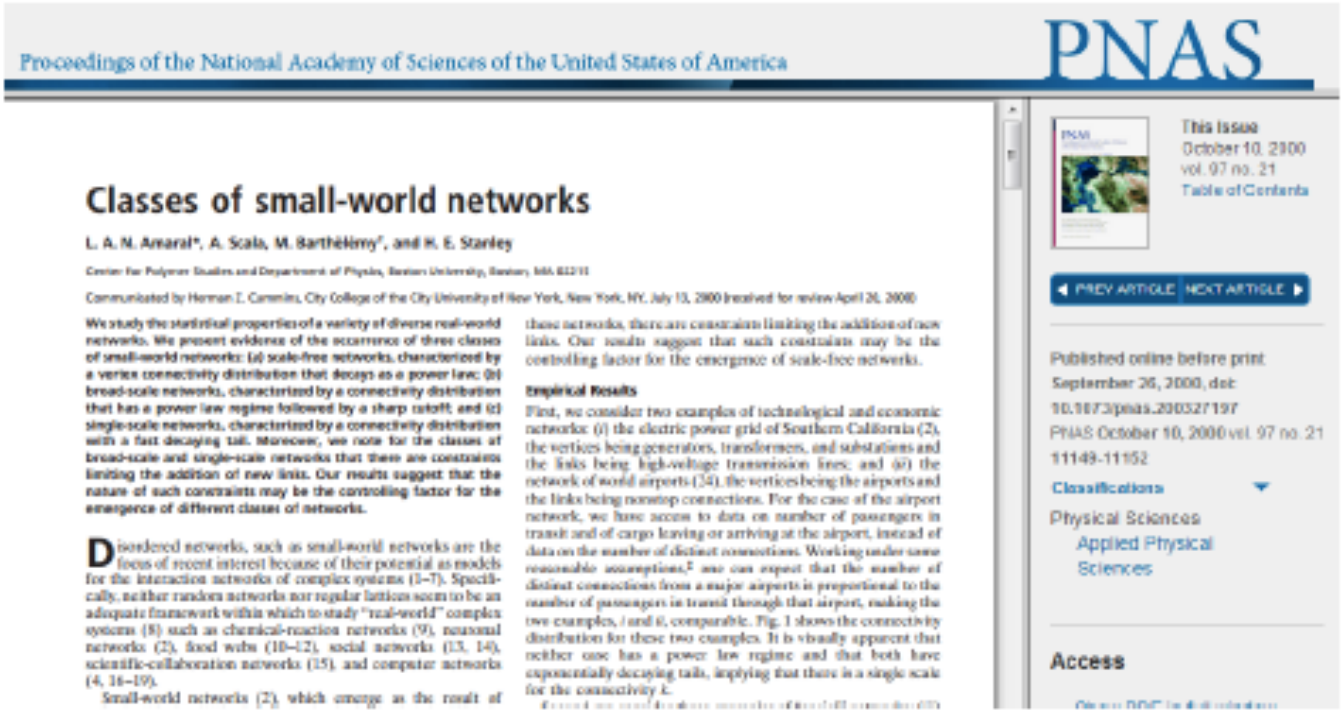


random failure

targeted attack

Source: Error and attack tolerance of complex networks. Réka Albert, Hawoong Jeong and Albert-László Barabási

Hubs Don’t Grow Forever



Proceedings of the National Academy of Sciences of the United States of America

PNAS

Classes of small-world networks

L. A. N. Amaral*, A. Scala, M. Barthélémy†, and H. E. Stanley

Communicated by Herman Z. Cummins, City College of the City University of New York, New York, NY, July 13, 2000 (received for review April 26, 2000)

We study the statistical properties of a variety of diverse real-world networks. We present evidence of the occurrence of three classes of small-world networks: (a) scale-free networks, characterized by a vertex connectivity distribution that decays as a power law; (b) broad-scale networks, characterized by a connectivity distribution that has a power law regime followed by a sharp cutoff; and (c) single-scale networks, characterized by a connectivity distribution with a fast decaying tail. Moreover, we note for the classes of broad-scale and single-scale networks that there are constraints limiting the addition of new links. Our results suggest that the nature of such constraints may be the controlling factor for the emergence of different classes of networks.

Disordered networks, such as small-world networks are the focus of recent interest because of their potential as models for the interaction networks of complex systems (1–7). Specifically, neither random networks nor regular lattices seem to be an adequate framework within which to study "real-world" complex systems (8) such as chemical-reaction networks (9), neuronal networks (2), food webs (10–12), social networks (13, 14), scientific-collaboration networks (15), and computer networks (4, 16–19).

Small-world networks (2), which emerge as the result of

these networks, there are constraints limiting the addition of new links. Our results suggest that such constraints may be the controlling factor for the emergence of scale-free networks.

Empirical Results

First, we consider two examples of technological and economic networks: (*i*) the electric power grid of Southern California (2), the vertices being generators, transformers, and substations and the links being high-voltage transmission lines; and (*ii*) the network of world airports (24), the vertices being the airports and the links being nonstop connections. For the case of the airport network, we have access to data on number of passengers in transit and of cargo leaving or arriving at the airport, instead of data on the number of distinct connections. Working under some reasonable assumptions,[‡] one can expect that the number of distinct connections from a major airports is proportional to the number of passengers in transit through that airport, making the two examples, *i* and *ii*, comparable. Fig. 1 shows the connectivity distribution for these two examples. It is visually apparent that neither one has a power law regime and that both have exponentially decaying tails, implying that there is a single scale for the connectivity *k*.

This Issue
October 10, 2000
vol. 97 no. 21
Table of Contents

◀ PREV ARTICLE NEXT ARTICLE ▶

Published online before print
September 26, 2000, doi:
10.1073/pnas.200327197
PNAS October 10, 2000 vol. 97 no. 21
11149-11152

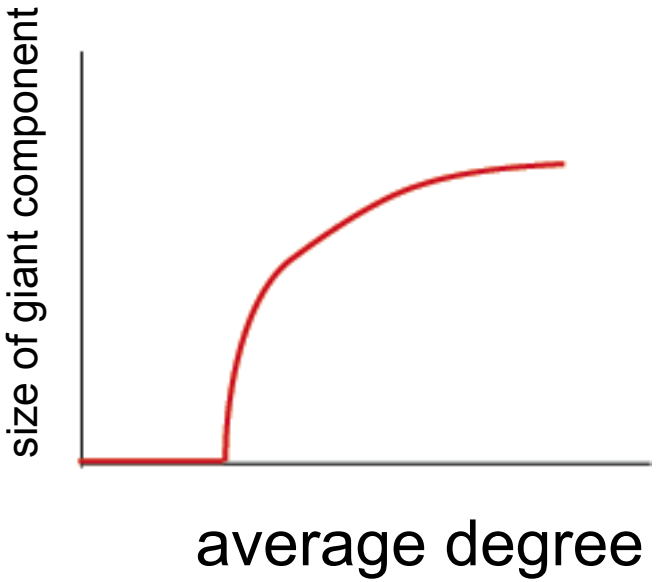
Classifications

Physical Sciences
Applied Physical Sciences

Access

Models of growing networks better fit real data when nodes have the added parameter of aging, or the added parameter of max physical capacity.

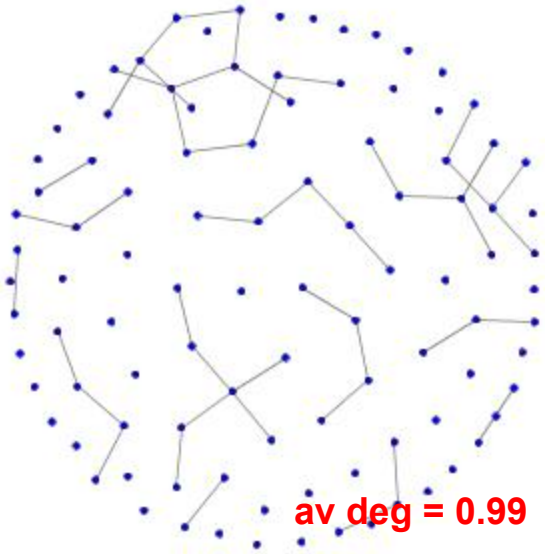
Percolation threshold in Erdös-Renyi Graphs

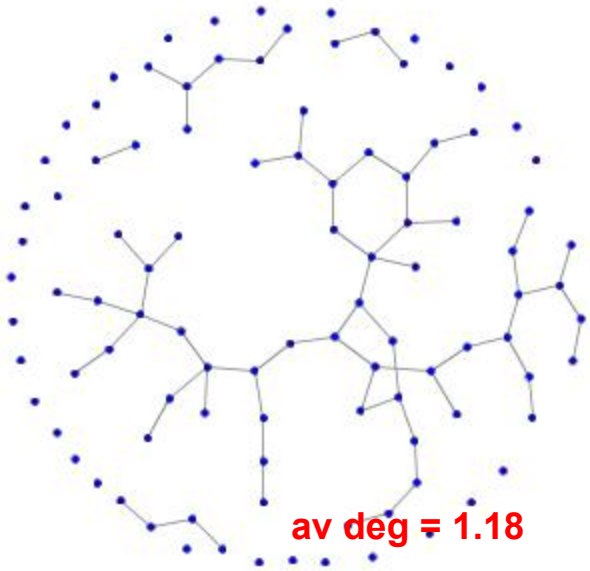


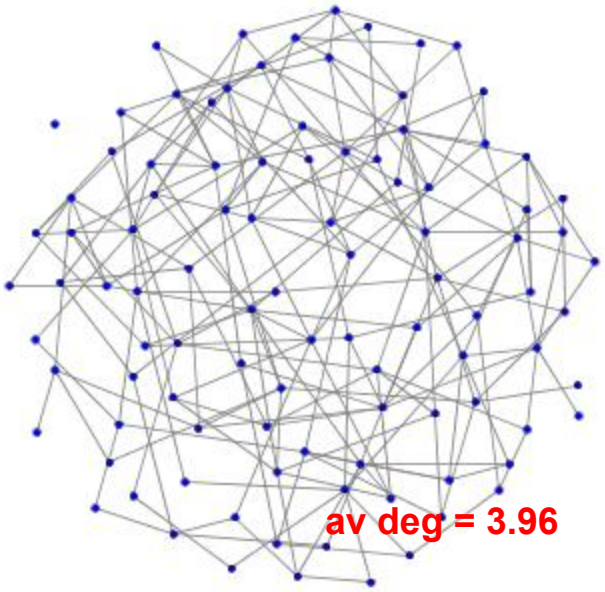
Percolation threshold: the point at which the giant component emerges

As the average degree increases to z = 1, a giant component suddenly appears

Edge removal is the opposite process
As the average degree drops below 1 the network becomes disconnected

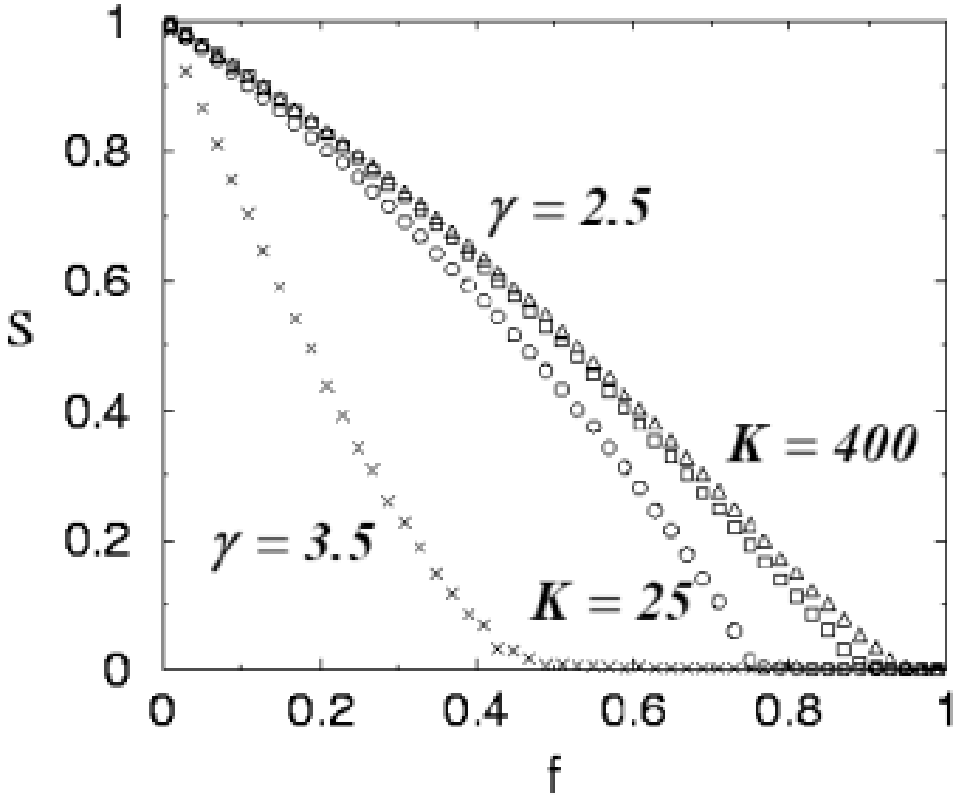






Percolation Threshold scale-free networks

- What proportion of the nodes must be removed in order for the size (S) of the giant component to drop to 0?

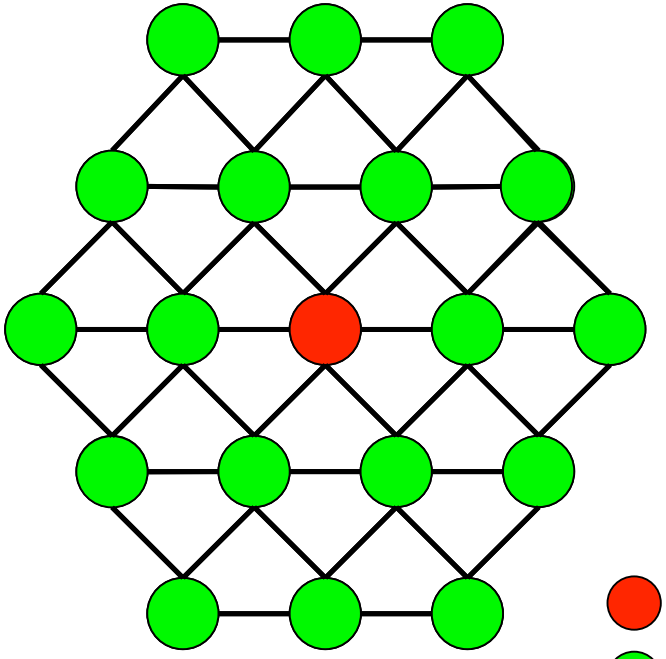


- For scale free graphs there is always a giant component
  - the network always percolates

Source: Cohen et al., Resilience of the Internet to Random Breakdowns

Comparing network and mass-action epidemics

- On a network, the set of interactions is **restricted** and **fixed**.
- The spread of infection is reduced:
  - Slowed "spatially" - there are path lengths > 1.
  - Blocked locally - local pool of susceptibles is reduced by local infection even when global prevalence is low.
- Key point: on a network interactions are **long-lasting**.



Extreme example:
- Highly infectious "index case" . .
- Infects six others .
- Despite still being infectious, cannot infect anyone else.

= infected
= susceptible

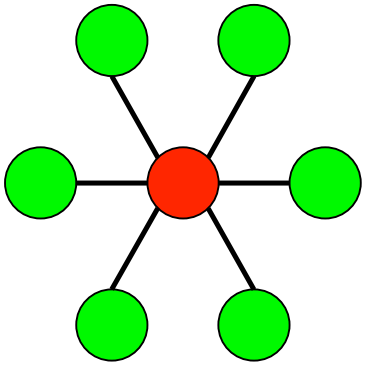
Networks and R_0

- Previously defined R_0 as "the mean number of secondary cases if the population was entirely susceptible".

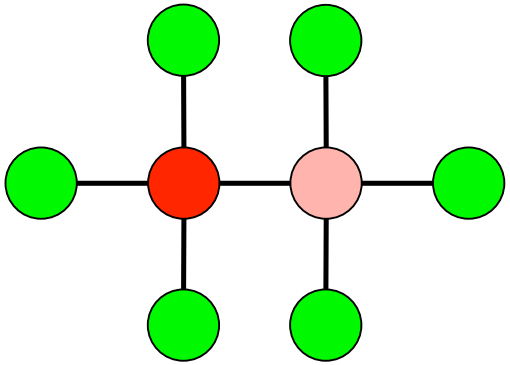
Recall: mass-action approximation gives $R_0 = \frac{\beta N}{\gamma}$

N=population size;
β=transmission parameter

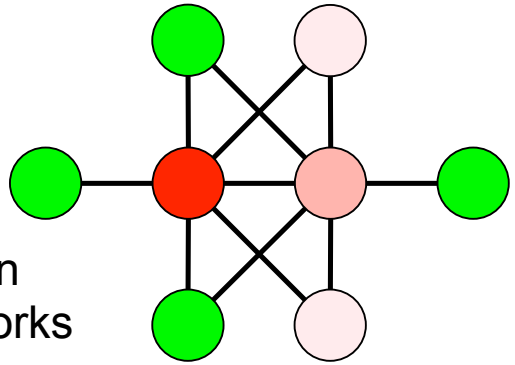
- R_0 can be increased indefinitely by increasing the transmission rate, infectious period, or the population size.
- The same cannot happen on a network - limited by neighbourhood size.



- Even the first two cases interfere with each other.
- Early cases do not behave independently.
- Infection is quick to become **locally saturated**.



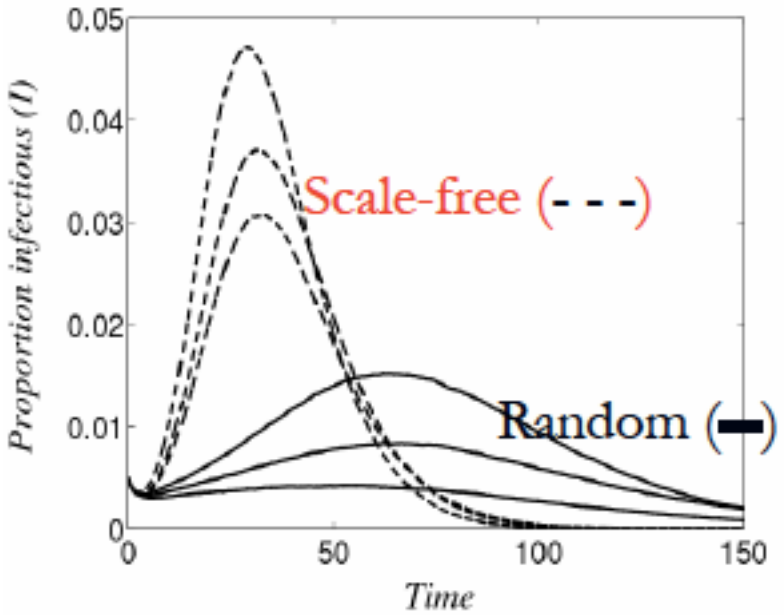
Greater effect in clustered networks



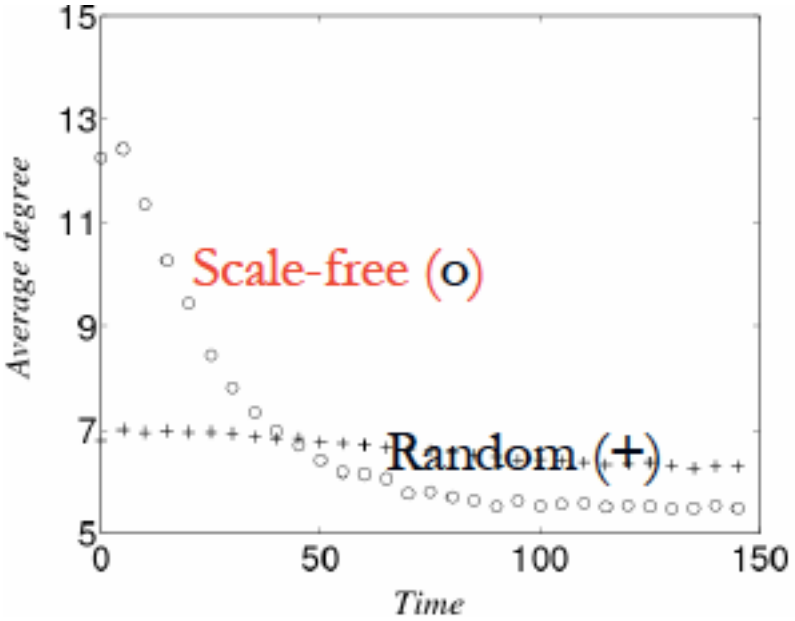
- Defining a sensible R_0 on a network is not obvious.
- If we define R_0 using the early epidemic growth rate then, if each individual has degree k, on an unclustered network, we can show that "R_0"=(k-2) β / γ.
- Clustering further slows epidemic spread.

Disease dynamics on networks

I. Z. Kiss, D. M. Green and R. R. Kao (2006) Infectious disease control using contact tracing in random and scale-free networks. J. R. Soc. Interface 3, 55 – 62.

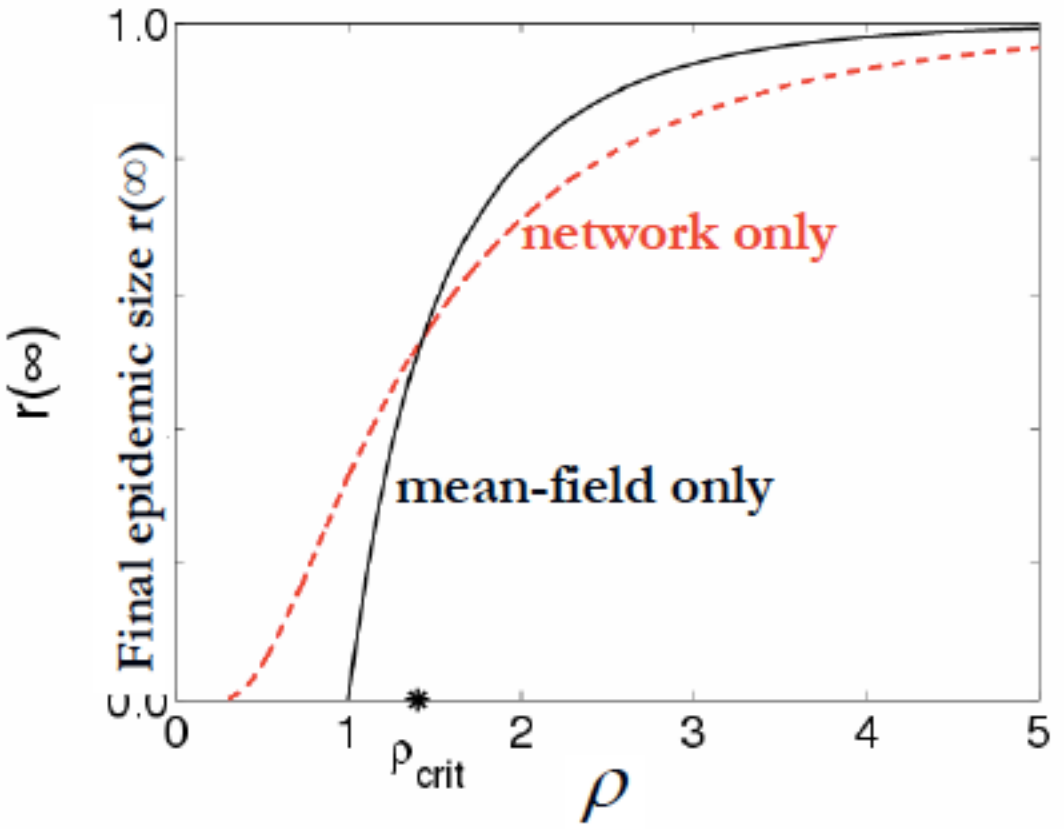


Time evolution of the proportion of infectious nodes for random (—) and truncated scale-free networks (- - -) ($p(k) = Ck^{-\gamma}e^{-k/L}$, $\gamma = 2.5$, $L = 100$ and $k \geq 3$) with $N = 2000$, $\langle k \rangle = 6$. Epidemics with infection rates per link $\tau = 0.067, 0.0735, 0.08$ and infectious period is $1/g = 3.5$.



Average degree of new infectious nodes for random (+) and truncated scale-free (o) networks ($p(k) = Ck^{-\gamma}e^{-k/L}$, $\gamma = 2.5$, $L = 100$ and $k \geq 3$) with $N = 2000$, $\langle k \rangle = 6$. Epidemics with infection rate per link $\tau = 0.15$ and infectious period $1/g = 3.5$

Final epidemic size (scale-free versus ODE model)



- High heterogeneity = quick initial spread
- Depletion of highly connected nodes
- Limited further spread

I. Z. Kiss, D. M. Green and R. R. Kao (2006) Infectious disease control using contact tracing in random and scale-free networks. J. R. Soc. Interface 3, 55 – 62.

Key learning points

- The usefulness of networks in modelling
- How to use networks to make sense of contact data (e.g., livestock movement)
- Key network properties
  - Why are these important?
  - What is their impact?
- Network models
- Simulating disease transmission on networks

- Network simulations (Netlogo / R)
- Epidemics_on_networks.R

MCQ_1

1) What is the connectivity distribution of Erdös-Renyi random graphs?

(a) Power-law

(b) Scale-free

(c) Poisson

(d) Log-normal

2) Small-world networks have (compared with random networks):

(a) Low clustering and relatively long average path length

(b) High clustering and roughly the same average path-length (roughly the same average path-length contrasted with the difference between the clustering coefficients)

(c) High clustering and relatively long average path length

(d) Low clustering and relatively short average path length

3) When nodes preferentially attach to high degree nodes, the diffusion over the network is

(a) faster

(b) slower

(c) unaffected

MCQ_2

Network Epidemics

- SIR Model (igraph)
- You need to perform epidemic SIR model on different types of networks: Try different parameters for network generation

  ?barabasi.game

  ? erdos.renyi.game

  ?watts.strogatz.game

- Plot mean degree to check if it is roughly the same

  ?mean

  ?degree

- Your goal is to perform a research on epidemics: Use different values of parameters listed below

  #Setting the infection rate to be β = 0.5, and the recovery rate, to be γ = 1.0,

  beta <- 0.5

  gamma <- 1

  # we use the function sir in igraph to produce

  ntrials <- 500

  ?sir

  sim <-

- At least 3 different versions, for example:
  - beta (4 6 8)
  - gamma (8 6 2)
  - niter (100 500 1000)
- For some reason beta and gamma parameters should not be set below 0 and 1. Looks like they are somehow normalised during simulation.
- You need to plot three values on the graphics: Number of infected, number of suseprible, number of recovered - all depends on time. As a result of this task, you need to provide 12 plots (one for each network with 3 diferent parameters) with explanation.The code below can help you with plotting
  - plot(sim$er)
  - ?plot.sir

```
# The output from each simulation is an sir object, containing information about
# the times at which changes of states occurred and the values of the processes
# NS(t),NI (t), and NR(t) at those times. Plot the total number of
# infectives NI(t) for each of these three networks
```

- Plot only the median curves of all 3 networks in one plot to compare them (with explanation).
  - #differences may be better seen plotting the median of the curves NI(t), for each
  - #graph, on one.

- gl <- list()
- gl$ba <- barabasi.game(250, m=5, directed=FALSE)
- gl$er <- erdos.renyi.game(250, 1250, type=c("gnm"))
- gl$ws <- watts.strogatz.game(1, 100, 5, 0.05)

- ?lapply

```
x.max <- max(sapply(sapply(sim, time_bins), max))
y.max <- 1.05 * max(sapply(sapply(sim, function(x)
  + median(x)[["NI"]]), max, na.rm=TRUE))
```

- Networks: An Introduction by M.E.J. Newman
- Watts, D. J. and Strogatz, S. H., Collective dynamics
- of 'small-world' networks, Nature 393, 440–442 (1998)
- Keeling MJ & Eames KT (2005). Networks and epidemic models. J. R. Soc. Interface 2, 295-307. (This is a very nice review of the field.)
- Liljeros F et al. (2001). The web of human sexual contacts. Nature 411, 907-908. (A brief introduction to sexual networks.)
- Gómez-Gardenes J et al. (2008). Spreading of heterosexually transmitted diseases in heterosexual populations. Proc. Natl. Acad. Sci. USA 105, 1399-1404. (With a lot of data on sexual networks.)
- Szendrői B & Csányi G (2004). Polynomial epidemics and clustering in contact networks. Proc. R. Soc. Lond. B 271, S364-S366. (On the general effect of network structure on epidemic spreading.)
- Eubank S et al. (2004). Modelling disease outbreaks in realistic urban social networks. Nature 429, 180-184. (A tour de force of predicting epidemic spreading in huge populations.)
- Halloran ME et al. (2002). Containing bioterrorist smallpox. Science 298, 1428-1432. (To show that such models can advise policy.)
- Barabási A-L & Albert R. (1999). Emergence of scaling in random networks. Science 286, 509-512. (A seminal paper on scale-free networks.)
- de Sousa JD et al. (2010). High GUD incidence in the early 20th century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. PLoS ONE 5(4), e9936. (with an R model and parameters on HIV epidemics in Africa)
- Kolaczyk, Eric D.. Statistical Analysis of Network Data with R.
- Graphs and Graph Theory on Wikipedia.
- Homepage of the igraph package.