Network Models

Dr Daniele Rotolo

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

Network Analysis
& Infographics

Week 8

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

Network analysis [recap]

Network analysis [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)

Network-level measures

- Diameter
- APL
- Density
- Components
- Cutpoints and bridges
- Point/Line connectivity
- Cliques
- Inclusiveness
- Reachable pairs
- Transitivity

Node-level measures

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Bonacich's centrality
- Weighted centrality
- Brokerage roles
- Effective network size/efficiency
- Constraint

Network analysis [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)

Modelling and inference of networks

- To generate networks with properties that we observe in the real world
- To test the significance of certain characteristics in a given network (e.g. is the observed network expected?)
- To identify which factors predict the formation of ties (e.g. endogenous and exogenous factors)

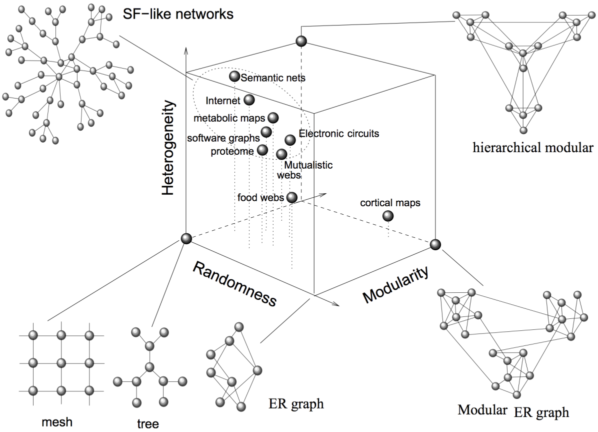


FIG. 3 A zoo of complex networks. In this qualitative space, three relevant characteristics are included: randomness, heterogeneity and modularity. The first introduces the amount of randomness involved in the process of network's building. The second measures how diverse is the link distribution and the third would measure how modular is the architecture. The position of different examples are only a visual guide. The domain of highly heterogeneous, random hierarchical networks appears much more occupied than others. Scale-free like networks belong to this domain.

Source: [Solé and Valverde, 2004]

We can define a model for a network as a collection of networks $\mathbb{G}$ [Kolaczyk and Csárdi, 2014]:

$$\{\mathbb{P}_\theta(G), G \in \mathbb{G} : \theta \in \Theta\}$$

- $\mathbb{P}_\theta$ is the probability distribution of $\mathbb{G}$ — its definition has a fundamental role in defining the model
- θ is a vector of parameters the values of which range in Θ

Mathematical models

Mathematical models

- Random graph models assume $\mathbb{P}_\theta(G)$ to be a uniform distribution
  - Erdós-Rényi random graph model
  - Bernoulli random graph model
  - Generalised random graph models
- Models based on mechanisms that aim to mimic certain properties observed in the real world (they are also called *deterministic models*)
  - Small-worlds models
  - Preferential attachment models
- Mathematical network models are often not capable to fitting observed data – to address this issue statistical network models are used

In the case of the Erdós-Rényi random graph model networks with N nodes and E edges are equally likely

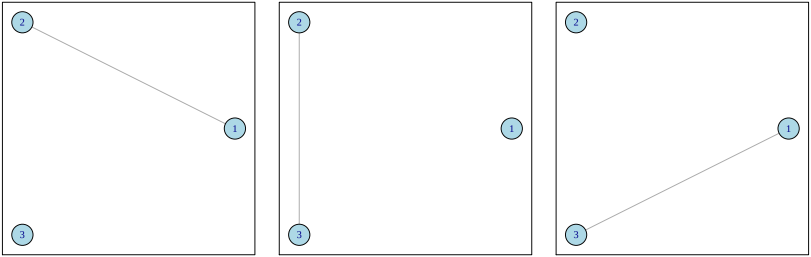
$$\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$$

- N, number of nodes
- E, number of edges
- $\binom{N_G}{E}$ are all possible networks of N nodes and E edges (i.e. permutations)
- $N_G = \binom{N}{2}$
- binomial coefficient: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- factorial: $n! = n \times (n-1) \times (n-2) \times \ldots \times 1$

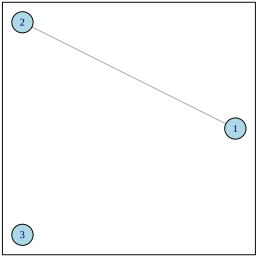
Example of a random network of 3 nodes and 1 edge, $G(3, 1)$

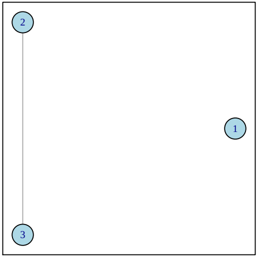
Mathematical models
Erdós-Rényi random graph model

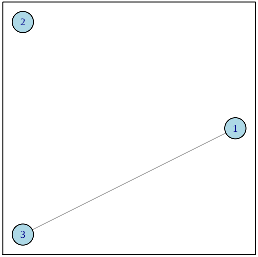
Example of a random network of 3 nodes and 1 edge, $G(3,1)$



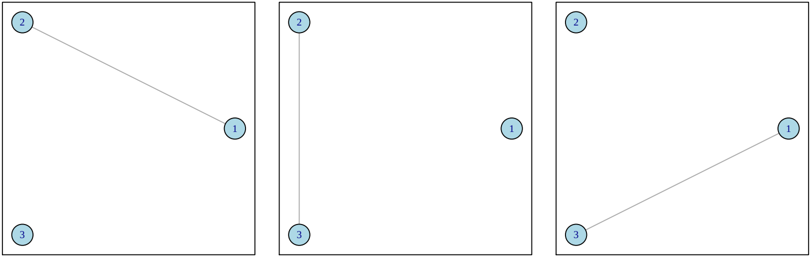
Example of a random network of 3 nodes and 1 edge, $G(3,1)$





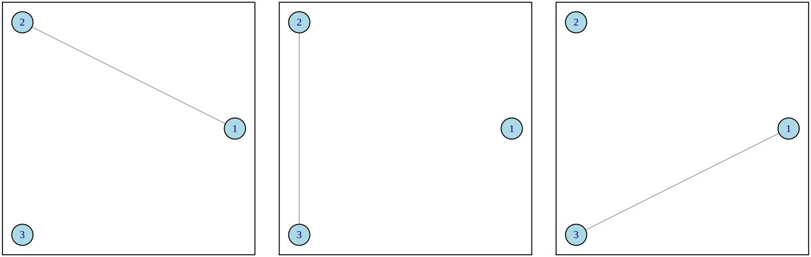


Example of a random network of 3 nodes and 1 edge, $G(3,1)$



1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$

Example of a random network of 3 nodes and 1 edge, $G(3,1)$

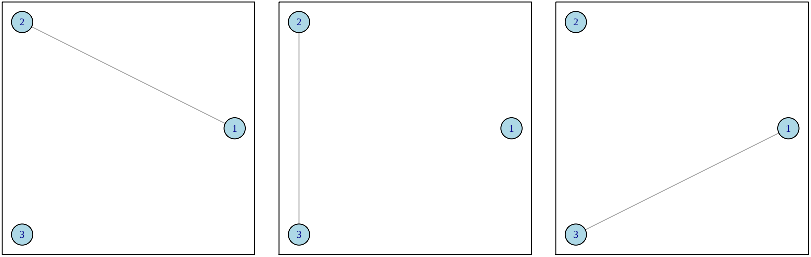


1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$
2. $N_G = \binom{N}{2} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3\times2\times1}{(2\times1)\times(1)} = 3$

Mathematical models

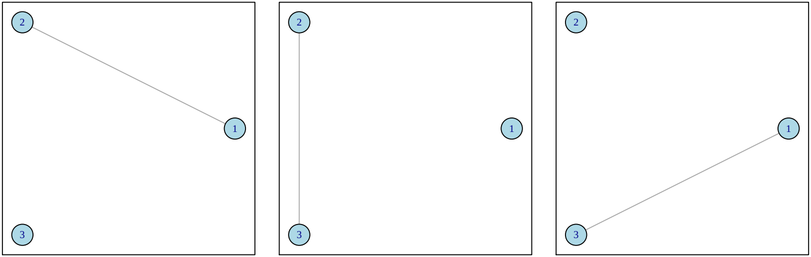
Erdós-Rényi random graph model

Example of a random network of 3 nodes and 1 edge, $G(3,1)$



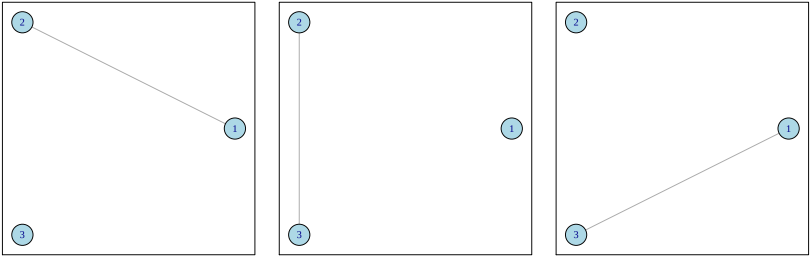
1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$
2. $N_G = \binom{N}{2} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3\times2\times1}{(2\times1)\times(1)} = 3$
3. $\binom{N_G}{E} = \binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3\times2\times1}{(1)\times(2\times1)} = 3$

Example of a random network of 3 nodes and 1 edge, $G(3,1)$



1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$
2. $N_G = \binom{N}{2} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3\times2\times1}{(2\times1)\times(1)} = 3$
3. $\binom{N_G}{E} = \binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3\times2\times1}{(1)\times(2\times1)} = 3$
4. 3 possible permutations of $G(3,1)$

Example of a random network of 3 nodes and 1 edge, $G(3,1)$



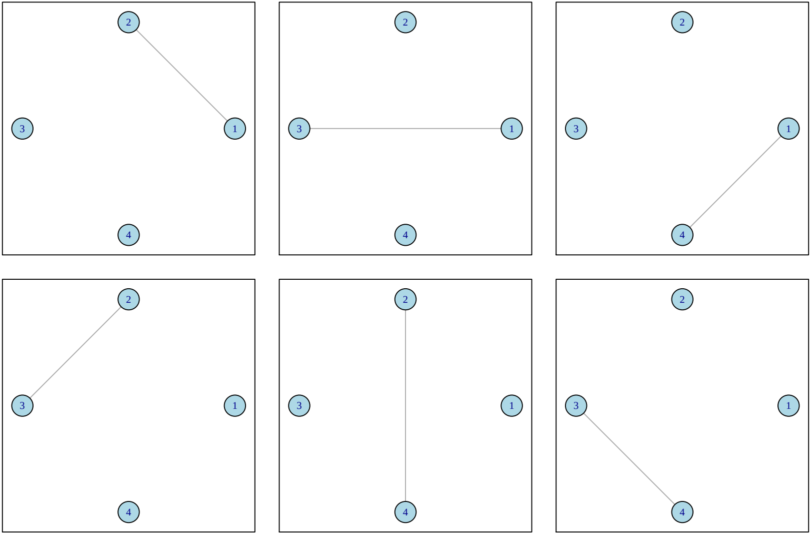
1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$
2. $N_G = \binom{N}{2} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3\times2\times1}{(2\times1)\times(1)} = 3$
3. $\binom{N_G}{E} = \binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3\times2\times1}{(1)\times(2\times1)} = 3$
4. 3 possible permutations of $G(3,1)$
5. The probability of observing each network is $\binom{N_G}{E}^{-1} = \frac{1}{3} = 0.33$

Example of a random network of 4 nodes and 1 edge, $G(4, 1)$

1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$
2. $N_G = \binom{N}{2} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4\times3\times2\times1}{(2\times1)\times(2\times1)} = 6$
3. $\binom{N_G}{E} = \binom{6}{1} = \frac{6!}{1!(6-1)!} = \frac{6\times5\times4\times3\times2\times1}{(1)\times(5\times4\times3\times2\times1)} = 6$
4. 6 possible permutations of $G(6, 1)$
5. The probability of observing each network is $\binom{N_G}{E}^{-1} = \frac{1}{6} = 0.17$

Mathematical models

Erdós-Rényi random graph model



The number of possible permutations can rapidly increase—for example let's consider a random network of 10 nodes and 30 edges, $G(10,30)$

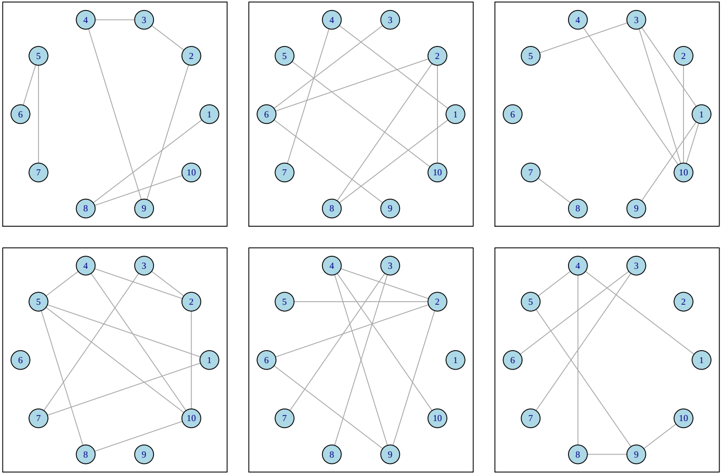
1. $\mathbb{P}_\theta(G) = \binom{N_G}{E}^{-1}$
2. $N_G = \binom{N}{2} = \binom{10}{2} = \frac{10!}{2!(10-2)!} = 45$
3. $\binom{N_G}{E} = \binom{45}{30} = \frac{45!}{30!(45-30)!} = 344,867,425,584$
4. About 345 billion of possible permutations of $G(10,30)$
5. The probability of observing each network is $\binom{N_G}{E}^{-1} = \frac{1}{344,867,425,584} = 2.9 \times 10^-12$

The Bernoulli random graph model considers all graphs of N nodes that are generated by independently assigning ties to nodes with a probability $p \in (0,1)$—networks with no edges or fully connected are not likely outcomes, but they are possible

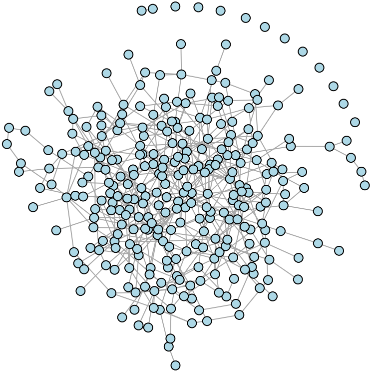
$$\mathbb{P}_\theta(G) = p^E(1-p)^{N_G - E}$$

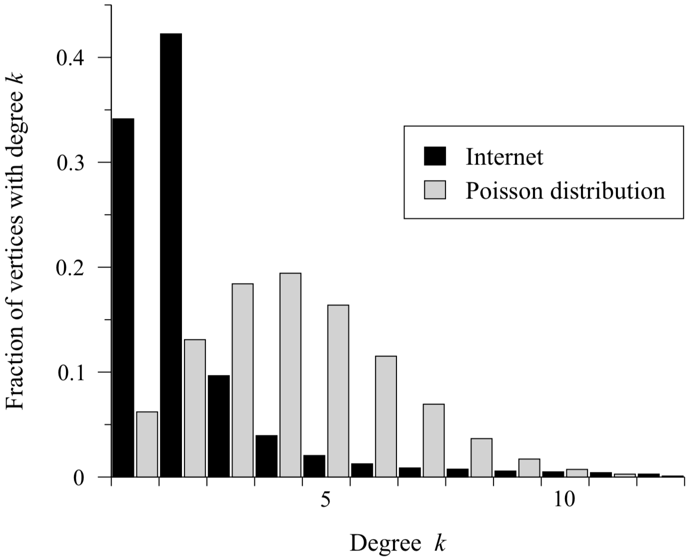
- N, number of nodes
- E, number of edges
- $N_G = \binom{N}{2}$
- binomial coefficient: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- factorial: $n! = n \times (n-1) \times (n-2) \times \ldots \times 1$

Example of a random network of 10 nodes and ties between nodes with $p = 0.20$



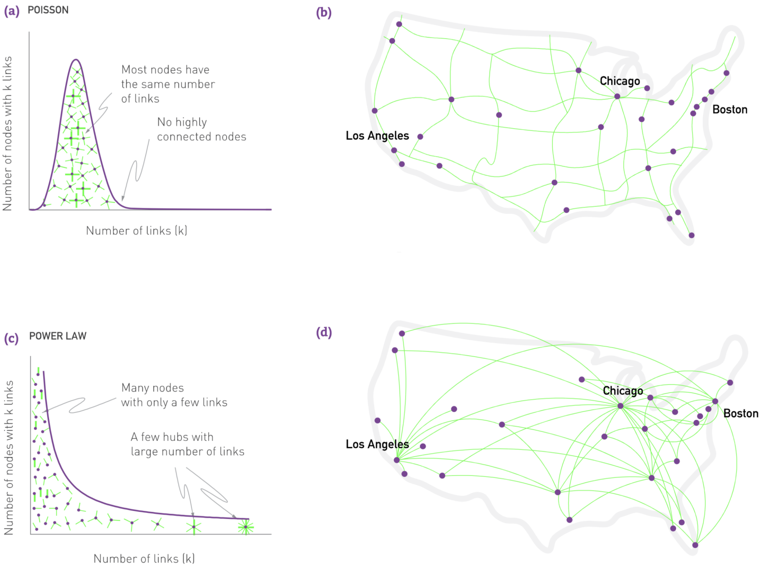
- Network generated by using Bernoulli random graph models are likely to have a giant component when $p > 1/N$
- Erdós-Rényi and Bernoulli random graph models are equivalent for large N and when $E \sim pN^2$





Source: Distributions with the same averages of degree [Newman, 2010]

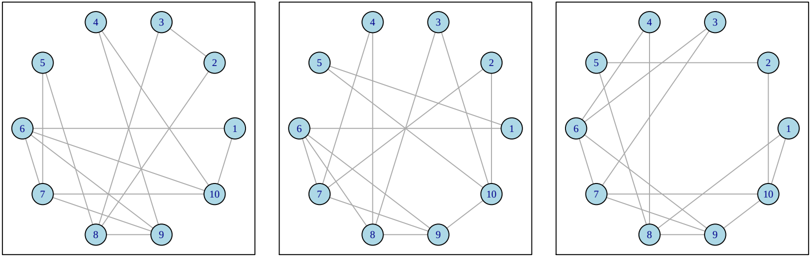
Mathematical models: Erdós-Rényi and Bernoulli random graph models



Source: Random networks (top, road connections) and scale-free networks (bottom, flight connections) [Barabási, 2016]

Generalised random graph model extends the Erdós-Rényi random graph model to a collection of graph of N nodes that have certain properties

- Two stubs are chosen uniformly at random and connected, with the number of stubs for a node equal to the defined degree of the node
- The sum of degree values must be even
- Networks do not appear with the same probability
- Example: $N = 10$ and $d(2, 2, 2, 2, 2, 4, 4, 4, 4, 4)$



Deterministic models attempts to generate two main characteristics that are observed in real networks, but that are not adequately reproduced by random graph models

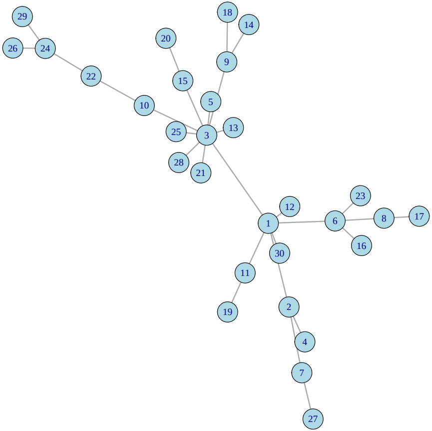
- Scale-free networks (i.e. degree distribution following a power law)
  $\Rightarrow$ **Preferential attachment models**
- High levels of transitivity
  $\Rightarrow$ **Small-world network**

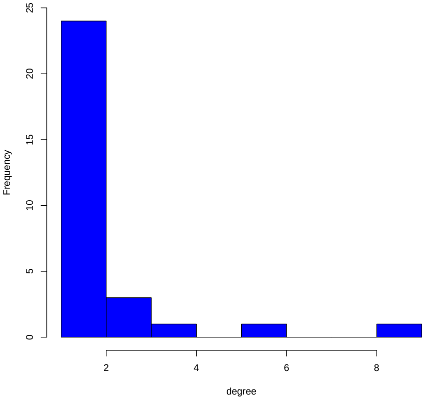
Preferential attachment models: "the rich get richer" (used to model accumulation effect in science) [Barabasi and Albert, 1999]

1. Network at $t = 0$
2. The network is modified at $t = 1$ by adding a node j with a degree of $m > 0$
3. The probability that each edge of the new node is attached to an existing node is proportional to the degree of the existing node: $\frac{C_D(n_i)}{\sum_k C_D(n_k)}$
4. This model is likely to generate hubs, i.e. nodes with high degree

Mathematical models

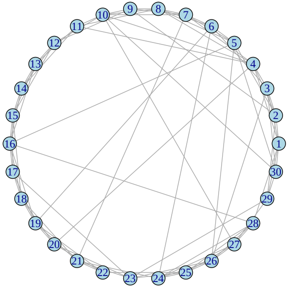
Models based on mechanisms: Preferential attachment models





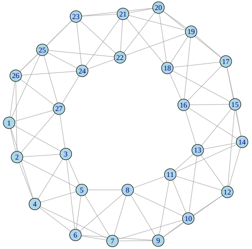
Small-world models generate networks with high clustering and small distances between nodes [Watts and Strogatz, 1998]

- Clustering coefficient (or transitivity): $C = \frac{\text{number of triangles x 3}}{\text{number of connected triples}}$
- High clustering in relation to c/N
- Many networks in the real world have these characteristics
- 'Six degree of separation'
- Brain connectivity (neurons)
- Power grids
- ...
- Specialisation and efficiency
- Do not mimic scale-free networks

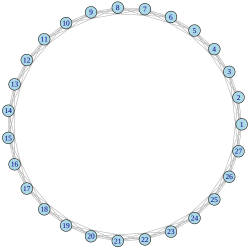


Source: Kevin Bacon's six degree of separation (www.youtube.com/watch?v=n9u-TITxwoM)

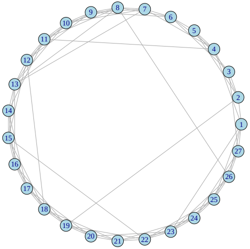
1. Creation of a network of N nodes with lattice structure: each node is connected to r of its neighbours
2. Each edge, independently and with probability p, is moved to be incident to another node, which is chosen uniformly (no loops and multi-edges)



lattice structure
$N = 27,\ r = 3$



lattice structure (circle layout)
$N = 27,\ r = 3$



small-world network
$N = 27,\ r = 3,\ p = 0.1$

Statistical network models

Statistical network models

Mathematical network models are often not capable to fitting observed data – to address this issue statistical network models are used

- Exponential Random Graph Models (ERGM)
  - Consider the presence/absence of a tie as the response variable
  - The response variable is dependent on a number of endogenous (e.g. transitivity) and exogenous (e.g. attributes) characteristics [Robins et al., 2007]
- Stochastic Actor-Oriented Models (SAOM): The co-evolution of a network structure and attributes is modelled as a stochastic process [Snijders et al., 2010]
- Network Block Models model the propensity to establish a tie between two nodes as dependent on the 'class' membership of the two nodes (e.g. nodes in the same class are more likely to establish a tie) [Doreian et al., 1984]
- ... new models every years (this is an emerging area)

Limitations of random graph models and the key idea of ERGM

- In the case of random graph models, we fix a network property in absolute values — for example, the number of edges
- We then generate a (uniform) distribution of networks with such a property making assumption that ties between nodes occur at random
- We are, however, excluding the possibility that the network we observe may have been different — for example:
  - we could have observed networks with a different number of edges
  - social forces may influence the formation of ties (e.g. preferential attachment)
- A better approach would be:
  - to fix the average value of the network property, i.e. average number of edges
  - to generate a distribution of networks where networks with a number of edges that is close to the desired value have higher probabilities than network with a lower/higher number of edges

Statistical network models

The key idea of ERGM

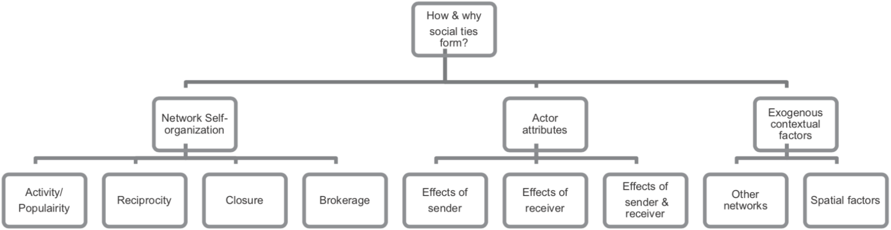
- ERGM gives flexibility by fixing multiple network properties:
  - Number of edges or mean degree
  - Degree of individual vertices
  - Number of triangles or clustering coefficient
  - ...
- Physicists and statisticians have demonstrated that the exponential distribution is the 'best choice' (minimum assumptions except for those imposed by the properties we select)
- We do not enter into the details of the estimation

The general formulation of ERGM is reported below:

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y}) = \frac{e^{\sum_k \theta_k s_k(\mathbf{y})}}{\sum_y \left(e^{\sum_k \theta_k s_k(\mathbf{y})}\right)}$$

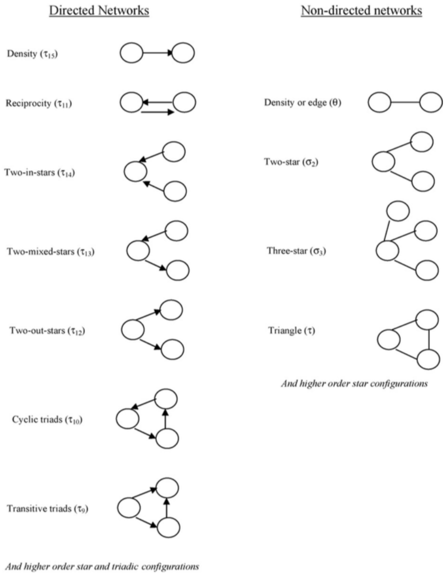
- $\mathbf{Y} = [Y_{ij}]$, is the random adjacency matrix
- $\mathbf{y} = [y_{ij}]$, is the particular realisation of $\mathbf{Y}$
- θ, vector of k parameters to estimate
- $s_k(\mathbf{y})$, independent variables (endogenous and/or exogenous)
- In few words
  - The numerator is the probability of observing the specific network $\mathbf{y}$
  - The denominator is the sum of all probabilities of the networks in the distribution (this is, in general, very difficult to estimate: simulation with Markov Chain Monte Carlo Maximum Likelihood Estimation, MCMCMLE)

Statistical network models
The formulation of ERGM



Source: Selection of independent variables (endogenous and/or exogenous) [Amati et al., 2018]

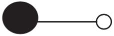
Statistical network models
The formulation of ERGM



Source: Some of the $s_k(\mathbf{y})$ endogenous network statistics [Robins et al., 2007]

Statistical network models

The formulation of ERGM

Configuration		Statistic	Parameter
Binary attributes			
1		$\sum_{i<j} x_{ij}(y_i + y_j)$	Attribute-based activity
2		$\sum_{i<j} x_{ij} y_i y_j$	Homophily (interaction)
Continuous attributes			
3		$\sum_{i<j} x_{ij}(y_i + y_j)$	Attribute-based activity
4		$\sum_{i<j} \lvert y_i - y_j \rvert x_{ij}$	Homophily (difference)

Source: Some of the $s_k(\mathbf{y})$ exogenous attributes [Lusher et al., 2012]

- When estimating ERGM, we try to find a vector of θ such that the network we observe in the distribution of networks is the most likely
- We can reformulate the ERGM as the odds of drawing the network $\mathbf{y}$ over the network $\mathbf{x}$ given θ

$$\frac{\mathbb{P}_\theta(\mathbf{Y}=\mathbf{y})}{\mathbb{P}_\theta(\mathbf{Y}=\mathbf{x})} = \frac{e^{\sum_k \theta_k s_k(\mathbf{y})}}{e^{\sum_k \theta_k s_k(\mathbf{x})}}$$

$$\log\left(\frac{\mathbb{P}_\theta(\mathbf{Y}=\mathbf{y})}{\mathbb{P}_\theta(\mathbf{Y}=\mathbf{x})}\right) = \frac{\sum_k \theta_k s_k(\mathbf{y})}{\sum_k \theta_k s_k(\mathbf{x})}$$

- If we assume **y** and **x** to differ only by one tie y_{ij}, we can transform the formulation as follows [Hunter et al., 2008]

$$logit\left(\mathbb{P}(Y_{ij} = 1 |\text{ n nodes}, Y_{ij}^{c})\right) = \sum_{k} \theta_k \delta_k(\mathbf{y})$$

$$\mathbb{P}(Y_{ij} = 1 |\text{ n nodes}, Y_{ij}^{c}) = logistic\left(\sum_{k} \theta_k \delta_k(\mathbf{y})\right)$$

- $\delta_k(\mathbf{y})$, change in the network statistic, i.e. $s_k(\mathbf{y}) - s_k(\mathbf{x})$
- when θ_k is significant the given variable δ_k play a role in generating the network

- ERGM may suffer of degeneracy, i.e. the model produces that networks are fully connected or with no connections at all
- [Snijders et al., 2006, Hunter, 2007] proposed to include the following terms
  - Geometrically Weighted Degrees (GWD)
  - Geometrically Weighted Edgewise Shared Partnership (GWESP)
  - Geometrically Weighted Dyadwise Shared Partnership (GWDSP)
- It is important you run some test (Goodness of Fit, GoF) to check that the model is generating a distribution of networks that fits your data

- Standard statistical approaches (e.g. OLS) assume that observations are independent one of another, but this is not the case of networks where the focus is on ties between nodes (dependency) [Robins et al., 2012, Snijders, 2011]
- One observation may be associated with another through network links, thus the errors are correlated with each other
- Example: disease contagion (sources of infection: related to the individual and network transmission of the disease)
- If using regression, you should account for this effect (e.g. multi-level models, autocorrelation)

Questions

Next time ...

Next time ...

- **Seminar: Network models**
  - Network mathematical models in igraph

- **Lecture: Innovation networks**
  - Use of network analysis to map science and technology

References I

Amati, V., Lomi, A., and Mira, A. (2018).
Social Network Modeling.
Annual Review of Statistics and Its Application, 5(1):343–369.

Barabási, A.-L. (2016).
Network science.
Cambridge university press, Cambridge, MA.

Barabasi, A. L. and Albert, R. (1999).
Emergence of scaling in random networks.
Science, 286(5439):509–512.

Doreian, P., Teuter, K., and Wang, C.-H. (1984).
Network Autocorrelation Models: Some Monte Carlo Results.
Sociological Methods & Research, 13(2):155–200.

Hunter, D. R. (2007).
Curved exponential family models for social networks.
Social Networks.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008).
Goodness of fit of social network models.
Journal of the American Statistical Association.

Kolaczyk, E. and Csárdi, G. (2014).
Statistical analysis of network data with R, volume 65.
Springer-Verlag, New York, NY, USA.

Lusher, D., Robins, G., and Koskinen, J. (2012).
Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications.
Cambridge University Press, Cambridge.

References II

Newman, M. (2010).
Networks. An introduction.
Oxfor University Press, Oxford, UK.

Robins, G., Lewis, J. M., and Wang, P. (2012).
Statistical Network Analysis for Analyzing Policy Networks.
Policy Studies Journal, 40(3):375–401.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007).
An introduction to exponential random graph (p*) models for social networks.
Social Networks, 29(2):173–191.

Snijders, T. A. B. (2011).
Statistical Models for Social Networks.
Annual Review of Sociology, 37(1):131–153.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006).
New specifications for exponential random graph models.

Snijders, T. A. B., van de Bunt, G. G., and Steglich, C. E. G. (2010).
Introduction to stochastic actor-based models for network dynamics.
Social Networks, 32(1):44–60.

Solé, R. V. and Valverde, S. (2004).
Information Theory of Complex Networks: On Evolution and Architectural Constraints BT.
In Ben-Naim, E., Frauenfelder, H., and Toroczkai, Z., editors, *Complex Networks*, pages 189–207. Springer Berlin Heidelberg, Berlin, Heidelberg.

Watts, D. J. and Strogatz, S. H. (1998).
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442.