An Introduction to Social Network Data Analytics

Dr. Şebnem Er

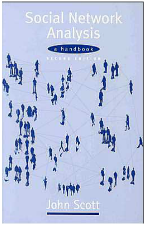
University of Cape Town
Statistical Sciences Department
Sebnem.Er@uct.ac.za
https://sebnemer.github.io/

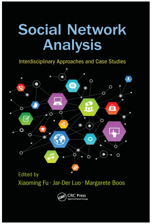
August 30, 2019

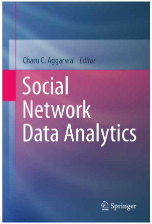
Structure of the Workshop

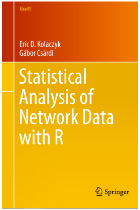
- What is social network analysis?
  - Network data: examples and data collection.
  - foundational concepts of social networks: graph theory
- Types of social network analysis.
  - visualization of social networks,
  - descriptive analysis of social networks,
  - statistical network models.
- R packages.

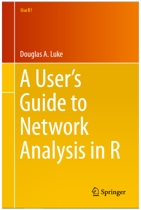
Resources (Books)

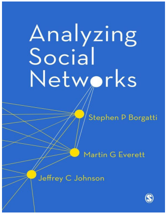












What is Social Network Analysis?

- ⋆ A social network is defined as a network of interactions or relationships between actors.
- ⋆ Social networks are mainly represented with a graph in which a set of vertices/nodes (actors) are connected to each other with edges representing some sort of a relationship (i.e. friendship, trade, flights etc.).
- ⋆ Many different relationships can be represented with networks, i.e.
  - between friends representing friendship levels,
  - between academics representing collaboration levels,
  - between cities representing flight routes,
  - between companies representing trade,
  - between neurons in our brain representing neural connection etc.
- ⋆ Social networks have become very popular in recent years.
- ⋆ Online social networks

Examples


University of Cape Town

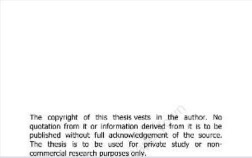
Faculty of Science

Department of Environmental and Geographical Sciences

The capacity of local governments to build flood resilience in informal settlements: a social networks approach

Bouwer, Roy (2017)

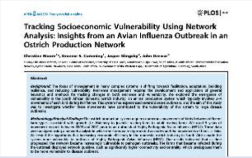
adaptation strategies for long-term resilience. Drawing on **social network analysis** this research examines the role of **social** ties and **social** capital in facilitating knowledge sharing to improve the capacities of local governments to deal with flood...


The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Understanding Highly Pathogenic Avian Influenza outbreaks in Western Cape Ostrich industry: did network dynamics enhance vulnerability? Christine Moore.

Moore, Christine (2012)

utilized **network analysis** to understand how the movement of ostrich stock between farm locations in the Western Cape, South Africa may have contributed an epidemic outbreak of Highly Pathogenic Avian Influenza (HPAI) within the ostrich industry in 2011....


Tracking Socioeconomic Vulnerability Using Network Analysis: Insights from an Avian Influenza Outbreak in an Ostrich Production Network

Abstract

Tracking socioeconomic vulnerability using network analysis: insights from an avian influenza outbreak in an ostrich production network

Moore, Christine; Cumming, Graeme S; Slingsby, Jasper; Grewar, John (Public Library of Science, 2014)

to an outbreak of Highly Pathogenic Avian Influenza (H5N2). These data were analyzed using a **network analysis** in which the farms were represented as nodes and the movements of birds as links. We tested the hypothesis that increasing economic efficiency...

Examples

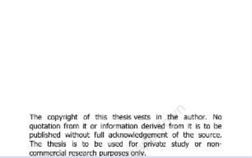
Understanding pathogen transmission dynamics in waterbird communities: At what scale should interactions be studied?

Introduction

Understanding pathogen transmission dynamics in waterbird communities: At what scale should interactions be studied?

MacGregor, Lindy H; Cumming, Graeme S; Hockey, Philip A (Academy of Science of South Africa, 2011)

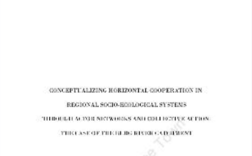
transmission. In this paper we ask whether these data can be used to infer fine-scale transmission patterns. We tested for non-randomness in waterbird assemblages and explored waterbird interactions using **social network analysis**. Certain **network** parameter...



The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Human **networks** of tetrapod translocations in the Western Cape, South Africa: trends and potential impacts on biodiversity

Goss, Jeremy R (2012)



CONCEPTUALIZING HORIZONTAL COOPERATION IN REGIONAL SOCIO-ECOLOGICAL SYSTEMS

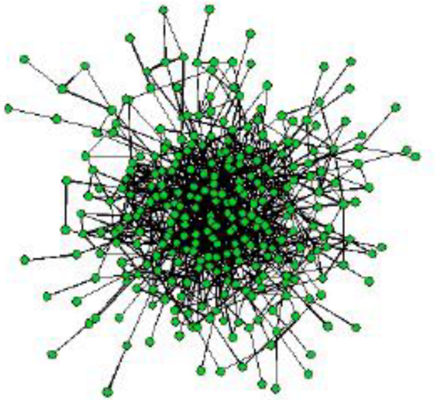
Conceptualizing horizontal cooperation in regional socio-ecological systems through actor **networks** and collective action: the case of Berg River catchment

Methner, Nadine (2014)

in the SES. A formal **social network analysis** is employed to describe and analyze the management of the Berg River catchment. The **analysis** focuses on (i) cohesion and (ii) heterogeneity, which are two **network** characteristics that affect learning and collective...

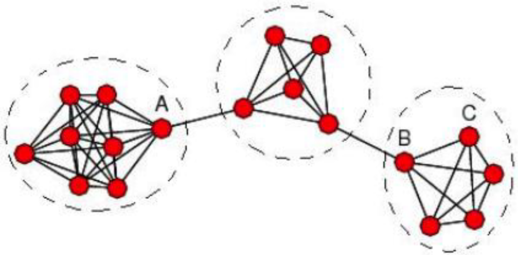
Community structure and centrality effects in the South African company network by Durbach, Katshunga, Parker 2013

Investigating community structure in the South African company network in which the vertices/nodes are South African companies listed on the JSE (at March 2008), and two companies are connected if they share one or more directors on their respective boards.



An analysis of corporate board networks in South Africa by Durbach, Parker 2009

In this paper we analyse the networks created from directors sitting on the boards of companies in South Africa. We consider two projections of this network: a director network, in which only directors are present and two directors are linked if they sit together on one or more common boards; and a firm network, in which only firms are present and an edge indicates that the two firms share one or more directors. We describe these networks in terms of the statistical properties that they possess, and compare them to theoretical values obtained under various random network models.



Who are the researchers that are collaborating with industry? An analysis of the wine sectors in Chile, South Africa and Italy by Giuliani, et.al, 2010

This paper provides new empirical evidence and, in particular, looks at the importance of researchers individual characteristics and their institutional environments in explaining the propensity to engage in different types of U-I linkages. Based on an original dataset, we present new evidence on three wine producing areas Piedmont, a region of Italy, Chile and South Africa that have successfully responded to recent structural changes in the industry worldwide. Empirical findings reveal that researchers individual characteristics, such as centrality in the academic system, age and sex, matter more than publishing records or formal degrees. Institutional specificities at country level also play a role in shaping the propensity of researchers to engage with industry.

Sources of Network data

Network data can be collected from either

- primary (we directly ask people questions or observe their behavior) or
- secondary (we gather data that already exists somewhere, whether in paper records, or electronic databases)

sources.

There is also increasing amount of longitudinal studies.

Types of Nodes and Edges

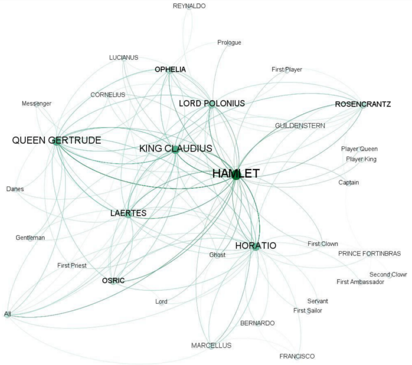
Most network studies involve persons as the nodes and interpersonal relations as the ties. However, the nodes can be all kinds of entities

- firms,
- cities,
- countries,
- monkeys,
- neurons,
- or even the characters in a book and so on.

And the type of node obviously has a major impact on what kinds of relationships are collected and how they are collected. These decisions are interlinked and must to some extent be considered together.

Example: Hamlet - Evans (2017)

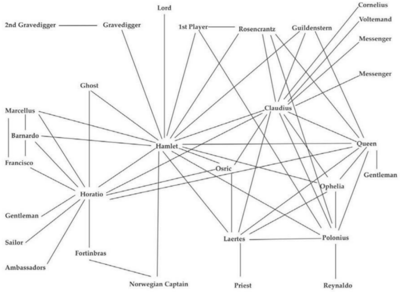
Generation and Analysis of a Social Network: Hamlet (2017) by Preston Evans, University of Arkansas Computer Science and Computer Engineering Undergraduate Honours Thesis. Nodes and edges were created between each speaking character in every act of Hamlet, i.e. edge exists if two appear together:



https://pdfs.semanticscholar.org/077f/1d99df7c4aa9d25a0c82d69c55957ee9f117.pdf

Example: Hamlet - Moretti (2011)

Network Theory, Plot Analysis (2011) by Moretti, F.
Basically, two characters are linked if some words have passed between the actors: an interaction is a speech act.



https://newleftreview.org/issues/II68/articles/franco-moretti-network-theory-plot-analysis

Sampling and Bounding

One of the most vexing problems for those just starting out in network research is the problem of bounding the set of nodes to be included in the study.

- Often the whole network is very large, and we only observe a portion of it.
- One needs to be careful because when you sample from the large network, the sub sample might not hold the original network properties.
- Sometimes a large network (ie. twitter) may be available but might be too complex.
- Sometimes we are interested in a rare feature of the observations, then random sampling is not efficient at all.

Sampling and Bounding

There are two main sampling types to achieve these:

- induced subgraph sampling: sample vertices at random
- snowball sampling: interview any qualifying actor with a tie to any actor already selected, up to K waves or until quotas or cost limits reached.

Network Basics - Preliminaries

Graphs are commonly used to represent network data, and the graph shows the existence of some sort of a relationship between the individuals.

$$G = (V, E, W)$$

is the mathematical form where $V = \{v_1, v_2, \ldots, v_n\}$ represents the set of vertices/nodes (individuals), and $E = \{e_1, e_2, \ldots, e_L\}$ represents the set of edges (interconnections between individuals/vertices), where W represents the edge weights, or in other words the strength of the connection. There are n vertices and L edges. In our friendship example, friend names are the vertices and the friendship levels are the edges.

Graphs can be either

- directed (digraph)
- undirected (no ordering of the vertices defining an edge)

Network Data Representation

Network Data can be represented in 2 ways:

- Edge list,
- Adjacency matrix.

Simple example - Edge list with no information

2 column list, each row represents an edge between 2 nodes. All possible node points joined by an edge are represented by a list. Most common way of collecting data.

```
library(igraph)
# example in data frame
a=c("A","A","A","A","A","E","F")
b=c("B","C","D","E","F","F","G")
df = cbind(a,b)
df = as.data.frame(df)
df
```

```
> df
  a b
1 A B
2 A C
3 A D
4 A E
5 A F
6 E F
7 F G
```

Simple example - Adjacency matrix with no information

Represented in a matrix form ($N_V x N_V$ matrix) where the nodes are in the columns and rows of the matrix and the elements of the matrix are 0/1 representing if a connection exists or not:

$$A_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

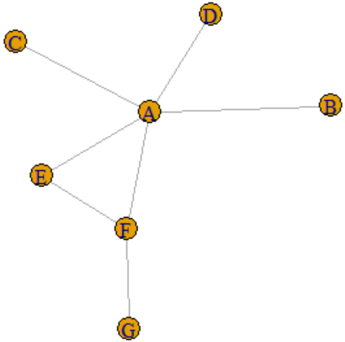
When data is large, then very sparse and difficulties arise.

```
g = graph.edgelist(as.matrix(df[,1:2]),directed =
    FALSE)
get.adjacency(g)
```

```
7 x 7 sparse Matrix of class "dgCMatrix"
  A B C D E F G
A . 1 1 1 1 1 .
B 1 . . . . . .
C 1 . . . . . .
D 1 . . . . . .
E 1 . . . . 1 .
F 1 . . . 1 . 1
G . . . . . 1 .
```

Simple example... Plot

```
plot(g)
```



Simple example... information on nodes

We normally have more information about nodes and edges rather than only existence. We might have information on the ages of the friends and how many times they call each other during the day.

```
# network attributes
## vertex/node attributes
g = set.vertex.attribute(g, "age", value = c(20,
    25, 21, 23, 24, 23, 22))
g
```

```
IGRAPH c099c46 UN-- 7 7 --
+ attr: name (v/c), age (v/n)
+ edges from c099c46 (vertex names):
[1] A--B A--C A--D A--E A--F E--F F--G
```

```
vertex.attributes(g)
```

```
$name
[1] "A" "B" "C" "D" "E" "F" "G"

$age
[1] 20 25 21 23 24 23 22
```

Simple example... information on edges

```

## edge attributes
g=set.edge.attribute(g, "frequency", value = c
  (2,1,1,1,3,2,4)) #how many times friends call
  each other during day
g
```

```
IGRAPH c099c46 UN-- 7 7 --
+ attr: name (v/c), age (v/n), frequency (e/n)
+ edges from c099c46 (vertex names):
[1] A--B A--C A--D A--E A--F E--F F--G
```

```
edge.attributes(g)
```

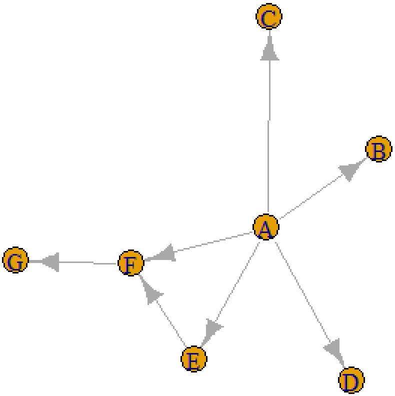
```
$frequency
[1] 2 1 1 1 3 2 4
```

Simple example... alternatively as a df - directed

```
name=c("A","B","C","D","E","F","G")
age=c(20, 25, 21, 23, 24, 23, 22)
vertices.df = data.frame(name, age)

from = c("A","A","A","A","A","E","F")
to = c("B","C","D","E","F","F","G")
frequency = c(2,1,1,1,3,2,4)
edges.df = data.frame(from,to,frequency)

g = graph_from_data_frame(d = edges.df, vertices =
    vertices.df, directed = TRUE)
get.adjacency(g)
```

```
7 x 7 sparse Matrix of class "dgCMatrix"
  A B C D E F G
A . 1 1 1 1 1 .
B . . . . . . .
C . . . . . . .
D . . . . . . .
E . . . . . 1 .
F . . . . . . 1
G . . . . . . .
```

Simple example... directed graph



Visualizing and Characterizing: Graph Layouts

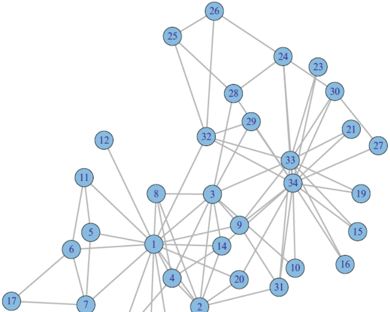
Techniques for displaying network graphs are the focus of the field of graph drawing or graph visualization. Such techniques typically seek to incorporate a combination of elements from mathematics, human aesthetics, and algorithms. There are uncountably many ways.

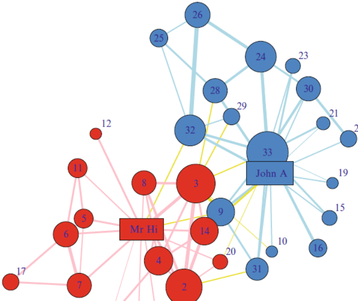
Efficient algorithms have been developed and these can be called with the `plot` command in `igraph` package.
At the heart of graph visualization is the graph layout, i.e., the placement of vertices and edges in space:

- circular layout,
- spring-embedder (pulling connected vertices closer to one another) methods: Fruchterman and Reingold layout
- grid layout,
- Kamada-Kawai algorithm,
- tree layout etc...

Visualizing and Characterizing: Decorating Graph Layouts

- node shape
- node size
- node colour
- node label
- edge width
- edge colour
- edge type





Visualizing - rules of thumb

Network graphics are easier to understand if they follow as much as possible the following five guidelines:

- Minimize edge crossings.
- Maximize the symmetry of the layout of nodes.
- Minimize the variability of the edge lengths.
- Maximize the angle between edges when they cross or join nodes.
- Minimize the total space used for the network display.

Descriptive Analysis

How large the network is, how densely connected it is, whether the network is made up of one or more distinct groups, how compact it is, and how clustered are the network members?

- Nodal degree
  - Mean nodal degree
  - Variance of nodal degrees
  - Distribution of nodal degrees
- Nodal centrality
- Diameter
- Network cohesion

Descriptive Analysis: Nodal Degree

Nodal degree: Degree of a node is the number of edges which involve that node.

```
d.g = degree(g)
d.g
vcount(g) # total number of vertices
```

```
A B C D E F G
5 1 1 1 2 3 1
7 # there are seven vertices
```

For directed graphs, indegree-column sum, outdegree-row sum of the adjacency matrix.

- Mean nodal degree: a statistic that reports the average degree of nodes.

$$\bar{d} = \frac{\sum_{i=1}^{g} d(v_i)}{g} = \frac{5+1+1+1+2+3+1}{7} = 2$$

Descriptive Analysis: Nodal degree

- Variance of nodal degrees

$$s_D^2 = \frac{\sum_{i=1}^{g}(d(v_i) - \bar{d})^2}{g}$$

The nodal degrees are an important property of a graph, and we will often want to control for or condition on the set of nodal degrees in a graph when we use statistical models to study the tendencies towards higher order network properties.

For directed graphs, there are two measures: in-degree which is the column sum, and out-degree which is the row sum of the adjacency matrix.

Descriptive Analysis: Nodal degree

- Distribution of nodal degrees The distribution of nodal degrees (k different nodes) is represented with a probability distribution $P(k)$ defining the fraction of nodes with a particular degree level.
  Hence in our example there are 0 node with 0 edge, 4 nodes with 1 edge, 1 node with 2 edges, 1 node with 3 edges, 0 node with 4 edges and 1 node with 5 edges.

```
      dd.g = degree.distribution(g)
      dd.g; sum(dd.g); table(degree(g))/7
```

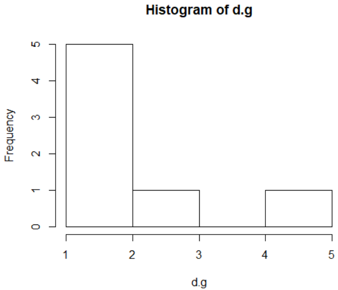
```
[1] 0.0000000 0.5714286 0.1428571 0.1428571 0.0000000 0.1428571
[1] 1
        1         2         3         5
0.5714286 0.1428571 0.1428571 0.1428571
```

Large majority of networks are right-skewed, more nodes with low degree, few nodes ("hubs") with high degree.

Descriptive Analysis: Nodal degree

- Distribution of nodal degrees

```r
hist(d.g)

```



Large majority of networks are right-skewed, more nodes with low degree, few nodes ("hubs") with high degree.

Descriptive Analysis: Nodal Centrality

Nodal Centrality: Vertex importance in the network. Which node is most central? Which actors in a social network seem to hold the 'reins of power'? There are a vast number of different centrality measures:

- Closeness centrality: An important vertex is central. Centrality is measured with the inverse of the total shortest path distance of a vertex from all others. The highest C_C measure tells us that node is tied with highest closeness.

$$C_C(v_i) = \left[\sum_{j=1}^{g} d(v_i, v_j) \right]^{-1}$$

```
cc = closeness(g)
cc
```

```
     A      B      C      D      E      F      G
0.1429 0.0833 0.0833 0.0833 0.1000 0.1111 0.0714
```

Descriptive Analysis

- Nodal Centrality
  - Betweenness centrality: The importance of a vertex is related to where it is located with respect to the paths in the network graph.
  - Connectivity: The smallest number of edges whose removal results in a disconnected graph.
  - Spectral decomposition: Eigenvalues and eigenvectors of the adjacency matrix also provide another centrality measure.

Descriptive Analysis: Diameter

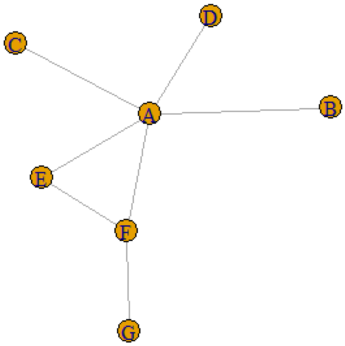
The diameter of a network is a useful measure of this compactness. A path is the series of steps required to go from node A to node B in a network.

The shortest path is the shortest number of steps required. The diameter then for an entire network is the longest of the shortest paths across all pairs of nodes. This is a measure of compactness or network efficiency in that the diameter reflects the ‘worst case scenario’ for sending information (or any other resource) across a network.

Although social networks can be very large, they can still have small diameters because of their density and clustering.

Simple example... Plot

```
plot(g)
```



Descriptive Analysis: Diameter

In total there are $C_2^7 = 21$ pairs of nodes in this example:

```
[A->B,C,D,E,F]   1   [A->G]  2
[B->C,D,E,F]     2   [B->G]  3
[C->D,E,F]       2   [C->G]  3
[D->E,F]         2   [D->G]  3
[E->F]           1   [E->G]  2
[F->G]           1
```

The longest of the shortest path length is 3.

```
diameter(g)
```

```
[1] 3
```

Descriptive Analysis: Network Cohesion

Network cohesion: The extent to which subsets of vertices are cohesive or 'stuck together' with respect to the relation defining edges in the network graph.

Cohesive subgroups are sets of actors that are tied together through frequent, strong, and direct ties.

Eg. Do friends of a given actor in a social network tend to be friends of one another as well?

There are many ways (i.e. subgraphs and censuses) that we can summarise network cohesion, we will focus on cliques and density.

- Cliques

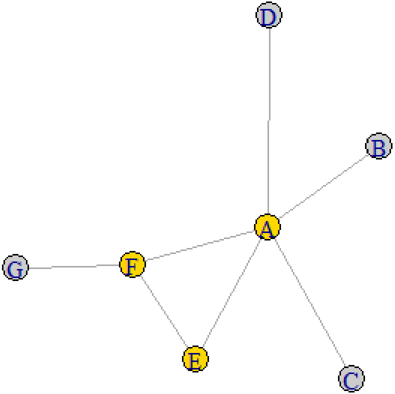
Descriptive Statistics: Network Cohesion

- Cliques: Cliques are one of the simplest types of cohesive subgroups, and one of the easiest types to understand. A clique is a maximally complete subgraph; that is, it is a subset of nodes that have all possible ties among them.

```
clique.number(g)  # 3, not the number of
    cliques, but the size of the largest clique
    .
cliques(g, min = 2)
largest.cliques(g)
```

```
[[1]]
+ 2/7 vertices, named, from c27eb32:
[1] F G

[[2]]
+ 2/7 vertices, named, from c27eb32:
[1] A F

[[3]]
+ 2/7 vertices, named, from c27eb32:
[1] A D

[[4]]
+ 2/7 vertices, named, from c27eb32:
[1] A C

[[6]]
+ 2/7 vertices, named, from c27eb32:
[1] E F

[[7]]
+ 3/7 vertices, named, from c27eb32:
[1] A E F

[[8]]
+ 2/7 vertices, named, from c27eb32:
[1] A E

[[5]]
+ 2/7 vertices, named, from c27eb32:
[1] A B
```

Descriptive Statistics



Descriptive Analysis: Network cohesion

- Density of graphs: Ego-centric and Socio-centric. Degree is about the edges with each vertex in a graph. We can also consider the proportion of edges as a whole. The *density* of a graph is the proportion of possible edges that are actually present in the graph. It is the proportion of the number of edges present, L, to the maximum possible number of edges, denoted as Δ (for undirected graphs):

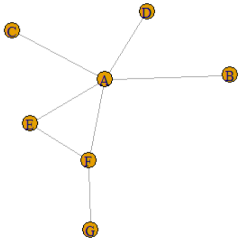
$$\Delta = \frac{L}{n(n-1)/2}$$

and $0 \leq \Delta \leq 1$

Descriptive Analysis: Network cohesion

- Clustering of graphs: One of the fundamental characteristics of social networks is the presence of clustering, or the tendency to formed closed triangles.
  - Local clustering coefficient: Proportion of neighbours of a node v, which are neighbours themselves. The local clustering coefficient shows how locally dense a graph is.
  - Global clustering coefficient
  - Expected clustering coefficient

Descriptive Analysis: Local Clustering Coefficient



Eg. A has 5 connections, (B, C, D, E, F) and we are looking for which one of those connections are also connected.

There are C_2^5 possibilities that can happen: (B-C, B-D, B-E, B-F, C-D, C-E, C-F, D-E, D-F, E-F). Out of these 10 connections only 1 of them exists. Therefore the local clustering coefficient for A is 1/10. For C it is zero, doesn't even have two friends.

```
transitivity(g, type = "local")
```

```
[1] 0.1000000       NaN       NaN       NaN 1.0000000 0.3333333       NaN
```

Descriptive Analysis: Local Clustering Coefficient

Therefore the average local clustering coefficient will be:

```
(0.100+1.000+0.333)/7
```

```
[1] 0.2747
```

Descriptive Analysis: Global Clustering Coefficient

- The process of closure can be measured in a social network by examining its transitivity. Transitivity is defined as the proportion of closed triangles (triads where all three ties are observed) to the total number of open and closed triangles (triads where either two or all three ties are observed).
- Thus, like density, transitivity is a ratio that can range from 0 to 1.

```
transitivity(g)
```

```
[1] 0.2142857 # moderate level of clustering
```

Network Models

So far we looked at the summary statistics such as density, centrality, transitivity etc.

With network models we can move beyond simple description to build and test hypotheses about network structures, formation processes, and network dynamics.

Several mathematical and statistical models have been proposed for this purpose. These provide insight into fundamental properties of social networks, serve as baseline or comparison models for empirical social networks, and act as building blocks for more complex network simulations.

Before we move on... What is a real network like?

In real life networks, given the number of edges and nodes, we observe

- a higher clustering coefficient than expected,
- a shorter average shortest path length than expected,
- mode nodes with high degrees than expected (heavy tails)

Then how can we approach to these with theoretical models?

Models of Network Structure and Formation

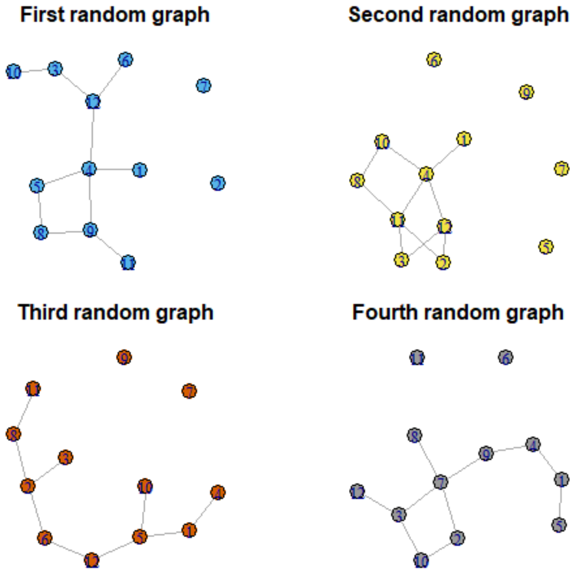
- Erdős Rényi Random Graph Model: earliest and still one of the most important mathematical models.
- Small-World Model: Also known as Watts and Strogatz model
- Scale-Free Models: Also known as Barabasi-Albert model

Erdős Rényi Random Graph Model

The model is quite simple, $G(V,E)$, where a random graph G is defined with n vertices and L edges among those vertices chosen randomly. An equivalent model that is easier to work with is, where instead of specifying the number of edges, L, we define the probability, p, each edge appears in the graph.

```
op <- par(mar=c(0,1,3,1),mfrow=c(2,2))
plot(erdos.renyi.game(n=12,10,type='gnm'),
     vertex.color=2,main="First random graph")
plot(erdos.renyi.game(n=12,10,type='gnm'),
     vertex.color=4,main="Second random graph")
plot(erdos.renyi.game(n=12,0.1515152,type='gnp'),
     vertex.color=6,main="Third random graph")
g = erdos.renyi.game(n=12,0.1515152,type='gnp')
plot(g,
     vertex.color=8,main="Fourth random graph")
```

Erdős Rényi Random Graph Model



Erdős Rényi Random Graph Model

Let us take a look at the 4th model

```
vcount(g) # we specified to have 12 nodes
```

```
[1] 12
```

```
ecount(g) # we specified to have 10 expected edges
    with a probability of .15 (C(10,2)*.1515152)
```

```
[1] 10
```

```
degree(g) # number of edges per node
mean(degree(g)) # average nodal degree
graph.density(g) # density of the graph
```

```
[1] 2 2 3 2 1 0 4 1 2 2 0 1
[1] 1.666667
[1] 0.1515152
```

Erdős Rényi Random Graph Model

Properties of Erdős Rényi Random Graph Model:

- Number of edges:

$$L = C_2^n x p$$

- Average degree size:

$$\bar{d} = (n-1) x p$$

- For larger networks, one can also explore the degree distribution, and in fact the degree distribution follows a Poisson distribution (only for large networks!)
- Random graphs become entirely connected for fairly low values of average degree.

Small World Model

Erdős Rényi random graph model has one major limitation in that it does not describe the properties of many real-world social networks.

In particular, fully random graphs have degree distributions that do not match observed networks very well, and they also have quite low levels of clustering (transitivity).

Small-World Model

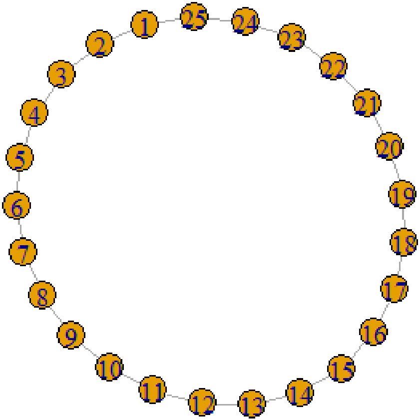
Steps of Small-World Models:

- The small-world model starts with a circle of nodes, where each node is connected to its c immediate neighbors (forming a formal lattice structure).

```
g=watts.strogatz.game(dim=1, size=25, nei=1, p
    =0)
# 25^1 nodes, connected with 1 neighbour
# c=1 (1 path between two nodes along the
    circle)
# degree(g) will give us 2, each node has 2
    connections
g
plot(g)
```

```
IGRAPH 350f814 U--- 25 25 -- Watts-Strogatz random graph
+ attr: name (g/c), dim (g/n), size (g/n), nei (g/n), p (g/n), loop
| (g/l)
+ edges from 350f814:
[1]  1--2   2--3   3--4   4--5   5--6   6--7   7--8   8--9   9--10
 10--11 11--12 12--13 13--14  14--15 15--16 16--17 17--18 18--19
  19--20 20--21 21--22 22--23  23--24 24--25  1--25
```

Small-World Model



Small-World Model

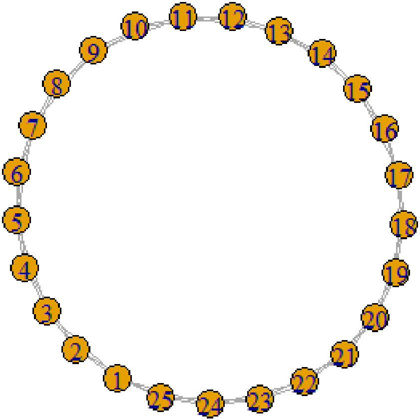
Steps of Small-World Models:

- Another example with $c = 2$.

```
g=watts.strogatz.game(dim=1, size=25, nei=2, p
    =0)
# 25^1 nodes, connected with 2 neighbours
# c=2 (2 paths between two nodes along the
    circle)
# degree(g) will give us 4, each node has 4
    connections
g
plot(g)
```

```
IGRAPH f7beb58 U--- 25 50 -- Watts-Strogatz random graph
+ edges from f7beb58:
[1]  1--2   2--3   3--4   4--5   5--6   6--7   7--8   8--9   9--10
10--11 11--12 12--13 13--14 14--15 15--16 16--17 17--18 18--19
19--20 20--21 21--22 22--23 23--24 24--25
1--25  1--3 1--24 2--25  2--4  3--5  4--6  5--7  6--8  7--9
8--10  9--11 10--12 11--13 12--14 13--15 14--16 15--17 16--18
17--19 18--20 19--21 20--22 21--23 22--24 23--25
```

Small-World Model



Small-World Model

Steps of Small-World Models:

- Then, a small number of existing edges are rewired, where they are removed and then replaced with another tie that connects two random nodes. Number of edges do not change but links do change.
  - If the rewiring probability is 0, then we end up with the original lattice network (previous examples). 0 number of edges get rewired.
  - When p is 1, then we have an Erdős Rényi random graph. All of the edges get rewired.

The main interesting discovery of Watts and Strogatz (and others), is that only a small fraction of ties needs to be rewired to dramatically reduce the diameter of the network.

Small-World Model

The main interesting discovery of Watts and Strogatz (and others), is that only a small fraction of ties needs to be rewired to dramatically reduce the diameter of the network.

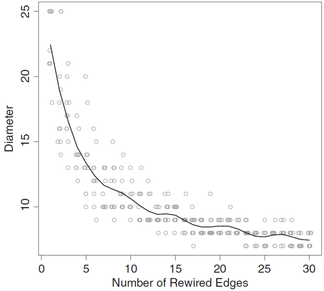


Fig. 10.6 Relationship of rewiring probability to network diameter for the small-world model

This plot is for 100 nodes. If we specify $c = 2$, we will have 200 edges. When we examine the relationship of rewiring probability (number of edges rewired/200) and the diameter of the network, we see that to reduce the diameter of the network, we just need to rewire 15 edges ($p = 15/200 = 0.075$).

Scale-Free Models

Previous two mathematical network models produce graphs with degree distributions that are not representative of many real-world social networks.

Numerous studies, in fact, have shown that a wide variety of observed networks have heavy-tailed degree distributions that approximately follow a power law distribution. These are typically called scale-free networks.

That is, as networks grow, new nodes are more likely to form ties with other nodes that already have many ties, due to their visibility in the network. This 'rich-gets-richer' phenomena has been shown to lead to the power-law distribution in networks.

One particular example of this type is World-Wide-Web.

Scale-Free Models

Barabási and Albert (1999) discovered that World-Wide-Web had degree distribution of:

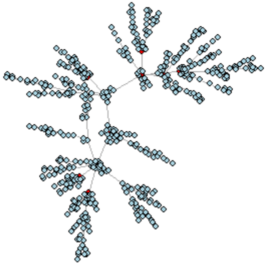
$$Prob(\text{a random vertex has degree k}) \sim Ck^{-\gamma}$$

for $k \to \infty$, C constant. This is called power-law behaviour and the γ is the power-law exponent.

The default behavior of the algorithm is that as each new node is added to the network, it is connected to another node in the network, with probability proportional to the degree of that node. Thus, some nodes in the network will end up with many more ties than most of the other nodes.

Scale-Free Models

```
g <- barabasi.game(500, directed = FALSE)
V(g)$color <- "lightblue"
V(g)[degree(g) > 9]$color <- "red"
plot(g, vertex.label = NA, vertex.size = 5)
```



```r
table(degree(g))
```

```
1     2   3   4   5   6   7   8   9  10  11  12  17  18  22
307  94  45  17  10   9   5   2   1   2   1   4   1   1   1
```

Others

- Stochastic block model (Erdős Rényi mixture model, latent block model)
- Configuration model (Bernoulli random graphs with fixed degree sequence)
- Chung-Lu model (sticky model)
- A duplication and divergence model (Bebek et al., 2006)

Comparing Random Models to Empirical Models

- By comparing the characteristics of the network models with the empirical network, we can highlight the interesting or important characteristics of the empirical network that might be worth further exploration by using, for example, the types of statistical modeling and simulation approaches.
- However, we need to remember that the observations are not independent!

Network Modeling and Inference

- Mathematical Models for Network Graphs
- Statistical Models for Network Graphs:
  - Exponential Random Graph Models - we will focus on this (which have turned out to be the most powerful, flexible, and widely used modeling approach for building and testing statistical models of networks.)
  - Network Block Models
  - Latent Network Models

ERGM

- ERGMs are generative statistical models of network structure and characteristics.
- Characteristics of the individual elements in the network (i.e., actors) and local structural properties can be used to predict properties of the entire network (e.g., diameter, degree distribution, etc.).
- ERGMs are flexible and can handle many different types of predictors and covariates, four broad categories:
  - node-level predictors,
  - dyadic predictors,
  - relational predictors, and
  - local structural predictors
- ERGMs are fit using Monte Carlo Markov Chain maximum-likelihood estimation (MCMC).
- ERGMs are implemented in the `ergm` package.

ERGM - Predictors

- The first type of predictor is node-level characteristics, where having a particular characteristic is hypothesized to affect the likelihood of observing a tie.
  - For example, if you hypothesize that girls are more likely to make friends than boys in middle school, then you could use actor gender as a node-level predictor.
- Dyad-level predictors are used when you hypothesize that the characteristics of both actors in a dyad may influence the probability of observing a tie between those two actors.
  - For example, if you think that friendships are more likely to be formed within the same grade in a middle school (assortative mixing), then you could use grade as a dyad-level predictor.
- The third type of predictor in ERGMs is a powerful option to use information about other relationships or ties when predicting the observed ties in a network.
  - That is, you can use one type of network tie to predict a second type of tie (as long as they are both collected on the same set of network members).
- Finally, information about local structural properties of the observed network can be used as model covariates.
  - This, for example, allows the network model to be conditioned on the observed degree distribution, or on the level of transitivity (closed triangles) that is observed.

Process of ERGMs

- Build ERGMs,
  - Build a null model - no covariates, only edges,
  - Include node attributes,
  - Include dyadic predictors,
  - Include relational terms (network predictors),
  - Include local structural predictors (dyad dependency).
- Examine the build model,
  - Model interpretation,
  - Model fit,
  - Model diagnostics,
  - Simulating networks based on fit model.

ERGM: 1. Build a null model

A null model typically only has one term, edges, which produces a random graph model that has the same number of edges as the observed network.

- Start the modeling process by building a null model, no substantive or structural predictors.
- Use as a baseline model to judge how much subsequent models are improving.

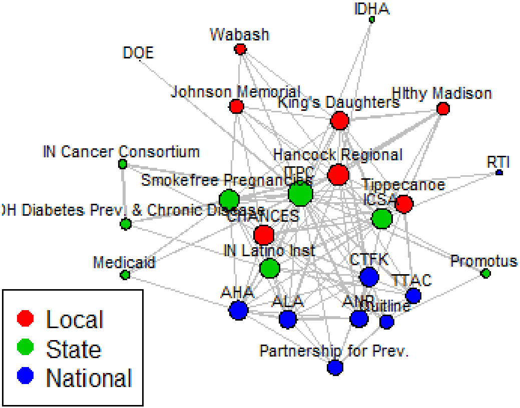
ERGM example: UserNetR data

Network data that describe interorganizational relationships among 25 agencies within the Indiana state tobacco control program in 2010. These data include three different types of interorganizational ties: frequency of contact, level of collaboration, and whether each pair of agencies communicated with one another about a particular evidence-based guideline published by the Center's for Disease Control and Prevention (CDC), called Best Practices for Tobacco Control. This latter relationship is conceptualized as a type of dissemination tie.

```
library(UserNetR)
detach("package:igraph") # because sna package and
    igraph clash!
library(sna)
data(TCnetworks)
TCcnt <- TCnetworks$TCcnt
TCcoll <- TCnetworks$TCcoll
TCdiss <- TCnetworks$TCdiss
TCdist <- TCnetworks$TCdist
gden(TCdiss)
```

```
[1] 0.3433333
```

UserNetR Network Plot



ERGM: 1. Build a null model

```
library(ergm)
DSmod0 <- ergm(TCdiss~edges, control=control.ergm(
    seed=40))
class(DSmod0)
summary(DSmod0)
```

```
==========================
Summary of model fit
==========================

Formula:   TCdiss ~ edges

Iterations:  4 out of 20

Monte Carlo MLE Results:
Estimate Std. Error MCMC % z value Pr(>|z|)
edges  -0.6485     0.1216      0  -5.333   <1e-04 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Null Deviance: 415.9  on 300  degrees of freedom
Residual Deviance: 385.9  on 299  degrees of freedom

AIC: 387.9    BIC: 391.6    (Smaller is better.)
```

This ensures that the simulated networks have the same number of edges as the observed network.

ERGM: 2. Include node attributes - Model hypothesis

Node attributes are as follows and here we can test the following hypothesis related to the node attributes:

- Lead agency: It might be reasonable to assume that agencies are more likely to be connected to the lead agency.
- Tobacco years: It also would make sense that agencies with a longer history of tobacco control experience would be more likely to be connected to other agencies.

ERGM: 2. Include node attributes - Model build

There are two main ergm model terms for testing node characteristic main effects:

- nodefactor: used for a categorical attribute (e.g., lead agency)
- nodecov: used for quantitative characteristics (such as tob yrs)

```
DSmod1 <- ergm(TCdiss~edges + nodefactor('lead_
    agency') + nodecov('tob_yrs'), control=control.
    ergm(seed=40))
summary(DSmod1)
```

```
Monte Carlo MLE Results:
Estimate Std. Error MCMC % z value Pr(>|z|)
edges                     -1.67835    0.32931      0  -5.097  < 1e-04 ***
nodefactor.lead_agency.1  17.93665  933.55514      0   0.019  0.98467
nodecov.tob_yrs            0.05994    0.02283      0   2.625  0.00867 **
---
Null Deviance: 415.9  on 300  degrees of freedom
Residual Deviance: 323.5  on 297  degrees of freedom

AIC: 329.5    BIC: 340.6    (Smaller is better.)
```

ERGM: 2. Include node attributes - Model interpretation

- The number of years in tobacco control is positively and significantly associated with the likelihood of observing a tie between the two agencies.
- The effect of being the lead agency is is not significant. With only one lead agency in a small network of 25 members, there may not be the power to detect this effect.
- The AIC (Akaike information criterion) for this model with two predictors is lower than the AIC for the null model. This shows that this model is doing a better job of explaining the data than the baseline model.

ERGM: 3. Include dyadic predictors -

A rich source of hypotheses for network structures derive from questions about

- homophily (are ties more or less likely between network members who are similar to each other on some characteristic)
- heterophily (or dissimilar)

This is a type of dyadic interaction predictor, and ergm includes a number of these terms.

ERGM: 3. Include dyadic predictors -

The raw frequencies of observed ties among and between different types of actors in the network can be displayed with the `mixingmatrix()` function.

```
TCnetworks$TCdiss
mixingmatrix(TCdiss,'agency_lvl')

```

```
Note:  Marginal totals can be misleading
for undirected mixing matrices.
   1  2  3
1 13 24 14
2 24 16 23
3 14 23 13
```

Here, for example, we see that the most frequent dissemination ties (24) are observed between local and state-level agencies. It can be somewhat complicated to interpret these raw frequency patterns; however, they can generate hypotheses about dyadic interrelationships that can be formally tested in the ERGM.

ERGM: 3. Include dyadic predictors -

In the following models, the non-significant lead agency predictor is dropped. Three version of Model 2 are estimated, showing different options for including the dyadic comparison of agency level as a predictor.

```
DSmod2a <- ergm(TCdiss~edges + nodecov('tob_yrs')
  + nodematch('agency_lvl'), control=control.
  ergm(seed=40))

DSmod2b <- ergm(TCdiss~edges + nodecov('tob_yrs')
  + nodematch('agency_lvl',diff=TRUE), control=
  control.ergm(seed=40))

DSmod2c <- ergm(TCdiss~edges + nodecov('tob_yrs')
  + nodemix('agency_lvl',base=1), control=
  control.ergm(seed=40))
```

ERGM: 3. Model2a

```
summary(DSmod2a)
```

```
Monte Carlo MLE Results:
Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -2.48077    0.34131      0  -7.268   <1e-04 ***
nodecov.tob_yrs       0.11331    0.02006      0   5.648   <1e-04 ***
nodematch.agency_lvl  0.68751    0.27698      0   2.482   0.0131 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Null Deviance: 415.9  on 300  degrees of freedom
Residual Deviance: 342.2  on 297  degrees of freedom

AIC: 348.2    BIC: 359.4    (Smaller is better.)
```

Model 2a uses the basic nodematch term to include one network predictor that assesses the effect on the likelihood of a dissemination tie when both organizations are the same level (e.g., both are local organizations). This is a homophily hypothesis, where we are testing if the same types of organizations are more likely to communicate with one another. The positive and significant parameter indicates that there is a homophily effect here.

ERGM: 3. Model2b

```
summary(DSmod2b)
```

```
Monte Carlo MLE Results:
Estimate Std. Error MCMC % z value Pr(>|z|)
edges                  -2.77924    0.36851      0  -7.542  < 1e-04 ***
nodecov.tob_yrs         0.13314    0.02174      0   6.123  < 1e-04 ***
nodematch.agency_lvl.1  1.61450    0.49835      0   3.240  0.00120 **
nodematch.agency_lvl.2 -0.21484    0.39735      0  -0.541  0.58873
nodematch.agency_lvl.3  1.30157    0.44220      0   2.943  0.00325 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Null Deviance: 415.9  on 300  degrees of freedom
Residual Deviance: 329.9  on 295  degrees of freedom

AIC: 339.9    BIC: 358.4    (Smaller is better.)
```

Model 2b shows how to test a hypothesis of differential homophily. Here, instead of one dyad term, there are three; one each for the three levels of the agency level characteristic. Thus, this model now has three homophily terms, one for local agencies, one for state, and one for national. The results suggest that the overall homophily effect is seen mainly at the local and national levels.

ERGM: 3. Model2c

```
summary(DSmod2c)
```

```
Monte Carlo MLE Results:
Estimate Std. Error MCMC % z value Pr(>|z|)
edges                -1.17567    0.53718      0  -2.189  0.02862 *
nodecov.tob_yrs       0.13405    0.02225      0   6.026  < 1e-04 ***
mix.agency_lvl.1.2 -1.58049    0.55005      0  -2.873  0.00406 **
mix.agency_lvl.2.2 -1.83543    0.60280      0  -3.045  0.00233 **
mix.agency_lvl.1.3 -1.53626    0.56587      0  -2.715  0.00663 **
mix.agency_lvl.2.3 -1.70730    0.54454      0  -3.135  0.00172 **
mix.agency_lvl.3.3 -0.31105    0.61186      0  -0.508  0.61119
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Null Deviance: 415.9  on 300  degrees of freedom
Residual Deviance: 329.7  on 293  degrees of freedom

AIC: 343.7    BIC: 369.6    (Smaller is better.)
```

Finally, Model 2c shows how to include the most detailed tests for homophily and heterophily. The nodemix term includes dyadic comparisons for all the possible patterns of a categorical node attribute. The base option sets a reference category for the effects, otherwise all possible effects are included (which may lead to model stability problems). Here the reference category is (1,1), indicating the ties among the local-level agencies. Note that for smaller networks and categorical attributes with a large number of values, the mixing matrix may have empty cells (as well as use up many more degrees of freedom). These can cause problems for the model estimation and interpretation.

ERGM: 4.Including Relational Terms (Network Predictors)

The third type of predictor that can be included in ERGMs is a relational predictor, where information about ties among the network members is used to predict the likelihood of the dependent variable tie.

Why R?

Software: Use R! Why R?

- vast integrated system of thousands of packages and functions
- R is free and open nature
- There is a vast R user and developer community

Packages: Several different packages:

- `igraph`: Network Analysis and Visualization to create, decorate, assess basics of a network.
- `sand`: Statistical Analysis of Network Data with R. Data sets for the book 'Statistical Analysis of Network Data with R'.
- `sna`: Tools for Social Network Analysis.
- `statnet`: Software Tools for the Statistical Analysis of Network Data.
- `UserNetR`: R Package provides datasets for Douglas A. Luke's "A User's Guide to Network Analysis in R" book (2015).

Other Programmes: UCINet, Structure, Pajek, Negopy, SocNetV, Gephi, StOCNET, NetMiner and many more...

Further Topics

- Hierarchical structure in networks
- Dynamic structure of networks
- Multilayer networks

- **Textbooks:**
  * Scott, J. (2007). Social Network Analysis: A Handbook (2nd Ed.). Newbury Park, CA: Sage.
  * Fu, X., Luo, J-D., Boos, M. (2017). Social Network Analysis - Interdisciplinary Approaches and Case Studies. CRC Press.
  * Aggarwal, C. (2011). Social Network Data Analytics. Springer.
  * Kolaczyk, E.D., Csardi, G. (2014). Statistical Analysis of Network Data with R. Springer.
  * Luke, D.A. (2015). A User's Guide to Network Analysis in R. DOI 10.1007/978-3-319-23883-8 1
  * Borgatti, S.P., Everett, M.G., Johnson, J.C. (2013). Analyzing Social Networks. Sage Publications. * Knoke (2008). Social Network Analysis. (2nd Ed.). Sage.
  * Huisman, M., van Duijn, M. A. J. (2005). Software for social network analysis. Models and methods in social network analysis. Carrington, P. J., Scott, J., Wasserman, S. (eds.). New York: Cambridge University Press, p. 270 - 316.
- **Links:**
  * `https://toreopsahl.com/`
  * `https://youtu.be/QT2xj9k00q0`
- **Articles:**
  * Durbach, I., Katshunga, D., Parker, H. (2013). Community structure and centrality effects in the South African company network. S.Afr.J.Bus.Manage.2013,44(2).
  * Durbach, I., Parker, H. (2009). An analysis of corporate board networks in South Africa. S.Afr.J.Bus.Manage.2009,40(2).
  * Giuliani, E., Morrison, A., Pietrobelli, C., Rabellotti, R. (2010). Who are the researchers that are collaborating with industry? An analysis of the wine sectors in Chile, South Africa and Italy. Research Policy Vol. 39, pp.748761.