# Further Topics in Social Network Analysis
# Small-world phenomenon

Dominik Batorski Michał Bojanowski Bartosz Chroł Kamil Filipek

Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw

WARSAW
SCHOOL of
DATA
ANALYSIS

## Contents

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

ICS
RUG / UU / RU

UNIWERSYTET
WARSZAWSKI

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Project co-financed by the European Union under the European Social Fund

# 1 Introduction - global properties of networks

In recent years, an increasing part of the study of social networks are conducted on very large networks, i.e. networks consisting of thousands or millions of nodes. This is primarily due to the significant increase in opportunities to collect data on large networks, resulting from the development of digital technologies, as well as telecommunications networks and other technologies to collect information about the behavior of people and relationships and communication between them.

Many especially large networks has some special properties. Those global properties of network structure are important because of the impact on the dynamics of processes that are taking place within the network.

Firstly, it is worth noting that such large networks are generally very sparse, i.e., the density of the network is low. Most of the nodes is not adjacent to each other. On the other hand, some of the large networks are characterized by high local density. Two people who know each other usually have many mutual friends. In other words, the probability that two people who have a common friend also know each other is very large. In contrast, the two individuals randomly selected from a large population will very rarely know each other. Such networks, even though they have little global density, have high local clustering.

The average node is connected only with a small fraction of other nodes as compared to how many nodes are in the network. However surprisingly, the average distance between the nodes turns out to be very small. The first illustration of occurrence of short paths were experiments conducted in the 1960s by Stanley Milgram (Milgram 1967; Travers and Milgram 1969; Korte and Milgram 1970). These studies showed that the average distances between different persons in the United States through a chain of acquaintances are very short.

An ingenious study conducted by Milgram was based on letters that were passed through a chain of acquaintances. People chosen for the study received a letter with a name, surname and place of residence of the addressee, who lived on the other side of the US. They were asked to pass a letter to one of their acquaintances, who had to pass it to the next person, so that the letter reached the addressee in the fewest number of steps. At the same time, each person who received and forwarded the letter were invited to send information to researchers about when and from whom she received a letter and to whom and when it was passed. As a result, it was possible to track how messages were passed. To the surprise of many, the letters sent in this way reached its destination very quickly - in 6 steps on average. The result of Milgram's experiments was surprising, but probably every one of us experienced that "the world is small", when we discovered to have mutual friends with people previously unknown and in no way related to us.

Properties of the various networks structure have been studied by mathematicians for a long time. One of the simplest network models are random graphs, i.e. networks in which the probability of a direct link between each pair of nodes is the same. Such graphs have many interesting characteristics. As was shown by Solomonoff and Rapoport (1951) and independently by (Erdos and Renyi 1959), in random graphs in which there are more links than nodes, most of the nodes are connected (indirectly) to form one large component. A connected component is a maximal subgraph in which any two nodes are connected to each other by paths. Moreover, the size of the largest component, is highly dependent on the number of connections held by the nodes.

Another property of random graphs is that the distances between the nodes belonging to the largest component are small (Erdos and Renyi 1959). This is true even if the graph is large, and the number of connections possessed by each node is much lower than the total number of nodes in the graph. It turns out, that the average length of paths connecting nodes in a random network is $log(N)/log(k)$, where $N$ is the number nodes, and $k$ the average number of links owned by one node. This means that the paths in the random graphs are very short.

However, social networks are very far from being random. Large local clustering is a good proof of that fact. Circles of friends of two people who know each other overlap to a very large extent and random graphs do not have this property. In other words there is large transitivity in social networks.

A model of network with both a high local clustering and short paths between the nodes has been proposed by Watts and Strogatz (1998). They created a class of networks, which are a combination of regular and random graphs. The network is constructed from highly structured graph, for example $N$ nodes distributed around the circle, each with $k$ connection with the nearest nodes. Then, with probability $p$ the other end of each connection is assigning to another randomly selected node. Watts-Strogatz model thus consists of close connections with high local density for small $p$ and with a fraction $p$ of random connections. The $p$ takes values from the interval $[0,1]$ and determines the degree of disorder of the network - from a regular grid, when $p = 0$ to a completely random system for $p = 1$.

Watts and Strogatz had shown that even for small values of $p$, when a network has high local clustering, the average distance between nodes is short. Such graphs are called a small world networks.

A graph is considered small-world, if its average local clustering coefficient $\bar{C}$ is significantly higher than a random graph constructed on the same set of nodes, and if the graph has approximately the same mean-shortest path length as its corresponding random graph. In the next section, we show an example how to verify in *R* if a given graph is a small-world network.

# 2 Example - global properties of coauthoship network

We use `coauthorship` network from `isnar` R package.

```r
library("igraph")
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```r
library("isnar")
# Load data on co-authorship
data(coauthorship)
# number of nodes
N <- vcount(coauthorship)
# average number of edges per node
k <- ecount(coauthorship)/N
```

This network has $N = 10114$ nodes and the average number of relations per node $k$ is 4.8580186.

## 2.1 Connected components of a graph

In the first step, it is worth checking whether the network is connected, and if not, how many components there are, and what is the size of the largest one.

```r
is.connected(coauthorship, mode="strong")
```

returns value `FALSE` therefore we should divide a network into strongly connected components and compute the number of components.

```r
coaut.Clusters <- clusters(coauthorship, mode="strong")
coaut.Clusters$no # the number of clusters
```

The network has 600 components. The size of the largest one is:

```r
max(coaut.Clusters$csize) #the size of the largest component
```

There are 6796 nodes in the largest component, this is 67.2% of the graph.

Then we choose a subgraph of the entire network, whith only the nodes from the largest component. It can be done with `subgraph` function from the `igraph` package.

```r
coaut.LargestComponent <- subgraph(coauthorship, coaut.Clusters$membership==1)
```

### 2.2 Local density

We use `transitivity` function with `type="local"` to compute the average of the local clustering coefficients.

```r
mean(transitivity(coaut.LargestComponent, type="local"), na.rm=T)
```

Now we can compare the obtained value 0.8452537 with the expected value of the average local clustering coefficient $\bar{C}$ for a random graph constructed on the same set of nodes. The empirically observed value is significantly higher than $k/N$ expected for a corresponding random graph, that is $4.80326 \times 10^{-4}$ in this network.

### 2.3 Shortest paths between nodes

In order to compute the average distance between nodes in the largest component one can use `average.path.length` function from `igraph` package.

```r
average.path.length(coaut.LargestComponent)
```

The average short paths for the largest conected component of coauthorship network is around 10.725493. Whereas the average geodesics for a corresponding random graph equals $log(N)/log(k)$ that is 5.834175 in our example. This can be interpreted that, despite the high local density, the network is not a small world because average path lenght between nodes is not short enough.

We can also compute the diameter $d$ of a network, that is the greatest distance between any pair of nodes.

```r
max(shortest.paths(coaut.LargestComponent))
```

Here $d$ equals 43.

## 3 Degree distribution

Duncan Watts argued that the phenomenon of small worlds is in the decentralized nature of the network. A network can be small world even in an absence of nodes that would connect with a significant number of other nodes. However Watts and Strogatz didn't analyze the number of relationships held by individual nodes. In their model, initially each node had the same number of relationships. Then as a result of small random changes some nodes lost and some gained new connections. However, not only the average number of relationships is an important characteristics of the network. The distribution of the number of relations owned by the nodes is no less important. This property was systematically analyzed by Albert-László Barabási and his collaborators (Barabasi and Albert 1999; Barabasi, Albert, and Jeong 1999).

Let's define $P(k)$ as the frequency of the nodes that have $k$ relationship, and thus the probability that a randomly selected node will have exactly $k$ relationship. For random graphs the distribution of the number of relationships have a Poisson distribution. But as it turns out, the real networks generally have a different, strongly right-skewed distribution of the size of the nodes. Thus this is yet another property that distinguishes a real-world networks from random graphs. Networks having this property are called scale-free networks.

A scale-free network is a network whose degree distribution follows a power law, at least asymptotically. That is, the fraction $P(k)$ of nodes in the network having $k$ connections to other nodes goes for large values of $k$ as

$$P(k) \sim k^{-\gamma}$$

where $\gamma$ is a parameter whose value is typically in the range $2 < \gamma < 3$,

Some networks have exponential rather than power law distribution. For this kind of networks the number of nodes with a given number of connections falls more rapidly. Exponential distribution can be described by the formula:

$$P(k) \sim e^{-k/\kappa}$$

In both cases, there is a large number of nodes that have low number of relationship, while only some nodes have a lot of connections.

## 3.1 Degree distribution in R

Degree for each node can be computed using `degree` function.

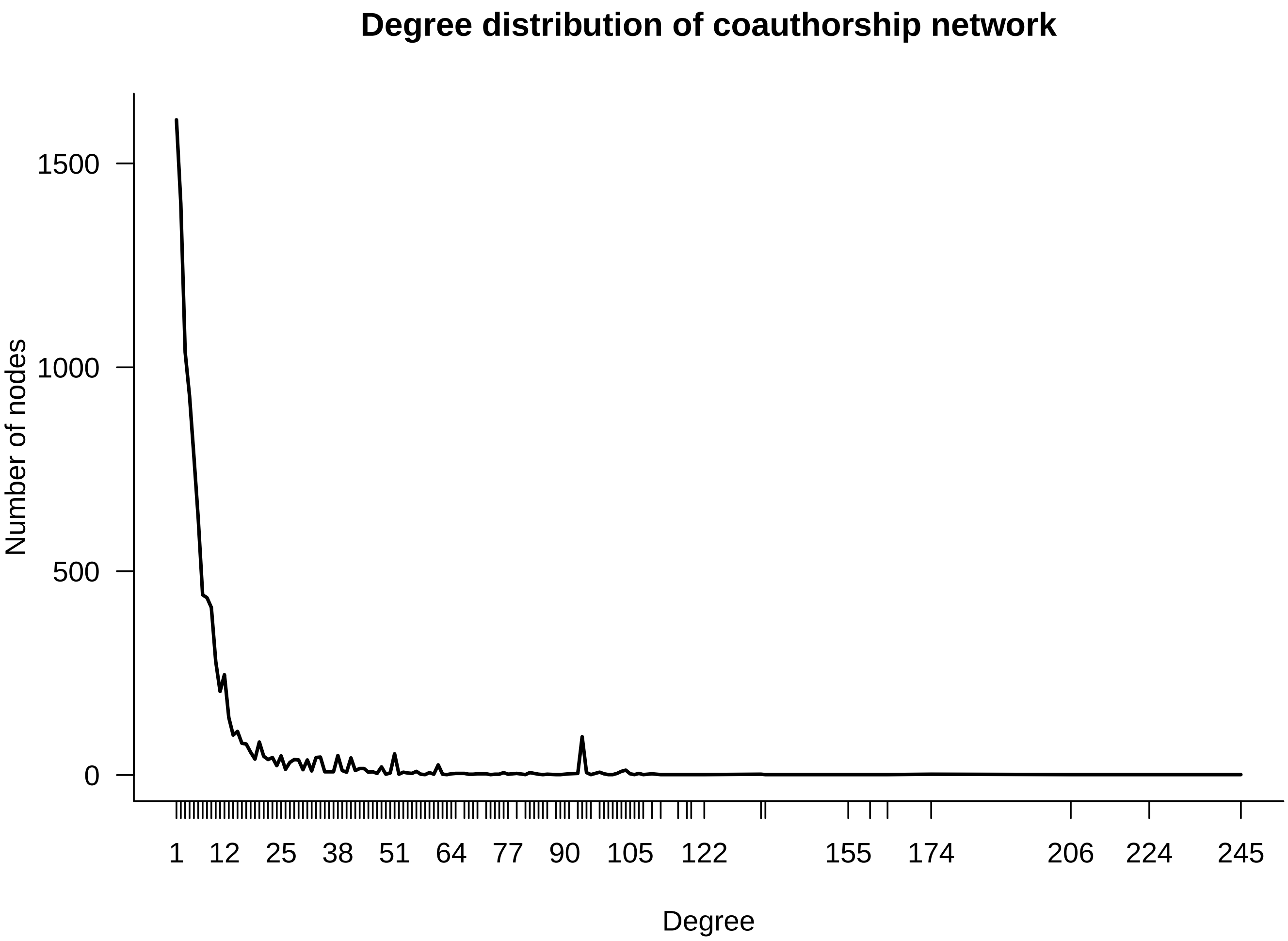
```
degree(coauthorship)
```

Then the degree distribution can be computed:

```
degreeDist <- table(degree(coauthorship))
```

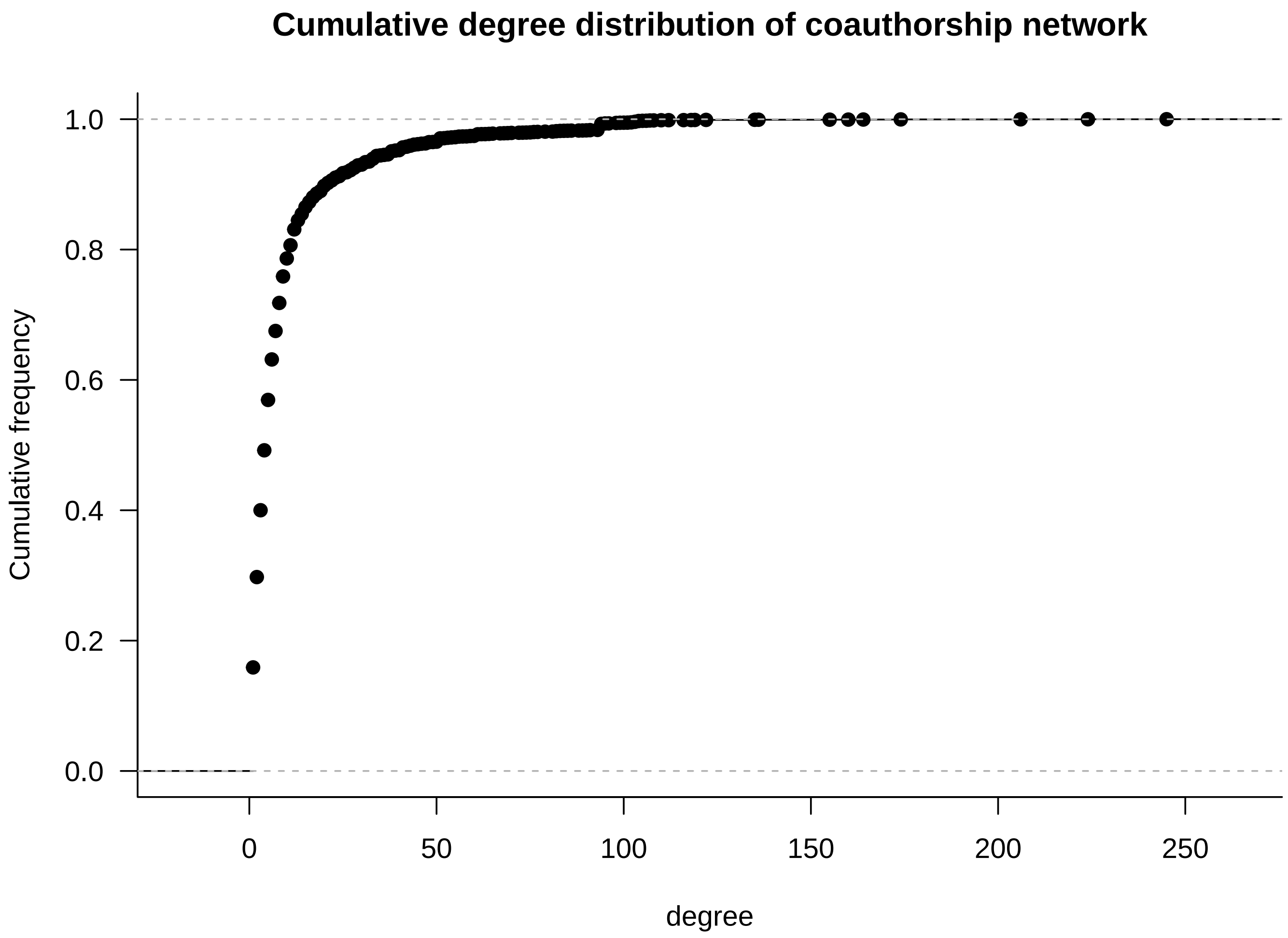
As one can see on the plot below, the degree distribution of coauthorship network is highly right-skewed.

```
plot(degreeDist, type='l',bty='l', las=1, main='Degree distribution of coauthorship network',
     xlab='Degree',ylab='Number of nodes')
```



In the analyzes of the degree distribution cumulative distribution are often used, as they are much more comprehensible.

```
plot(ecdf(degree(coauthorship)), main='Cumulative degree distribution of coauthorship network',
     bty='l', las=1, xlab='degree',ylab='Cumulative frequency')
```



## 4 Simulations - generating a network with a given properties

Sometimes one needs the ability to generate network with the given properties. The igraph package allows you to create a network according to various theoretical models. In this section we show three basic ones, which were discussed earlier in this chapter.

### 4.1 Random graphs

Random graphs are generated with erdos.renyi.game function

```
erdos.renyi.game(n, p.or.m, type=c("gnp", "gnm"),directed = FALSE, loops = FALSE, ...)
```

where:

n is the number of nodes in the network.
p is the probability for drawing an edge between two arbitrary vertices (G(n,p) graph)
m is the number of edges in the graph (for G(n,m) graphs).
The type of the random graph can be either gnp (G(n,p) graph) or gnm (G(n,m) graph) depending on which value $p$ or $m$ was given. One can also specify if the graph will be directed and whether to add loops or no (both parameters with defaults set to FALSE).

## 4.2 Small-world networks

Generate a graph according to the Watts-Strogatz network model with `watts.strogatz.game` function.

```
watts.strogatz.game(dim, size, nei, p, loops = FALSE, multiple = FALSE)
```

where:

`dim` is the number of dimensions of the starting lattice.
`size` is the size of the lattice along each dimension.
`nei` the neighborhood within which the vertices of the lattice will be connected.
`p` is the rewiring probability (value between zero and one).
One can also decide if `loops` and `multiple` edges are allowed in the generated graph.

## 4.3 Scale-free network

In order to generate a scale-free network use `barabasi.game` function:

```
barabasi.game(n, power = 1, m = NULL, out.dist = NULL, out.seq = NULL,
    out.pref = FALSE, zero.appeal = 1, directed = TRUE,
    algorithm = c("psumtree", "psumtree-multiple", "bag"),
    start.graph = NULL)
```

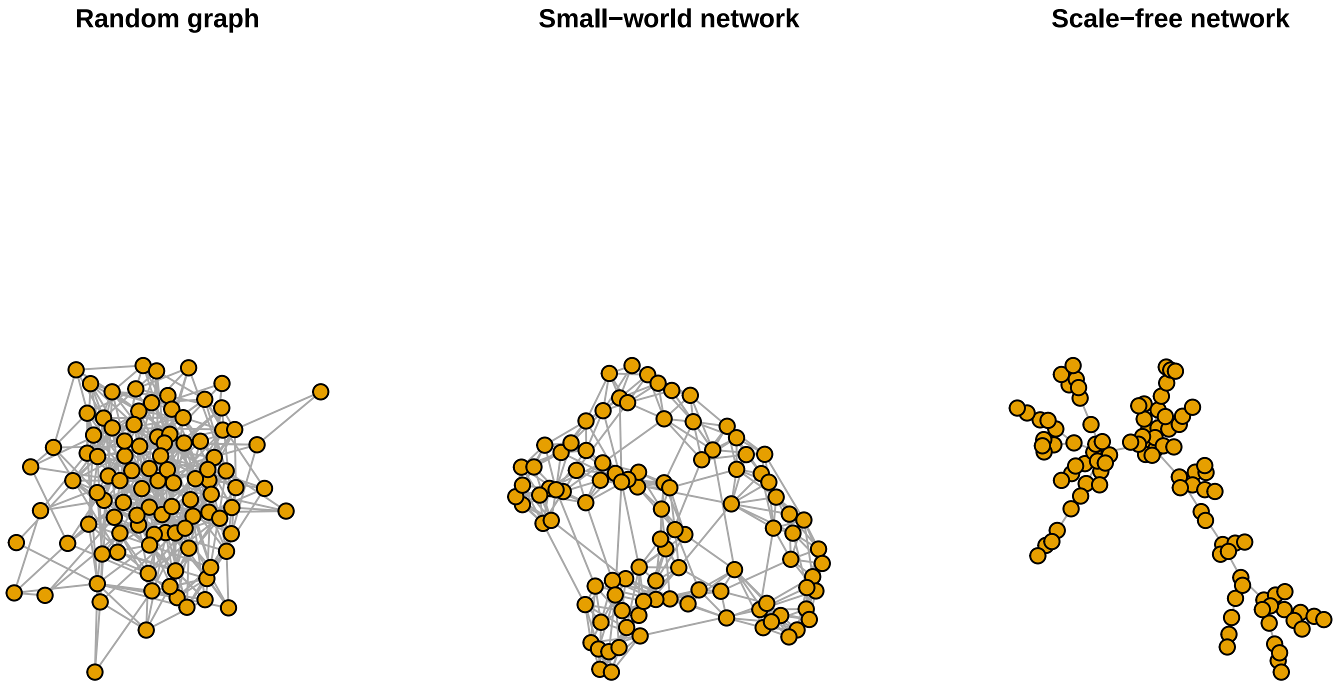
where:

`n` is a number of vertices.
`power` is the power of the preferential attachment, the default is one, ie. linear preferential attachment.
`directed` whether to create a directed graph.
other arguments explanation can be found using `?barabasi.game` call.

## 4.4 Example

Now generate a sample networks, visualize them and see how they differ.

```
n=100
SFNet <- barabasi.game(n, power=1, directed=F)
SWNet <- watts.strogatz.game(dim=1, size=n, nei=4, p=.03)
RNet <- erdos.renyi.game(n, p.or.m=400, type="gnm")
par(mfrow=c(1,3))
plot(RNet,vertex.label=NA,vertex.size=10,edge.arrow.size=.4, main='Random graph')
plot(SWNet,vertex.label=NA,vertex.size=10,edge.arrow.size=.4, main='Small-world network')
plot(SFNet,vertex.label=NA,vertex.size=10,edge.arrow.size=.4, main='Scale-free network')
```



## 5 Implications

These global properties of networks are important because they have impact on processes that take place within the network structure and in relations between the nodes.

## References

Barabasi, A.L., R. Albert, and H. Jeong. 1999. "Mean-Field Theory for Scale-Free Random Networks." *Physica A*, no. 272: 173–87.

Barabasi, A.L., and R. Albert. 1999. "Emergence of Scaling in Random Networks." *Science*, no. 286: 509–12.

Erdos, P., and A. Renyi. 1959. "On Random Graphs." *Publicationes Mathematicae* 6: 290–97.

Korte, C., and S. Milgram. 1970. "Acquaintance Networks Between Racial Groups: Application of the Small World Method." *Journal of Personality and Social Psychology* 15 (2): 101–18.

Milgram, S. 1967. "The Small World Problem." *Psychology Today* 1: 60–67.

Solomonoff, R., and A. Rapoport. 1951. "Connectivity of Random Nets." *Bulletin of Mathematical Biophysics* 13: 107–17.

Travers, J., and S. Milgram. 1969. "An Experimental Study of the Small World Problem." *Sociometry* 32 (4): 425–43.

Watts, D.J., and S. Strogatz. 1998. "Collective Dynamics of Small-World Networks." *Nature* 393: 440–42.