Introduction to network models

(Chapter 3 & 4 of the manual)

Alberto Caimo

Dublin Institute of Technology
Ireland

SSNAR 2017
JUNE 21-23, 2017
Birkbeck, University of London

Networks as random graphs

Networks as random graphs

Definitions and notation

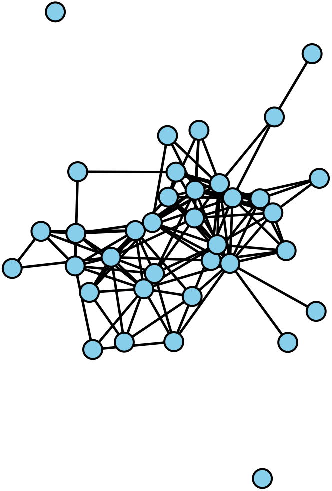
- Networks are generally represented by graphs of nodes (actors) and edges (relations)
- N number of nodes (fixed).
- D number of dyads (pair of nodes) in a N-node network (fixed).
- Y random $N \times N$ adjacency matrix where:
  - $Y_{ij} = 0$, if i and j are not connected;
  - $Y_{ij} = 1$, if i and j are connected;
  - $Y_{ii} = 0$, (self-loops are not allowed).
- y realisation of Y (observed adjacency matrix);
- $E = s_1(y)$ number of edges in the network (= number of 1's in y).

Networks as random graphs

Example - Lazega network:
corporate law partnership in a Northeastern US corporate law firm

```
library(statnet)
load(url("https://acaimo.github.io/lazega.RData"))
y <- network(Y, directed = FALSE)
set.seed(11)
plot(y,
     vertex.col = "skyblue",
     vertex.cex = 2)
```

Networks as random graphs



Basic info: undirected network, 36 nodes (partners and associates of the firm), 115 edges (co-work relations).

Networks as random graphs

Let's calculate some other descriptives:

- We know that $N = 36$;
- What about the number of dyads D?
  In an undirected graph, we have:

$$D = \frac{N^2 - N}{2} = \frac{36^2 - 36}{2} = 630;$$

- In the Lazega network there are 115 edges, so $s_1(y) = 115$:

```
s1 <- summary(y ~ edges); s1
```

- Let's calculate the density of the network, i.e., the proportion of connected dyads:

$$density(y) = \frac{s_1(y)}{D} = \frac{115}{630} \approx 0.1825 \approx 18.25\%.$$

The random graph model

Definition of the model

$$\Pr(Y = y) = \eta^{s_1(y)}(1-\eta)^{D-s_1(y)}.$$

- Describe the probability of observing y as a function of the parameter η;
- The parameter η represents the probability of observing an edge between any dyad;
- To estimate the model we just need to estimate η:

$$\hat{\eta} = \frac{s_1(y)}{D}.$$

- The parameter η corresponds to the **density** of the network.

The random graph model

Definition of the model – natural parametrisation

$$\Pr(Y = y) = \frac{\exp\{\theta s_1(y)\}}{c(\theta)}.$$

- The parameter θ is defined as the **logit** of η:

$$\theta = \log\left(\frac{\eta}{1-\eta}\right);$$

- $c(\theta)$ is a normalising constant;
- The random graph model belongs to the **exponential family** of models.

The random graph model

Parameter interpretation:

- $\theta = 0 \Rightarrow \eta = 0.5$ (50% of the dyads are not connected):

$$\Pr(Y_{ij} = 1 | \theta = 0) = \frac{\exp\{0\}}{1 + \exp\{0\}} = \frac{1}{1+1} = \eta = 0.5.$$

- $\theta < 0 \Rightarrow \eta < 0.5$ (most of the dyads are not connected);
- $\theta > 0 \Rightarrow \eta > 0.5$ (most of the dyads are connected).

Estimation of θ using **statnet**:

```
RG.model <- y ~ edges
theta <- ergm(RG.model)$coef
theta
```

Network simulation

To simulate from the estimated random graph model:

- We can simply simulate D Bernoulli trials by assuming $Y_{ij} \sim Bernoulli(\eta)$;
- Then we arrange them into a $N \times N$ matrix:

```
y.sim.1 <- simulate(RG.model,
                    coef = theta)
plot(y.sim.1,
     vertex.cex = 2,
     vertex.col = 'skyblue',
     main = 'y.sim.1')
```

Network simulation

Suppose that:

- We simulate 50 networks $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{50}\}$ from the estimated model;
- $s_1(\tilde{y})$ is the vector containing the number of edges measured in each of the simulated networks;
- We calculate the average number of edges measured in the simulated networks as follows:

$$E(s_1(\tilde{y})) = \frac{1}{M} \sum_{i=1}^{M} s_1(\tilde{y}_i);$$

- We expect that the average number of edges measured in the simulated networks is close to the observed number of edges in the observed network $(s_1(y))$:

$$E(s_1(\tilde{y})) \approx s_1(y).$$

Network simulation

```
y.sim.1_50 <- simulate(RG.model,
                       coef = theta,
                       statsonly = TRUE, # returns only the
                                         # network statistics
                                         # in the model
                                         # (number of edges)
                       nsim = 50) # number of network simulated

mean(y.sim.1_50) # ~115
```

Exponential random graph models (ERGMs)

Basic assumptions

- The observed network y is generated by a stochastic process in which edges are created because of the presence or absence of other edges (and possibly node-level attributes).
- Local effects (represented by **network statistics** $s(y)$) that generate dyadic relations and these processes may depend on the surrounding social environment.

For example:

- We can assume that actors with similar attributes are more likely to form friendship edges (**homophily**);
- If two unconnected actors were connected to a third actor, at some point they are likely to form a friendship link between them (**transitivity**).

Exponential random graph models

Definition of the model

Exponential family representing the probability distribution of y given a vector of parameters θ:

$$\Pr(Y = y|\theta) = \frac{\exp\{\theta^T s(y)\}}{c(\theta)}.$$

- Describe the probability of observing y as a function of the parameter θ;
- $s(y)$ is a vector of network statistics (e.g. number of edges, number of triangles, etc.) associated to effects of interest;
- θ is the vector of parameters associated to the network statistics $s(y)$;
- $c(\theta)$ is a normalising constant which **cannot** be computed for not trivially small networks.

Dependence assumptions and network statistics

- Dyadic dependence (as in the random graph model) is an unrealistic assumption in many circumstances;
- Network statistics involving more than a dyad imply **dependence** between dyads: an edge between node i and j is assumed to be dependent on the presence of other edges.

For example:

- **Stars** statistics assume that an edge between i and j is contingent on any possible edge involving node i and j (i.e. on the degrees of i and j).
- **Triadic** statistics assume that an edge between i and j is contingent on any possible edge involving any node of the network connected to both i and j.

Parameter interpretation

- The parameter θ associated with the network effects expressed by the network statistics $s(y)$ provide insights about the contribution of each network statistic to edge formation.
- ERGMs allow to establish a relationship between presence/absence of an edge and a set of network statistics.

Parameter interpretation

For example, suppose $s(y)$ includes the number of edges ($s_1(y)$) and the number of 2-stars ($s_2(y)$).

- $\theta_1 < 0|\theta_2 \Rightarrow$ sparse network;
- $\theta_1 > 0|\theta_2 \Rightarrow$ dense network;
- $\theta_2 > 0|\theta_1 \Rightarrow$ edges tend to connect nodes with high degree (i.e. presence of high-degree nodes);
- $\theta_2 < 0|\theta_1 \Rightarrow$ absence of high-degree nodes;