Introduction to multi-edge network inference in R using the ghypernet-package

Laurence Brandenberger, Giona Casiraghi

EUSN 2019

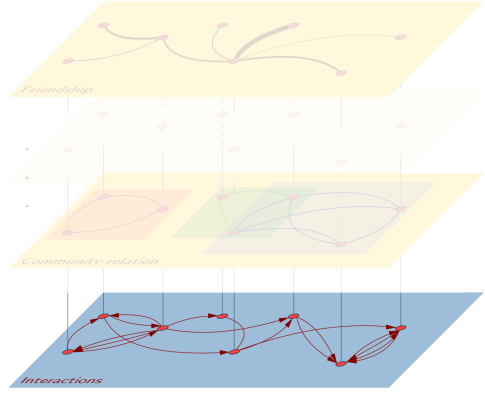
Outline of this tutorial

1. gHypEG
2. Change Statistics

Table of Contents

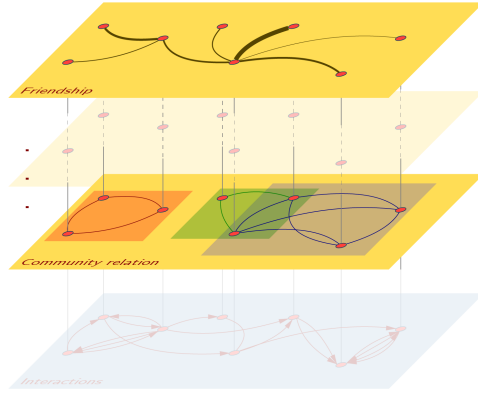
What kind of data are we dealing with?

- Often, network data consists of **repeated interactions**
  - face-to-face networks
  - contact networks
  - ecological and biological interactions
  - ...
- $\Rightarrow$ Optimally represented as **multi-edge networks**



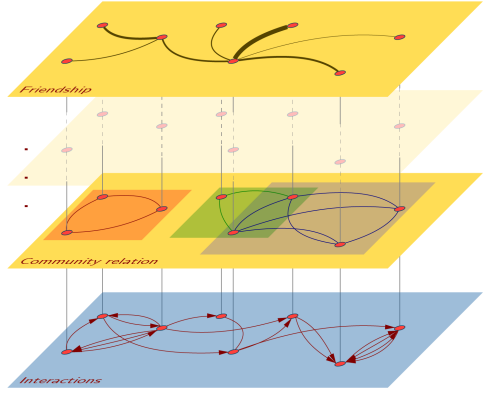
What kind of data are we dealing with?

- Often, network data consists of **repeated interactions**
  - face-to-face networks
  - contact networks
  - ecological and biological interactions
  - ...
- $\Rightarrow$ Optimally represented as **multi-edge networks**
- **Additional data** available, e.g.:
  - distance between nodes
  - friendship links between individuals
  - clusters and groupings of nodes



What kind of data are we dealing with?

- Often, network data consists of **repeated interactions**
  - face-to-face networks
  - contact networks
  - ecological and biological interactions
  - ...
- $\Rightarrow$ Optimally represented as **multi-edge networks**
- **Additional data** available, e.g.:
  - distance between nodes
  - friendship links between individuals
  - clusters and groupings of nodes
- $\Rightarrow$ Information combined in a **multiplex network**



Network Models 101: Inference

- To perform **statistical inference** with network models we usually start from a model

$$\Pr(N|\boldsymbol{\theta}) = \mathcal{F}\left[\boldsymbol{h}(N), \boldsymbol{\theta}\right]$$

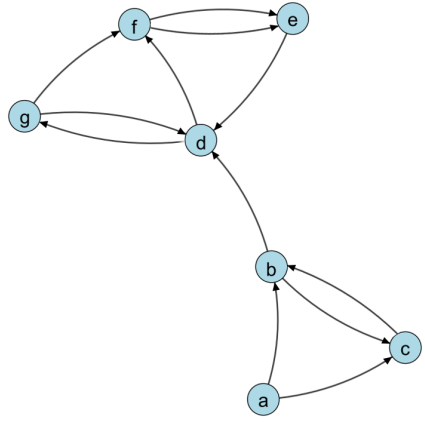
  - $\boldsymbol{h}(N)$ denotes a collection of (endogenous or exogenous) **properties** of the network N that take the role of **covariates** in the model
  - $\boldsymbol{\theta}$ a vector of **parameters** whose estimation allows for the fitting of the model (usually MLE)
- the interpretation of the model needs always to be related to some **null-model**, against which **model fit** and **parameters interpretation** needs to be compared

Network Models 101: Null-models

- Simplest network null models:
  - the $G(n, m)$ model

$$\Pr(N|m) = \left(\binom{n}{2} + m - 1 \atop m \right)^{-1}$$

hard to generalise to allow introduction of covariates

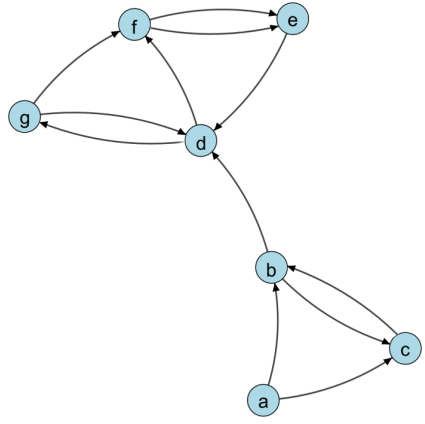


Network Models 101: Null-models

- Simplest network null models:
  - the $G(n, p)$ model

$$\Pr(N|p) = p^{\sum_{ij} A_{ij}} e^{-p} \left(\prod_{ij} A_{ij}! \right)^{-1}$$

easy to generalise to incorporate covariates in $p \Rightarrow$ null-model for various types of models (e.g., SBM, ERGM)



Network Models 101: Null-models

- Simplest network null models:
  - the $G(n, p)$ model

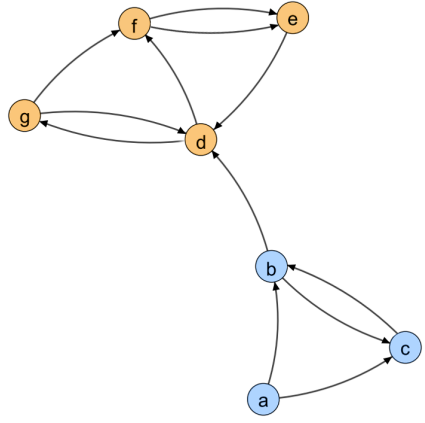
$$\Pr(N|p) = p^{\sum_{ij} A_{ij}} e^{-p} \left(\prod_{ij} A_{ij}! \right)^{-1}$$

easy to generalise to incorporate covariates in $p \Rightarrow$ null-model for various types of models (e.g., SBM, ERGM)

  - Example: **Stochastic Block Model**

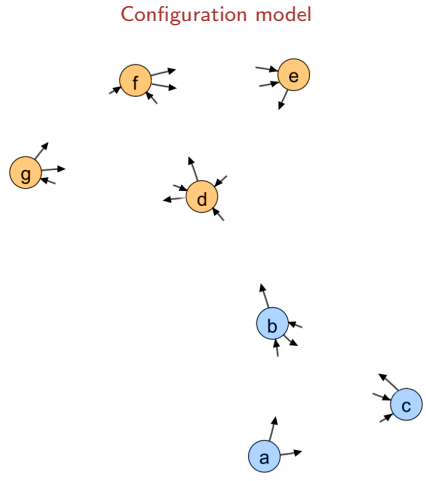
$$P(N|\boldsymbol{\theta}) = \prod_{B_{lm} \in B} \frac{\theta_{lm}^{B_{lm}/2} \exp[-\frac{1}{2} n_l n_m \theta_{lm}]}{\prod_{i,j \in \mathbf{B}_{lm}} A_{ij}!}$$

where B_{lm} is the number of edges within block lm, and θ_{lm} is the corresponding model parameter

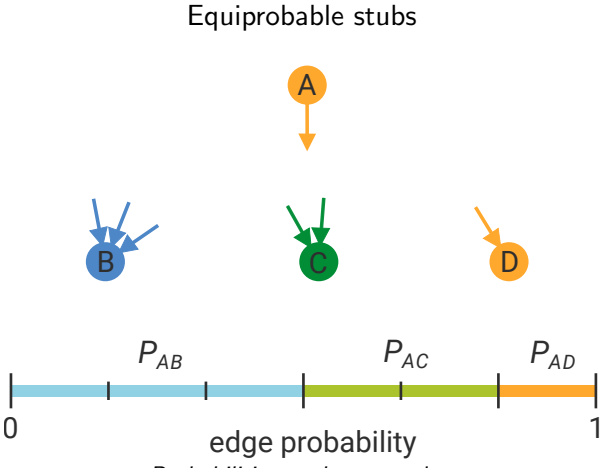


Network Models 101: the Configuration Model

- all networks with a **given degree sequence** $s(N)$ are **equiprobable**
- extension of $G(n, m)$
- Network realisation obtained by **randomly wiring stubs**
- common null-model when **preserving degree sequences**, or degree distributions, is important
- e.g., null-model for computation of **modularity**
- **How can we extend it to incorporate covariates?**



Generalising the Configuration Model

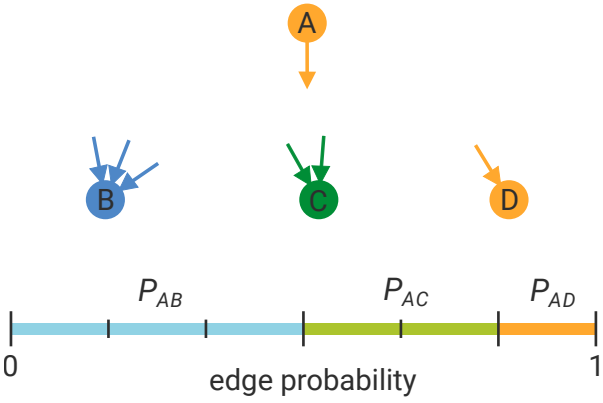


Probabilities to draw an edge:

$p_{AB} = 3/6$, $p_{AC} = 2/6$, $p_{AD} = 1/6$

Generalising the Configuration Model

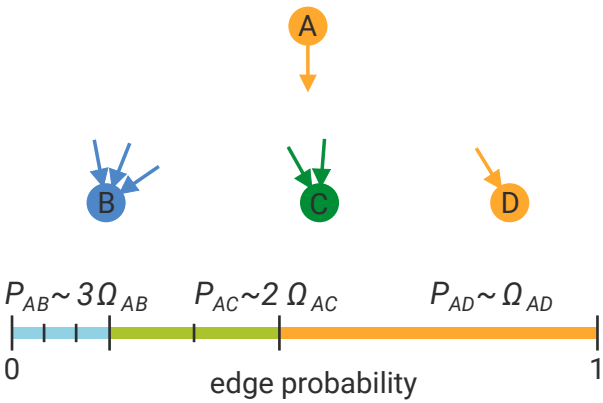
Equiprobable stubs



Probabilities to draw an edge:

$p_{AB} = 3/6$, $p_{AC} = 2/6$, $p_{AD} = 1/6$

Weighted edge propensities Ω_{ij}:



$\Omega_{AB} \ll \Omega_{AC} < \Omega_{AD}$

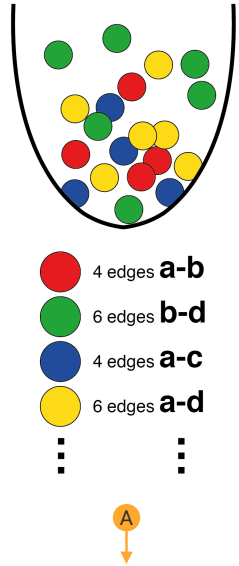
Ω_{ij}/Ω_{ik} *is the odds to draw an edge* ij *versus an edge* ik

Hypergeometric Configuration Model

- possible number of **stubs combinations** between nodes i, j:

$$\Xi_{ij} = k_i^{\text{out}} \cdot k_j^{\text{in}}$$

- if preserving **degrees in expectation** $\Rightarrow$ configuration model mapped to **urn problem**
- sampling of stubs combinations, i.e., edges, without replacement follows **hypergeometric distribution**



Generalized Hypergeometric Ensemble of Random Graphs

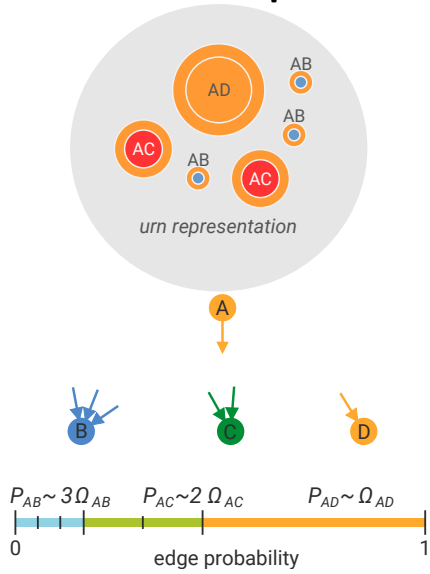
Probability to observe N given edge propensities Ω_{ij}s:

Wallenius' non-central hypergeometric distribution

$$\Pr(N|\mathbf{\Omega}) = \left[\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} dz$$

$$S_\Omega = \sum_{i,j}\Omega_{ij}(\Xi_{ij} - A_{ij})$$

Ω_{ij}/Ω_{ik} is the odds to draw an edge ij versus an edge ik



Inferential Model with gHypEG

- infer of **relative propensity** Ω_{ij} of nodes to be connected in terms of covariates $\boldsymbol{h}_{ij}(N)$:

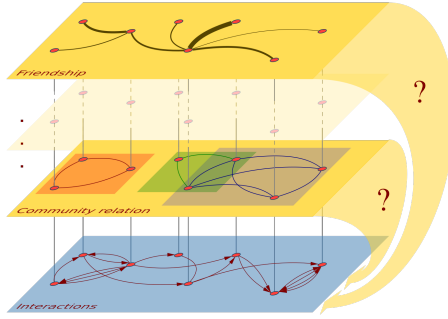
$$\Omega_{ij} = \exp\{\boldsymbol{\theta}^T \log[\boldsymbol{h}_{ij}(N)]\}$$

- Treat N as realisation of the **generalised hypergeometric ensemble** with adjacency $\mathbf{A}$:

$$P(N, \boldsymbol{\theta}) = \prod_{ij} \binom{\Xi_{ij}}{A_{ij}} \int_0^1 \prod_{ij} (1 - z^{\frac{\Omega_{ij}}{S_{\boldsymbol{\Omega}}}})^{A_{ij}} dz$$

with $S_{\boldsymbol{\Omega}} = \sum_{i,j} \Omega_{ij}(\Xi_{ij} - A_{ij})$

- **Maximum likelihood estimation** of parameters β



Model Selection and Model Evaluation

- **How good is a gHypEG model?**
- We provide 3 separate ways to assess the fit of models obtained by introducing new covariates
- Let $\mathcal{L}(\boldsymbol{\theta}_M|N)$ be the likelihood of model M
  - Cox and Snell pseudo R-squared

$$R^2_{\text{CS}} = 1 - \left(\frac{\mathcal{L}(\boldsymbol{\theta}_0|N)}{\mathcal{L}(\boldsymbol{\theta}_M|N)} \right)^{\frac{2}{m}}$$

  - McFadden pseudo R-squared

$$R^2_{\text{McF}} = 1 - \frac{\log \mathcal{L}(\boldsymbol{\theta}_0|N)}{\log \mathcal{L}(\boldsymbol{\theta}_M|N)}$$

  - AIC: Akaike information criterion

$$AIC = 2k - 2 \log \mathcal{L}(\boldsymbol{\theta}_M|N)$$

- AIC can further be used to compare different models and **perform model selection**.

Case Study: High School Contacts

- SocioPatters high school contacts (2015)
  - 327 students, 5 days record
  - 188508 **active contacts** during 20-second intervals (edges in *interaction layer*)
  - **Class membership** (colors in figure)
  - **Topic grouping** (visible in fig.)
  - Gender
  - Self-reported **friendship relations** (668 unweighted directed edges)
  - **Facebook connections** (4515 unweighted undirected edges)

Which of these effects are relevant?
Does triadic closure play a role?

SocioPatterns – High school contacts

Case Study: Inference

- Parameter estimates show strong effect of
  - **triadic closure** (in accordance with social theories)
  - **class homophily** (due to physical separation into classes)
  - **friendship relations**
- Effect of $\mathcal{R}_F$, $\mathcal{R}_T$, and $\mathcal{R}_G$ on $\mathcal{I}$ is **small**
- **Facebook connections**, **topic grouping**, and **gender homophily** are not good predictors for interactions between students $\Rightarrow$ effect included in other layers

	gHypEG		
	Estimate	ΔAIC	relative improv
closure	$0.633^{(*)}$	712208	0.5283
+class	$0.831^{(*)}$	40691	0.0641
+friend	$0.570^{(*)}$	42705	0.0719
+fb	$0.257^{(*)}$		
β_ε	$0.303^{(*)}$	6169	0.0112
+topic	$0.461^{(*)}$	4150	0.0052
+gender	$0.076^{(*)}$	1297	0.0024
+1/2friend	-0.014	2	0.0000

$^{(*)}$ indicates p-value < 0.001

Table of Contents

Endogenous network statistics in social networks

To answer the question:

Which social patterns guide social interactions?

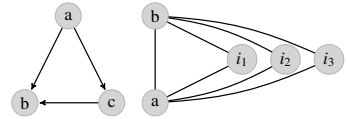
we often use:

- inferential network models
- data on social interactions, i.e., networks
- to ensure unbiased results, we need to (at least) control for **endogenous network patterns**
  - also called:
  - relational mechanisms*
  - social patterns
  - network statistics
  - endogenous model terms
  - ..

*See e.g., Rivera, M. T., Soderstrom, S. B., and Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. annual Review of Sociology, 36:91–115.

An old problem

- **quantifying relational mechanism is an old problem**
- → and solved for `ERGM`, `SAOMs`, `latent space models`, etc.
- Problem: these models use binary network data
- binary data → bad representation for (repeated) social interaction data



Change statistics

- One way of quantifying endogenous processes are through change statistics
- also called *change scores*

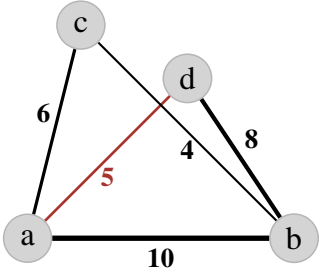
$$S^{(ij)_r}(N) = h_r(N^+_{(ij)}) - h_r(N^-_{(ij)}) \tag{1}$$

The change in the h-statistic (= endogenous network term) if the dyad (i,j) is set to 1 and 0

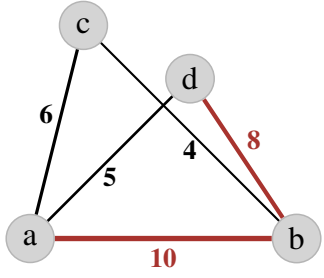
Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. Sociological methodology, 36(1):99–153.
Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. Journal of the American Statistical Association, 103(481):248–258.
Krivitsky, P. N., Handcock, M. S., and Morris, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. Statistical methodology, 8(4):319–339.
Leifeld, P., Cranmer, S. J., and Desmarais, B. A. (2018). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. Journal of Statistical Software, 83(6):1–36.

Change statistic for triadic closure (1/2)

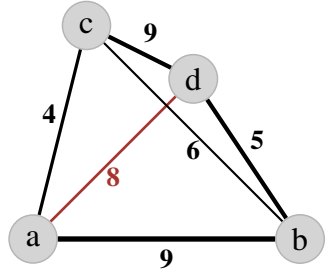
For each dyad:



Find all shared partners:

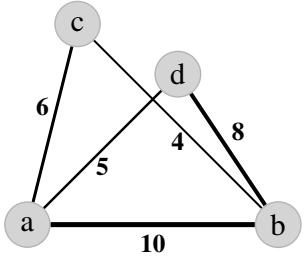


Dyad value in change statistic:



Change statistics (2/2)

Weighted Network



Binary change statistic

	a	b	c	d
a	-	2	1	1
b	2	-	1	1
c	1	1	-	2
d	1	1	2	-

Weighted change statistic

	a	b	c	d
a	-	9	4	8
b	9	-	6	5
c	4	6	-	9
d	8	5	9	-

Brandenberger, L., Casiraghi, G., Nanumyan, V., and Schweitzer, F. (2019). Quantifying triadic closure in multi-edge social networks. arXiv preprint arXiv:1905.02990.

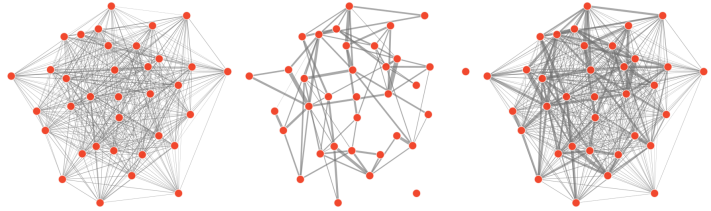
Inferential network models for multi-edge networks

- count ERGM[†]
  - part of the ERGM-family of models
  - models the probability of observing a network N over all possible permutations $\mathcal{N}$
- hypernets regression[‡]
  - based on generalized hypergeometric ensembles
  - models the probability of observing the given network N by sampling its edges from an urn containing all possible combinations of edges

[†] Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. Electronic journal of statistics, 6:1100.

[‡] Casiraghi, G. (2017). Multiplex Network Regression: How do relations drive interactions? arXiv preprint arXiv:1702.02048.

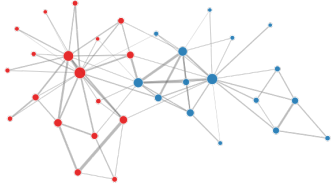
Synthetic example



	(a) random		(b) triangles		(c) both	
ERGM						
triadic closure	0.001	(0.002)	0.03*	(0.002)	0.025*	(0.002)
nonzero	-0.07	(0.14)	-1.67*	(0.16)	-1.67*	(0.16)
sum	0.53*	(0.12)	0.13	(0.09)	0.13	(0.09)
hypernets						
triadic closure	.30	(.19)	1.13*	(.04)	1.76*	(.09)

Real life data example 1

- Zachary's Karate Club network
- No triadic closure



	ERGM			hypernets		
nonzero	−3.281	***	(0.267)	–		
sum	−1.166	***	(0.297)	–		
degree dist.	0.028	***	(0.004)	–		
triadic closure	−0.016		(0.012)	−0.160	.	(0.086)
faction	1.123	***	(0.178)	1.090	***	(0.104)
AIC	−869.4			674.7		
Null AIC	0			869.1		

Data: Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. Journal of anthropological research, 33(4):452–473.