Inferential Network Analysis (and Big Data): Challenges and Opportunities

Lisa Lechner

2020-04-21

University of Innsbruck

Why (not) networks?

Gains and sacrifices of network analysis

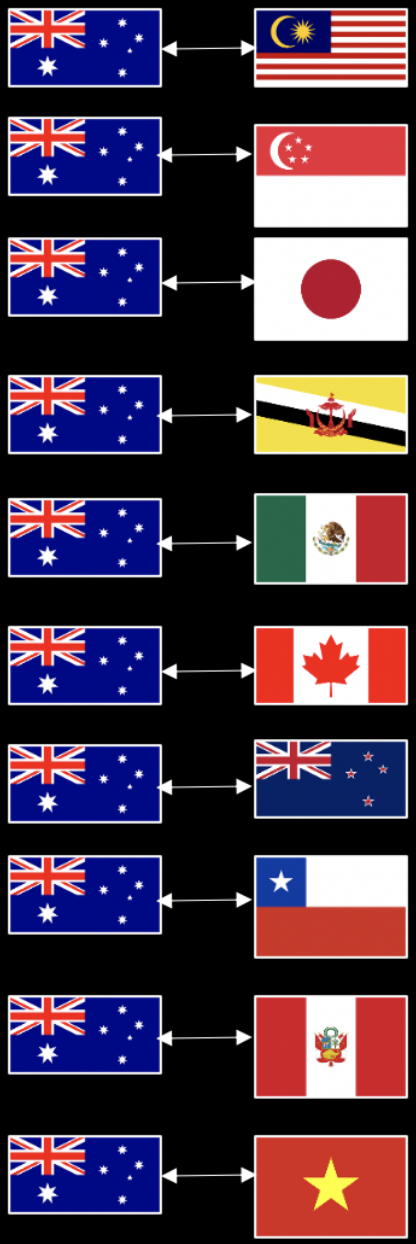
Gains

- Honest perspective
- Better measures
- Lower Typ I Error

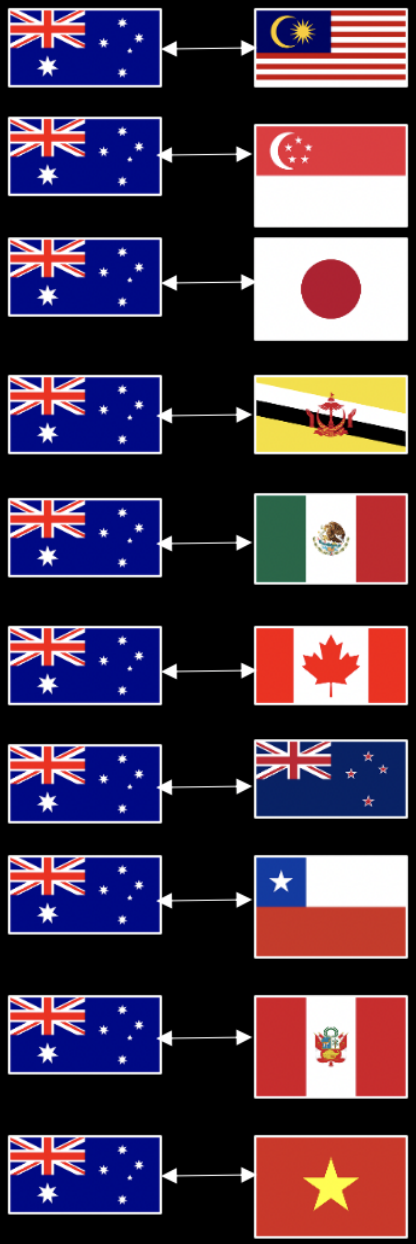
Sacrifices

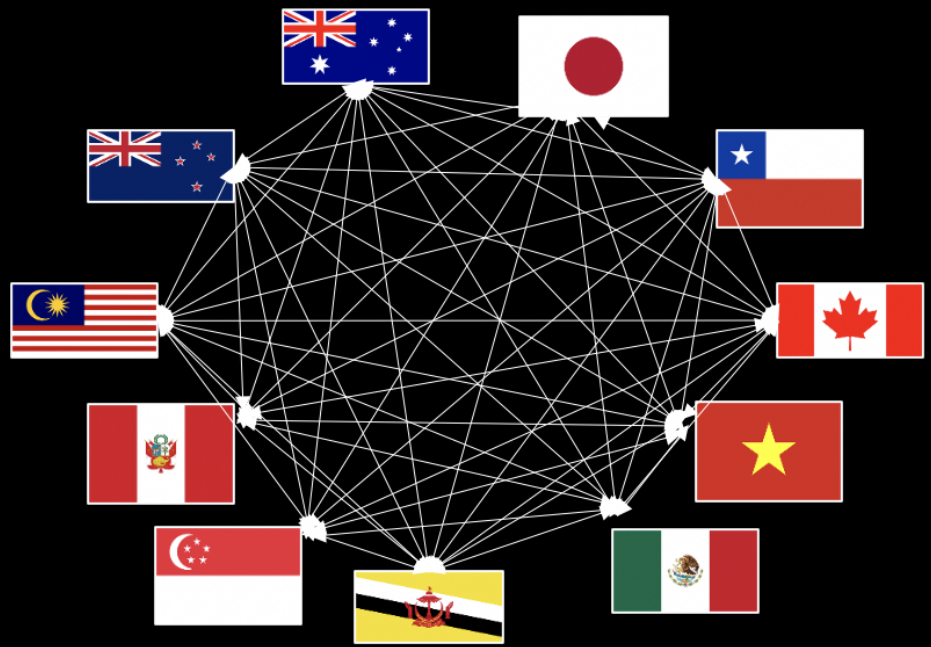
- Time
  - for good theory
  - for estimation
- Computational Power

Honest perspective

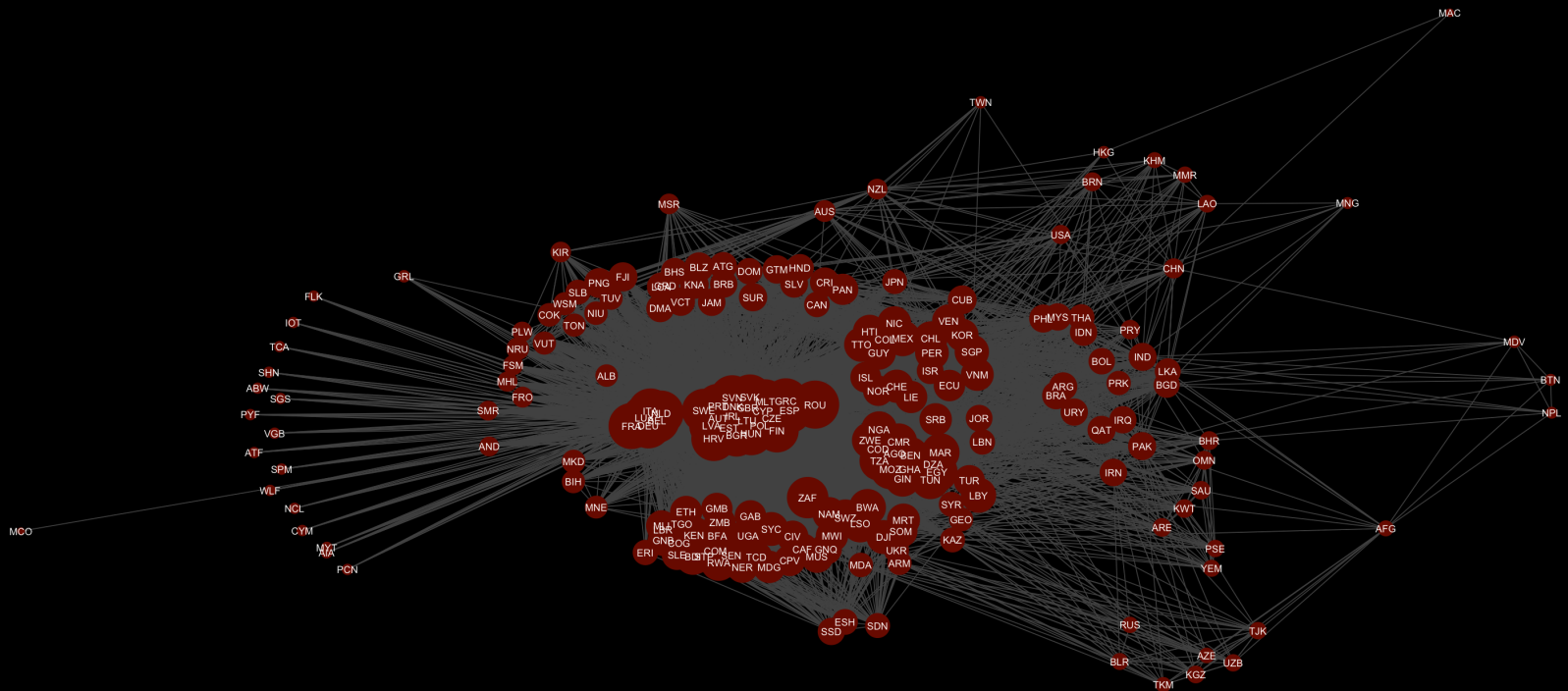


Honest perspective





Honest perspective



Better measures

Lower Typ I Error (Dorussen et al., 2016)

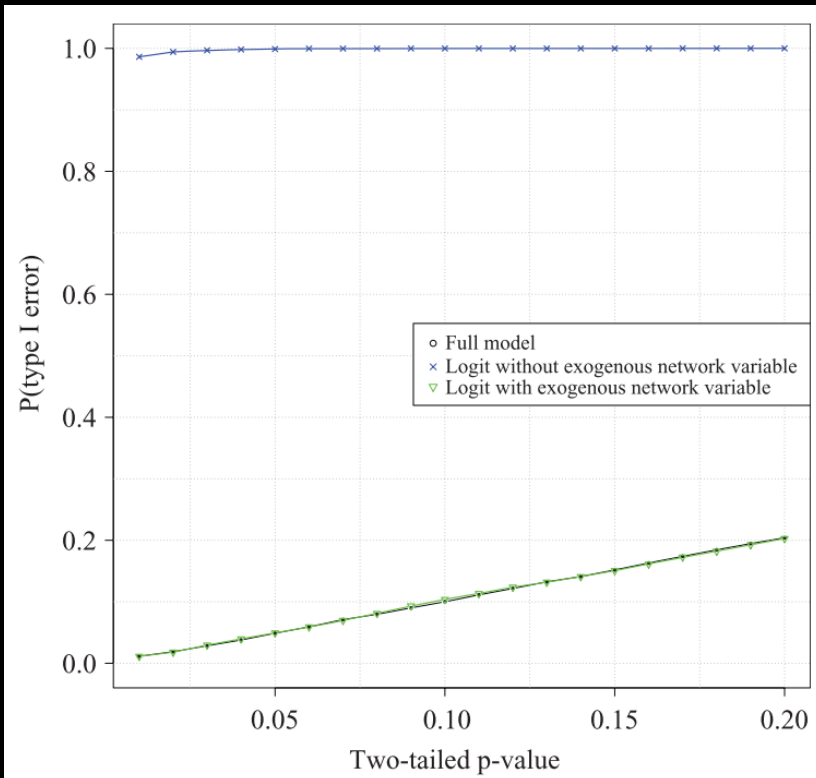


Figure 1. Type I error rate covariate correlated with exogenous network variable

Calculations based on 5,000 repetitions.

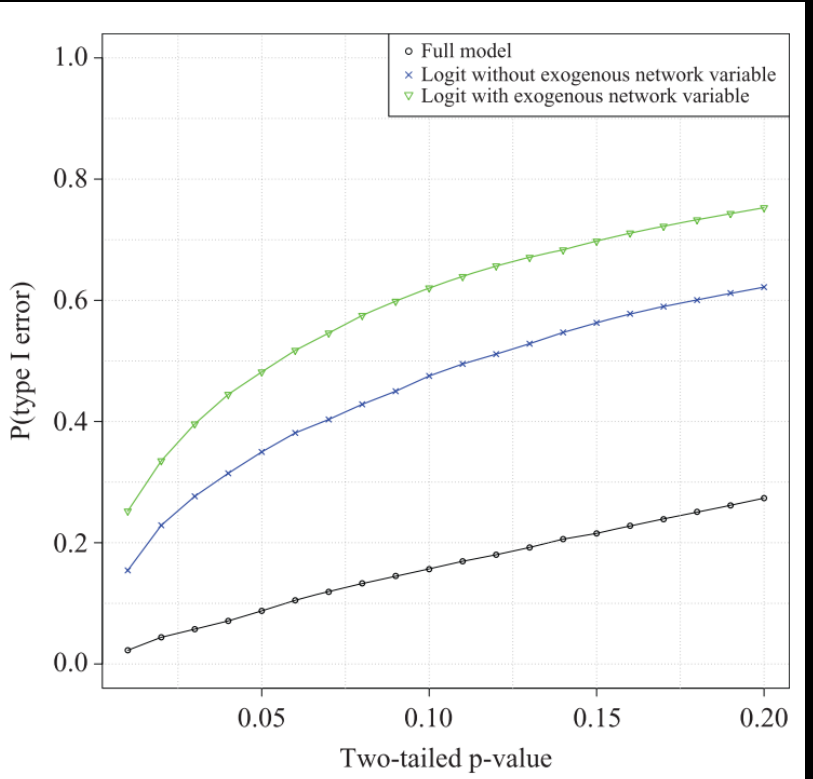


Figure 2. Type I error rate covariate correlated with endogenous network interdependence variable

Calculations based on 5,000 repetitions.

Navigation through network tools

Basic distinctions

- Descriptive versus inferential network analysis
- Complete versus ego-centric network analysis

Inferential network analysis options (See Cranmer et al. (2016))

- Latent Space Approaches (Hoff et al., 2002)
- Quadratic Assignment Procedure (Hubert and Schultz, 1976)
- Exponential Random Graph Models (Wasserman and Pattison, 1996)

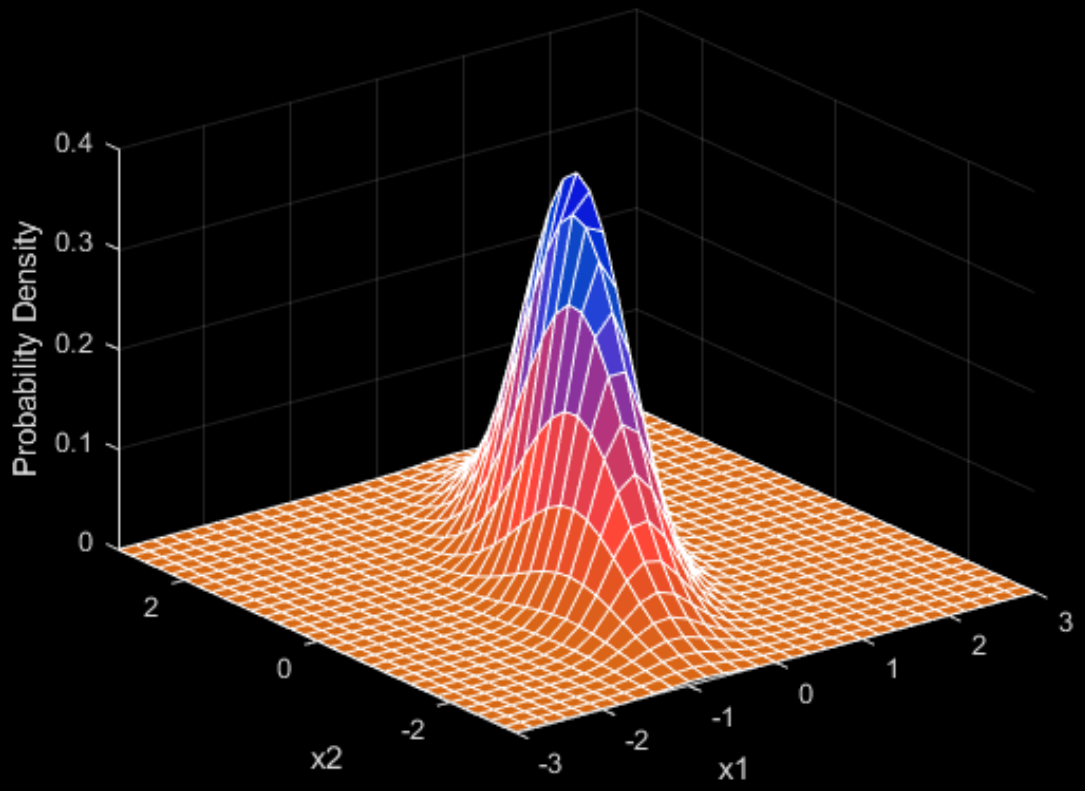
Inferential network analysis options (See Cranmer et al. (2016))

TABLE 2 Comparison of Three Models for Inferential Network Analysis

Criterion	QAP	ERGM	LSM
Operationalization of relational theories	−	+	−
Easy to specify and interpret	+	−	○
Parsimonious (= few parameters needed)	+	○	−
Reports standard errors	−	+	+
Avoids problems with oversensitivity of dependencies	+	+	−
Simulations do not suffer from omitted variable bias	−	+	+
Avoids numerical instability (= degeneracy)	+	−	+
Unbiased under arbitrary empirical distributions	−	+	+
Flexible with respect to outcome distribution	+	+	+
Temporal specifications are available	−	+	+
Spatial visualization and model-based clustering	−	−	+
Full-fledged statistical model	−	+	+
Availability in standard statistical software	+	+	+

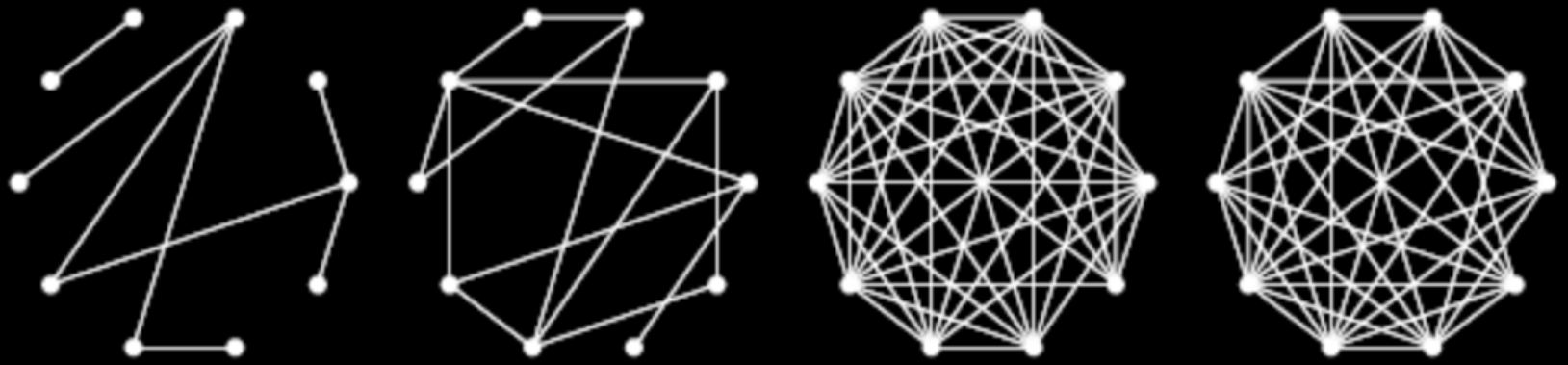
ERGM (the idea)

DGP from multivariate distribution.



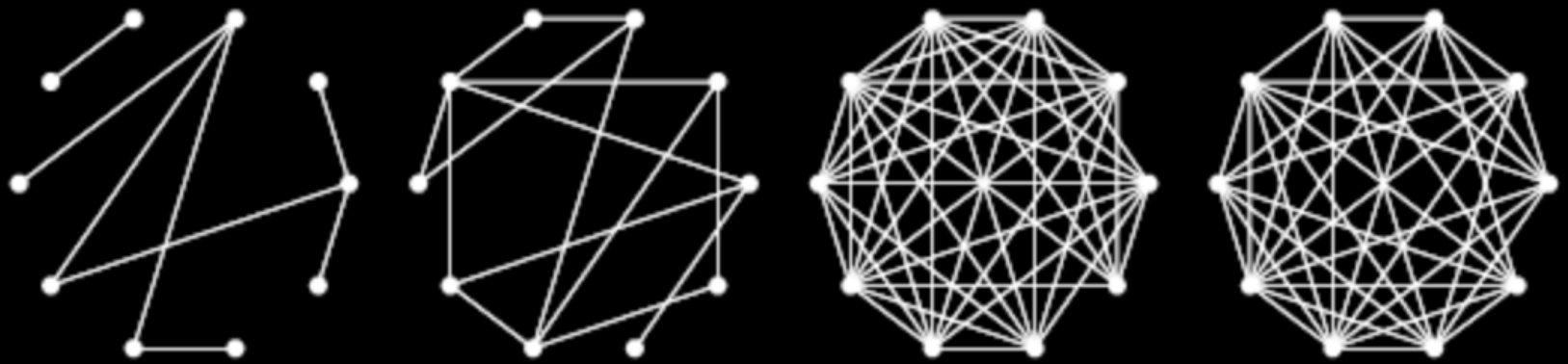
ERGM (the idea)

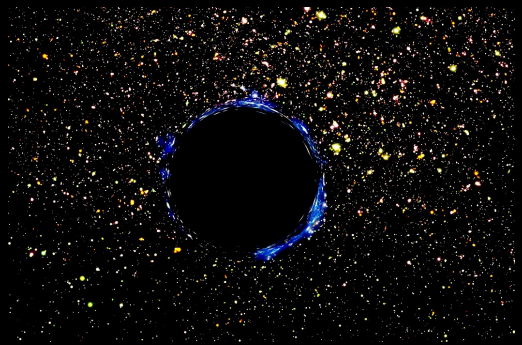
SAMPLE SPACE



ERGM (the idea)

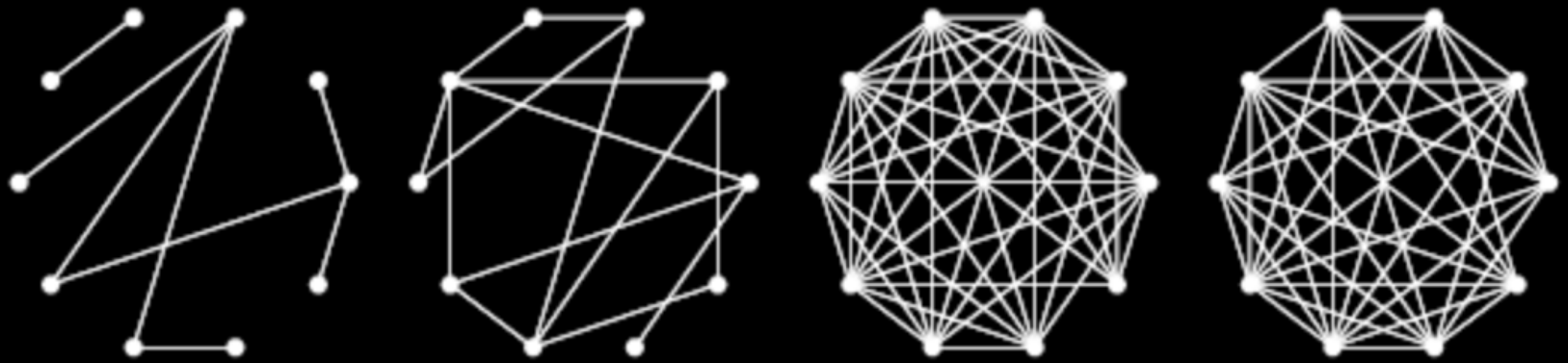
SAMPLE SPACE

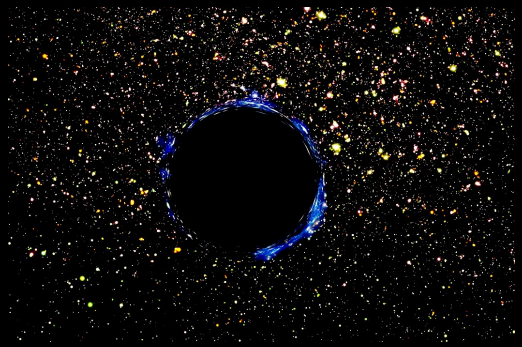




ERGM (the idea)

SAMPLE SPACE





$\rightarrow$ Approximation through maximum pseudolikelihood and Markov Chain Monte Carlo (MCMC) maximum likelihood

ERGM (the idea)

OBSERVED VERSUS POSSIBLE

$$P(G_i) = \frac{\exp(-\sum_{j=1}^{k} \Gamma_{ij}\theta_j)}{\sum_{m=1}^{M} \exp(-\sum_{j=1}^{k} \Gamma_{mj}\theta_j)}$$

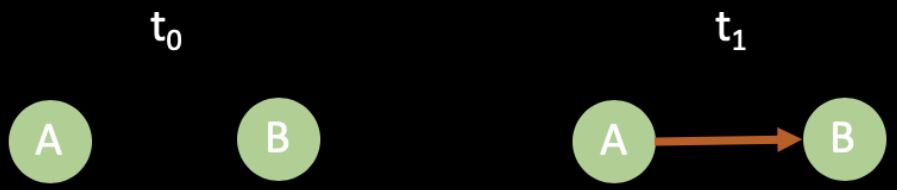
ERGM (the idea)

ERGM - parameters (see for instance Morris et al. (2008))

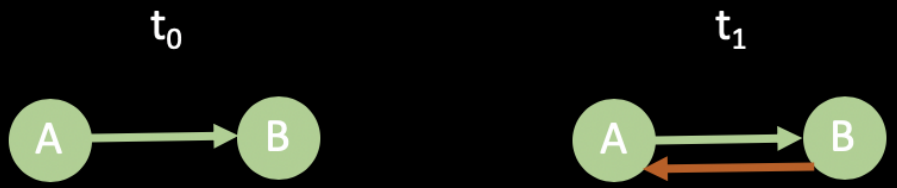
- Node variables
- Edge variables
- Structure variables

ERGM - parameters (see for instance Morris et al. (2008))

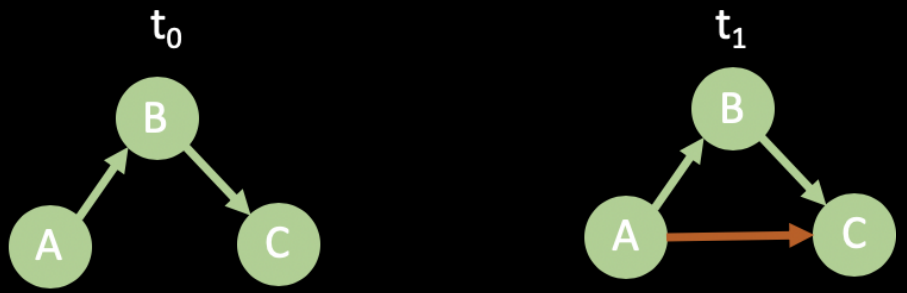
edge



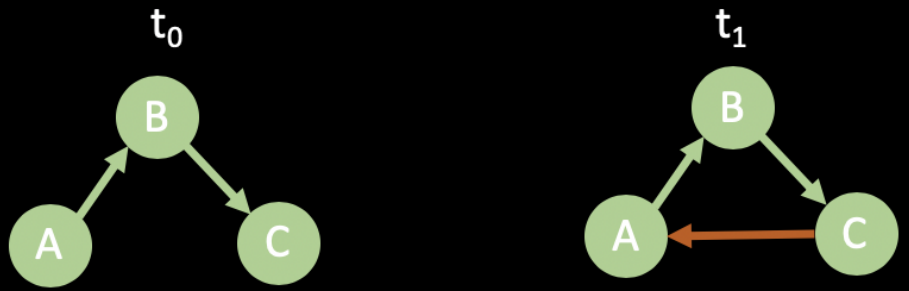
reciprocity



transitive closure

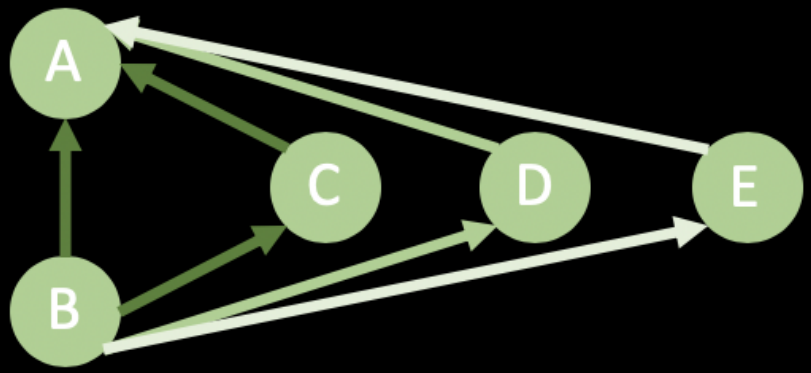


three cycles



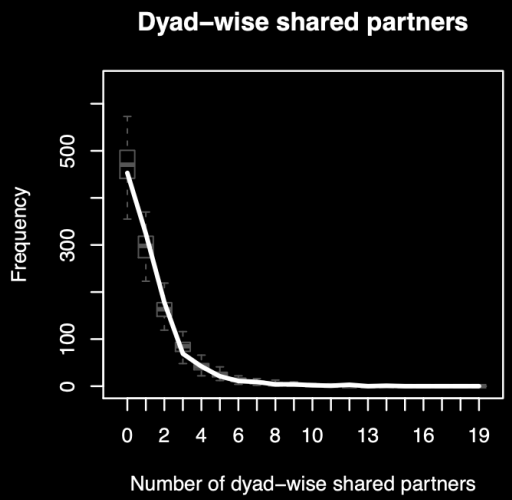
ERGM - parameters (see for instance Morris et al. (2008))

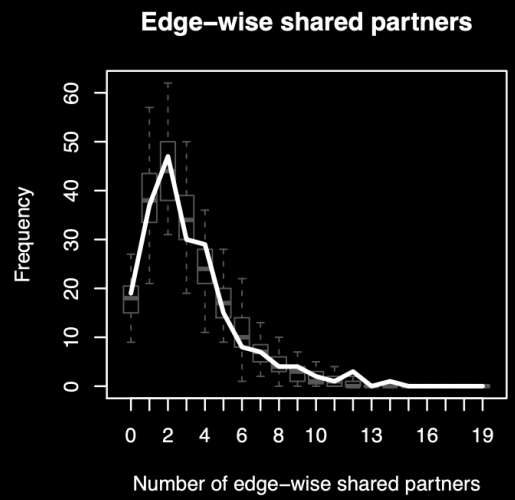
GWESP

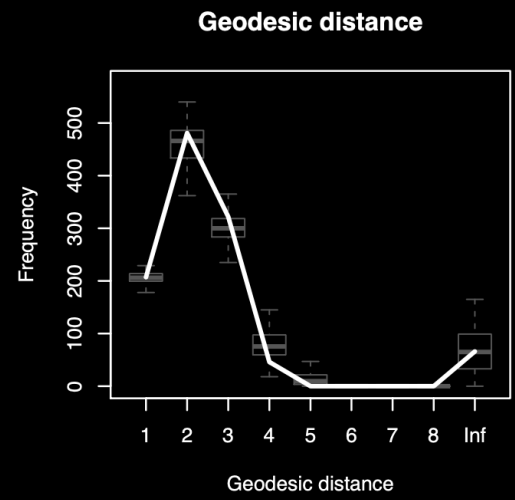


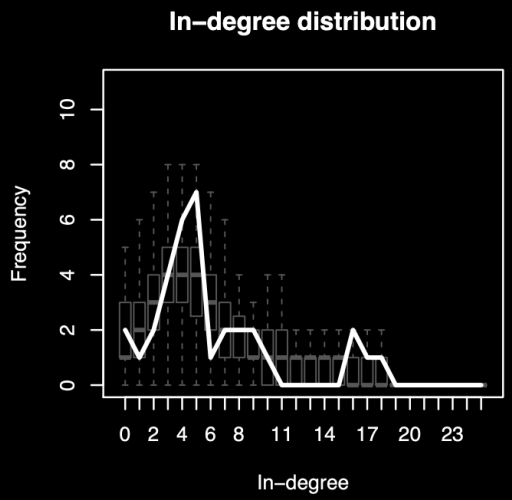
See `help("ergm-terms")`

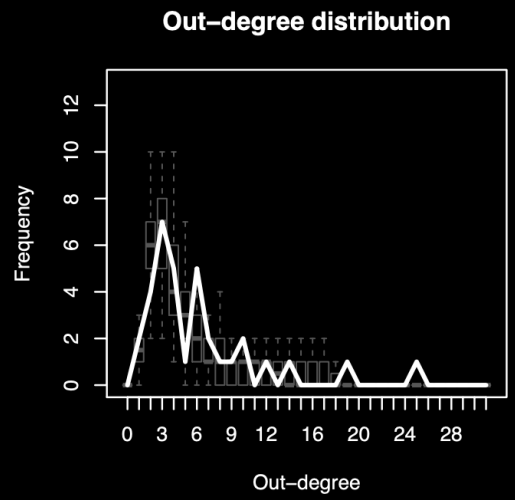
Model fit

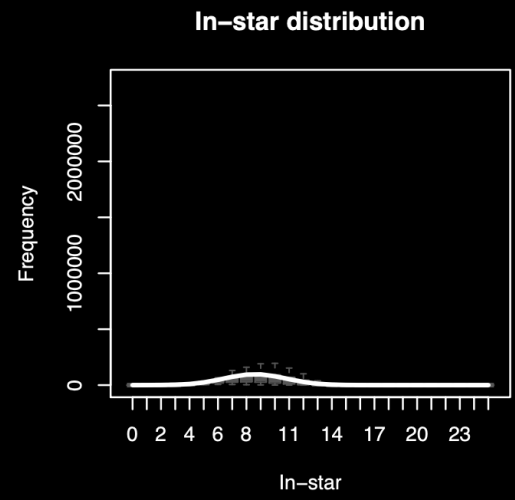

Dyad−wise shared partners
Frequency
0 100 300 500
0 2 4 6 8 10 13 16 19
Number of dyad−wise shared partners


Edge−wise shared partners
Frequency
0 10 20 30 40 50 60
0 2 4 6 8 10 13 16 19
Number of edge−wise shared partners


Geodesic distance
Frequency
0 100 200 300 400 500
1 2 3 4 5 6 7 8 Inf
Geodesic distance


In−degree distribution
Frequency
0 2 4 6 8 10
0 2 4 6 8 11 14 17 20 23
In−degree


Out−degree distribution
Frequency
0 2 4 6 8 10 12
0 3 6 9 12 16 20 24 28
Out−degree

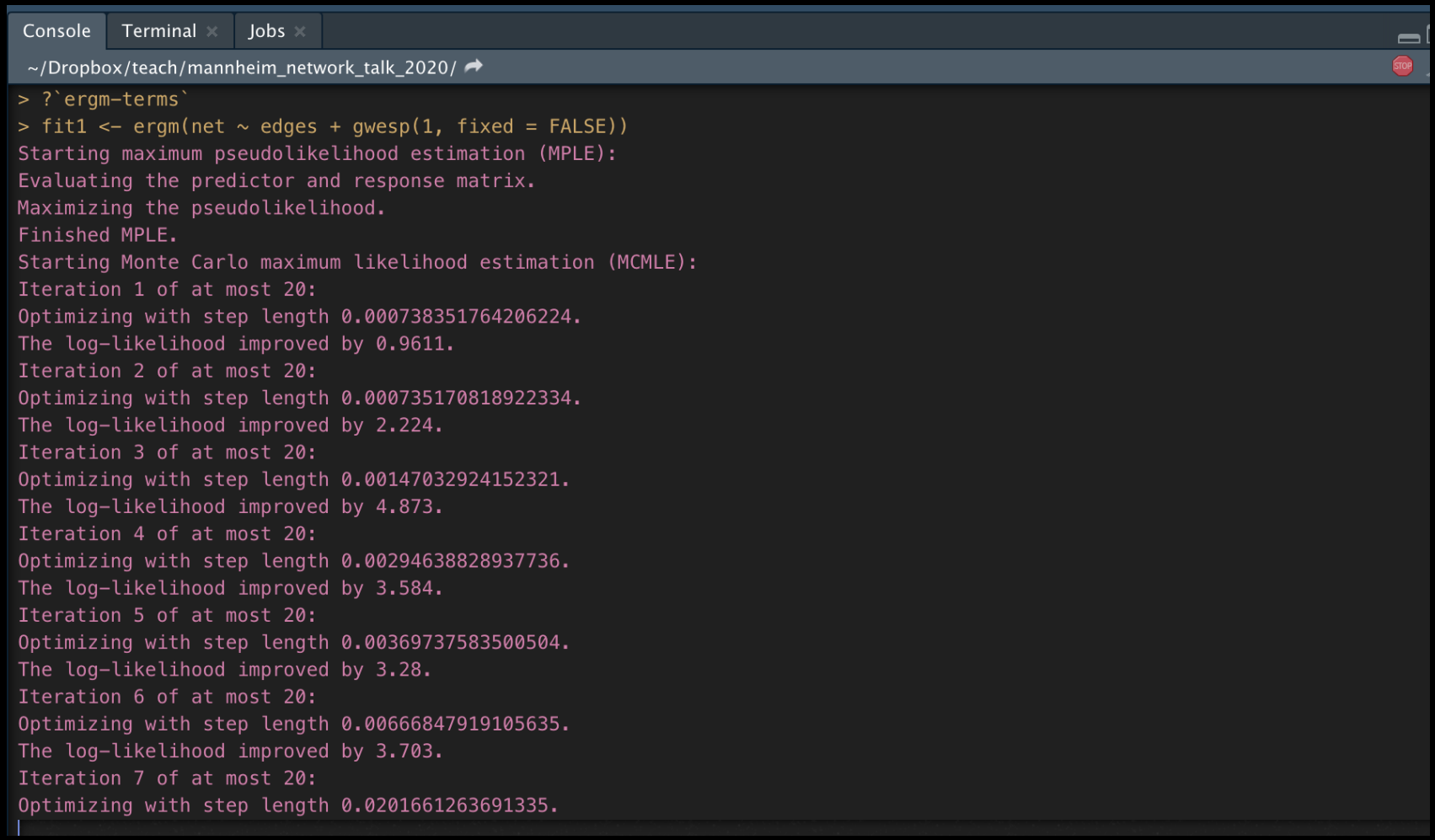

In−star distribution
Frequency
0 1000000 2000000
0 2 4 6 8 11 14 17 20 23
In−star

Drawback I: Degeneracy

```
Error in ergm.MCMLE(init, nw, model, initialfit =
(initialfit <- NULL), : Number of edges in a
simulated network exceeds that in the observed by a
factor of more than 20.
```

```
This is a strong indicator of model degeneracy or a
very poor starting parameter configuration. If you
are reasonably certain that neither of these is the
case, increase the MCMLE.density.guard control.ergm()
parameter.
```

Drawback I: Degeneracy



Drawback II: Missing Data

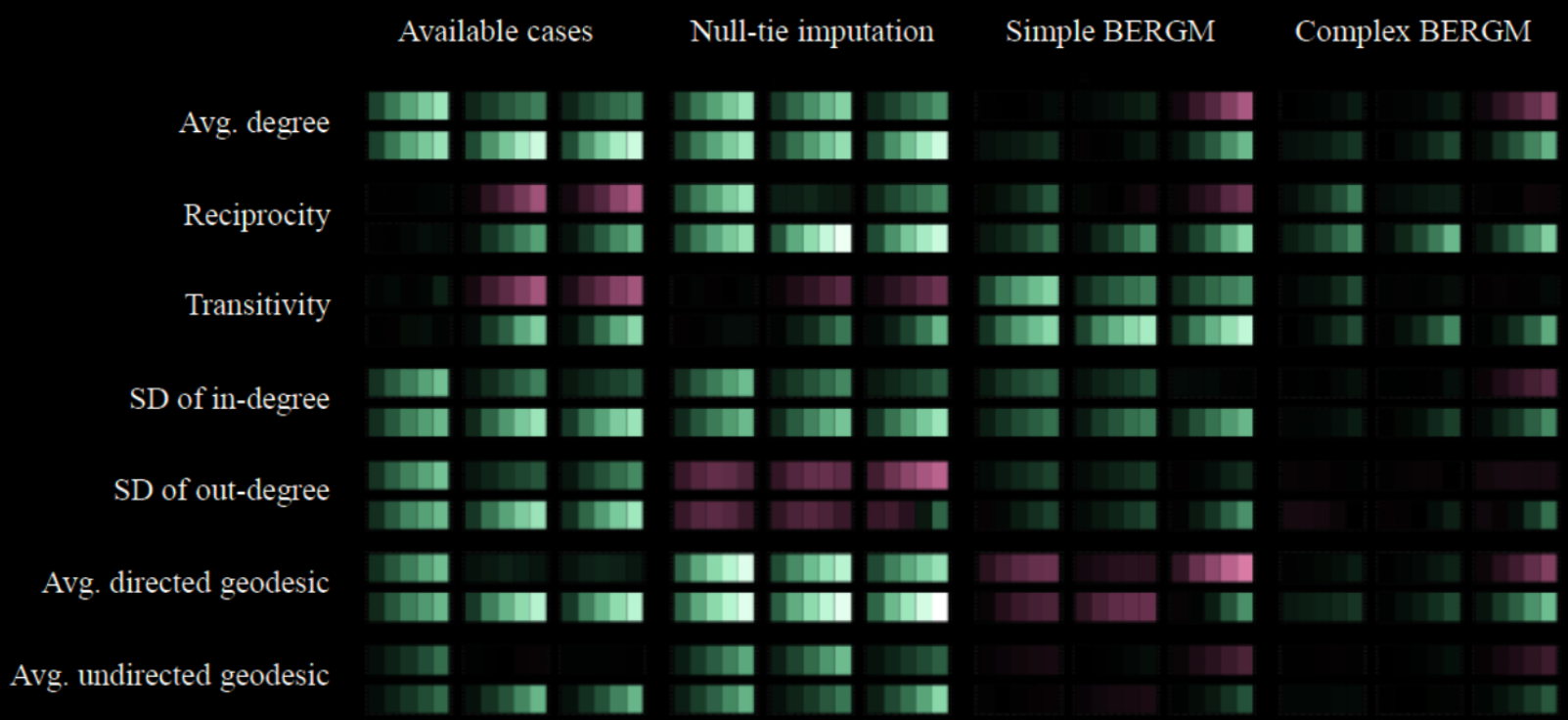


Fig. 2. Average relative bias for each descriptive statistic by treatment method. For interpretation of the plots, consult Fig. 1.

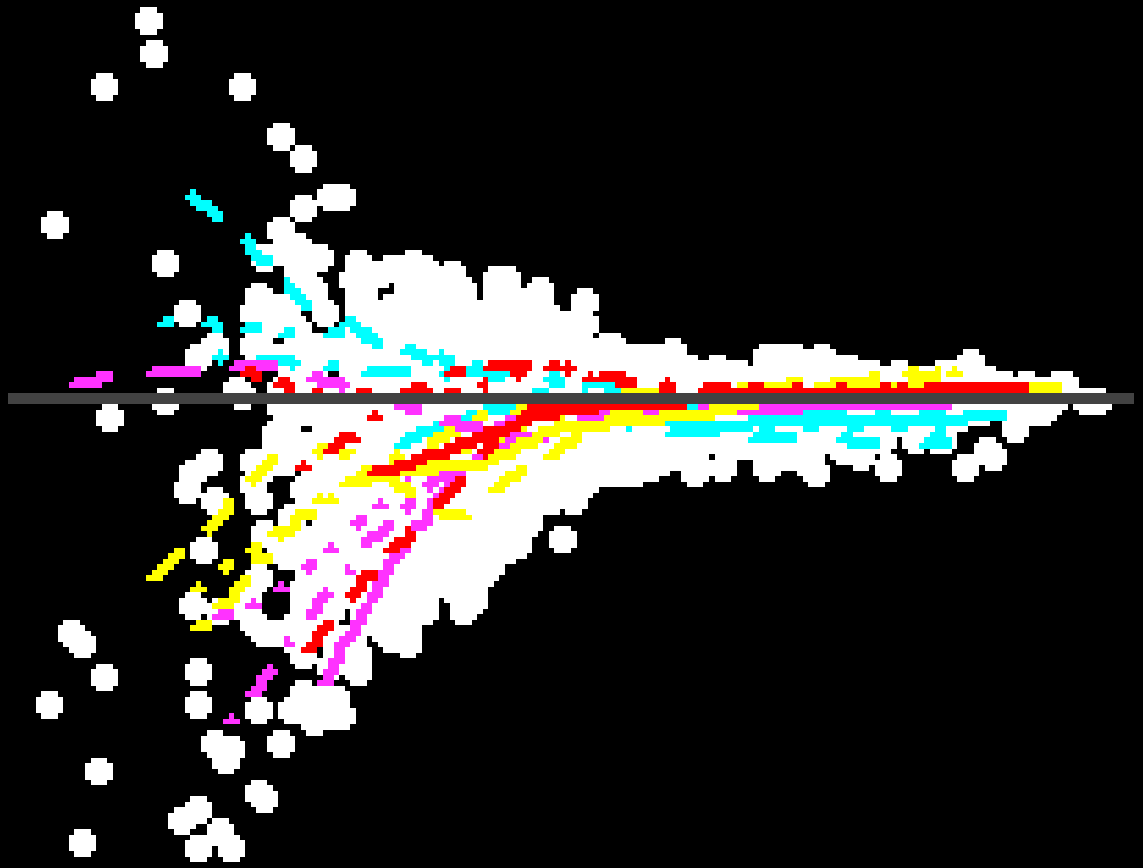
(Krause et al., 2018)

ERGM Family

TERGM, FERGM, mERGM, GERGM, BERGM, ...

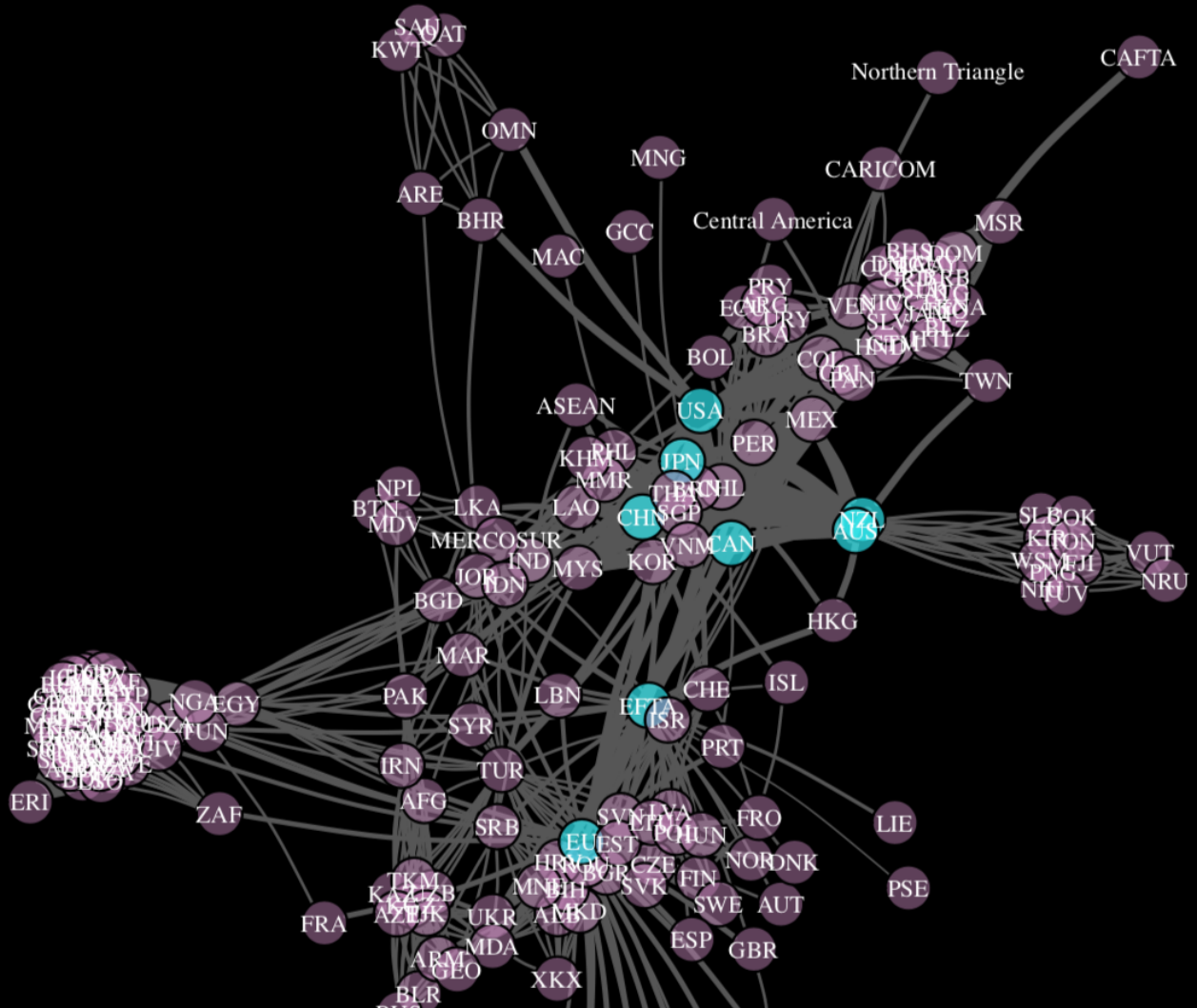
ERGM Family - Unobserved Hetoreogeneity

- FERGM (Box-Steffensmeier et al., 2018) - `install.packages("fergm")`
- mixed ERGM (Kevork and Kauermann, 2019)



ERGM Family - Weigthed NEtwork

- Weighted edges: GERGM (Wilson et al., 2017) - `devtools::install_github("matthewjdenny/GERGM")`



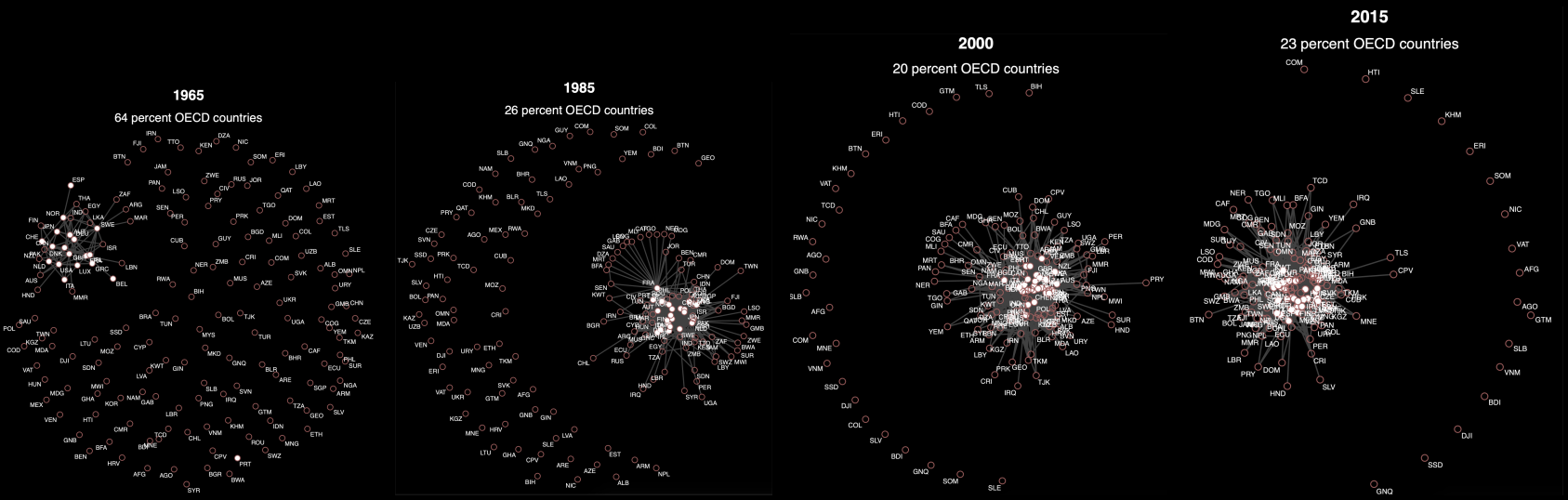
ERGM Family - Faster convergence $\rightarrow$ a baysian approach

- BERGM (Caimo and Friel, 2011, 2012) - `install.packages("Bergm")`

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{\exp\{\boldsymbol{\theta}^\top s(\boldsymbol{y})\}}{z(\boldsymbol{\theta})}\ \frac{p(\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

ERGM Family - Temporal aspect

- Temporal model: TERGM (Hanneke et al., 2010; Leifeld et al., 2018) - `install.packages("btergm")`
- Stochastic Actor-Oriented Model (SAOM or SIENA)



ERGM Family - More models

- TNAM
- REM for event models

Hands-on

Challenges (theoretical and methodological)

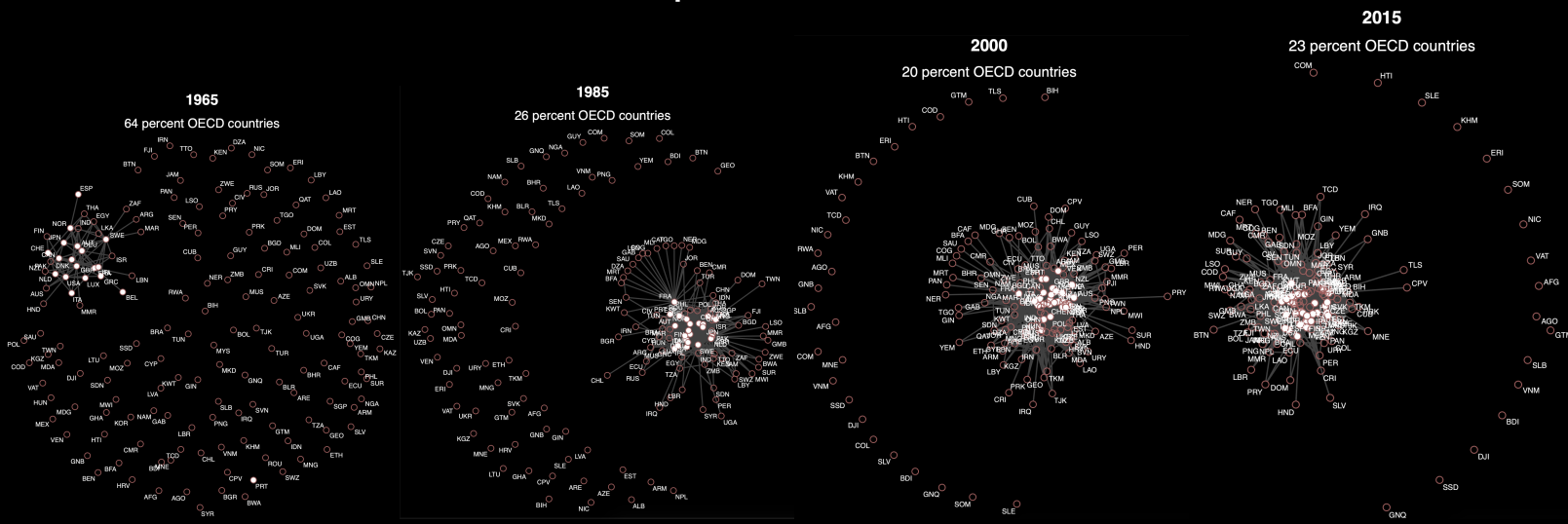
Theoretical challenge

Draw a line from each word to the correct part of the body.

- Nose
- Eyebrow
- Forehade
- Eyelids
- Teeth
- Mouth

Methodological challenge example

- Joint project with Vincent Arel-Bundock on convergence of tax regime
  - 192 countries and 74 time steps



  - network analysis combined with text analysis, where preprocessing matters $\rightarrow$ we had 64 estimations per model.

Methodological challenge example

Dilemma:

1. Rent expensive cloud computing
2. Wait ages until computing is done and start to work on other projects
3. Computing is done, but caught up in other things
4. Slowly get back to project
5. Estimate again and wait...

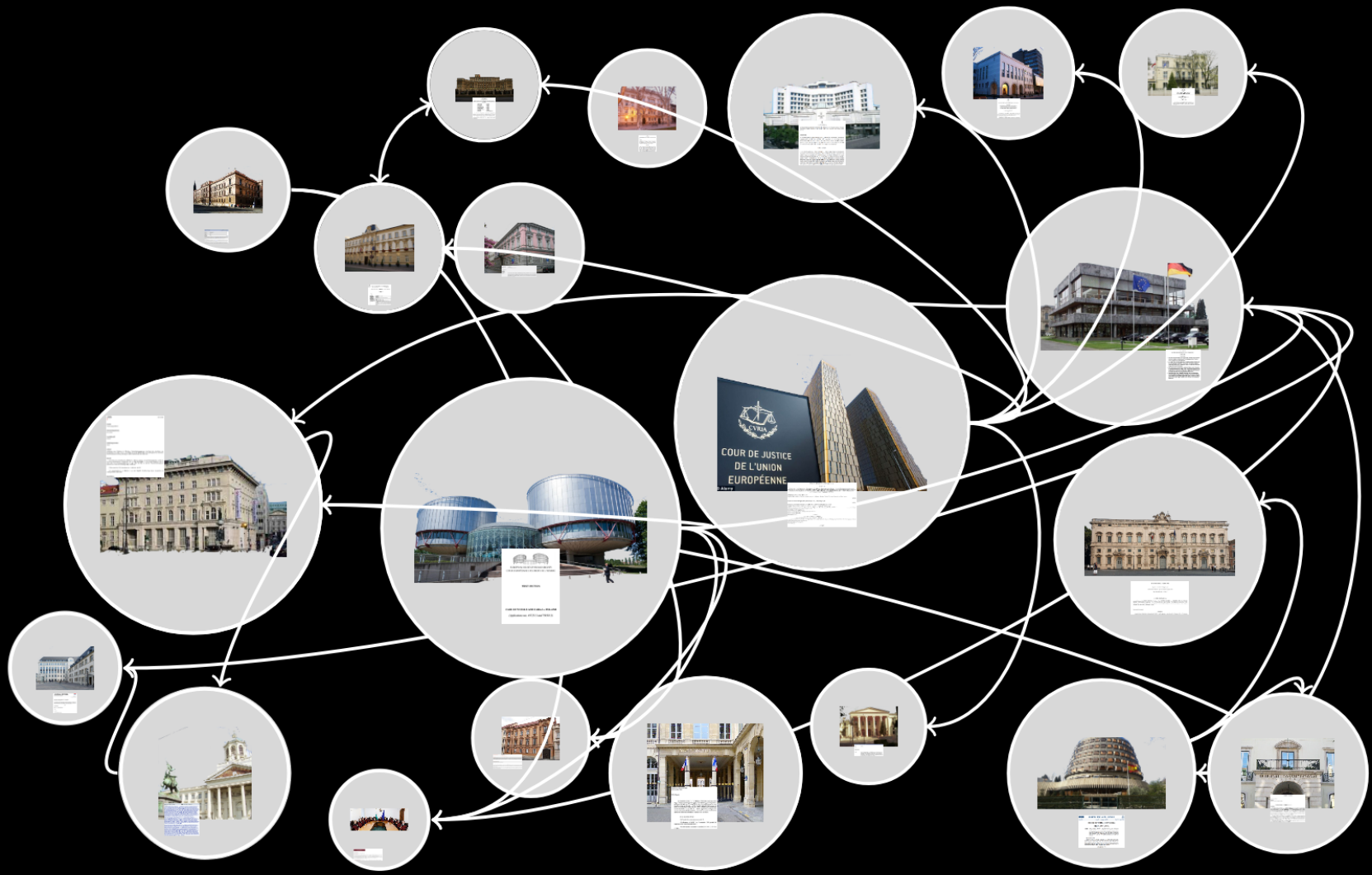
Methodological challenge example

Dilemma:

1. Rent expensive cloud computing
2. Wait ages until computing is done and start to work on other projects
3. Computing is done, but caught up in other things
4. Slowly get back to project
5. Estimate again and wait...

$\rightarrow$ very inefficient research process

There are much bigger networks



There are much bigger networks

- Survey data
- Text data
- Panel data

ERGM(s) challenges with big data

1. Loading big network data requires a lot of memory
2. Struggles processing long vectors (which are often used for big networks)
3. Assumption of MCMLE of actors having full knowledge of remaining network becomes unrealisitc

Big(ger) data options

A. Rent cloud computing (expansive and annoying)

B. Model full network, but move beyond MCMLE

C. Structural reduction of network

Modeling complete network

- Reparametrization (see Statistical ERGM - SERGMs by Chandrasekhar and Jackson (2012))

Modeling complete network

- improve MCMLE by dealing more effectively with isolates

Modeling complete network

- improve MCMLE by dealing more effectively with isolates
- speed up MCMLE by parallelization (either for the full chain - but this is hard (Calderhead, 2014), or for small subnetworks)

Modeling complete network

- improve MCMLE by dealing more effectively with isolates
- speed up MCMLE by parallelization (either for the full chain - but this is hard (Calderhead, 2014), or for small subnetworks)
- Pseudo maximum likelihood estimation (Strauss and Ikeda, 1990)

Modeling complete network

- improve MCMLE by dealing more effectively with isolates
- speed up MCMLE by parallelization (either for the full chain - but this is hard (Calderhead, 2014), or for small subnetworks)
- Pseudo maximum likelihood estimation (Strauss and Ikeda, 1990)
- Graph limit maximum likelihood estimation (He and Zheng, 2015)

Modeling complete network

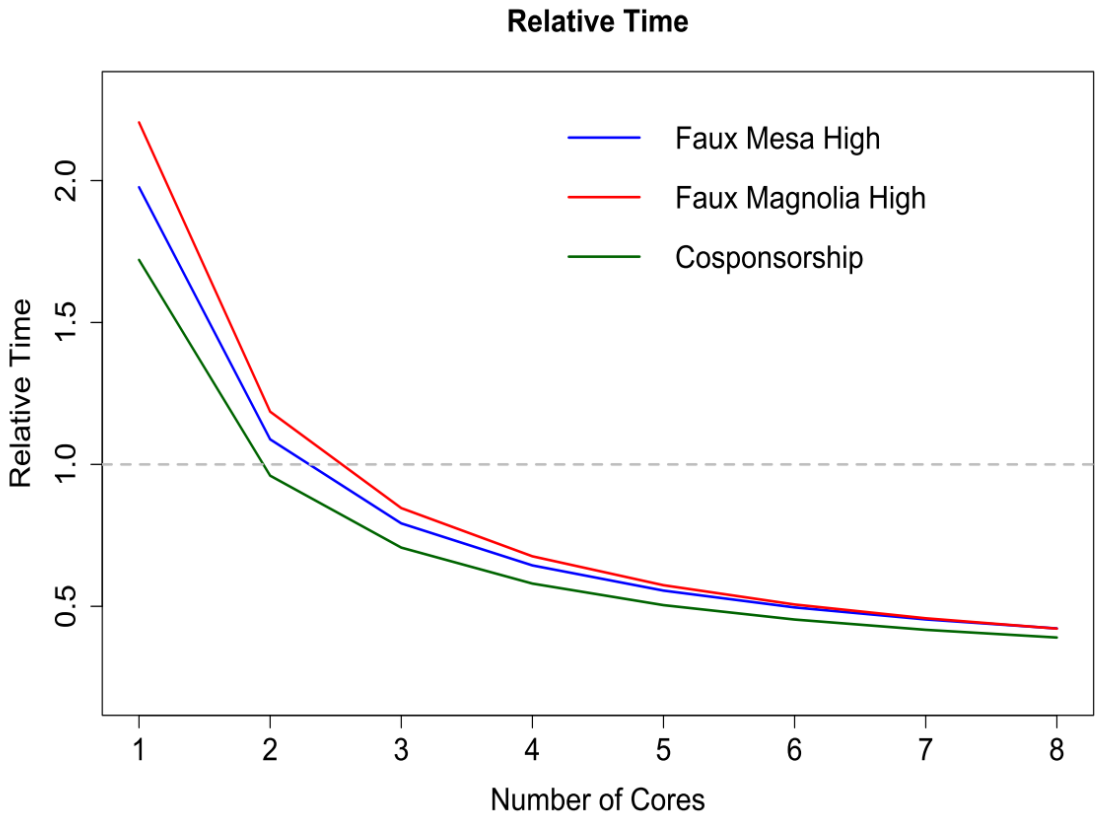
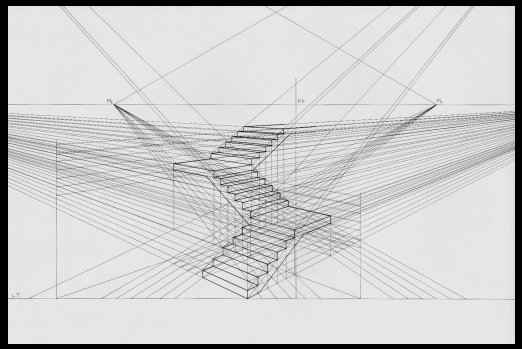
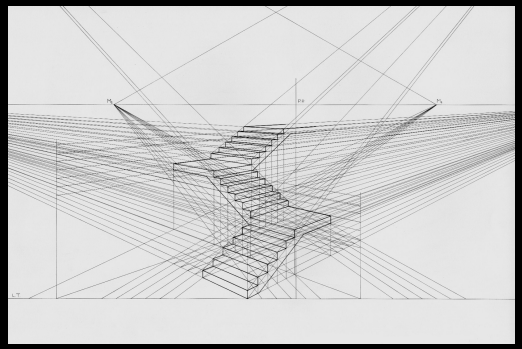


Fig. 4. The y-axis gives the ratio of the bootstrapped MPLE time to that of the MCMLE time. Values below 1 indicate that the bootstrapped MPLE requires a shorter computing time.

Structural reduction of the network (An, 2016)

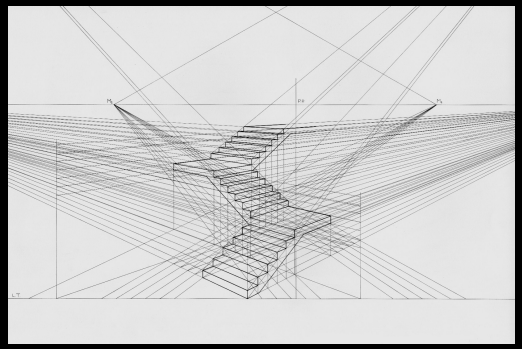


Structural reduction of the network (An, 2016)



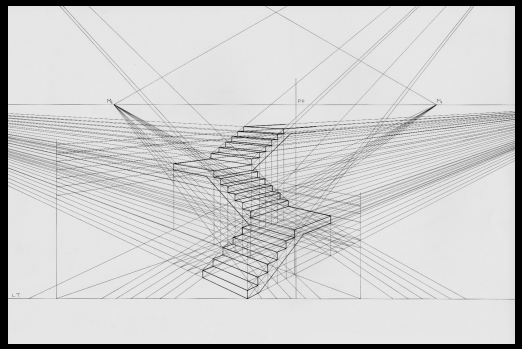
- Aggregating nodes

Structural reduction of the network (An, 2016)



- Aggregating nodes
- Model only core of the network

Structural reduction of the network (An, 2016)



- Aggregating nodes
- Model only core of the network
- Meta network analysis
  - Blocking
  - Bridging
  - Stacking

Conclusion

- Methods try to catch up with our requirements.
- Network analysis is not merely a technological attraction with no theoretical content. It is rather a necessity to advance theoretically.
- Big data is great and often network data, but challenges us in finding efficient estimation strategies.

References

An, W. (2016). Fitting ERGMs on big networks. *Social Science Research*, 59:107–119.

Box-Steffensmeier, J. M., Christenson, D. P., and Morgan, J. W. (2018). Modeling Unobserved Heterogeneity in Social Networks with the Frailty Exponential Random Graph Model. *Political Analysis*, 26(01):3–19.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55.

Caimo, A. and Friel, N. (2012). Bergm: Bayesian Exponential Random Graphs in R. *Journal of Statistical Software*, 61(2).

Calderhead, B. (2014). A general construction for parallelizing Metropolis-Hastings algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, 111(49):17408–17413.

Chandrasekhar, A. G. and Jackson, M. O. (2012). Tractable and Consistent Random Graph Models. *SSRN Electronic Journal*.

Cranmer, S. J., Leifeld, P., McClurg, S. D., and Rolfe, M. (2016). Navigating the Range of Statistical Tools for Inferential Network Analysis. *American Journal of Political Science*, 61(1):237–251.

Dorussen, H., Gartzke, E. A., and Westerwinter, O. (2016). Networked international