

ERGMs 2.1

RECENS / BCE PhD course

Lecturers: Takács Károly, Vörös András, Samu Flóra, Néray Bálint, Boda Zsófia

With thanks to Garry Robins, Paola Zappa, Vörös András and Boda Zsófia

MTA TK, Budapest 2016. április 27 – május 6.

A projekt az MTA TK „Lendület” RECENS Kutatócsoport támogatásával
valósult meg | <http://recens.tk.mta.hu>



What happened yesterday?

Interest in social networks(?)

- In social science, we might be interested in various types of social networks between various types of actors (e.g.: friendship among school kids ...)
- Then we can ask the question, why do certain students have more friends than others?
- In this case the network (or one of its properties) is the outcome variable of interest: the existence of the friendship tie
- (actor attributes/behaviour can also be explained by the network – you will learn about this later in the course)
- Given the data of friendship ties and certain attributes what method could you use to answer this question?

What's wrong with regression?

- ?

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),
 - Dyadic dependence,

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),
 - Dyadic dependence,
 - Markov dependence,

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),
 - Dyadic dependence,
 - Markov dependence,
 - Social circuit dependence and there is also

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),
 - Dyadic dependence,
 - Markov dependence,
 - Social circuit dependence and there is also
 - Higher-order dependence assumptions (Pattison et al, 2011)

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),
 - Dyadic dependence,
 - Markov dependence,
 - Social circuit dependence and there is also
 - Higher-order dependence assumptions (Pattison et al, 2011)
- What is the possible consequence not to deal with this conditional dependence?

What's wrong with regression?

- Regression (and non-network statistical models) typically assume that the observations are independent
- We have good theoretical reasons to believe that this is not true (e.g. ?)
 - Bernuolli dependence (independence),
 - Dyadic dependence,
 - Markov dependence,
 - Social circuit dependence and there is also
 - Higher-order dependence assumptions (Pattison et al, 2011)
- What is the possible consequence not to deal with this conditional dependence?
- Think about the value of the “Edge parameter” from different models ... unreliable estimates, likely over-estimation of parameters

How to deal with the dependence?

- We might not be interested in the specific structure of dependences between network ties – just want to get **reliable estimates** for some parameters by using **(logistic) regression**
- In this case we can **control for the lack of independence** by correcting **biased standard errors**
 - clustering of residuals in the model: dealing with heteroskedasticity
 - 2-way clustering method proposed by Lindgren (2010)
 - calculating heteroskedasticity-consistent standard errors (White, 1980)
- However, parameter estimates can be still biased due to **omitted variables**:
 - Those capturing network structure
 - Sometimes difficult, if at all possible, to include these in the model (especially higher order effects)

How to deal with the dependence?

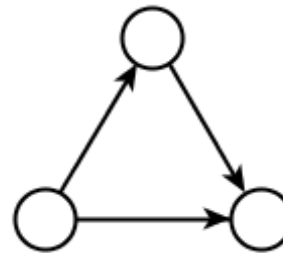
- Regardless of our scientific interest in the specific structure of dependences between network ties, we can **model this dependence** by using statistical models for social networks such as **ERGMs**, SOAMs – or MRQAP models (Krackhardt 1988, Dekker et al. 2007), Latent Space Models (see e.g. in Snijders 2011), Gravity Model (see e.g. in Broekel et al. 2014)
- How do we do this?
- Remember:
 - We would like to know the likelihood of a friendship tie
 - We have learnt that a model of independence (Bernoulli model) overestimates this likelihood
 - We have good reasons to believe that there are dependences between network ties on a dyadic level, or within triangles, sometimes in 4-cycles, or even within more complex structures
 - But **all we have is an empirical network** and **we do not know the theoretical model** that would help us describing the specific structure of dependences between network ties in the empirical network

How do we model then?

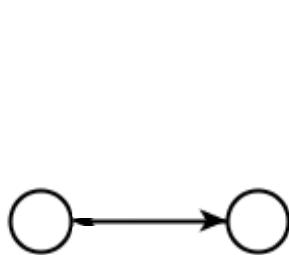
- We generate random network graphs that is similar to the observed one from the aspects that we think are important: e.g. size, density, proportion of reciprocated ties, transitive triangles (this is both a theoretical and empirical work and has to be validated later on)
- We assume that the observed network is a random draw from a population of networks described by our probability model
- This way we generate a null distribution for the observed statistic of interest
- Hence, we can count e.g. the number of triangles in each of the simulated networks – this gives us the null distribution and p value for the triangle count
- This gives us a clue about the importance (weight) of triangulation processes in the empirically observed network

Dealing with multiple processes

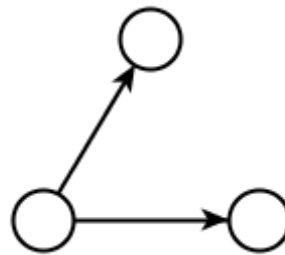
- There is one more thing to remember
- There is not only triangulation in the network, but **multiple processes operate simultaneously**
- The multiple dependence among network configurations is accounted for by a so called **dependence matrix** (see the book for more details, p.78).



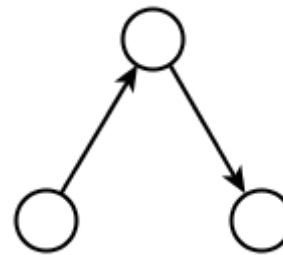
Transitive triad



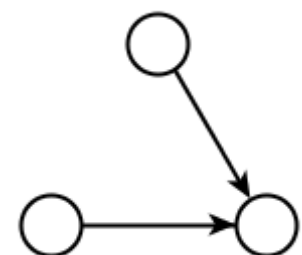
Arc



out-2-star



2-path



in-2-star

Extending multiple processes

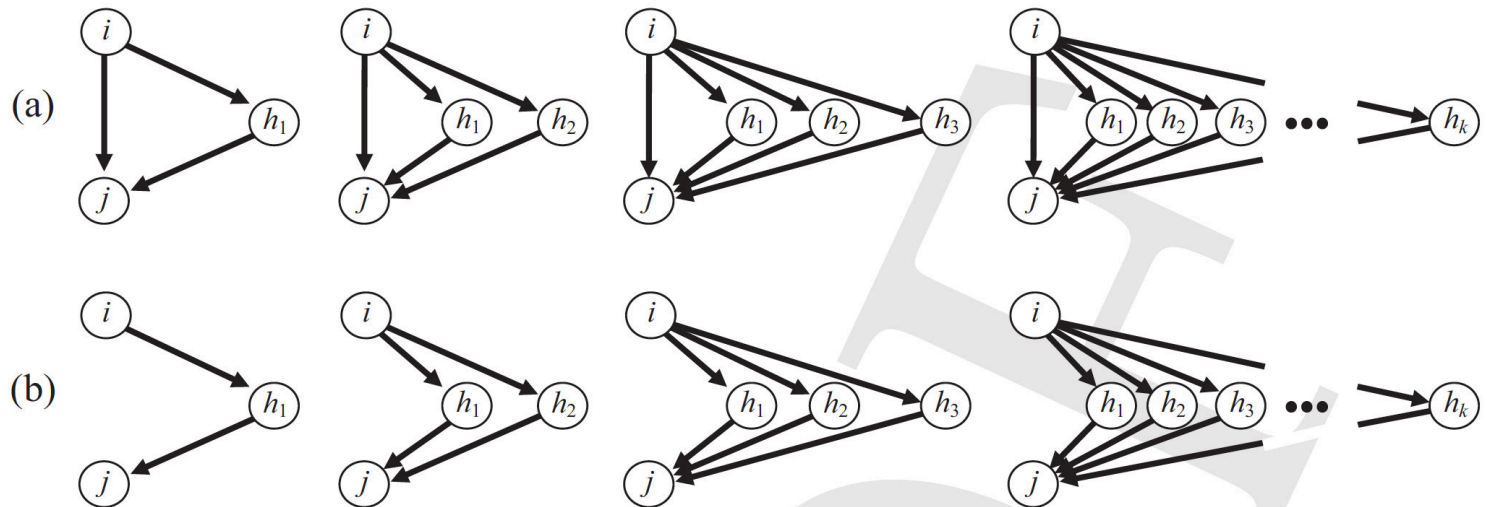


Figure 6.10. Configurations for directed graphs in alternating forms (a) AT-T and (b) A2P-T.

Alternating transitive triangles (AT-T) and alternating two-path effects (A2P-T)

Extending multiple processes

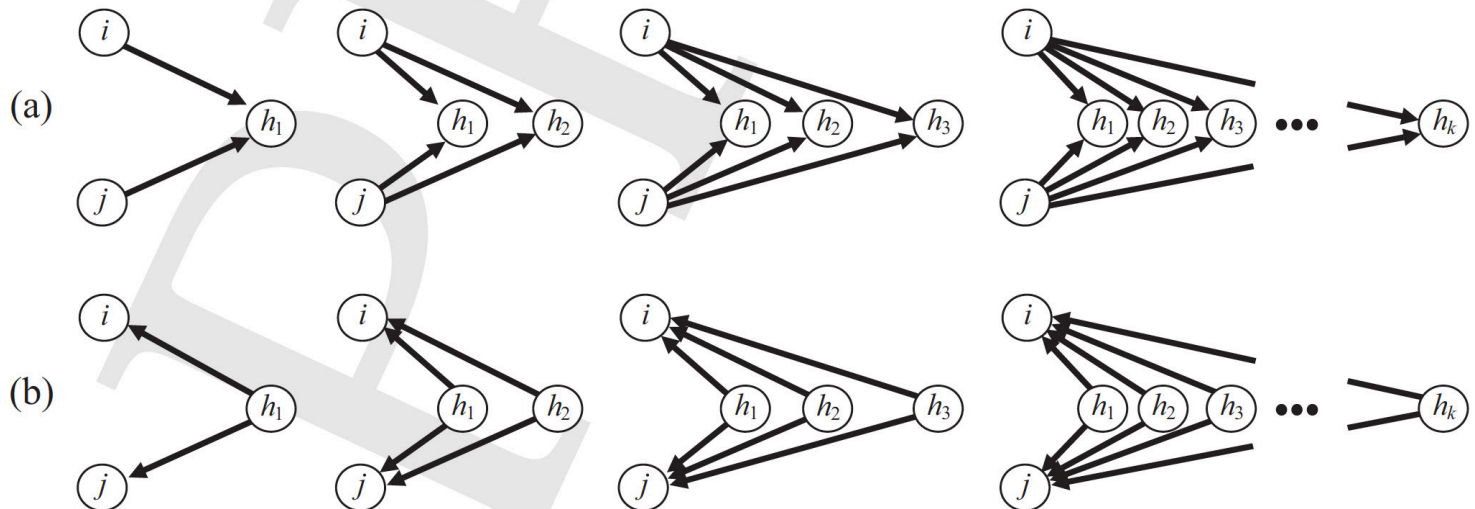


Figure 6.12. Additional 2-path configurations for directed graphs in alternating forms (a) A2P-U and (b) A2P-D.

Alternating two-path up (A2P-U) and alternating two-path down (A2P-D)

Extending multiple processes

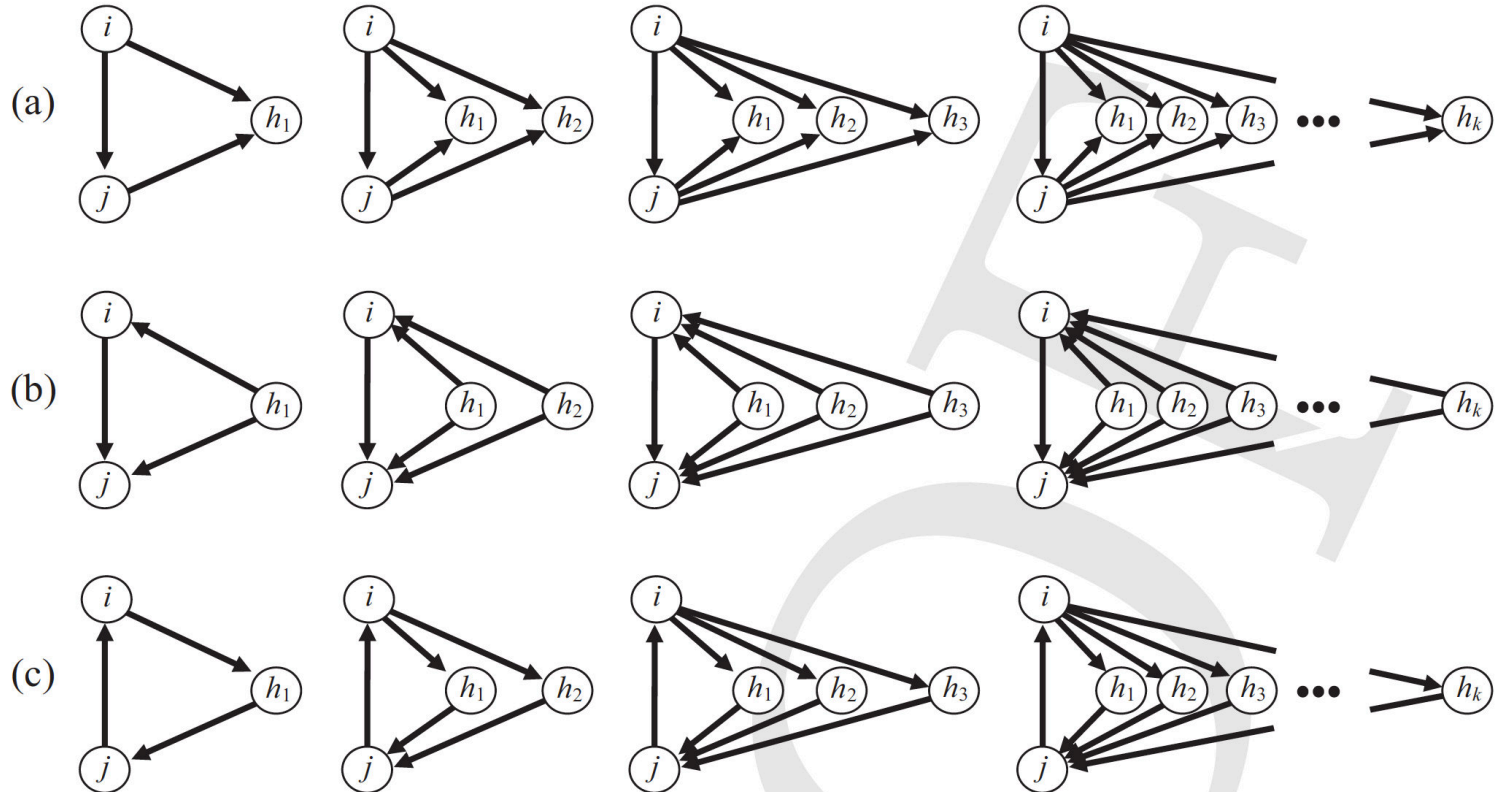


Figure 6.11. Additional triadic configurations for directed graphs in alternating forms (a) AT-U, (b) AT-D, and (c) AT-C.

Alternating triangle up (AT-U), alternating triangle down (AT-D) and alternating cyclic triangle (AT-C)

Making sense of parameters

- markov 2-path controls for the correlation between in and outdegree
- activity closure tells us whether 2 fans liking the same rock stars become friends
- popularity closure tells us whether 2 rock stars who are liked by the same fans become friends
- path closure (or transitive closure) leads to local hierarchisation

4. Technical details

Simulating random graph distribution

1. Start from a random graph (defined by our network configurations)
2. For each step, propose to change **one edge at a time** (random walk). If the probability of the graph increases, make the change, if the probability decreases, do not make the change)
3. Throw away the early iterations so the starting graph has no effect on the distribution – “**burn-in**”
4. **Sample as many graphs as needed** (e.g. every 1000th), controlled by the gaining factor
5. **Stop after a suitable number of iterations** that is controlled by the “multiplication factor”
6. **Change the parameter values** by comparing the distribution of graphs against the observed graph
7. **Repeat until** the parameter estimates stabilize: **convergence**
8. If hard to get convergence, try with **bigger multiplication factor** and **number of iterations**
9. If close to convergence, can use a **smaller gaining factor**

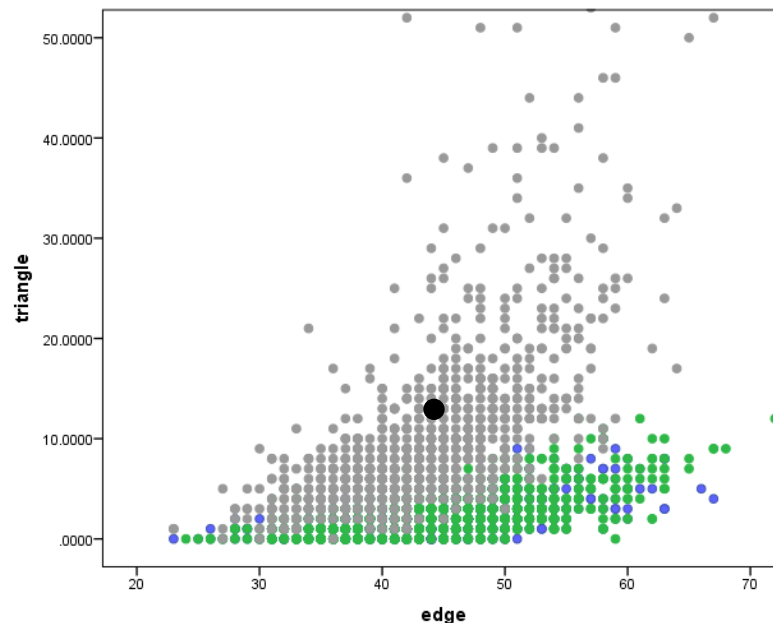
5. Interpreting the model

The logic of interpretation

- The **dependent variable** is the presence of a tie x_{ij} that is either present (1) or absent (0), hence, similarly to binary logistic regression, we are estimating a **binary outcome**
- The **constant** of the model is the **arc** (directed tie) or **edge** (undirected tie) parameter
- It is **conditionally dependent** on the other **network configurations** in the model that account for dependences among ties within the network, and (might be) conditional on **individual** and **dyadic attributes**
- Note that for purely structural (i.e., endogenous) network effects:
 - **Negative parameter** = less of such substructures (than expected by chance)
 - **Positive parameter** = more of such substructures (than expected by chance)

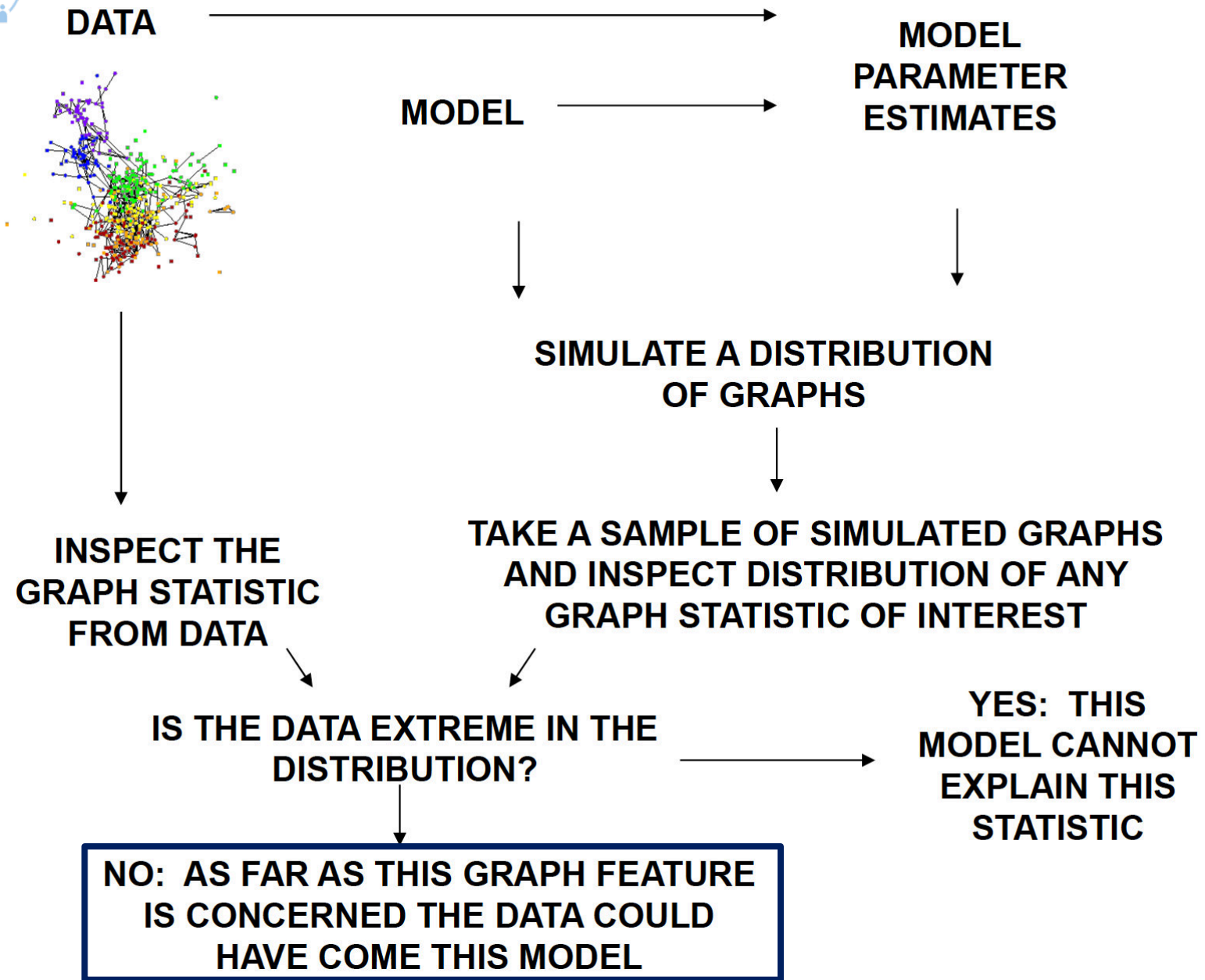
Parameter and model fit

- Parameter t-ratio should be below 0.1 which indicates the fit of the parameter
- The sample auto-correlation function (SACF) describes the correlation between values of the simulation process at different times. For better results we want this less than 0.5
- In order to get a smaller SACF you can increase the multiplication factor.
- Significance of the parameter is calculated from the parameter estimate and the SE



5. Goodness of fit (GoF)

Goodness of fit



Why do we need the GoF?

- If we are confident that our empirical data could have come from the model we used, we have to ask the question:
- Is this really the appropriate theoretical model that accounts for the important interdependencies in the network or are we missing something?
- In order to answer this question we have to check whether it is possible to improve the model?
- This is where GoF comes to the pictures

How do we do the GoF test?

- Again, we estimate parameters
- Simulate a distribution of graphs using these parameters
- This time, from the simulation, we collect graph statistics of any sort
- Compare the observed data with the collected statistics:
 - For all the included parameters in the model the, t-ratio should be under 0.2
 - For the parameters that are not included in the model, the t-ratio should be under 2.0
- Header in the GoF output: statistic's name, observed value, mean value, SD, t-ratio

Exercise 5

Exercise 5

- Fitting Bernoulli and Social Circuit models to
- the Fishermen's network
- And examining Goodness of Fit

6. Actor attributes: Social selection models

Social selection



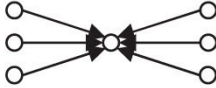
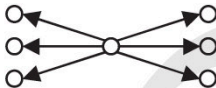
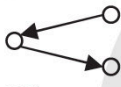
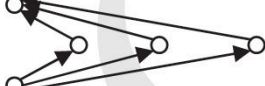


- Actors select network partners based on actor attributes
- An other process of tie formation
- Possible mechanisms:
 - **Homophily**: actors of similar attributes tend to form ties (McPherson et al, 2001).
 - Homophily in itself cannot explain the emergence of hierarchy in relations (so difference may also be important)
- Also actor main effects
 - **Sender effects**: Actors with certain attributes may send out more ties (more active or expansive)
 - **Receiver effects**: Actors with certain attributes may received more ties (more popular)

Dyadic covariates

- Some other relationship among nodes that could influence the network structure
- Examples:
 - Formal organisation structure
 - Geography
 - Another network








Example model

Purely structural effects (endogenous)

Arc		-1.96 (0.73)*
Reciprocity		2.88 (0.46)*
Popularity (in-degree)		-0.27 (0.32)
Activity (out-degree)		-0.34 (0.34)
Simple 2-path ³		-0.06 (0.08)
Multiple 2-paths		-0.06 (0.09)
Transitivity (transitive path closure of multiple 2-paths)		1.22 (0.19)*
Cyclic closure (cyclic closure of multiple 2-paths)		-0.37 (0.17)*

Actor relation effects (exogenous)

(black nodes indicates actor with attribute)

Sender (seniority)		-0.56 (0.29)
Sender (projects)		0.01 (0.02)
Receiver (seniority)		0.08 (0.23)
Receiver (projects)		-0.02 (0.02)
⁴ Homophily (seniority)		0.64 (0.26)*
Heterophily (projects)		-0.08 (0.02)*
Homophily (office)		-0.01 (0.17)

Covariate network (exogenous)

Advice entrainment (covariate arc)		1.76 (0.30)*
------------------------------------	--	--------------

Exercise 6

Exercise 6

- Estimating social circuit models on the communication network with
 - attributes, and
 - dyadic covariates
- attributes should be grouped based on their type (e.g. binary in one, categorical in an other, continuous again in a different one)
- if we put more attributes in one file we can separate them with tabs
- we have to tell MPNet how many attributes we have in one file
- if there are multiple dyadic variables in one file those have to be underneath each other and separated by tabs
- and finally: GoF

Exercise 7

Exercise 7

- Select a class from the RECENS data
- Select a network of your interest
- And at least 2 attributes that could, in your opinion, effect network formation processes

- Try to fit a Markov-model
- Run a GoF test

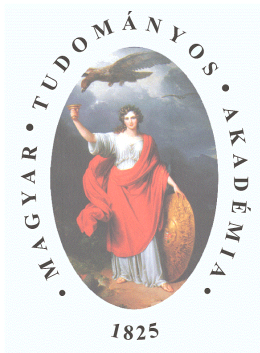
- Try to a SC model (even if the previous one was successful)
- Run a GoF test

If had more time ...

Additional topics

- Working with structural zeros
- Bipartite networks
- Multilevel networks
- ERGM in R

Thank you!



European
Research
Council

