Misspecification in statistical models for networks

Viviana Amati
Chair of Social networks, ETH Zurich

3rd THE PIIK Swiss Symposium on Network Science
Zurich, 2 October 2018

Statistical model for networks

Definition and use

A *statistical model for networks* is a mathematical expression that describes the process that is assumed to have generated the observed network data

The dependent variable is a network

It is used for making inference: testing hypotheses on the mechanisms that might have generated the network data

Statistical model for networks

Definition and use

A *statistical model for networks* is a mathematical expression that describes the process that is assumed to have generated the observed network data

The dependent variable is a network

It is used for making inference: testing hypotheses on the mechanisms that might have generated the network data

Statistical model for networks

Fundamentals of model specification

A model should

1. represent substantive theory for the analysed phenomenon
   (theory-driven)
2. represent research questions and tested hypotheses
   (no error of measurements in the explanatory variables)
3. be consistent with the observed data
   (model assumptions are matched)
4. be parsimonious
   "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances." Occam's razor

Statistical model for networks

Misspecification

When a model does not fulfil one of the previous principles, the model is *misspecified*

Issues of model misspecification have received so far scant attention

Little is known about the robustness and sensitivity of network models to misspecification

A few methods are available for detecting some forms of misspecification

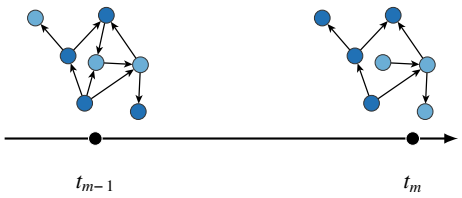
We focus on misspecification in stochastic actor-oriented models

Intermezzo: Stochastic actor-oriented models

SAOMs

Model for panel network data:
repeated observations of a network (X) and a behaviour (Z) over time

$$(x, z)_{t_0}, (x, z)_{t_1}, \quad \ldots \quad , (x, z)_{t_M}$$



Intermezzo: Stochastic actor-oriented model

The model is a continuous-time Markov chain assuming that the observed networks is the outcome of a sequence of micro-steps:

at each micro-step one actor randomly gets an opportunity to make a change
$\Longrightarrow$ modelled by the rate function

$$\lambda_i(x, z; \alpha) = \rho^{[V]} \quad V = X, Z$$

The selected actor chooses the "most rewarding" change
$\Longrightarrow$ modelled by a multinomial model

$$p^{[X]}_{(x',z)} = \sum_{x''} \frac{\exp\left(f_i^{[X]}(\beta^{[X]}, x', z)\right)}{\exp(f_i^{[X]}(\beta^{[X]}, x'', z))} \text{ and } p^{[Z]}_{(x,z')} = \sum_{z''} \frac{\exp\left(f_i^{[Z]}(\beta^{[Z]}, x, z')\right)}{\exp\left(f_i^{[Z]}(\beta^{[Z]}, x, z'')\right)}.$$

where

$$f_i^{[V]}(x, z; \beta) = \sum_k \beta_k^{[V]} s_{ik}^{[V]}(x, z; \beta) \quad V = X, Z$$

is called evaluation function

Misspecification of SAOMs

A model should

1. represent substantive theory for the analysed phenomenon

   Omission of a relevant effect in the evaluation function
   (partially addressed in Wang (2007))
   Inclusion of an irrelevant effect in the evaluation function

2. represent research questions and tested hypotheses

   "Right" choice of the effects

3. be consistent with the observed data

   Assumptions of time heterogeneity (addressed in Lospinoso et al. (2011))
   Assumptions of multinomial logit models

4. be parsimonious

   Overfitting model and model selection
   (partially addressed in Schweinberger (2012))

Misspecification of SAOMs

A model should

1. represent substantive theory for the analysed phenomenon

   Omission of a relevant effect in the evaluation function
   (partially addressed in Wang (2007))
   Inclusion of an irrelevant effect in the evaluation function

2. represent research questions and tested hypotheses

   "Right" choice of the effects

3. be consistent with the observed data

   Assumptions of time heterogeneity (addressed in Lospinoso et al. (2011))
   Assumptions of multinomial logit models

4. be parsimonious

   Overfitting model and model selection
   (partially addressed in Schweinberger (2012))

Investigating the consequences of misspecification

Investigation based on simulations

Consider a fully specified SAOMs
⇒ "true model"

Given a form of misspecification, define a misspecified model
⇒ "postulated model"

Generate a large number of network panel data

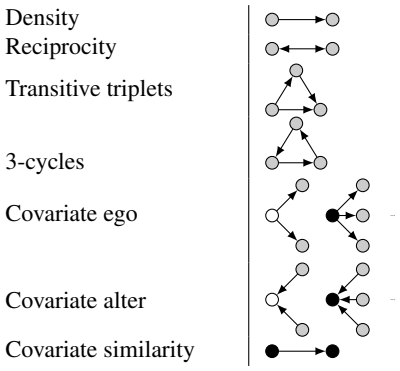
Estimate the postulated model for each simulated panel data

Compare the estimates of the postulated model with those
of the true model

Omitting a relevant variable

Simulation:example

True Model:



Postulated model: model without the covariate similarity

Omitting a relevant variable

Consequences: example (MoM)

	true	mean	power	bias
rate p.1	4.0	3.911	0.062	1.000
rate p.2	5.0	4.935	0.083	1.000
Density	-2.0	-1.970	0.062	1.000
Reciprocity	2.0	2.012	0.062	1.000
Transitive triplets	0.3	0.268	0.146	0.646
3-cycles	-0.3	-0.326	0.042	0.250
Covariate ego	0.1	-0.094	0.354	0.146
Covariate alter	-0.1	0.081	0.229	0.062
Covariate similarity	0.4	omitted variable		

Omitting a relevant variable

Consequences

Omitting a relevant variable affects

- the estimate of the parameters
- the significance of the parameters

that are strongly correlated to the omitted variable

GMoM and ML estimators are affected in a similar way

MoM and GMoM are less sensitive to model misspecification than the ML

Detection (under investigation): correlation between residuals and covariate

Model misspecification in SAOMs

Choice among the effects for modeling selection and influence

Selection statistics

- Covariate Similarity
- Covariate difference (and its variants)
- Covariate ego × Covariate alter

Influence statistics

- Total similarity
- Average similarity
- Total alter
- Average alter
- Maximum alter
- Minimum alter

"Good theory is by far the most important guide we have in avoiding misspecification. But given that social science theories are often weak or tentative, empirical tests of proper specification become more important."

(Lewis-Becket al. (2003). The Sage encyclopedia of social science research methods)

Choice among the effects

Simulation

Average alter vs. total alter

	true	mean	test bias	power
Net. rate p1	6.50	6.40	1.00	0.02
Net. rate p2	7.00	6.99	1.00	0.04
Outdegree	-2.10	-2.09	1.00	0.01
Reciprocity	1.90	1.90	1.00	0.05
Trans. triplets	0.30	0.30	1.00	0.04
3-cycles	-0.30	-0.31	0.93	0.05
Beh. similarity	1.50	1.53	0.99	0.01
Beh. rate p1	1.00	0.92	0.87	0.15
Beh. rate p2	1.50	1.77	0.67	0.11
Linear	-0.20	-0.23	0.05	0.01
Quadratic	0.30	0.35	0.39	0.01
Average alter	0.60			
Total Alter		0.04	0.00	0.99

Choice among the effects

Consequences

Similar results for GMoM and MLE

Similar results for other pairs of influence or selection effects

The “postulated” statistic highly correlates with the true statistic

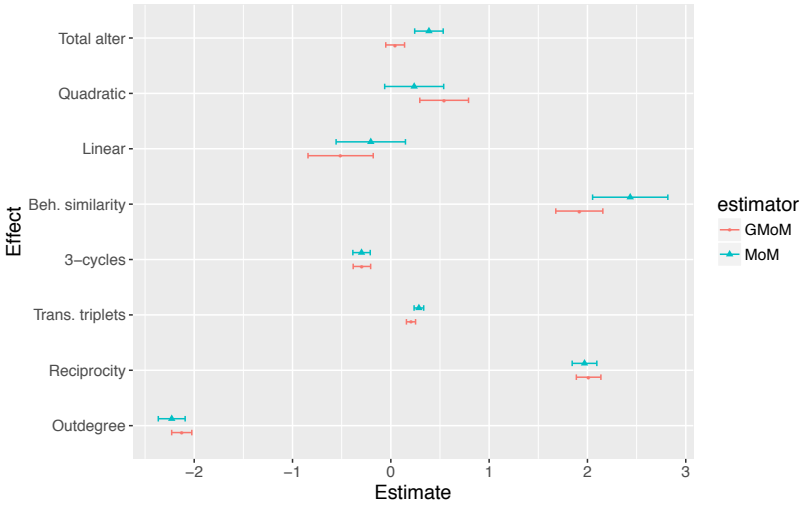
Those results were obtained on a small network (30 actors)

For larger networks (100 nodes) and more waves:

- biased parameters in both the selection and inference part
- different estimation method lead to differences in the estimated parameters

Choice among the effects

Detecting misspecification: An example



Choice among the effects

Detecting misspecification

Comparison of two estimators based on two sets of moment conditions
(which may or may not include elements in common)

Hypotheses

H_0: correct model specification vs. H_1: model misspecification

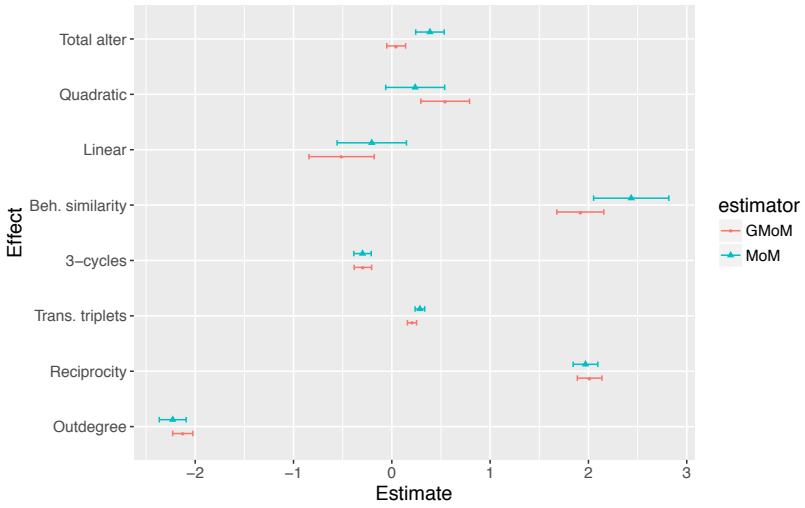
Test statistic

$$H = (B_2 - B_1)^T (\Sigma_1 - \Sigma_2)^{\dagger} (B_2 - B_1) \sim \chi^2_k$$

Large values of the stastics provide evidence against H_0

Choice among the effects

Detecting misspecification: An example



power 0.82

Misspecification of Network Models

Like any other statistical model, statistical models for networks are affected by model misspecifications:
omission of relevant variables, measurement error and model assumptions

In SAOMs

- omission of relevant variables lead to estimates that are biased and low power of the significance test
- Measurement errors for selection and influence process does not affect the estimates
  problem of network size?

Methods for detecting misspecification borrowed from standard statistical models and econometric theory

References

Lospinoso, J. A., Schweinberger, M., Snijders, T. A., and Ripley, R. M. (2011). Assessing and accounting for time heterogeneity in stochastic actor oriented models. *Advances in data analysis and classification*, 5(2):147–176.

Schweinberger, M. (2012). Statistical modelling of network panel data: Goodness of fit. *British Journal of Mathematical and Statistical Psychology*, 65(2):263–281.

Wang, J. (2007). Simulation studies of power and robustness in models for network dynamics. *MSc dissertation*.