Treatment of Missing Data in Longitudinal Network Studies

Mark Huisman

University of Groningen

SIENA workshop, January 12, 2009

Outline

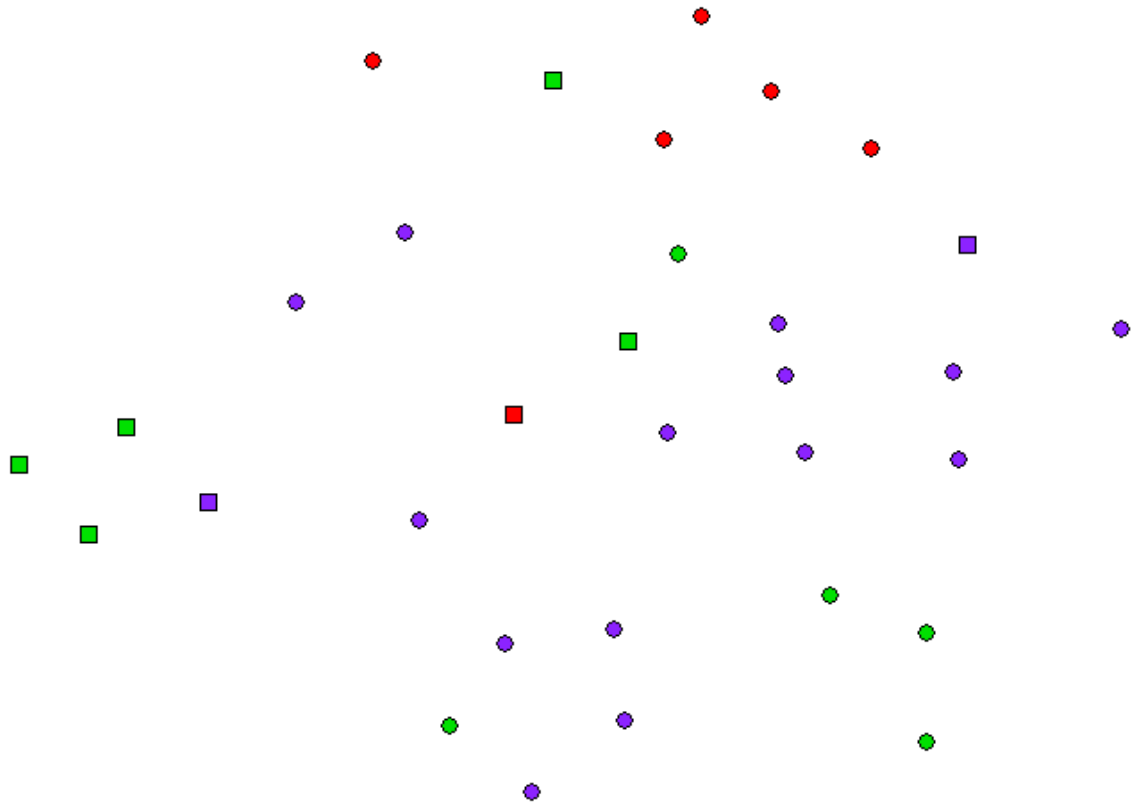
1. Example - friendship network
2. Missing data in social networks
   - effects on network structure
   - longitudinal network data
   - treatments?
3. Missing network panel data
4. Simulation studies
   - Composition changes *vs.* missing data
   - Simple missing data procedures

1. Example data

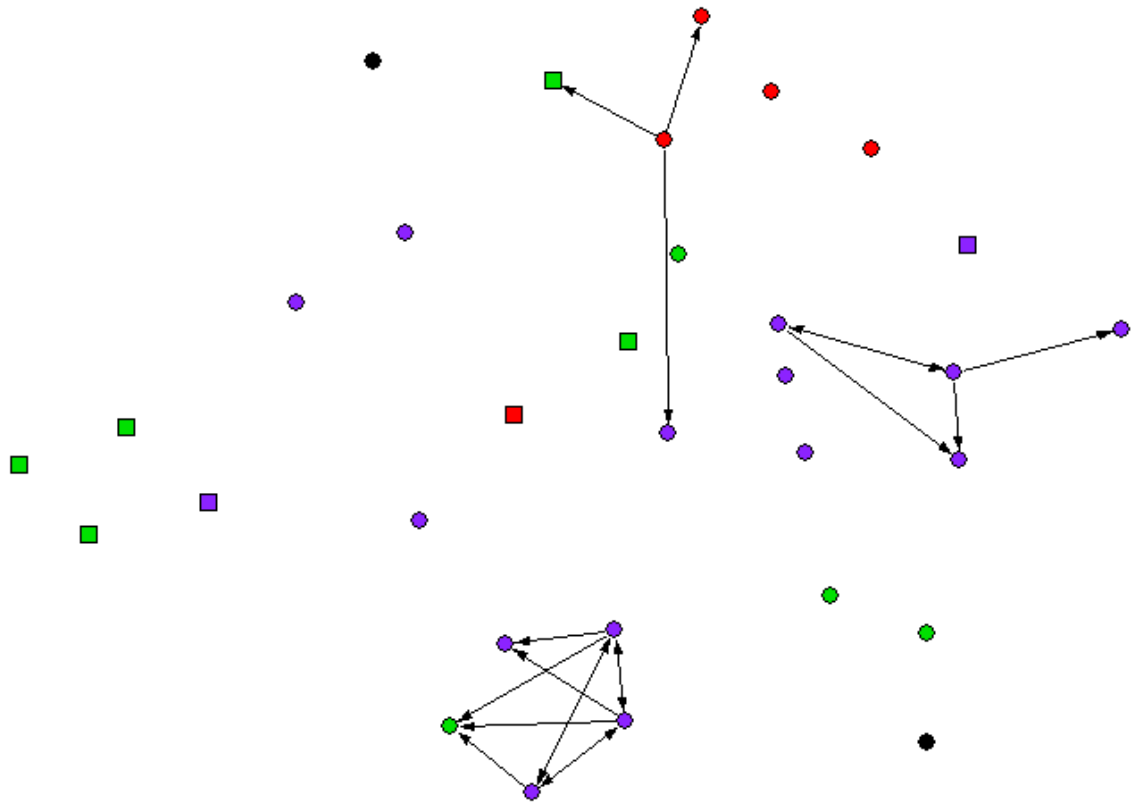
Van de Bunt's friendship network

- 32 actors: freshman university students
- Tie = *friendship*:
  (best) friend *vs.* friendly relation/known/unknown
- Actor covariates: program (color), gender (shape), smoking
- 7 measurements, several weeks/months apart

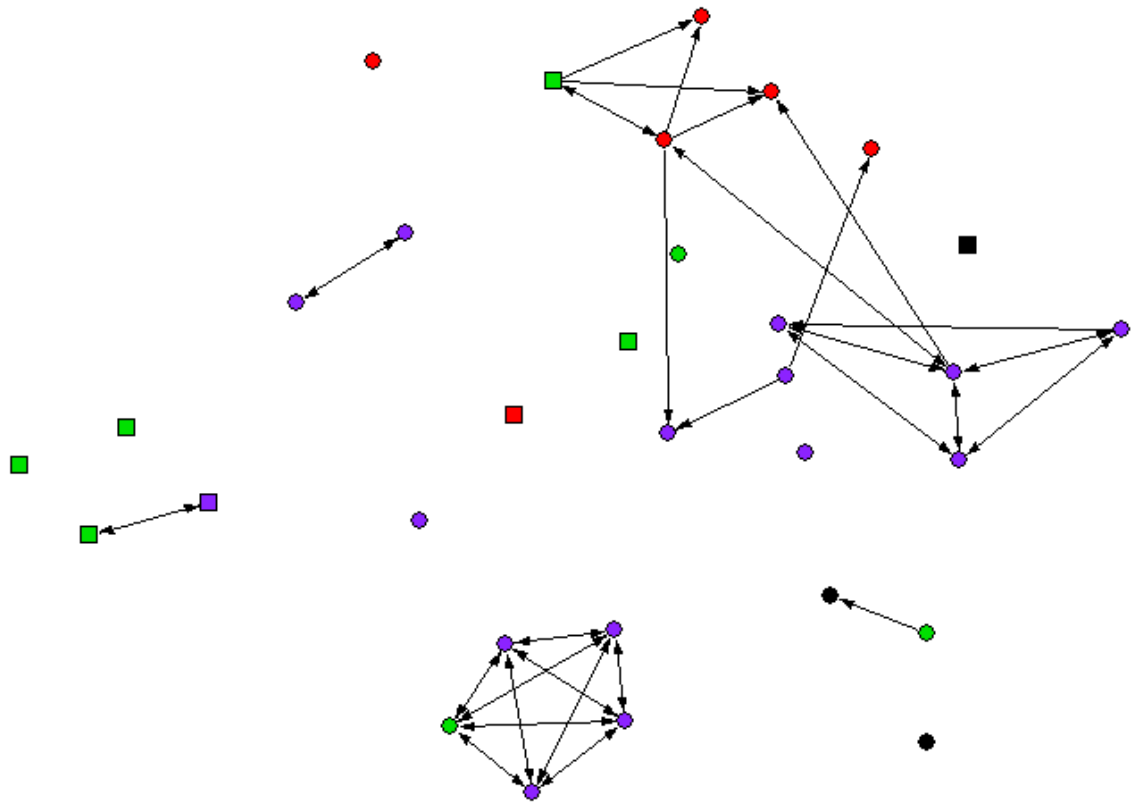
Van de Bunt, Van Duijn, & Snijders (1999), *Computational & Mathematical Organization Theory*, *5*, 167–192.



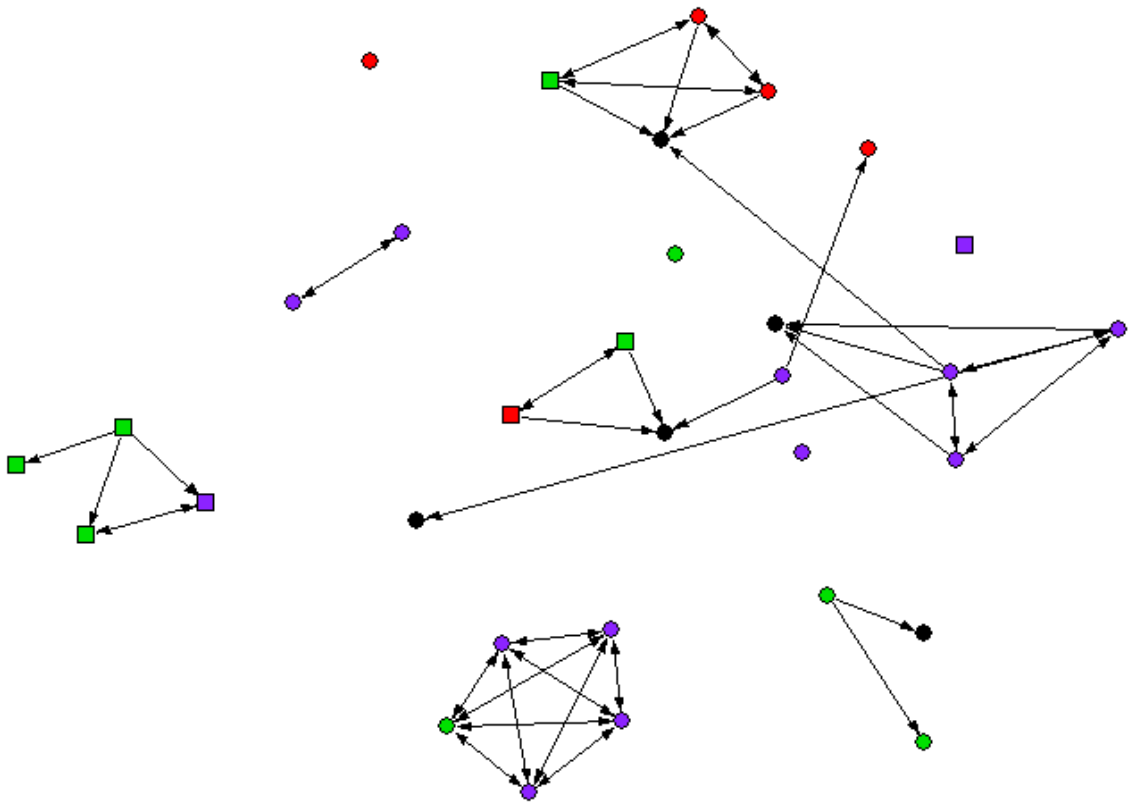
Friendship network time 0
Average degree 0.0, missing fraction 0.00



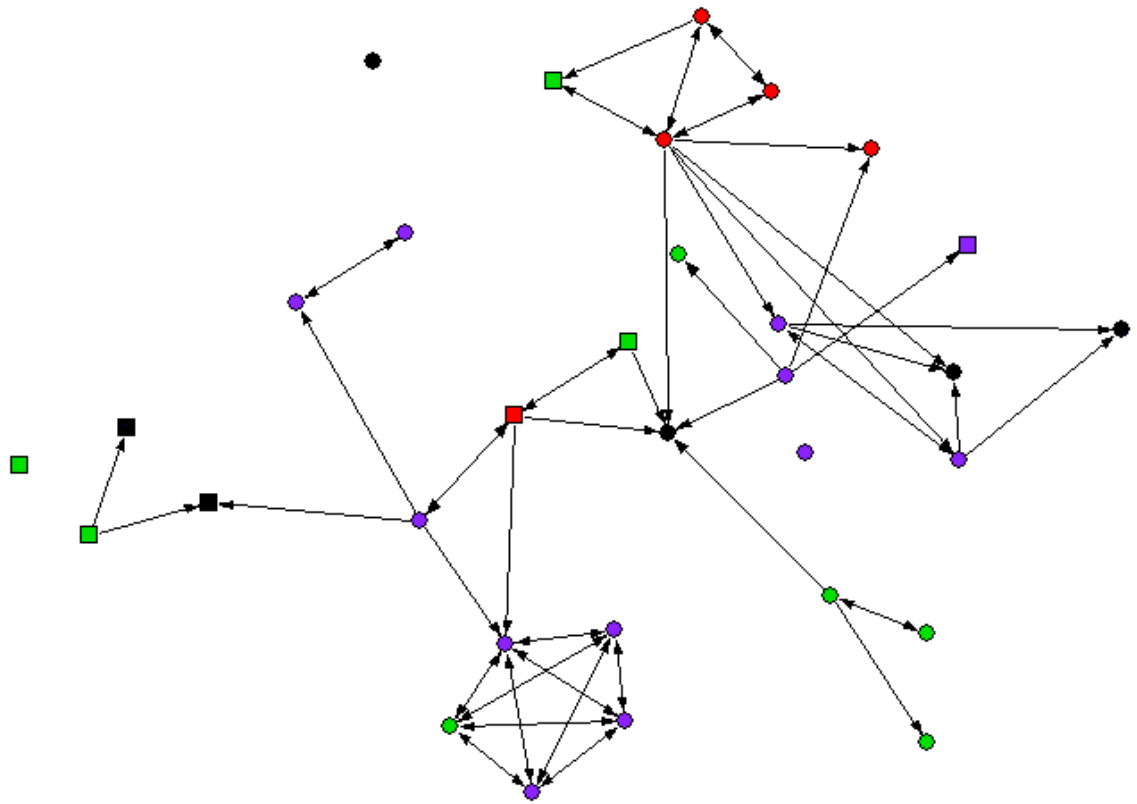
Friendship network time 1
Average degree 0.6, missing fraction 0.06



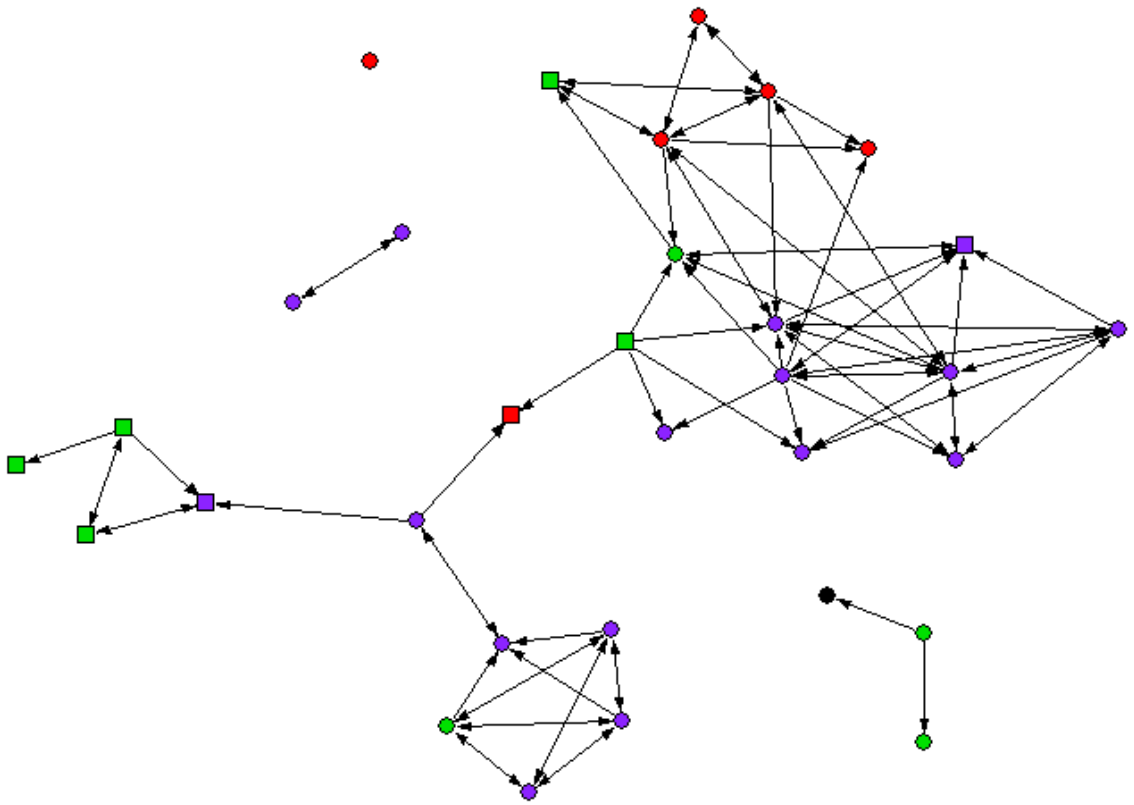
Friendship network time 2
Average degree 1.6, missing fraction 0.09



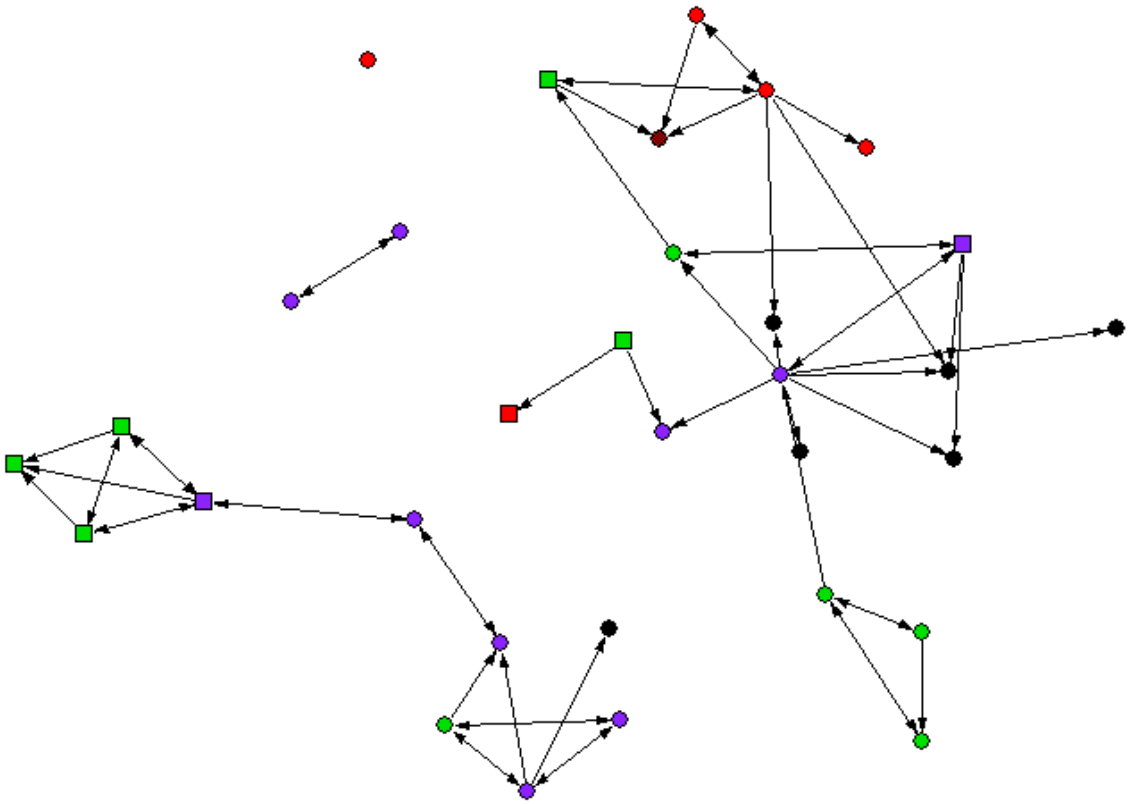
Friendship network time 3
Average degree 2.0, missing fraction 0.16



Friendship network time 4
Average degree 2.4, missing fraction 0.19



Friendship network time 5
Average degree 2.8, missing fraction 0.04



Friendship network time 6
Average degree 2.2, missing fraction 0.22

Statistical modeling

Purpose: To investigate network evolution as function of

1. structural (network) effects like reciprocity or transitivity
2. explanatory actor variables like program, gender, smoking behavior
3. (possibly explanatory dyadic variables)

All effects control for each other effect

By controlling adequately for structural effects, it is possible to test hypothesized effects of variables on network dynamics (without such control these tests would be unreliable)

2. Missing Data in Networks

"*Missing data are a doubly curse to survey network analysis.*"

(Burt, 1987)

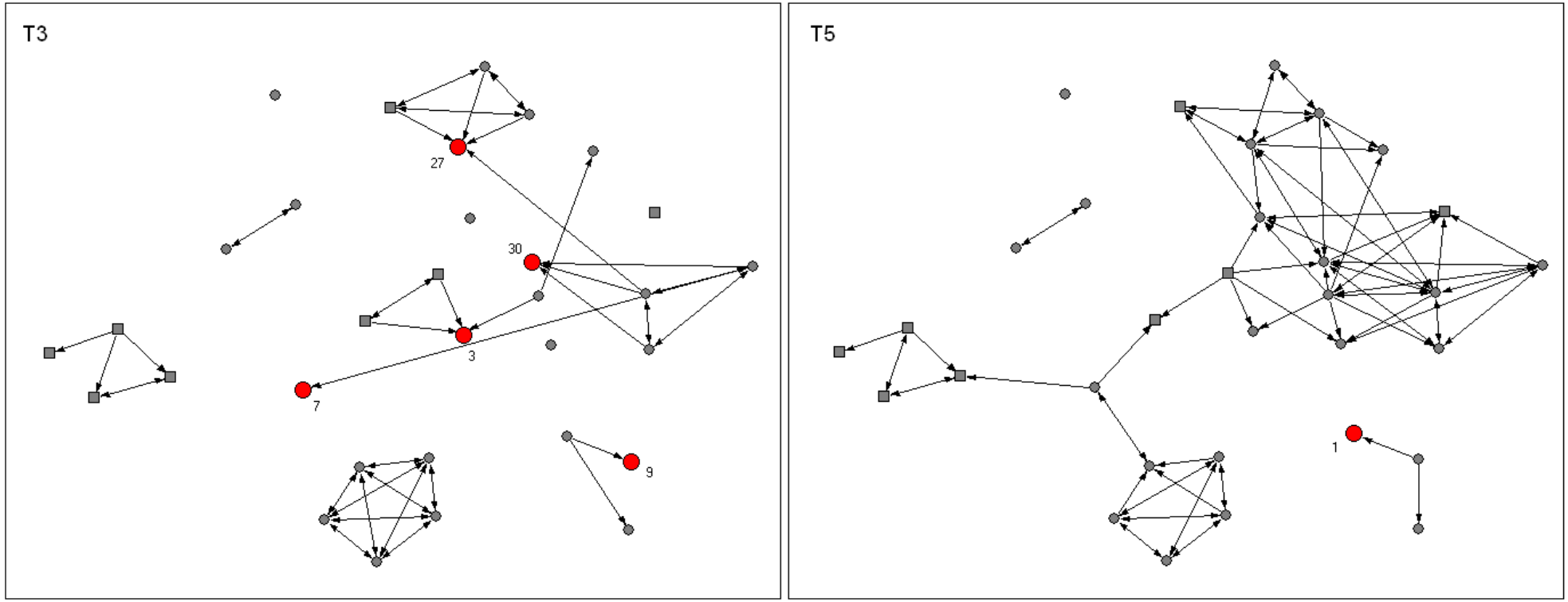
1. Network items *complex* and network surveys seem more likely to generate missing data
2. Network analysis is especially *sensitive* to missing data

If a network *tie*, or worse, an *actor* is missing, there is

- limited capacity to describe the network context of the actors whose ties are missing
- lack of information on the network context of neighboring actors

Third curse?

Longitudinal data *especially sensitive* to missing data



Freshmen friendship network (van de Bunt, 1999).

Types and Mechanisms

- *unit non-respons*: actor does not participate
- *item non-response*: particular outgoing ties are unavailable

Are missing network data randomly missing? The well-known definitions of Rubin should be applied here: *MCAR*, *MAR*, *MNAR*

Is missingness related to *network* and/or *actor characteristics*?

* type of relation: friendship, trust, bullying, trade, etc.
* postion of actors in the network, network patterns
* strength of the ties
* actor attributes
* ...

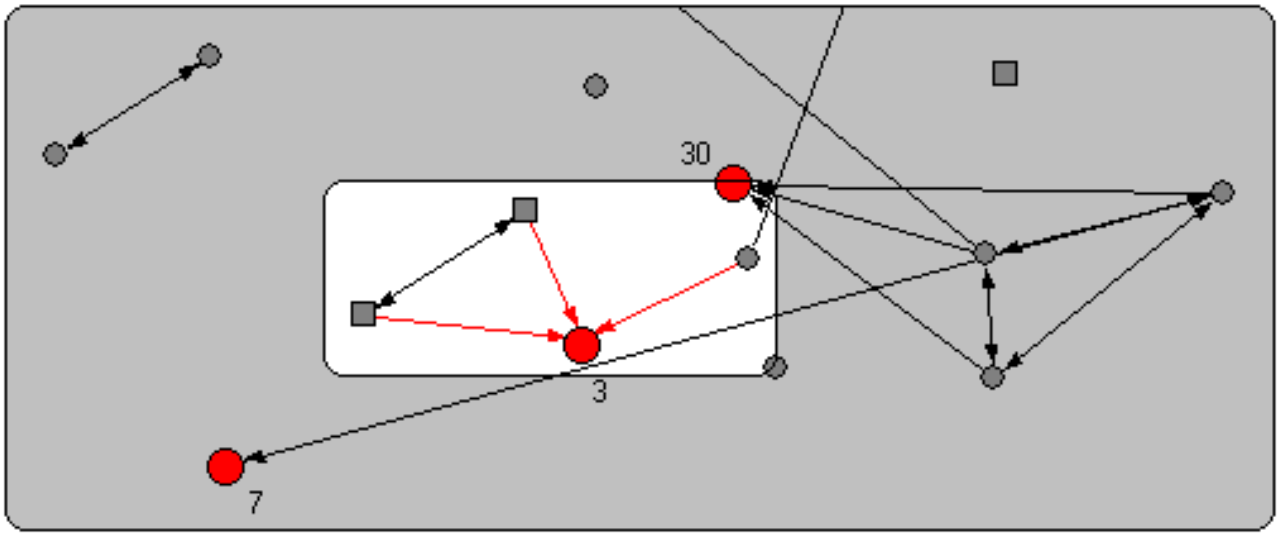
Effects

Negative effects of missing data on network structure:

- *Strength of relations* is underestimated (Burt, 1987)
- *Degrees* are underestimated (Kossinets, 2006; Costenbader & Valente, 2003; Huisman, 2009)
- *Closure* is underestimated (Kossinets 2006; Huisman, 2009)
- *Centrality measures* become unstable (Costenbader & Valente, 2003)

Some of the studied measures are reasonably robust for *small* proportions of missing data, especially those based on *incoming ties* (in-degrees)

Advantage of networks:



⇒ Measures based on incoming ties are robust

⇒ 'Repair' missing outgoing ties with reported incoming ties? *reconstruction* (Stork & Richards, 1992)

⇒ Directed or undirected graphs?

Missing network panel data

Longitudinal network data suffers from actors missing at one or more observation moments

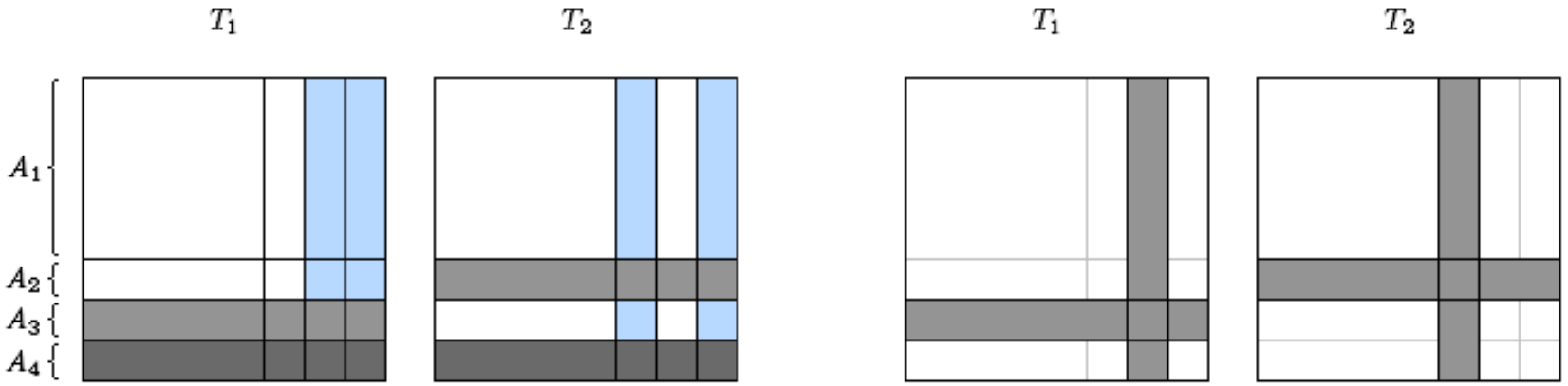
Types

- *unit non-respons*: actor does not participate
- *item non-response*: particular outgoing ties are unavailable
- *wave non-response*: time dependency (unit non-response on one wave), *panel mortality*, *attrition*, *drop-out*
- *network composition change*: actors join or leave the network

Effects

Effects on network dynamics: large (Huisman & Snijders, 2003; Huisman & Steglich, 2008)

Patterns of missingness (two waves T_1 and T_2)



Set of Actors:

A_1 observed at both time points

A_2 observed at T_1, missing at T_2 (left) or leaving (right)

A_3 missing at T_1, observed at T_2 (left) or joining (right)

A_4 missing at both time points (left: dark grey)

Treatments

Burt (1987): missingness is associated with strength of ties.
$\Rightarrow$ "*The implication is that the missing network data can be replaced with quantitative data indicating a weak relation.*"

Known treatments for single networks

- *Complete case analysis*: analyze ties of observed actors only
- *Impute zeros*: treat missing ties as absent
- *Impute by reconstruction* (Stork & Richards, 1992):
  replace missing ties with incoming ties of missing actors
  (needs imputation of relations between missing actors)

- *Hot Deck Imputation* (Burt, 1987; Goldstein, 1999): replace missing ties with ties of donor actors
- *Imputation* of missing ties using *latent space models* of network structures (Ward, Hoff, & Lofdahl, 2003)
- *Model-based procedures* using all available data based on *exponential random graph models* (Robins, Pattison, & Woolcock, 2004; Gile & Handcock, 2006; Handcock & Gile, 2007; Koskinen, 2007)

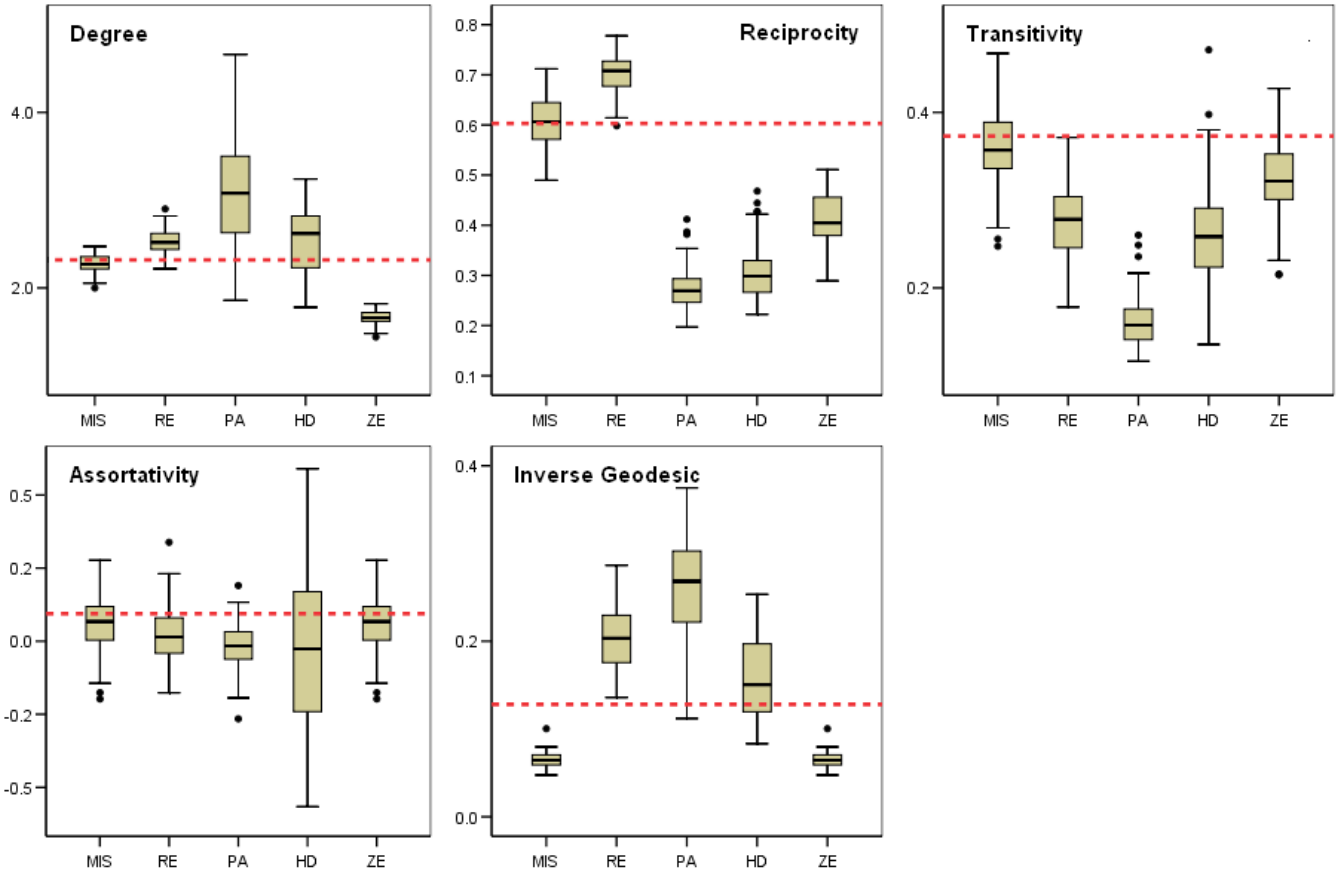
Simple imputation methods

Huisman (2009) investigated the effect of non-response and imputation on network properties:
outdegree, *reciprocity*, *transitivity*, *assortativity*,
and *geodesic distance*

- *Treat missing ties as absent*: impute 0's
- *Imputation based on density*
- *Reconstruction* (Stork & Richards, 1992)
- *Imputation based on degrees*: preferential attachment
- *Hot Deck imputation*

Conclusions (Huisman, 2009)

- *Ignoring missing data* can have a large effect on descriptive analysis of social networks
- Effects *unit and item non-response* are similar except for mean degree
- Effects for *undirected and directed networks* are similar
- Missingness mechanisms show different effects
- *Ignoring missing data* gives at least as good and often better results than naive imputation
- *Reconstruction* is 'best-of-the-rest' imputation method
- *Simple imputation methods are not very successful*



Unit non-response, and item-nonresponse related to outdegree and covariate

Missing fraction 0.36

Missing network panel data

Within the framework of *dynamic network models*

- *Imputation procedure* implemented in SIENA software:
  can handle missing actors as well as individual ties
  (Huisman & Steglich, 2008)
- *Procedure for network composition change*:
  analysis based on available cases and imputation
  For missingness due to actors joining or leaving the network
  (Huisman & Snijders, 2003)

Simulation studies to compare the procedures with other
(ad hoc) procedures

3. Missing network panel data

How to treat missing longitudinal network data?

Subquestions

1. **How to model observed longitudinal network data?**
   How to model network evolution? $\Rightarrow$ SIENA
2. How to treat incomplete data due to *wave non-response*?
3. How to treat incomplete data due to *composition change*?
4. Can we use (other) simple missing data treatments?

Actor-driven models for network evolution
(Snijders, 1996, 2001, 2005)

Data: *repeated measurements of a social network*

Principles: Regard the observations as discrete observations of a process *developing in continuous time*, where actors can make *unobserved changes* between the observation moments, being each others' changing environment

Model the data by construction of a *continuous-time Markov chain* of so-called *micro steps*

Micro steps

Model the process as a series of *micro steps* taken by the actors:
a sequence of small unobserved changes resulting in
differences between two observed networks

Condition on the first observation and refrain from modeling it

The micro steps consist of the change of one tie variable Y_{ij}
between two actors i and j
(on $\Rightarrow$ off, off $\Rightarrow$ on, or leave unchanged)

This change is modeled as maximization by actor i of his/her
evaluation function plus a random component

Model specification

The *evaluation function* is a weighted sum of network and covariate effects

$$f_i(\beta, x) = \sum_{k=1}^{L} \beta_k \, s_{ik}(x) \,,$$

where the weights β_k are statistical parameters indicating the strength of effect $s_{ik}(x)$.

Effects are *structural effects* (endogenous):
(e.g., outdegree, reciprocity, transitivity, distance 2)
and *covariate effects* (exogenous)

Simulation of networks

The distribution of *waiting times* between consecutive changes
is modeled using a *rate function* λ_i:
indicating the rate at which an actor may take a micro step

The rate and objective function together define
the *continuous-time Markov chain* of micro steps

Starting from the first observation,
given some estimated values of the parameters,
a *Markov chain of networks* can be simulated

The model is estimated with the *method of moments*,
based on the simulation of Markov chains

Stochastic actor-oriented models

Each actor "controls" his/her outgoing relations

At any given moment, with a given current network structure,
the actors act independently (i.e. take *micro steps*),
one-at-a-time

No strategy: objective function reflects
short-term goals, opportunities, constraints

The stochastic moments are governed by the *rate function*
The changes (micro-steps) are governed by the
objective function

How to treat missing longitudinal network data?

Subquestions

1. How to model observed longitudinal network data?
   $\Rightarrow$ Simulate the evolution of the network
2. **How to treat incomplete data due to wave non-response**?
3. **How to treat incomplete data due to composition change**?
   $\Rightarrow$ *Use simulation of the evolutionary process*: Markov chains
4. Can we use (other) simple missing data treatments?

Wave non-response: imputation

Initial imputations at T_1: replace missing ties with *zeros*
(assume there is *no tie*; modal value for sparse networks)

In the *simulation phase* of the estimation process **all** *actors* are allowed to take micro steps

$\Rightarrow$ the imputed ties may change
$\Rightarrow$ missing actors have *indirect influence* on network evolution

At T_2 the missing ties are not replaced

Parameter estimation is based on the actors observed at both observation moments T_1 and T_2, ignoring the missing actors

For more than two observations moments
initial imputations are based on *last observation carry forward*
if earlier observation is present (otherwise impute zero's)

For dependent *behavior variables* same principle is used:
impute either *last observation* or *mode of observed values*

For missing covariate scores *average values* at the current
observation moment are imputed

Network composition change

Two ways:

- using structural zeros
- method of Huisman & Snijders (2003)

Structural zeros

Some values in the digraph are *structurally determined*

Structural zero: it is certain that there is no tie from i to j

Relations of joiners and leavers can given structural zeros

Use reserved codes in input data file: 10 for structural zeros

Huisman-Snijders method

Not for maximum likelihood estimation

Joining and/or leaving of actors are treated as *exogeneous events* in the simulation of micro steps

It is handled separately from missing data treatment

Needs specification of *times of composition change*

Example times of change

5 observation moments

5 actors

1	5		
2.6	5		
1	3.4		
1.7	4.2		
1	2.6	3.8	5

Simulation of micro steps

The specification of the rate function λ_i
and the evaluation function f_i define
the *continuous-time Markov chain* of micro steps in (t_m, t_{m+1}):

- all actors present at t_m and/or t_{m+1} are included
- exogenous events occur at *fixed time points* t_C:
  times of change
- Markov chain starts anew from last simulated state before t_C:
  only actors present in current state of network can change their relations

Times of change are expressed as a fraction of the length of the period: divides the period in several parts with different sets of *active actors*

Example times of change

5 observation moments
5 actors

1	5		
2.6	5		
1	3.4		
1.7	4.2		
1	2.6	3.8	5

Before joining

Ties are *fixed* on 0 (use missing value code), or observed relations are used (prior information)

After leaving

Ties are *fixed* on last observed value (use missing value code), or observed relations are used (additional information)

Use missing value codes, however, ties of joiners/leavers are not treated as regular missings

4. Simulation Studies

How to treat missing longitudinal network data?

Subquestions

1. How to model observed longitudinal network data?
   $\Rightarrow$ Simulate the evolution of the network
2. How to treat incomplete data due to *wave non-response*?
   $\Rightarrow$ Imputation in the simulation of Markov chains
3. How to treat incomplete data due to *composition change*?
   $\Rightarrow$ Use structural zeros
   $\Rightarrow$ Model changes as exogeneous events
4. **Can we use (other) simple missing data treatments?**
   $\Rightarrow$ In some situations yes, but *generally NO*

Missing data or composition change

Small simulation study by Huisman & Snijders (2003):
missing data *vs.* composition change

- Differences in parameter estimates: extra information
- Composition change has higher change rates
- Differences are largest for covariate effects
- Different sets of times of change give different estimates

Simple missing data procedures

Simulation study by Huisman & Steglich (2008)

1. *Complete case analysis* (*listwise deletion*: ignore incoming ties of missing actors)
2. Imputation by *reconstruction* (Stork & Richards, 1992)
3. Imputation based on *preferential attachment* on indegree
4. Imputation procedure implemented in SIENA

Ad hoc imputation methods

1. *Reconstruction* (Stork & Richards, 1992; Huisman, 2009)
   - impute *incoming ties* for missing outgoing ties
   - randomly impute of ties between missing actors proportional to (observed) density

2. *Preferential attachment* based on indegree (Barabasi & Albert, 1999; Huisman, 2009):

   The probability that a missing actor i will be connected to another (observed or missing) actor j is proportional to k_j, the (observed) indegree of actor j:

$$\Pi_i(k_j) = \frac{k_j}{\sum_{j \neq i} k_j}$$

   For each missing actor i

   - randomly draw an outdegree from the observed *distribution of outdegrees*: d_i
   - impute d_i ties based on the *preferential attachment probabilities* to actors with observed indegrees

The ad hoc imputations are independent of the dynamic model

Design of the simulation study:

- Generate *complete simulated data* under known model
- Generate *missing data by* erasing fraction of actors
- *Re-estimate* evolution model on data treated for missingness
- Investigate *sensitivity of parameter estimates*

Data: Two waves of *friendship network* in one year group of a secondary school in Glasgow, consisting of 50 actors

Covariate data:
alcohol consumption, from 1 (not) to 5 (more than once a week)

Estimated true model

Effect	True Parameter
Constant rate	6.87
Density	−2.01
Reciprocity	2.11
Transitivity	0.27
Indirect relations	−0.79
Alcohol similarity	0.92

Simulate complete network data:

First observation at T_1 is initial state

Simulate evolution process according to 'true' model (500 times)

$\Rightarrow$ Simulated end-networks are second observations at T_2

Generate missing data:

Fractions missing actors in each wave: *0.2*, *0.4*, *0.6*

Generate missing data

Four different types of *missing data mechanisms*:

1. random deletion (MCAR)
2. deletion proportional to score on alcohol covariate (MAR)
3. deletion proportional to indegree (MAR?)
4. deletion proportional to outdegree (MNAR)

$\Rightarrow$ Analyze $500 \times (3 \times 4) = 6000$ incomplete data sets
under each missing data treatment,
plus 500 completely simulated data sets as reference

Results

With respect to *convergence*:

- High percentage of missing actors resulted in convergence problems
- Least problems for two ad hoc imputation methods

With respect to *parameter bias*:

- Mechanisms based on structural effects do not necessarily give worse results
- The *complete case methods* performs worst especially with medium to high percentages of missing actors
- The *rate parameter* is underestimated by complete case the model-based method and *overestimated* by the other two methods

- Reciprocity is overestimated by reconstruction but also by other methods
- Largest bias occurs when the parameter is sensitive to the mechanism (alcohol, outdegree)

Conclusion

- Ad hoc imputations *underestimate uncertainty levels*
- Complete case en SIENA procedures do not, but *convergence problems* arise
- The *SIENA approach is the best method* to use: *biases* are not too large for small and medium fractions missing and *standard errors* are not underestimated

References

Barabasi, A-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Burt, R.S. (1987a). A note on missing network data in the general social survey. *Social Networks*, *9*, 63–73.

Burt, R.S. (1987b). Social contagion and innovation: Cohesion versus structural equivalence. *The American Journal of Sociology*, *92*, 1287–1335.

Costenbader, E. and Valente, T.W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, *25*, 283–307.

Gile, K. and Handcock, M.S. (2006). Model-based assessment of the impact of missing data on inference for networks. CSSS Working paper no. 66, University of Washington, Seattle. (`http://www.csss.washington.edu/Papers/wp66.pdf`)

Goldstein, J.R. (1999). Kinship networks that cross racial lines: The exception or the rule? *Demography*, *36*, 399–407.

Handcock, M.S. and Gile, K. (2007). Modeling social networks with sampled or missing data. CSSS Working paper no. 75, University of Washington, Seattle. (`http://www.csss.washington.edu/Papers/wp75.pdf`)

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. Journal of Social Structure, 10.1 (February 4, 2009). http://www.cmu.edu/joss/

Huisman, M. and Steglich, C.E.G. (2008). Treatment of non-response in longitudinal network studies. *Social Networks*, *30*, 297–308.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, *28*, 247–268.

Koskinen, J. (2007). Fitting models to social networks with missing data. Paper presented at Sunbelt XXVII, the International Sunbelt Social Network Conference, May 1–6, 2007, Corfu, Greece.

Robins, G., Pattison, P., and Woolcock, J. (2004). Missing data in networks: exponential random graph ($p*$) models for networks with non-respondents. *Social Networks*, *26*, 257–283.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Snijders, T.A.B. (2005). Models for longitudinal network data. In Carrington, P.J., Scott, J., and Wasserman, S. (Eds.), *Models and Methods in Social Network Analysis*, pp. 215–247. Cambridge University Press, Cambridge.

Stork, D. and Richards, W.D. (1992). Nonrespondents in communication network studies. *Group & Organization Management*, *17*, 193–209.

Van de Bunt, G.G., Van Duijn, M.A.J., & Snijders, T.A.B. (1999). Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, *5*, 167–192.

Ward, M.D., Hoff, P.D., and Lofdahl, C.L. (2003). Identifying international networks: Latent spaces and imputation. In Breiger, R., Carley, k., and Pattison, P. (Eds.), *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, pp. 345–360. Washington: The National Academic Press.