Absent Data in RSiena

Tom A.B. Snijders

University of Oxford
University of Groningen

March 2023

1. Absent Data in RSiena

In three different ways, data about tie variables may be absent:

1. Network delineation is such that, at time t_m, actor i or j is not part of the network.
2. Actors i and j are indeed part of the network at time t_m, but information about whether there is a tie is missing. This is the classical case of *missing data*.
3. Actors i and j are indeed part of the network at time t_m, but the tie $i \rightarrow j$ at time t_m is impossible. Analogous to the phenomenon of 'structural zeros' in contingency tables.

2. Composition change

The case
⇒ ‘Network delineation is such that, at time t_m,
actor i or j is not part of the network.’

Note that network delineation is basic in network analysis,
but also rather artificial / hypothetical:
it is assumed (almost) **that actors outside the network do not exist**.

Often, network delineation is very practical:
students in a classroom, employees in a company department,
firms in an industrial sector in a geographical region.

When memberships change, this can be accommodated in RSiena
by using the function *sienaCompositionChange*.

The function *sienaCompositionChange* defines, for all actors,
the periods where they are part of the network.
This must be given by a text file with one line for each actor.

E.g., for an actor who was there from wave 1 to wave 2, then left,
and came back at 30% of the period between waves 3 and 4,
the times given are

```
1    2    3.3    5
```

An actor who always was there will have the line

```
1    5
```

Knowledge of the times of leaving/entering will of course,
in practice, be approximate.

This can be used for *unconditional estimation* (`cond=FALSE` in *sienaAlgorithmCreate*).

See the help page for *sienaCompositionChange* for options defining how the ties of actors who are absent are processed for estimation by the Method of Moments.

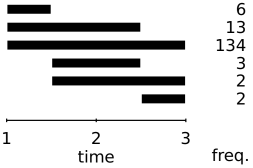
Example: Glasgow data

Example: Changing composition in the Glasgow data set.

Of the 160 students in the Glasgow data set,
only 134 were present at all three waves.
Some came in after the first wave, or left before the last.
No detailed information was available for the precise timing,
and therefore midpoints 1.5 and 2.5 were used.

End of the file used for *sienaCompositionChange*:

```
1 3
1.5 3
1.5 3
1.5 2.5
1.5 2.5
1 3
1.5 2.5
2.5 3
2.5 3
```



Patterns of presence in the school for the Glasgow network, with frequencies

3. Missing Data in RSiena

The case

⇒ ‘Actors *i* and *j* are indeed part of the network at time t_m, but information about whether there is a tie is missing.’

The internal treatment of missing tie values in RSiena is simple:

- Impute missing tie variables in wave 1 by 0.
- Impute missing tie variables in later waves by Last Observation Carried Forward.
- Exclude these imputed values from the calculation of the statistics used for estimation in the MoM.

This procedure is reasonable;
see Huisman and Steglich (*Social Networks*, 2008).

It can be improved if you have more knowledge of the data and also if you are willing to take more effort.

Missing Data: improvements

- ⇒ Sometimes there is enough information to make some imputations,
  based on knowledge of the data,
  with a high degree of confidence.
  If possible, do this!
- ⇒ Option *imputationValues* in *coCovar*, *varCovar* and *sienaDependent* for behavioral variables :
  these values will be used for imputation of missings for the simulations,
  but (like always happens for missings) they are not taken into account for the statistics used for estimation.

  Can be used if there are reasonable, not completely reliable values for imputation.

Missing Data: further improvements

Another possibility is to use *multiple imputation*.

This means that several (e.g., 20) hypothetical
complete data sets are constructed,
each is analysed separately, and the results are combined.

This is a lot of work, but it can help
if the proportion of missing data is very high.

This is treated later on in this set of slides.

4. Structural Values

The case
⇒ 'Actors i and j are indeed part of the network at time t_m, but the tie $i \rightarrow j$ at time t_m is impossible.'

This information can be given to RSiena by specifying the tie value as a **structural zero**,
which is represented (arbitrarily) by the value $x_{ij}(t_m) = 10$.

This will specify that the simulated $X_{ij}(t) = 0$ for $t_m \leq t < t_{m+1}$, and omit the values for $x_{ij}(t_m)$ and $x_{ij}(t_{m+1})$
from the calculation of estimation statistics for the MoM.

The same technique can be used to represent that actor i is absent from the network from wave m until just before wave $m+1$.

The logical counterpart is that the tie $i \rightarrow j$ at time t_m must be present from wave m to wave $m+1$,
$X_{ij}(t) = 1$ for $t_m \leq t < t_{m+1}$.

This information can be given to RSiena by specifying the tie value as a **structural one**,
which is represented (arbitrarily) by the value $x_{ij}(t_m) = 11$.

5. Multiple imputation: principle

A good statistical method for treating missing data is multiple stochastic imputation, developed by Don Rubin.

For a given incomplete data set,
the missing data is imputed independently D times
by drawing from the **conditional distribution of the missing data given the observed data**.

This leads to D complete data sets,
that differ only with respect to the imputed values.

For each complete data set the desired analysis is executed;
standard errors of parameters are a combination
of the within-data set standard errors,
and the variability of estimates between the data sets.

[illegible]
the larger will be the variability between imputed data sets.

How to combine the multiple imputations

The parameter of interest is denoted θ.

Suppose that the d'th randomly imputed data set leads to estimates $\hat{\theta}_d$ and estimated variances W_d (‘Within’),

$$W_d = \text{var}\{\hat{\theta}_d \mid \text{ data set } d\} \; .$$

Note that W_d underestimates true uncertainty,
because it treats imputed data as real data.

The combined estimate is the average

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d \; .$$

Combine multiple imputations....

Compute the average within-imputation variance

$$\overline{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d ,$$

and the between-imputation variance

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}\Big(\hat{\theta}_d - \bar{\theta}_D\Big)^2 .$$

Estimated total variability for $\bar{\theta}_D$ is

$$T_D = \widehat{\mathrm{var}}\Big(\bar{\theta}_D\Big) = \overline{W}_D + \frac{D+1}{D} B_D ,\ s.e.\Big(\bar{\theta}_D\Big) = \sqrt{T_D} .$$

The ratio of standard errors within the completed data sets to final standard errors can be used to define 'missing fraction' m.f.:

$$\frac{\mathrm{diag}(W_D)}{\mathrm{diag}(T_D)} = 1 - \mathrm{m.f.} .$$

This will differ across the parameters.

This procedure will perform well
if data are missing at random (‘MAR’), which means that
the probability of missingness depends on observed variables.

The number *D* of imputed data sets
has to depend on the proportion of missing data.
This number will be in the range of 20–50
for good estimation of standard errors.

How to obtain the multiple imputations

This all is good and well, but how should one obtain
random draws from the conditional distribution of the missing tie and/or behavior variables, given the observed data and given the estimated parameter ?

With respect to the unknown parameter,
it is OK to estimate this provisionally,
e.g., with default treatment of missings.

With respect to the conditional distribution,
we can use RSiena :
the function `siena07` with `maxlike=TRUE`
can be used to provide such simulated values.

This is explained in the following

6. Multiple Imputation for RSiena

Here we describe Multiple Imputation for RSiena
as proposed in Krause, Huisman and Snijders,
'Multiple imputation for longitudinal network data',
Italian Journal of Applied Statistics (2018).

Here the first wave is treated differently from the later waves,
because for the first wave there is no previous wave.

For the later waves the ML option in RSiena will give
a model-based simulation of the missings in this wave,
conditional on the data of the preceding wave.

Intermezzo: Maximum Likelihood (ML) estimation in RSiena

Differences in MCMC simulations in RSiena between
the 'traditional' Method of Moments' (MoM) estimator
and the ML estimator, for waves at t_0 and t_1:

MoM simulations go from $x(t_0)$
to any network with (on average) the right statistics
for amount of change and components of objective function;

ML simulations go from $x(t_0)$ to $x(t_1)$ itself.
ML simulations take much more time!
Efficiency of MoM in many situations quite good.

For missings in $x(t_1)$, the ML simulations in Phase 3 provide a
sample from their distribution, conditional on available data.

Now back to using ML simulations
for missing data imputation in RSiena.

ML simulations are used for getting
model-based longitudinal imputations in the later waves,
according to the steps on the next page.

1. If the first wave has any missings,
   estimate an ERGM or stationary SAOM
   and impute the missings in the first wave using this.
   This produces a completed data set for wave 1.
2. For each wave m, $m = 1, \ldots, M-1$:
   given the completed data set for wave m, produce a
   model-based random draw from the missings in wave $m+1$
   from an ML simulation.
   This produces a completed data set for waves up to $m+1$.
   This is not as time-consuming as full ML estimation,
   because only one simulation is required.
3. Use this complete data set to obtain one estimate $\hat{\theta}_d$.
4. Repeat this procedure D times and use Rubin's rules
   for combining the estimates and standard errors.

This is treated in R. Krause, M. Huisman, & T. Snijders (2018), ‘Multiple imputation for longitudinal network data’, *Italian Journal of Applied Statistics*:
impute first wave (for which there is no help from earlier observations!) by Bayesian ERGM or stationary SAOM, and further waves by likelihood-simulation of SAOM.

This assumes ’missingness at random’: i.e., observed data are sufficient for randomly generating missing data.

However, parameters have to be estimated provisionally, and this may need to depend on the completed data sets.

The main remaining disadvantage is that
the future values are not used for the imputations.

On the Siena scripts page, there is available
a script `multipleImputation_for_RSiena.R`,
with explanation: `AdSUMMissingDataMD.html` .

Example

Waves 2-3-4 of the van de Bunt students data.

Wave 0 is complete, so no ERGM imputation is needed!

Number of missing actors in waves 0–4 are
0; 2; 3; 5; 6, out of 32.

Impute wave 1 – then 2 – then 3 – then 4.

	default		multiple imputation		
Effect	par.	(s.e.)	par.	(s.e.)	m.f.
Rate 1	4.207	(0.640)			
Rate 2	5.063	(0.668)			
outdegree	–1.728***	(0.317)	–1.804***	(0.343)	.16
reciprocity	2.024***	(0.233)	2.100***	(0.260)	.18
trans. trip.	0.324***	(0.048)	0.329***	(0.049)	.12
indeg. - pop.	0.002	(0.038)	0.024	(0.039)	.16
outdeg. - pop.	–0.132***	(0.027)	–0.155***	(0.031)	.11
outdeg. - act.	0.014	(0.014)	0.013	(0.014)	.09
sex alter	0.409*	(0.200)	0.323	(0.204)	.08
sex ego	–0.386†	(0.208)	–0.282	(0.218)	.13
same sex	0.379*	(0.189)	0.362*	(0.193)	.07
program sim.	0.604**	(0.205)	0.687**	(0.213)	.09

par. = estimate; s.e. = standard error; m.f. = missing fraction;
† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;

convergence t ratios all < 0.06; overall maximum convergence ratio 0.08.

Note:
in waves 3 and 4 the proportion of missing actors is 0.15;
proportion missing information is of about this size.

Standard errors of the two approaches are similar;
estimates sometimes (3 cases) differ by about half s.e.,
in other cases differ hardly.

7. Multiple Imputation Networks and Behavior

For imputing behavior in the first wave,
the package MICE can be used.

This can use information from all waves
and also from the network.

On the Siena scripts page, there is available
a script `MultipleImputationNetworkAndBehaviorScript.R`
with the explanation
`MultipleImputationNetworkAndBehavior.html` .

For the later waves, the procedure is similar to the above.