Utrecht University

Master Thesis Applied Data Science
June 2024

Project-ID: F12

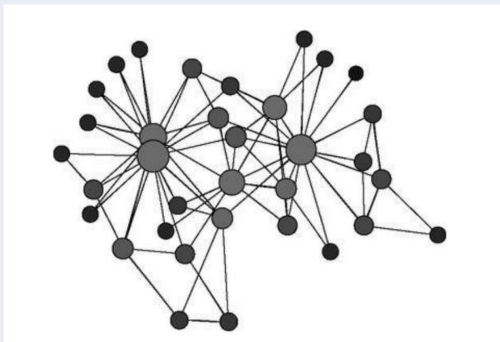
Handling Missing values in Relational Event History Data A Framework in Social Network Research

Myrthe Prins
6753566

Supervisors: Mahdi Shafiee Kamalabad & Gerko Vink

INTRODUCTION

- Social network: A relation defined on a collection of individuals.
- Nodes (actors/vertices) represent entities while the links (edges/ties) connecting them represent any form of interaction or connection between the entities.[1]



- Relational event history data (REH) contains detailed information what happened (message, email, etc.), when it happened (time), and who were involved (sender, receiver).[2]
- REH versus panel data:
  - The ties in REH data are short lasting
  - No unobserved tie changes
  - Relational events occur in exact moments in time

- REM model: a statistical model to analyze relational event history data.[2]
- Why REH data?
  - High precision
  - Increasingly available
  - Little research on missingness implications
- Missing data causes problems in analyses. The impact of missing data is larger when the data has a complex structure, which is the case in network data. REM is sensitive to missingness.

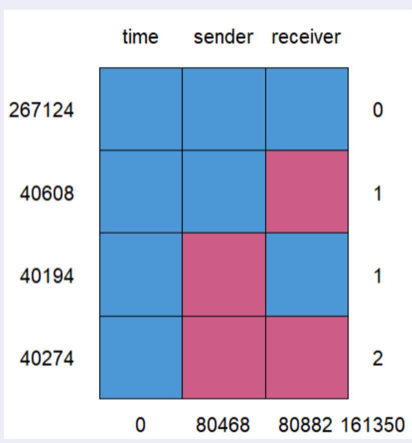
RESEARCH QUESTION

How effective is multiple imputation in handling missing data in Relational Event History data to produce valid inferences?

APPROACH

- The data used in this study is part of the Apollo 13 dataset. This consists of 38982 rows and three columns: time, sender and receiver.
- Types of missingness: Missing Completely at Random (MCAR) and Missing At Random (MAR).
- Proportion of missing data: 10%, 30%, 50%.
- REM model is applied on fully observed dataset and imputed datasets.
- Multiple imputation performance versus fully observed analysis on REH data is evaluated.

time	sender	receiver
11849.2	18	2
11854.2	2	18
11885.2	18	2
11890.2	2	18
...	...	...



RESULTS AND DISCUSSION

- The true model showed non-significant effects for reciprocity and outdegree receiver. The analysis on the simulated models showed for both MAR and MCAR significant effects for all four statistics.
- The bias increased significantly with a higher proportion of missingness. Increase of bias is less under MAR than MCAR.
- Coverage rate is not optimal for all statistics.
- **Multiple imputation can produce valid inferences when the percentage of missingness is not too high, however unexpected significant results are found.**

Table 1

REM results on a subset of the Apollo 13 dataset

Statistic	Estimate	Standard Error	p-value
Reciprocity	2.332e-2	1.856e-2	0.209
Indegree sender	4.314e-4	7.398e-5	< .001
Outdegree receiver	-9.023e-5	7.437e-5	0.225
Same location	-8.629e-1	3.217e-2	< .001

Note: The subset consists of the first 38982 rows of the Apollo 13 dataset

Table 2

Averaged REM results on the simulations under the MCAR assumption for each missingness proportion

Statistic	Prop	Estimate	Std. error	P-value	CI	CV	RB	PB	AW
Reciprocity	0.1	2.38e-2	5.34e-4	< .001	[2.23e-2, 2.53e-2]	0.93	5.11e-4	2.55	2.97e-3
	0.3	2.49e-2	9.75e-4	< .001	[2.22e-2, 4.47e-2]	0.84	1.60e-3	6.92	5.41e-3
	0.5	2.58e-2	1.37e-3	< .001	[2.27e-2, 2.82e-2]	0.77	2.51e-3	10.76	7.59e-3
Indegree sender	0.1	4.28e-4	5.34e-6	< .001	[4.13e-4, 4.42e-4]	0.93	-3.89e-6	1.24	2.97e-5
	0.3	4.20e-4	9.70e-6	< .001	[3.93e-4, 4.47e-4]	0.89	-1.16e-5	2.99	5.39e-5
	0.5	4.11e-4	1.31e-5	< .001	[3.78e-4, 4.46e-4]	0.76	-2.01e-5	4.79	7.26e-5
Outdegree receiver	0.1	-9.11e-5	2.01e-6	< .001	[-9.67e-5, -8.56e-5]	0.92	-9.09e-7	2.53	1.12e-5
	0.3	-9.36e-5	3.92e-6	< .001	[-1.04e-4, -8.27e-5]	0.93	-3.32e-6	5.30	2.18e-5
	0.5	-9.48e-5	5.39e-6	< .001	[-1.02e-4, -8.84e-5]	0.91	-4.55e-6	6.61	3.00e-5
Same location	0.1	-8.78e-1	4.77e-3	< .001	[-8.91e-1, -8.65e-1]	0.39	-1.53e-2	1.78	2.65e-2
	0.3	-9.04e-1	1.10e-2	< .001	[-9.34e-1, -8.73e-1]	0.24	-4.09e-2	4.74	6.12e-5
	0.5	-9.28e-1	1.71e-2	< .001	[-9.76e-1, -8.81e-1]	0.19	-6.54e-2	7.57	9.48e-2

Prop = Proportion of missingness; Std. error = Standard error; CI = Confidence interval; CV = Coverage, RB = Raw bias; PB = Percent bias; AW = Average width

Table 3

Averaged REM results on the simulations under the MAR assumption for each missingness proportion

Statistic	Prop	Estimate	Std. error	P-value	CI	CV	RB	PB	AW
Reciprocity	0.1	2.35e-2	3.85e-4	< .001	[2.24e-2, 2.46e-2]	0.83	1.95e-4	2.00	2.14e-3
	0.3	2.42e-2	7.57e-4	< .001	[2.22e-2, 2.64e-2]	0.85	9.35e-4	4.43	4.20e-3
	0.5	2.50e-2	1.20e-3	< .001	[2.17e-2, 2.83e-2]	0.91	1.70e-3	7.35	6.67e-3
Indegree sender	0.1	4.33e-4	4.34e-6	< .001	[4.21e-4, 4.45e-4]	0.92	1.47e-6	0.87	2.41e-5
	0.3	4.34e-4	8.46e-6	< .001	[4.11e-4, 4.58e-4]	0.97	2.73e-5	1.54	4.70e-5
	0.5	4.32e-4	1.13e-5	< .001	[4.01e-4, 4.64e-4]	0.99	8.18e-7	1.90	6.25e-5
Outdegree receiver	0.1	-8.93e-5	1.82e-6	< .001	[-9.43e-5, -8.42e-5]	0.87	9.68e-7	2.41	1.01e-5
	0.3	-8.89e-5	3.63e-6	< .001	[-9.90e-5, -7.88e-5]	0.93	1.34e-6	3.79	2.02e-5
	0.5	-8.90e-5	4.73e-6	< .001	[-1.02e-4, -7.59e-5]	0.96	1.24e-6	4.74	2.63e-5
Same location	0.1	-8.82e-1	5.90e-3	< .001	[-8.98e-1, -8.65e-1]	0.37	-1.88e-2	2.18	3.27e-2
	0.3	-9.13e-1	1.29e-2	< .001	[-9.49e-1, -8.77e-1]	0.16	-4.99e-2	5.79	7.14e-2
	0.5	-9.38e-1	1.80e-2	< .001	[-9.88e-1, -8.88e-1]	0.13	-7.52e-2	8.71	9.97e-2

Prop = Proportion of missingness; Std. error = Standard error; CI = Confidence interval; CV = Coverage, RB = Raw bias; PB = Percent bias; AW = Average width

Conclusion

Multiple imputation has the potential to be a valid method for handling missing data in REH data when the percentage of missingness is not too high

1. Newman, M. F. (2018). Networks. In Oxford University Press eBooks. https://doi.org/10.1093/oso/9780198805090.001.0001
2. Meijerink-Bosman, M. L. (2023). Who, when, and how long? Time-sensitive social network modeling using relational event data. Gildeprint.