

Applied Data Science

Utrecht University, the Netherlands

MSc Thesis Dominic Comerford (5972944)

TITLE: A Simulation Study of (Multiple) Imputation in Relational Event History (REH) data: Missingness in Time, Sender and/or Receiver

June, 2024

Supervisor:

Dr. Mahdi Shafiee Kamalaba

Second grader:

Dr. Gerko Vink

Word count: ----

**Abstract**

Handling missing data is crucial in statistical analyses, with (multiple) imputation being reliable but underexplored in dynamic social networks. This study focuses on relational event history (REH) data, known for its high resolution and growing availability, to address this gap. Using Apollo 13 mission data, it simulates missingness, imputes values, and compares relational event model (REM) analyses to true coefficients, evaluating bias, coverage, and confidence interval width in the statistics reciprocity, in-degree sender, out-degree receiver, and same location. Multiple imputation was employed for missingness in sender and receiver nodes, while time values were interpolated in various ways.

Results of the imputed analyses were compared against fully observed and complete case analyses, revealing biases in the statistics like. Imputed analyses showed reduced bias but incorrect statistical significance due to smaller standard errors. Limitations include single imputation for time, overlooking message content, and focusing on the NDD mechanism alone.

Despite these limitations, multiple imputation improved effect size estimation over complete case analysis, suggesting its potential in REH data while highlighting the need for refined methods in imputing time data.

*Key words*: Missing data, Relational Event History, Relational Event Model, social network analysis, Multiple Imputation

**Contents**

[Introduction 4](#_Toc169187793)

[Theoretical Background 5](#_Toc169187794)

[Missing data 5](#_Toc169187795)

[Social network analysis 7](#_Toc169187796)

[Relational event models 8](#_Toc169187800)

[This study 10](#_Toc169187801)

[Data & Methods 10](#_Toc169187802)

[Data 10](#_Toc169187803)

[Measures 11](#_Toc169187804)

[Analysis strategy 12](#_Toc169187809)

[Results 15](#_Toc169187815)

[Descriptive statistics 15](#_Toc169187816)

[Fully observed data 16](#_Toc169187817)

[Complete case analysis 17](#_Toc169187818)

[Imputed simulations 18](#_Toc169187819)

[Time imputed with spline and Stineman interpolation 20](#_Toc169187820)

[Discussion 21](#_Toc169187821)

[Limitations 22](#_Toc169187822)

[References 23](#_Toc169187823)

[Appendix A: Apollo 13 actors and IDs 25](#_Toc169187824)

[Appendix B: Sensitivity Analysis 26](#_Toc169187825)

[Appendix C: R Syntax 27](#_Toc169187878)

# Introduction

There are numerous options for handling missing (network) data. In many statistical software the default method overlooks missing values and merely uses the measured observations – referred to as complete case analysis. When missingness occurs randomly this method may produce reliable means, regression coefficients and correlations and simply result in overestimated standard errors (Van Buuren, 2018). Unfortunately, applying complete case analysis often does lead to a loss of information, reduction of statistical power, and more worryingly, bias in the coefficients (Van Buuren, 2018; Schafer & Graham, 2002). Other treatments of missing values involve weighting, likelihood-based procedures, and single-value - or multiple imputation and much is known about how these treatments affect the harm inflicted by missingness (e.g., Schafer & Graham, 2002; Newman, 2014). Consequently, it is also known that multiple imputation is a relatively reliable method to handle missingness in statistical analyses (Van Buuren, 2018). However, there still is a gap in the literature on the effects of missingness and treatment thereof in social network models in general, and in *dynamic* social network models specifically. Here, ‘dynamic’ refers to the timestamps some network data contains that allow for analyzing the changes in interaction over time. Researching to what extent missingness in this kind of data introduces bias and possible solutions is the focal point in this study.

Because of the widespread occurrence of missingness in social network data the necessity of addressing missing data problems in social network analysis (SNA) is generally accepted. However, current SNA often utilizes incomplete data, resulting in the exclusion of nodes (actors) or edges (associations) between nodes. Employing such incomplete data has been found to result in biased results in SNA, even when missingness is randomized (Kossinets, 2006; Huisman, 2009). Bias can be introduced, for instance, because nodes that occur infrequently may be removed from the network altogether at even small proportions of missingness. Typically, SNA originates from *static* social network data that merely allows for analyzing a snapshot of that network. In recent decades, however, statistical and computational advances have made it possible to model more complex network dynamics, through analyzing the network’s time-ordered interactions (Butts, 2008). This type of *dynamic* data is referred to as relational event history (REH) data.

There are three reasons that together legitimize analyzing the missing data problem - and possible solution - in REH data in favor to more traditional social network data. Firstly, REH data is one of the highest resolution and precise network data, which allows a deeper understanding of how social interactions evolve over time and for the modelling of more complex social phenomena. By maintaining the order of interactions, it is possible to include the past in the prediction of future interactions rendering it more informative than traditional SNA. Secondly, REH data is becoming increasingly available due to data being more and more recorded in a time series fashion (e.g., digital communication is often stored automatically, and updated when a new communication is sent). And thirdly, there is even less research on the missingness problem in REH compared to traditional social network data.

Further closing the divide between missing data treatment and REH data could improve the credibility of analyses based on REH data, advancing our knowledge on social interactions and their dynamics. The current study simulates and imputes such missingness in REH data in the shape of Apollo 13 mission data containing the timestamped, chronologically ordered communications sent among ground and space crew and compares resulting analyses to the true coefficients. By doing so this research aims to add to the literature by exploring and simulating 1) the effects of random missingness in *dynamic* social networks, or networks based on REH data, and 2) to what extent (multiple) imputation of random missingness leads to valid results in *dynamic* social networks, or networks based on REH data.

# Theoretical Background

## Missing data

There could be numerous reasons (social network) data are incomplete and these include, but are not limited to respondent inaccuracy, non-response, and technological failures (Kossinets, 2006; Kiang et al., 2021). For example, nodes might inaccurately portray the (non)existence of edges to other nodes, not respond at all, or data might go missing due to electronic malfunctioning. The mechanisms at which missingness occur can vary too and in the literature are described as Not Data Dependent (NDD), Seen Data Dependent (SDD), and Unseen Data Dependent (UDD) (Rubin, 1976; van Buuren, 2018).[[1]](#footnote-2)

NDD describes situations where the probability of missingness is equal across all cases, meaning that the research questions we pose to answer are unrelated to the distribution of the missing values. Consequently, beyond the loss of information, various complexities stemming from such missing data may be overlooked. In contrast, in situations where missingness is affected by either observed (SDD), or unobserved (UDD) characteristics of the data, the research questions of interest *are* related to the missingness. Hence, making inferences from subsequent analyses requires more critical evaluation than in a NDD context (van Buuren, 2018).

Because the current article is exploratory in terms of imputing missing values in a time-series variable in relational event history (REH) data, only NDD is further described as it serves as the benchmark to which imputation should be evaluated against. In other words, if imputation is not satisfactory in an NDD context it will certainly not be in one defined by the more problematic contexts of SDD or UDD. Mathematically, the NDD situation can be formulated as:

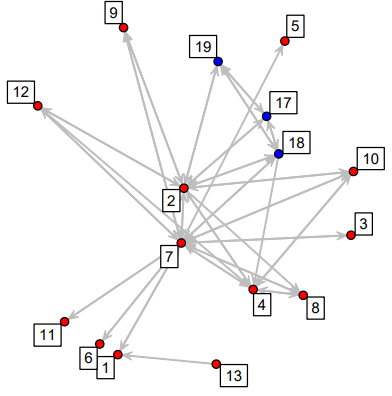
Pr(*R* = 0|*Y*obs, *Y*mis,*ψ*) = Pr(*R* = 0|*ψ*).

Here, *Y* is a matrix composed of *Y*obs and *Y*mis, or the observed and missing values, *R* represents a missingness matrix in which each cell indicates whether the aligning cell in *Y* is observed (0) or missing (1), while *ψ* encompasses the missing data model parameters. So, the probability of data being missing in an NDD context depends on *ψ,* the general missingness probability, as each value has an equal chance to be missing, rather than on *Y*obs or *Y*mis. In sum, NDD is a mechanism resulting in missingness to occur randomly across the data.

Most social network analyses ignore the problem of missingness by analyzing complete cases while some others transform missing edges between nodes to be non-existing edges, which can both lead to biased inferences (Gile & Handcock, 2017). A more truthful method to handle missing data is through multiple imputation as it acknowledges the uncertainty and variance surrounding missing values. By creating multiple versions of the data through replacement of a missing value by a plausible one, multiple imputation allows for analyzing each imputed dataset individually before merging the estimates (van Buuren, 2018). In the current article, the type of data that is simulated to be missing through an NDD mechanism and consequently imputed through multiple imputation stems from relational event history (REH) data, a specific type of social network data.

## Social network analysis

A (social) network can be defined as a collection of nodes connected through edges (Newman, 2018). The units of interest in social network analysis revolve around the relationships among nodes, such as the edges between individuals, communities, or other entities. An example of a social network is displayed in Figure 1, which shows the aggregated communication network among nodes in the Apollo-13 dataset. Some informative network characteristics can be derived from such network graphs as it clearly shows node 7 is a central unit that connects peripheral nodes and that the blue triangle nodes form a closely knit community within the network. Considering node 7 represents the flight director and the triplet constitutes the three astronauts, such an architecture seems plausible in this network (see Appendix A for the actor IDs and their roles).



**Figure 1.**

*Network graph (directed) of aggregated communication between nodes in Apollo-13 data.*

Note. Astronaut nodes in blue, ground control nodes in red.

Consequently, because of its emphasis on interaction among nodes, social network analysis (SNA) requires data on the edges between nodes, and these edges can take various forms. For example, SNA may focus on friendships between colleagues or on communication instances between astronauts and ground control. Based on the characteristics of these edges and the research objectives, edges can either be directed or undirected (Newman, 2018). Undirected edges encompass mutual ties such as shared affiliations while directed edges involve a certain flow or direction in the relationship such as a communication sent from ground control to astronaut at a certain time-point. This focus on edges contrasts with traditional studies that focus on individual attributes to understand behavior (McGloin & Kirk, 2014). To clarify the contrast between attributive and relational data; the degree one is communicative is an attribute, whereas the occurrence or intensity of communication between individuals represents a relational event.

Social network analysis based on relational events implicitly suggests that the quantity and type of connections among nodes can be informative explanatory factors in predicting future events (McGloin & Kirk, 2014). In other words, the relational history in the data may be predictive of further network dynamics. In the context of the Apollo-13 mission the past communications may be informative in predicting further communication dynamics among the Apollo 13 crew. For these *dynamics* to be analyzed a specific type of network model needs to be employed – the relational event model.

## Relational event models

Relational events can be understood as actions that occur as discrete events at a certain point in time where one node exhibits a behavior targeted at one or multiple other nodes in the network, potentially including themselves (Bates & Harvey, 1975; Butts, 2008). A sequence of those events is then described as relational event history (REH) data and encompasses at least the times or order of events, and dyads of sender and receiver nodes (Butts, 2008). Table 1 entails the first two and last two cases of the Apollo 13 REH data, as each row represents a discrete time-stamped event where a message is sent from a sender to a receiver node.

|  |  |  |
| --- | --- | --- |
| **Table 1.**  *Relational events of Apollo-13 communication.* | | |
| Time | Sender ID | Receiver ID |
| 11849.2 | 18 | 2 |
| 11854.2 | 2 | 18 |
| … | … | … |
| 50012.8 | 7 | 4 |
| 50014.8 | 4 | 7 |
| *Note.* Total number of recorded events is 3882 among 16 nodes. Time is in seconds from onset of the mission. | | |

The relational event model (REM) provides a framework for modelling the predictors, or statistics, that explain how REH data evolves by estimating the event rate , which determines which nodes will interact *and* when this interaction will occur (Butts, 2008; Meijerink-Bosman et al., 2023). To estimate the event rates, it is first necessary to construe a risk set entailing all possible events that might occur, resulting in a matrix of all conceivable dyads. In the context of *directed* edges of sender *s* and receiver *r*, the matrix *s* X *r* represents all possible relational events at time-point *t*. Thus, the Apollo-13 communication risk set comprises of *N* (*N* - 1), or (16 x 15 =) 240 potential events at each time-point.[[2]](#footnote-3)

Second, the likelihood of an event to occur between a sender and receiver at each time-point is equal to the occurrence rate of that event relative to the sum of rates for all possible events (Butts, 2008). This rule ensures more common events are assigned higher event rates compared the less common events, and can be defined as:

Finally, the event rate can then be modelled as the outcome, regressed on by predetermined statistics in a log-linear function (Meijerink-Bosman et al., 2023):

Here, *βp* refers to the parameter that represents the impact of predictor *xp* on the event rate. In REM literature, these predictors are called statistics and can be either exogenous or endogenous. Exogenous statistics entail characteristics such as ‘age’ or ‘location’ of individual nodes or edges and allow for researching to what extent certain attributes determine the event rate. Endogenous statistics encompass the likelihoods of potential subsequent events conditional on past events, such as a dyad’s prior communication. Consequently, by estimating the model parameters, *βp,*​ linked to exogenous and endogenous statistics inferences can be made on the occurrences and dynamics of communication within a network over time (Meijerink-Bosman et al., 2022)

## This study

This study aims to analyze how (multiple) imputation of missing values affects analyses of REH data. Missingness is simulated through the NDD mechanism in the Apollo-13 communication data and is allowed to occur in the time, sender, or receiver information or in any combination of these columns. Subsequently, missing values are imputed through (multiple) imputation and REMs of imputed datasets are compared to the analysis utilizing the full communication data and a complete case analysis. In doing so, the current research improves our understanding of 1) to what extent missing data occurring randomly, and specifically missingness in ‘time’, biases results in REMs, and 2) to what extent current techniques for imputing missing data may help reduce bias.

# Data & Methods

## Data

Following Kamalabad and colleagues (2023), communication data from the Apollo 13 mission was used for the empirical analyses, specifically from the time surrounding the iconic phrase “Houston, we’ve had a problem.”[[3]](#footnote-4) This ‘problem’ occurred fifty-six hours into the mission and referred to an exploded oxygen tank. At that moment, the mission turned from a routine journey destined to the moon into a mission to safely return the astronauts to Earth. Luckily, they did in the end - aided by clear communication within the network.

In the current article the REH data is analyzed, focusing on the sequence of communications that transpired within the network rather than on their contents. As such, the relational events are time-stamped directed communications from a sender to a receiver node (again: see Table 1 for an impression). Note that time is in seconds from the onset of the mission, and the sender and receiver columns contain the ID rather than the role of the nodes (again: see Appendix A for IDs and accompanying roles). After selecting the Apollo 13 communications around the time of the exploded oxygen tank 3882 relational events among 16 nodes remained.

## Measures

This study’s focus is on the impact of missing data and how to make valid inferences through multiple imputation in relational event models rather than analyzing the exact communications. Therefore, the content of the communications is excluded but like Kalamabad and colleagues (2023 the endogenous statistics; reciprocity, indegree sender, and outdegree receiver are included. Also, as exogenous statistic it is considered whether a sender and receiver of a communications are in the same or mixed locations. Including these statistics allows for modelling network dynamics based on past relational events as well as disentangle location effects on communication and compare the extent to which results from various models differ.

Reciprocity. Prior studies have shown that nodes that have received communications are likely to reciprocate these in the future (Stadfeldt & Block, 2017; Kalamabad et al., 2023). For example, the reciprocity statistic assumes there is a tendency for node 1 to reciprocate communication to node 2, if node 2 has contacted node 1 in the past.

In-degree sender. The in-degree of the sender statistic refers to the number of communications a node has received and assumes that those with higher in-degree have a higher likelihood of initiating contact in the future (Butts & Marcum, 2017). For instance, when node 1 receives relatively many communications up to a certain time-point it is expected that node 1 has a relatively high probability of initiating contact in the future compared to a node that has received fewer communications.

Out-degree receiver. The out-degree of the receiver statistic refers to the number of communications a receiver node has transmitted, and it assumed that nodes with a higher out-degree of communications have a higher likelihood of being contacted in the future (Butts & Marcum, 2017). For instance, when node 2 sends relatively many communications up to a certain time-point it is expected that node 2 has a relatively high probability of being contacted in the future themselves compared to a node that has sent fewer communications.

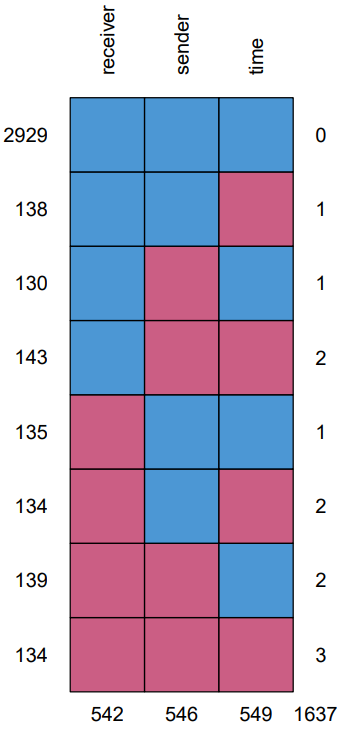
Same location. The same location statistic reflects an exogenous attribute that a sender and receiver’s rate of interacting is determined by whether the dyad shares the same location. In other words, whether the nodes are in the same group – astronauts or ground control. A binary variable was coded where intragroup communication among the space crew (node 17, node 18, and node 19) and ground control was assigned a 1, and intergroup communication a 0. A negative effect of ‘same location’ reflects that nodes who are in the same location initiate future communication with a lower event rate compared to nodes in different locations. Because there is a clear structure and hierarchy before sending a message from ground control to the astronauts it is expected ‘same location’ will have a negative effect on the event rate. Ground crew, as well as the astronauts, are likely to communicate ‘off the record’ before sending a coordinated message via the Apollo 13 channel.

## Analysis strategy

The analysis can be divided into three main components, and these are 1) missingness simulation, 2) missingness imputation, and 3) the REM analyses.

Missingness simulation. The first step in the analysis was to create random missingness through the NDD mechanism in the complete Apollo 13 data. To include every node in each simulated dataset the same 1500 relational events were drawn and preserved from the complete data and included in each simulation. Part of the data was preserved because some nodes occurred infrequently and would have been removed altogether in some simulations by amputation over all relational events. Therefore, preserving some communication resulted in the same number of nodes (16) in each simulation, and thus equal risk set sizes, a requirement for comparisons of different REMs.

The amputation was done one hundred times by creating incomplete versions with 40% of rows containing at least one missing value in the remaining data. This proportion allows for stable enough analyses under a substantial amount of missingness. Missing values were allowed to occur randomly in time, sender, or receiver data as well as in any combination of these variables. As an example, Figure 2 contains the missingness patterns of the first simulated dataset. The patterns show that in this simulation 2929 completely observed relational events remain and seven patterns of missingness exists that occur from 130 to 143 times each.[[4]](#footnote-5) Note that the last pattern where all columns are missing could occur in REH data. REH data might still contain the row of the communication even though all columns are missing as it was transmitted and logged as a relational event. Here, it is unknown what the time, sender and receiver of this communication is but *that* an event occurred is known.



**Figure 2**.

Missingness patters of the first simulated dataset

Missingness imputation. The second step of the analysis was to impute the generated missing values through (multiple) imputation. This required considerations regarding the number of imputations, the number of iterations, the methods employed and the specification of predictors for the imputation model. It was opted to set the number of imputations and iterations both to five for computational efficiency. Furthermore, research shows this number is often sufficient for reliable imputations (Van Buuren, 2018). The employed methods differed per missing variable, as ‘time’ was interpolated rather than multiple imputed.

Time was imputed by interpolation as the values before and after missing times were used for imputation. This was done in three ways, as time was interpolated linearly at first, but in two additional analyses time was interpolated using the spline and Stineman (Stineman, 1980) methods. Interpolation of time might be a realistic option for REH data with missingness in times as specific times may be lost but due to data being stored chronologically the order might still be preserved.[[5]](#footnote-6)

In contrast, imputations for sender and receiver data were estimated through predictive mean matching. Predictive mean matching interpolates from a pool of candidate donors that are most like the missing observation; one or more of which are used to impute the missing value (van Buuren, 2018). Consequently, such ‘regular’ predictive mean matching can result in an imputed node communicating to themselves whereas the Apollo 13 data does not allow for such communications. Therefore, a custom method was employed that prevents imputed data from containing nodes that communicate with themself. In all multiple imputation models, the target variables were regressed on both remaining variables. For instance, in the ‘sender’ imputation model both ‘time’ and ‘receiver’ were used as predictors of missing ‘sender’ values.

Analyses. The third step of the analysis was conducting the REMs and their comparison in performance. The statistics for the models were computed using the package ‘remstats’ (Arena & Meijerink-Bosman, 2024), while the REMs were conducted with ‘remify’ (Arena, 2024) using the cox proportional hazard function from the ‘survival’ package (Therneau, 2023). As the sample equals the population (i.e., the Apollo 13 from which data is sampled is the complete population) the variance around the statistics in the simulations is calculated in alignment with Vink and van Buuren (2014). This feature permits the use of the single observed Apollo 13 data as the truth, thereby removing sampling variance from the evaluation of model performance in the simulations.

First, the REM on the fully observed data will be discussed as this constitutes the ‘truth’. Thereafter, the REM will be conducted as complete case analysis pooled over the hundred simulations and compared to its fully observed counterpart. Finally, the REM will be discussed over the simulations as the pooled result where it will be compared to the previous analyses. Performance evaluation of the imputed simulations compared to the true coefficients was done based on Oberman and Vink’s (2023) recommendations. These recommendations entail analyzing the difference between the true coefficients and the estimated coefficients in the simulations, or the bias, and the coverage (CR). The proportion of bias (PB) in the coefficients should ideally be lower than five percent, while the coverage should be 95% and reflects how often the estimates’ 95% confidence intervals capture the truth. Moreover, the average width (AW) of the 95% confidence intervals is used for evaluation, where more narrow intervals in combination with sufficient coverage suggest less uncertainty than wider ones.

# Results

## Descriptive statistics

Table 2 shows descriptive statistics of the aggregated network of the Apollo 13 data. The number of nodes is 16 of which 3 are astronauts. Also, the number of communications is 3882 of which 1813 are sent to the same location and 2069 between locations. Furthermore, it entails network-level characteristics of the network’s architecture, as depicted in Figure 1. The density, or the proportion of connected nodes is .21, while the longest shortest path spans 4 nodes. The closeness centrality is .68 and the eigenvector centrality, or the average centrality based on neighbors’ influence is also high, reflected in the score of .61. The average shortest path in the network is 1.92, meaning that any dyad on average is separated by less than two nodes. Relatedly, the average betweenness, or the extent to which nodes fall on another dyad’s shortest path is .57, while transitivity, or the overall closure of triplets into connected triangles is .33.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2.**  *Network characteristics of complete Apollo 13 data.* | | | |
|  | Amount | Network-level | |
| Nodes | 16 | Density | .21 |
| Astronauts | 3 | Diameter | 4 |
| Ground control | 13 | Closeness | .68 |
| Communications | 3882 | Eigenvector | .61 |
| Same locations | 1813 | Average shortest path | 1.92 |
| Different locations | 2069 | Betweenness | .57 |
|  |  | Transitivity | .33 |

## Fully observed data

The results of the REM utilizing the fully observed Apollo 13 data can be found in Table 3. Firstly, it can be derived that the risk set, or the number of potential events, is 931.680, which amounts to the product of the number of events, 3882, multiplied by the number of possible pairs (16 x 15). Secondly, it shows that reciprocity has a small positive effect on a certain event occurring although this is not statistically significant (β = 2.332-02, p = .209). Nodes do not seem to return past communications in this network. Thirdly, a sender’s past in-degree does positively affect the likelihood of an event happening and this effect is statistically significant (β = 4.314-04, p < .001). Receiving a larger number of communications results in a higher likelihood of becoming a future sender. Fourth, the out-degree of the receiver has a small negative and statistically insignificant effect on a communication happening (β = -9.023-05, p = .225). This coefficient implies that the number of communications a node has sent, does not determine whether that node will be a receiver of future communications.

Finally, and perhaps unsurprisingly there is a tendency to engage contact with nodes that are in a different location. Whether a pair of nodes are in the same location proves to be a strong inhibitor of future communication, as those in the same location are less likely to engage in contact with each other through the Apollo 13’s channel (*β =* -.863, *p* < .001). Given that ground crew possibly talks to each other outside of the mission’s channel before a final message is sent to the astronauts, and vice versa, this is reflected in this strong relation of shared location and lower tendency to communicate.

|  |  |  |
| --- | --- | --- |
| **Table 3.**  *REM Results for the fully observed Apollo 13 data.* | | |
| Statistic | β | p-value |
| Reciprocity | 2.332-02  (1.856-02) | .209 |
| In-degree sender | 4.314-04  (7.398-05) | < .001 |
| Out-degree receiver | -9.023-05  (7.437-05) | .225 |
| Same location | -.863  (.032) | < .001 |
| *Note.* Number of possible events = 931.680, number of events = 3882. Standard errors in parentheses. BIC = 98241. Number of simulations is 100. | | |

## 

## Complete case analysis

Table 3 contains the results of the aggregated complete case analysis over the 100 simulations. Compared to their true counterparts the coefficients for reciprocity (*β =* 2.773-02, *p* = .196), in-degree sender (*β =* 5.679-04, *p* < .001) are somewhat overestimated whereas out-degree receiver (*β =* -1.208-04, *p* = .293) is somewhat underestimated. Also, same location seems to be relatively biased as its estimate is underestimated in the complete case analysis (*β* = -1.349, *p* < .001). Furthermore, and unsurprisingly for analysis utilizing fewer cases through an NDD mechanism all standard errors are larger which implies there is more uncertainty in the estimates. Although complete case analysis introduces bias in each coefficient, the statistical significance as reflected by the estimates’ *p*-values is not altered which speaks in favor of this method.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4.**  *Aggregated REM results for complete case analysis.* | | | |
| Statistic | β | p-value | Bias |
| Reciprocity | 2.773-02  (2.121-02) | .196 | 4.413-03 |
| In-degree sender | 5.679-04  (1.126-04) | < .001 | 1.364-04 |
| Out-degree receiver | -1.208-04  (1.129-04) | .293 | -3.054-05 |
| Same location | -1.349  (.039) | <.001 | -.486 |
| *Note.* Risk set size ranges from 690.720 to 717.360, and number of relational events ranges from 2878 to 2989 across simulations. Standard errors in parentheses. Average BIC is 71786. Number of simulations is 100. | | | |

## Imputed simulations

Table 4 shows the aggregated REM results over 100 simulations where ‘time’ was imputed via linear interpolation and ‘sender’ and ‘receiver’ through multiple imputation. Results indicate that reciprocity has a small positive effect on a certain event occurring, and this is statistically significant (*β =* 2.479-02, *p* < .001, 95% CI = [2.268-02, 2.690-02]). Similarly, a sender’s past in-degree now positively predicts an event happening and this effect is statistically significant (*β* = 4.183-04, *p* < .001, 95% CI = [3.920-04, 4.446-04]), while out-degree of the receiver has a small negative and statistically significant effect (*β* = -9.368-05, *p* < .001, 95% CI = [-9.851-05, -8.886-05]). The coefficient for sharing the same location remains the same after imputation up to the fourth decimal, because no missingness was created in this variable.[[6]](#footnote-7)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 5.**  *Aggregated REM results after imputation of missingness.* | | | | | | | | |
| Statistic | β | p-value | CI-95% [LB, UB] | CR | Bias | PB | AW | |
| Reciprocity | 2.479-02  (7.601-04) | < .001 | [2.268-02,  2.690-02] | .70 | 1.465-03 | 6.282 | 4.221-03 | |
| In-degree sender | 4.183-04  (9.476-06­) | < .001 | [3.920-04,  4.446-04] | .89 | -1.319-05 | 3.176 | 5.262-05 | |
| Out-degree receiver | -9.368-05  (1.738-06) | < .001 | [-9.851-05,  -8.886-05] | .78 | -3.450-06 | 3.833 | 9.652-06 | |
| Same location | -.863  (4.787-06) | < .001 | [-.863-01,  -8.629-01] | 1.00 | -9.700-06 | 3.450-04 | | 2.658-05 |
| *Note.* Number of possible events = 931.680, number of events = 3882. ‘Time’ imputed as single value by interpolation. ‘Sender’ and ‘Receiver’ imputed through multiple imputation in MICE. Standard errors in parentheses. Number of imputations is 5 in each simulation. Number of simulations is 100. | | | | | | | | |

There are both similarities and differences when comparing the aggregated simulation results to the fully observed REM. The effect sizes are like the fully observed REM, leading to only marginal absolute and acceptable relative bias. Considering the missingness mechanism was NDD, only a relatively small amount of bias was anticipated. The bias introduced by these simulations is smaller than in complete case analysis. Furthermore, as the relative biases range from 3.176% to 6.282% in the endogenous statistics, imputation seems to be a more viable method than complete case analysis.

However, the standard errors in the imputed simulations are substantially smaller than in the fully observed scenario, resulting in statistically significant effects for reciprocity and out-degree receiver statistics that were not found in the fully observed REM (and not in the complete case analysis either). Consequently, the confidence intervals of the statistics in the imputed data analysis are narrow as well, as showcased by the narrow average widths (AW) across the simulations. Unfortunately, the coverage rates (CR), or the proportion of times the 95% confidence intervals include the ‘true’ value are suboptimal. In only 70% of the confidence intervals for ‘reciprocity’ does the truth fall within the boundaries, while this reaches 89% and 78% for the ‘in-degree sender’ and ‘out-degree receiver’ statistics. Ideally, these rates should lay around 95% (Oberman & Vink, 2023). This underperformance implies that the current imputation procedure may lead to invalid inferences when caution is not preserved regarding the standard errors of the effect sizes even though absolute bias is smaller than in the complete case analysis.

## Time imputed with spline and Stineman interpolation

Two additional REMs were conducted where time was first interpolated through spline interpolation and then according to the Stineman algorithm (Stineman, 1980). Table 6 contains the results of these REMs and conclusions remain like the main analysis. Reciprocity and out-degree of the receiver become statistically significant while all standard errors become substantially smaller. The small standard errors also result in similar confidence intervals and their average widths, as well as comparable absolute and relative bias to the main analysis with linear interpolation. Using the spline method slightly improves the coverage rates for reciprocity and in-degree receiver but decreases it for out-degree receiver whereas the Stineman algorithm improves reciprocity’s coverage too but decreases somewhat in in-degree sender and out-degree receiver.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 6.**  *Aggregated REM results after imputation of time spline and Stineman interpolation.* | | | | | | | | | | |
| Statistic | | β | p | CI-95% [LB, UB] | CR | | Bias | PB | | AW |
| *Spline* | | | | | | | | | | |
| Reciprocity | | 2.479-02  (6.999-04) | < .001 | [2.285-02,  2.674-02] | .73 | | 1.471-03 | 6.312 | | 3.886-03 |
| In-degree sender | | 4.188-04  (9.960-06­) | < .001 | [3.911-04,  4.464-04] | .92 | | -1.269-05 | 3.057 | | 5.531-05 |
| Out-degree receiver | | -9.365-05  (1.708-06) | < .001 | [-9.839-05,  -8.891-05] | .75 | | -3.416-06 | 3.789 | | 9.482-06 |
| Same location | | -.863  (4.793-06) | < .001 | [-8.629-01,  -8.629-01] | .99 | | -1.646-06 | 3.712-04 | | 2.661-05 |
| *Stineman* | | | | | | | | | | |
| Reciprocity | 2.479-02  (7.560-04) | | < .001 | [2.269-02,  2.688-02] | .79 | 1.465-03 | | | 6.281 | 4.198-03 |
| In-degree sender | 4.187-04  (9.355-06­) | | < .001 | [3.928-04,  4.447-04] | .87 | -1.272-05 | | | 3.155 | 5.195-05 |
| Out-degree receiver | -9.364-05  (1.660-06) | | < .001 | [-9.824-05,  -8.903-05] | .72 | -3.406-06 | | | 3.774 | 9.219-06 |
| Same location | -.863  (4.419-06) | | < .001 | [-8.629-01,  -8.629-01] | .95 | -9.573-07 | | | 3.189-04 | 2.454-05 |
| *Note.* Number of possible events = 931.680, number of events = 3882. ‘Time’ imputed as single value by interpolation. ‘Sender’ and ‘Receiver’ imputed through multiple imputation in MICE. Standard errors in parentheses. Number of imputations is 5 in each simulation. Number of simulations is 100. | | | | | | | | | | |

# Discussion

This study aimed to explore the extent to which (multiple) imputation of missing values affects analyses of REH data. Missingness was simulated a hundred times through the NDD mechanism in Apollo 13 data and occurred in time, sender, and receiver information or in any combination of these columns. In subsequent imputation models missing values in sender and receiver nodes in the simulated dataset were imputed via multiple imputation, while missing time values were interpolated by the subsequent and prior value that was observed or interpolated if needed in this column. Subsequently, results of the REMs in the simulations were aggregated and compared to the fully observed analysis and complete case analysis based on the bias introduced in the statistics of reciprocity, in-degree sender, out-degree receiver and whether the dyad shared the same location.

In the fully observed analysis – the truth – the in-degree of the receiver and same location were statistically significant predictors of the event rate. The results of the subsequent analyses indicated that missingness in relational event history (REH) data, even in a randomized fashion, biased the estimates. This conclusion is in line with studies conducted by Kossinets (2006) and Huisman (2009) which utilized regular network data. In line with literature on the NDD mechanism and complete case analysis (e.g., Van Buuren, 2018), complete case analysis seemed moderately reliable in this context. Although coefficients for reciprocity and in-degree sender were overestimated, while out-degree receiver and same location were underestimated, statistical significance of the coefficients was not altered by complete case analysis in relation to the fully observed data.

In the analysis involving the imputed simulations coefficients were less biased compared to complete case analysis in how far from the ‘truth’ the coefficients were estimated to be. However, smaller standard errors of the coefficients led to the reciprocity and out-degree receiver statistics incorrectly being identified as statistically significant predictors of the event rate. Based on the criteria of relative bias of the estimates and narrow average width of the confidence intervals these simulated imputations seemed acceptable, but the suboptimal coverage rates for the endogenous statistics warrant caution for making inferences from these imputations. Additional analyses where time was interpolated through the spline and Stineman algorithms (Stineman, 1980) yielded somewhat different results in the statistics, but overall conclusions were alike. In sum, while analysis after imputation does estimate the effect size better than complete case analysis the former resulted in statistically significant results in reciprocity and out-degree of the receiver whereas complete case analysis did not.

## Limitations

The are several limitations in the current study that should be mentioned. First, missing times in the current study were interpolated from the timepoint before and after those missing values rather than through multiple imputation. This was done because the challenge of multiple imputations of time in REH data proved to be more complex than expected as it is essential the chronological order is preserved while adding enough noise to the various imputations to make reliable inferences. The noise in some imputations of missing times caused the order to shift in some cases resulting in a violation of one of REH’s assumptions – that there is at least an order in the relational events. Despite this limitation of employing single imputation, the single imputation of missing times could perhaps be worthwhile to explore further too as indicated by the to some extent varying results in the REMs where interpolation of time differed (regular, spline, and the Stineman method). Because REH data is stored chronologically, in some contexts missing values in ‘time’ may be imputed with relative certainty through single imputation as the order may have been retained even though some of the exact times are missing. However, when the order of social interaction in REH data is missing too, multiple imputation should most likely be the baseline for imputation of time. Therefore, correct methods for multiple imputation of time in REH data should be explored further as the element of time is so central to REH data.

Second, the content of the communications is overlooked in imputing missingness, but it is reasonable to assume the messages themselves could provide valuable information for more reliable imputation models. For instance, some communications may have a more positive sentiment whereas others have a more negative sentiment, and including such exogenous statistics could be an exciting direction for subsequent studies.

Third, the current study simulates missingness through the ‘simpler’ NDD-mechanism as it serves as a baseline to explore the missingness problem in REH data and a logical step would be to perform similar analyses in an SDD (or UDD) context too.

Despite these limitations the current study does indicate that multiple imputation of sender and receiver data and single imputation of time performs better than complete case analysis in an NDD-context when looking at effect sizes. However, caution is required when making further inferences from these imputation analyses based on the extremely small standard errors – and resulting statistically significant *p*-values.

# References

Apollo 13 Real-time. (n.d.). Retrieved from <http://apollo13realtime.org/>

Bates, F. L., & Harvey, C. C. (1975). *The Structure of Social Systems.* New York: Gardner Press.

Butts, C. T. (2008). A relational event framework for social action. *Sociological methodology*, *38*(1), 155-200. https://doi.org/10.1111/j.1467-9531.2008.00203.x

Gile, K. J., & Handcock, M. S. (2017). Analysis of networks with missing data with application to the National Longitudinal Study of Adolescent Health. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *66*(3), 501-519.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, *10*(1), 1-29.

Kamalabad, M. S., Leenders, R., & Mulder, J. (2023). What is the Point of Change? Change Point Detection in Relational Event Models. *Social Networks*, *74*, 166-181.

Kiang, M. V., Chen, J. T., Krieger, N., Buckee, C. O., Alexander, M. J., Baker, J. T., ... & Onnela, J. P. (2021). Sociodemographic characteristics of missing data in digital phenotyping. *Scientific reports*, *11*(1), 15408.

McGloin, J. M., & Kirk, D. S. (2014). An overview of social network analysis. *Advancing Quantitative Methods in Criminology and Criminal Justice*, 67-79.

Meijerink-Bosman, M., Leenders, R., & Mulder, J. (2022). Dynamic relational event modeling: Testing, exploring, and applying. PLoS One, 17(8), e0272309.

Meijerink-Bosman, M., Back, M., Geukes, K., Leenders, R., & Mulder, J. (2023). Discovering trends of social interaction behavior over time: An introduction to relational event modeling: Trends of social interaction. *Behavior Research Methods*, *55*(3), 997-1023.

Moritz S, Bartz-Beielstein T (2017). “imputeTS: Time Series Missing Value Imputation in R.” \_The R Journal\_, \*9\*(1), 207-218. doi:10.32614/RJ-2017-009

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372-411.

Newman, M. F. (2018). Networks. In Oxford University Press eBooks. <https://doi.org/10.1093/oso/9780198805090.001.0001>

Oberman, H. I., & Vink, G. (2023). Toward a standardized evaluation of imputation methodology. Biometrical Journal. <https://doi.org/10.1002/bimj.202200107>

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–590.

Salgado CM, Azevedo C, Proença H, Vieira SM. Missing Data. In: Secondary Analysis of Electronic Health Records. Springer, Cham (CH); 2016. PMID: 31314252.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Stadtfeld, C., & Block, P. (2017). Interactions, actors, and time: Dynamic network actor models for relational events. *Sociological Science*, *4*, 318-352.

Stineman, R. W. (1980). A consistently well-behaved method of interpolation. *Creative Computing*, *6*(7), 54-57.

Therneau, T. M. (2023, March 12). Survival Analysis [R package survival version 3.5-5]. https://cran.r-project.org/package=survival

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Boca Raton: CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*, 1-67.

Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv preprint arXiv:1409.8542*.

# Appendix A: Apollo 13 actors and IDs

* AFD: Assistant Flight Director from Flight directors (1)
* CAPCOM: Capsule Communicator from Flight directors (2)
* CONTROL: Control Officer from Flight directors (3)
* EECOM: Electrical, Environmental and Consumables Manager from Flight directors (4)
* All : Ground control team (without flight directores) (5)
* FDO : Flight dynamics officer (FDO or FIDO) (6)
* FLIGHT: Flight Director from Flight directors (7)
* GNC: The Guidance, Navigation, and Controls Systems Engineer from Flight directors (8)
* GUIDO: Guidance Officer from Flight directors (9)
* INCO: Integrated Communications Officer from Flight directors (10)
* NETWORK: Network of ground stations from Flight directors (11)
* TELMU: Telemetry, Electrical, and EVA Mobility Unit Officer from Flight directors (12)
* RECOVERY: Recovery Supervisor from Flight directorsc (13)
* PROCEDURES: Organization & Procedures Officer from Flight directors (14)
* FAO: Flight activities officer from Flight directors (15)
* RETRO: Retrofire Officer from Flight directors (16)
* CDR: Commander James A. Lovell Jr. crew (astronauts) (17)
* CMP: Command Module Pilot John (Jack) L. Swigert Jr. crew (astronauts) (18)
* LMP: Lunar module pilot Fred W. Haise Jr. crew (astronauts) (19)

# Appendix B: Sensitivity Analysis

In this analysis ‘same location’ was used in imputation models of sender and receiver as additional predictor. Conclusions are similar but note that coverage rates are substantially lower in this REM – except for reciprocity.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 6.**  *Aggregated REM results of 100 simulations after imputation of missingness with ‘same location’ as additional predictor.* | | | | | | | | |
| *Statistic* | *β* | *p-value* | *CI-95% [LB, UB]* | *CR* | *Bias* | *PB* | *AW* | |
| Reciprocity | 2.466-02  (7.153-04) | < .001 | [2.267-02,  2.664-02] | .72 | 1.338-03 | 5.750 | 3.972-03 | |
| In-degree sender | 4.079-04  (9.897-06­) | < .001 | [3.919-04,  4.445-04] | .60 | -2.356-05 | 5.460 | 5.496-05 | |
| Out-degree receiver | -9.440-05  (1.833-06) | < .001 | [-9.943-05,  -8.977-05] | .68 | -4.166-06 | 4.617 | 1.018-05 | |
| Same location | -.863  (4.555-06) | < .001 | [-9.943-05,  -8.977-05] | .97 | -1.537-06 | 3.546-04 | | 2.529-05 |
| *Note.* Number of possible events = 931.680, number of events = 3882. ‘Time’ imputed as single value by interpolation. ‘Sender’ and ‘Receiver’ imputed through multiple imputation in MICE. Standard errors in parentheses. Number of imputations is 5 in each simulation. | | | | | | | | |

# Appendix C: R Syntax

1. NDD, SDD and UDD are typically referred to as Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR), respectively. However, in the current study the ‘data dependent’ terminology from Hand (2020) is used as it directly conveys the missingness mechanism at play.  [↑](#footnote-ref-2)
2. For practical reasons it is assumed in the current study that a node cannot send message to themselves or to multiple other nodes simultaneously but is possible to model such interaction in a REM. [↑](#footnote-ref-3)
3. The complete communication transcript can be retrieved from <http://apollo13realtime.org/>. The used subset of Apollo 13 is web scraped and does not have any privacy or ethical limitations. [↑](#footnote-ref-4)
4. The *MICE* package in R (van Buuren & Groothuis-Oudshoorn, 2011) was used to create missingness as well as to impute missingness in the data in the next step. [↑](#footnote-ref-5)
5. Interpolation was conducted through the *imputeTS* package in R (Moritz & Bartz-Beielstein, 2017). [↑](#footnote-ref-6)
6. As a sensitivity analysis the REM was conducted on simulations where ‘same location’ was used as additional predictor in imputation models. This analysis ess yielded similar conclusions, but coverage was substantially lower for in-degree sender. See Appendix B. [↑](#footnote-ref-7)