# Introduction

There are numerous options for handling missing data. The simplest method is to overlook missing values and merely use the measured observations. Unfortunately, applying this method leads to loss of information, reduction of statistical power, and more worryingly, possible bias (Schafer & Graham, 2002). Other treatments involve weighting, likelihood-based procedures, and imputation and much is known about how these treatments reduce the harm inflicted by missingness (e.g., Schafer & Graham, 2002; Newman, 2014). Consequently, it is also known that multiple imputation in many circumstances is a relatively reliable method to cope with missingness in statistical analyses (Van Buuren, 2018). However, there is still a gap in the literature on the effects of missingness and treatment thereof in social networks 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 and dynamics within networks. The current study utilizes Apollo 13 data containing the timestamped messages sent among ground - and space crew which allows the modelling of communication dynamics among its members.

While the necessity of addressing missing data problems in social network analysis (SNA) is generally accepted, there has been relatively little work on how to handle social networks containing missing data. Current SNA therefore often utilizes incomplete data, resulting in the exclusion of nodes (actors) or edges (associations) between nodes. Employing incomplete data has been found to result in biased results in numerous SNA, even while missingness is randomized (Kossinets, 2006; Huisman, 2009). The current study aims to add to the few existing studies by exploring and simulating 1) the effects of missingness in *dynamic* social network models, and 2) to what extent multiple imputation of missingness results in valid results in a *dynamic* social network model. Further closing the divide between missing data treatment and SNA could improve the credibility of SNA, advancing our knowledge on social networks and their dynamics.

# 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 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 SNA, only NDD is further described as it serves as the baseline 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.

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 by a plausible value, 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 social network data.

## Social network analysis

A (social) network can be defined as a collection of nodes that are connected to each other 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 insightful initial network characteristics can be derived from such network graphs. The graph clearly shows node 7 is a central unit that connects peripheral nodes and that the triangle of 17, 18 and 19 form a closely knit community within the network. Considering node 7 represents the flight director and the triplet constitutes the three astronauts of the mission, such a network architecture is plausible (see Appendix A for the actor IDs and their roles).

Hence, social network analysis requires data on the edges between nodes, and these edges can take various forms, like connections between colleagues or 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.

**Figure 1.** Network graph (directed) of communication between nodes in Apollo-13 data.

A network of numbers and lines

Description automatically generated with medium confidence

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 exchange.

Social network analysis based on relational data implicitly suggests that the quantity and type of connections among nodes can be informative explanatory factors in predicting behavior (McGloin & Kirk, 2014). In the context of Apollo-13 communication, the edges between nodes may be informative in predicting further communication dynamics in the network. In other words, the relational history in the data may be predictive of further network dynamics and for such *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 and Harvey, 1975; Butts, 2008). A sequence of those events is then described as relational event history data and encompasses at least the times or order of events, and a pair of sender and receiver nodes (Butts, 2008). Table 1 entails the first two and last two cases of the Apollo 13 relational event history 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 event history 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 relational event history evolves by estimating the event rate λ, which determines which nodes will interact *and* when this interaction will occur (Butts, 2008; Meijerink-Bosman et al., 2023). First, it is necessary to construe a ‘risk set’ entailing all possible events that might occur at time-point *t*, resulting in a matrix of all conceivable pairs at each time-point. In the context of *directed* edges of sender *s* and receiver *r*, *s* x *r* is the matrix that represents all possible relational events. 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 pair of sender *s* and receiver *r* at time-point *t* 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 be modelled as the outcome, regressed on by chosen 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 the literature these predictors are referred to as statistics and can be either exogenous or endogenous. Exogenous statistics entail characteristics such as ‘age’ or ‘occupation’ of individual nodes and allow for researching to what extent certain attributes determine the event rate. Endogenous statistics, such as prior communication between a pair of nodes, consider the likelihood of potential subsequent events conditional on past events. By estimating the model parameters, *βp,*​ linked to the statistics, inferences can be made on the factors driving the evolution and dynamics of social interaction over time within a network.

## This study

The aim of this study is to analyze how multiple imputation of missing values affects analyses of relational events. 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 a combination of these columns. Subsequently, missing values are imputed through multiple imputation and relational event models of imputed datasets are compared to the analysis utilizing the complete communication data. In doing so, the current research improves our understanding of 1) to what extent missing data, and specifically missingness in ‘time’, impacts inferences in relational event models, and 2) to what extent sophisticated techniques for imputing missing relational event data may help reduce this impact.

# 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’ referred to an exploded oxygen tank fifty-six hours into the mission. At that moment, the mission turned from a relatively routine journey to the moon into a mission to let the astronauts return safely to earth, which they did in the end.

In the current article the relational event history is analyzed, focusing on the sequence of communication that transpired among ground and space crew rather than on the content of the communication. As such, the data contains relational events where each row represents a time-stamped directed communication from a sender to a receiver (again: see Table 1 for an impression). Note that time is in seconds from the start of the mission, and the sender and receiver columns contain the ID rather than the role of the nodes (see Appendix A for IDs and accompanying roles). After subsetting the Apollo 13 relational event history around the time of the exploded oxygen tank an edgelist containing 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 communication dynamics. Therefore, in contrast to Kalamabad and colleagues (2023, exogenous statistics such as a node’s role or location are ignored. However, the same endogenous statistics; indegree sender, outdegree receiver as well as reciprocity *are* included. Including these statistics allows for modelling network dynamics based on past relational events and compare the extent to which results from various models vary.

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.

Out-degree receiver. The out-degree of the receiver statistic refers to the number of communications a receiver node has transmitted and assumes that receivers with higher out-degree 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.

Reciprocity. Prior studies have shown that nodes that have received communications are more likely to reciprocate these in the future (Stadfeldt & Block, 2017; Kalamabad et al., 2023). The reciprocity statistic assumes there is a tendency for node 1 to reciprocate communication to node 2, based on node 2’s past communication to node 1.

…

## Analysis strategy

The analysis can be divided into three main components, and these are i) missingness simulation, ii) missingness imputation, and iii) analyses.

i) Missingness simulation. The first step in the analysis was to create missing values in the complete Apollo 13 data. For robustness, this was done by creating a hundred different incomplete versions through the NDD mechanism. Missing values were allowed to occur randomly in time, sender, or receiver data as well as in any combination of these variables.[[4]](#footnote-5) The proportion of missingness was set to 20% to allow for stable enough analyses under a substantial amount of missingness. Furthermore, to include every node in each simulated dataset, the same 1500 communications were drawn and preserved from the complete data and included in each simulated dataset. Some nodes occurred infrequently in the observed dataset and would be removed altogether in some simulations by naïve amputation. Therefore, preserving some communications resulted in the same number of nodes (16) in each simulation.

ii) 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 imputed in three ways to assess each method’s performance.

Firstly, time was imputed by interpolation as the values before and after missing times were used for imputation. Note that this method is applicable in relational event history as ‘time’ might be missing but the chronological order *is* recorded in such data. Also, this method pertains more to single imputation rather than multiple imputation but was included to serve as additional comparison. Secondly, time was also imputed through Bayesian linear regression (‘norm’) where imputations were squeezed to fall within the range of observed times. Finally, the third way was by creating a lag-1 regression were for all communications, except the first, the time before and after served as base to interpolate between. Note that this method is only applicable in simulations, as it requires the observed data, but again, was included to evaluate analyses too. {This still needs to be done}

In contrast, imputations for sender and receiver data were estimated in a single way 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). Such ‘regular’ predictive mean matching could 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 methods the target variables were regressed on both remaining variables. For instance, in the ‘sender’ imputation model both ‘time’ and ‘receiver’ were used as predictors.

3) Analyses. The third step of the analysis was the …. [Finish later when analyses to be done is clearer]

The Apollo 13 relational event history is a finite population as the sample equals the population (i.e., there are no communications in the population that are not in the data). In concordance with Vink and van Buuren (2014) this trait allows for treating the single observed Apollo 13 data as the comparative truth and exclude sampling variance from the evaluation of imputation performance.

Results fully observed Apollo 13 data

Complete case analysis REM, model evaluation stats

Analysis with incomplete data

Analysis with imputed data

* Time as for loop
* Time as lag-1 variable
* Time as exponential
* Time as post – squeeze

## Results

### Descriptive statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2.** Network characteristics of complete Apollo 13 data. | | | |
|  |  | Static |  |
| # of nodes | 16 | Density | .21 |
| # of communications | 3882 | Closeness | .68 |
| Mean degree | 242 | Eigenvector | .61 |
|  |  | Betweenness | .57 |
|  |  | Transitivity | .33 |
|  |  | Diameter | 4 |
|  |  | Average shortest path | 1.92 |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 3.** Results for the full observed Apollo 13 data. | | | |
| Statistic | *β* | *SE* | *p-value* |
| Reciprocity | 2.357-02 | 1.858-02 | .204 |
| In-degree sender | 4.314-04 | 7.400-05 | < .001/2 |
| Out-degree receiver | -9.115-05 | 7.437-05 | .220 |
| …. |  |  |  |
| *Note.* Number of possible events = 931680, number of events = 3882. | | | |

# 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.

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., 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.

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>

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.

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*.

# Layout

Introduction

Theoretical Backgroud

Missing data

Social network

Relational event models

This study

Data & Methods

Data

Measures

Analysis strategy

Results

Descriptive

Network/node level characteristics

Simulation

Discussion

Limitations & Strengths

References

Appendix

A: Apollo 13 actors and IDs

B: R Code

# 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: 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 article the ‘data dependent’ terminology from Hand (2020) is used as it conveys the missingness mechanism more clearly. [↑](#footnote-ref-2)
2. For practical reasons it is assumed that a node cannot send message to themselves or to multiple other nodes simultaneously. [↑](#footnote-ref-3)
3. The complete communication transcript can be retrieved from <http://apollo13realtime.org/> [↑](#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. [↑](#footnote-ref-5)