1. **Introduction**
   1. **Multiple imputation**
   2. **Missingness mechanisms**
   3. **Social networks**
   4. **Relational event history data**
   5. **Relational event model**
   6. **Current research**
2. **Data**
3. **Methods**
   1. **Missing data generation / amputation**
   2. **Multiple imputation**
   3. **Data analysis**
4. **Results and analysis**
5. **Discussion and conclusion**

**References  
Appendix**

**\*\* This is a draft version and the comments are just reminders for myself for the things I still need to do / look at**

1. **Introduction**

Missing data is a problem that is common across various research domains, affecting the accuracy of estimates and potentially introducing biases in parameter estimates. Thus, it can have a significant effect on the interpretation of findings. Most statistical analyses require complete data (Schouten et al., 2018). As a result, the presence of missing data may result in potentially wrong conclusions drawn from the data, which can lead to far-reaching consequences. Complete case analysis, a simple approach to handling missing data, involves removing the observations with missing values completely. However, this method causes loss of information, a decrease in statistical power and may introduce bias (Schafer & Graham, 2002). Multiple imputation (MI) offers a solution to this problem. MI is a method that creates multiple complete versions of the data by replacing missing values by plausible data values based on the patterns and relationships found in the observed data. By creating multiple complete versions, MI quantifies uncertainty in estimating missing values and does not causes loss of information. It therefore minimizes the risk of drawing incorrect conclusions (Vink & Van Buuren, 2014). Understanding the appropriate techniques for addressing missing data according to specific circumstances is crucial.

Missing data is also a common issue in social network data. Missing data can cause problems to the analysis of network data, as network measures are often based on the fully observed graph (Borgatti et al., 2006; Smith et al., 2017). Relational event history (REH) data is a specific type of social network data, which can be defined as time-ordered sequences of social interactions between a set of individuals or entities (Butts and Marcum, 2017, Meijerink-Bosman et al., 2022). This data is becoming increasingly available due to new technical developments and has the potential to greatly contribute to the understanding of dynamic social networks (Geldsite). Missing values in REH data can lead to invalid statistical inferences. However, research on the influence and how to handle missing values in social network data is limited (Huisman & Krause, 2017). Statistical tools for REH data are also currently underdeveloped (…). This paper focuses on using MI to address missing values within relational event history (REH) data.

**1.1 Multiple imputation**

MI is a method used to handle missing data by estimating and replacing each missing value multiple times. Each missing value is imputed m > 2 times, leading to *m* completed datasets. The *m* completed datasets are then analyzed independently and pooled using Rubin’s rules for combining estimates and standard errors (Rubin, 1987; Schafer & Graham, 2002). Rubin’s rules take into account both within- and between-imputation variability (van Buuren, 2018; Enders, 2022). MI differs from single-value imputation methods in that missing values are imputed many times, whereas single-value imputation methods estimate what each missing value might have been and replace it once with a single value. This method avoids creating false precision, as could be the case with single imputation. MI provides accurate estimates for the metrics of interest. It also minimizes the risk of drawing false-positive or false-negative conclusions (Peng et al., 2015).  
 MI consists of two stages. Firstly, imputations (replacement values) for missing values are generated. This results in many datasets with different replaced missing values, as is shown in the imputed data part in Figure 2. The imputations are generated based on statistical characteristics of the data. Secondly, the imputed datasets are analyzed (analysis results, Figure 2) and the results of these analyses are combined (Peng et al., 2015). This is the pooled result.  
Afbeelding met tekst, diagram, schermopname, Lettertype

Automatisch gegenereerde beschrijving

It is often assumed that data when using MI is sampled from infinite populations. However, if sampling from an infinite population is assumed while it happened to be sampled from a finite population, there may be an overestimate of the variance of the estimates. This leads to a loss of statistical efficiency as confidence intervals are longer than which is true according to the true data. In the case of pooling multiple imputations when the sample happens to be the population, simplified pooling rules that only account for the variation caused by the mechanism that created the missing data, need to be used (Vink & Van Buuren, 2014). This has to be considered when using MI.

**1.2 Missingness mechanisms**

Solving a missing data problem is challenging. Although there are techniques such as multiple imputation (Rubin, 1987; Little and Rubin, 2002), which are proven to be effective and intuitive, it is important to think about the nature of the missingness. The degree to which the observed and unobserved data are connected, may be of great influence on the validity of the imputation method. Inclusion of a variable that correlates either with the incomplete variable or with the missing values improves parameter estimates (Collins et al., 2001). This is the reason why predictor variables are often included in imputation methods.   
 A set of beliefs need to be formulated about the extent to which the observed data also applies for the missing parts of the data. It is important to know why the data is missing when handling missing data. We distinguish three different mechanisms: 1) Missing Completely At Random (MCAR); the probability of being missing is the same for all cases. 2) Missing At Random (MAR); the probability of being missing is the same only within groups defined by the observed data. 3) Missing Not At Random (MNAR); the probability of being missing varies for reasons unrelated to the observed data (Rubin, 1967; Van Buuren, …). These mechanisms can be further explained using a data matrix Y with 𝑦𝑖𝑗 either observed or missing, where *Yobs* is the observed data and Ymis is the missing data. Matirx *R* is considered the response indicator with Rij = 1 if *yij* is missing and Rij = 0 if *yij* is observed (Van Buuren 2012:30-35). Ψ are fixed parameters of the probability model (Van Buuren 2012:6).

Missing Completely At Random (MCAR)

With MCAR missingness, the observed data form a randomly obtained subset from the population. The probability of a variable being missing is independent from the observed and unobserved data. The missingness is thus not related to the data. This can be represented by the formula:

𝑃𝑟(𝑅=1|𝑌𝑜𝑏𝑠,𝑌𝑚𝑖𝑠,ψ)=𝑃𝑟(𝑅=1|ψ)

The missing values are solely induced by ψ and independent from the observed and unobserved data. The observed data and the missing data are thus exact representations of the true data model. When the data are MCAR, the remaining data can be considered a simple random sample of the full dataset (Book managing missing data in patient etc). MCAR missingness is also called *ignorable* missing data mechanisms, because bias is nonexistent and power is restored with modern treatment (Rubin, 1967; Little et al., 2014).

Missing At Random (MAR)

With MAR missingness, the probability of a value being missing depends on the values of the observed variable. This can be represented by the formula:

𝑃𝑟(𝑅=1|𝑌𝑜𝑏𝑠,𝑌𝑚𝑖𝑠,ψ)=𝑃𝑟(𝑅=1|ψ,𝑌𝑜𝑏𝑠)

The observed and missing data represent different parts of the population. However, with the right conditioning of the incomplete variables on the observed data, the results will still be statistically valid. MCAR missingness is also called *ignorable* missing data mechanisms as the bias is recoverable and power is restored with modern treatment (Rubin, 1967; Little et al., 2014).

Missing Not At Random (MNAR)

With MNAR missingness, the probability of a value being missing depends on the unobserved information. The missingness is related to events or factors which are not measured by the researcher (Book). This can be noted as:

𝑃𝑟(𝑅=1|𝑌𝑜𝑏𝑠,𝑌𝑚𝑖𝑠,ψ)=𝑃𝑟(𝑅=1|ψ,𝑌𝑜𝑏𝑠,𝑌𝑚𝑖𝑠)

The observed data alone is not enough to infer about the population. The missing data are called *nonignorable* (Rubin, 1967). The response and nonresponse represent different and unique parts of the true data.

When the variables in a dataset show low correlations, the distinction between MAR and MCAR missingness may become difficult. MI with MAR mechanisms would then primarily limit statistical power and increase variance without necessarily reducing the bias. This is also true for assuming MNAR missingness when data is highly correlated. Therefore, it is important to consider which mechanism to assume based on the observed data structure (Schouten & Vink, 2021). Imputation methods, such as multiple imputation, EM imputation and regression imputation, are all valid provided the missingness mechanism is not MNAR. The percentage of missingness should also not be to high (Peng et al., 2015; Sterne et al., 2019; Scheffer, 2002).

**1.3 Social Networks**

Networks are a collection of nodes (points) joined together in pairs by edges (lines). There are many examples of systems which are networks in the physical, biological and social sciences. These sciences all have different types of networks, which can be divided into four broad categories: technological networks, information networks and biological networks (Newman, 2018). Here, we will focus on social networks.

Social networks can be defined as any network in which the nodes (*actors*) represent individuals and the edges (*ties*) represent the relationships/connection between them, such as friendships or interactions (Newman, 2018). Social networks are based on the representation of social structure in terms of a set of social entities, such as people and organizations, that are connected via relationships (Wasserman & Faust, 1994; Carrington et al., 2005). The network is not static, relationships in social networks are dynamic, they can change over time. However, some social networks exhibit stability over time. This dynamic social network requires different models than static networks. (e.g. Snijders, 2001; Koskinen & Snijders, 2007; Almquist & Butts, 2014; Krivitsky & Handcock, 2014). Social networks have a high flexibility, many different definitions of an edge are possible and they can thus serve as a good representation of different phenomena (Butts, 2009; Newman, 2018). Most current models view relationships as changing over time, discrete or continuous. These changes are driven by mechanisms whose presence and strength that can be estimated from intertemporal network data. This models thus allow for better understanding of social networks. The representation of the social environment as patterns in relationships among interacting units is the reason why many researchers are interested in the network perspective (Wasserman & Faust, 1994).

Social network analysis can provide insight into the underlying relationships between individuals, which can reveal patterns (e.g. interactions, communication, relationships) that cannot be detected from the individual observations alone. Social network analysis is based on the assumption that relationships among interacting individuals are important (Wasserman & Faust, 1994; Serrat, 2017). It shows the formal and informal relationships between individuals and can be used to understand what facilitates or impedes the existence of ties between edges. SNA has gotten much more interest in social and behavioral sciences as the availability and technical tools of social network data increases (Carrington et al., 2005).

**1.4 Relational Event History data**

Social network data can come in different forms, for example network panel data.’

En wat is het…………

Relational Event History (REH) data is a type of social network data, which describes a time-ordered series of interactions between actors in a network. These interactions are also known as relational events. Minimally, the relational events contains information about the actors that are involved in the event and the time of the event (Meijerink-Bosman et al., 2023). The interactions between social entities in REH data are discrete instances. This is in contrast to the conventional social network setting in which the ties are temporally extensive. REH data focuses on individual interactions that occur at specific time moments, as in conventional social network data, the relationships are presented as ongoing over time. The interactions of REH data are considered as discrete events in continuous time, so they are chronologically ordered by time (Butts, 2008). Although REH data is not that rich and flexible because there is a duration limit as a dynamic network, it is particularly useful for studying the social relationships that drive the interactions.

A relational event can thus be defined as an action initiated by one entity and directed toward another entity within its environment (Butts & Marcum, 2017). A relational event thus consists of a sender, receiver and action, in which the sender is the sender of the action, the receiver the receiver of the action and the action the type of action at a given time point. Both the sender and receiver can consists of humans, animals, objects or a combination of multiple type of actors. Actions can consist of a variety of relationships between the actors. Multiple of these events combined and ordered by time given a time window, result in REH data (Marcum & Butts, 2015).

**1.5 Relational Event Model**

Analysis of REH data can help researchers answer questions, such as the most basic level question “what drives what happens next?” in a complex sequence of interdependent events (Marcum & Butts, 2015). This type of data is becoming more and more popular for the analysis of relational dynamics. Relational event dynamics are fundamentally about sequential relational structures, which differs from the social network analysis as the primary interest thereof is the simultaneous relational structure.   
 REH data is characterized by the inherent dependency between nodes and edges (Meijerink-Bosman et al., 2023). Consequently, this type of data is not handled very well by traditional statistical methods. Specialized tools are therefore needed for analysis. Relational Event Model (REM) is a statistical model that can take this into account (Butts, 2008). REM is built to analyze continuous, detailed, social interaction data, such as is found in REH data (Meijerink et al., 2023). It is used to understand communication structures based on observed social interactions in real-time (Butts, 2008).

REM can be used to examine the frequency and time to activation among relational events (pp). With the REM, endogenous and exogenous factors can be investigated that influence the evolution of the relational events over time (REH data). REM works by parameterizing the interaction rates between actors as a function of endogenous and exogenous statistics, based on the event history (Butts, 2008). The endogenous statistics contain the internal information until a given time point (e.g. past interactions), while exogenous statistics contain external information (e.g. attributes, age, sex). The REM model can be categorized into tie-oriented models and actor-oriented models. In tie-oriented models, the probability of a dyad interacting next is modeled in a single step (Butts, 2008). Actor-oriented models, first model the probability of a sender initiating an interaction and then the probability of the senders’ choice of receiver (Stadtfeld & Block, 2017). The outcome of the REM show the extent to which the specified statistics affect social interaction behavior in the network (Meijerink-Bosman et al., 2023).

**1.6 Current research**

Missing data is common in REH data. Missing data can cause problems to the analysis of network data, as network measures are often based on the fully observed graph (Borgatti et al., 2006; Smith et al., 2017). However, research on the impact and how to handle missing values in social network data is limited (Huisman, 2009; Huisman & Krause, 2017). Statistical tools for REH data are also currently underdeveloped (bron), although REH data is everywhere (e.g. Wiki edits, digital communication, video monitoring). Developing valid methodologies to address missingness in REH data is essential, as REH data is becoming increasingly available due to the development of technology and has the potential to greatly contribute to the understanding of dynamic social networks (bron + pp).

This paper gives more insight into the use of multiple imputation as an approach to missingness in REH data using REM. It aims to assess the effectiveness of MI to produce accurate estimates of missing data and evaluates its impact on the REM estimates. The research question of this paper is: *Can multiple imputation with REM be applied on REH data with missing values to produce valid inference?* The remainder of this paper has the following structure. First, I will provide information about the data used in the study. I will then outline the methods used in the study in the third section and present the results and analysis in the fourth section. In the fifth section, the conclusion and implications of these findings will be discussed.

1. **Data**

The data used in this study is part of the Apollo 13 dataset (……). Apollo 13 was the seventh mission in the Apollo program by NASA. The mission ended early due to an explosion in the oxygen tank. The communication between the flight and ground crew ensured a safe return on the ground. This unusual occurrence resulted in a well-documented dataset capturing the crisis communication between the flight and ground crew. The real-time playback of the events following the incidents is available on the Apollo 13 Real-time website (*Apollo 13 Real-time,* n.d.). The full Apollo 13 dataset is publicly accessible on GitHub (Tseng, 2020).

The part of the Apollo 13 dataset used, consists in total of 38982 rows and three columns: time, sender and receiver. Each row represents a single, directed communication event; the relational event. The sender column represents the actors that initiated the communication at a corresponding time point. The receiver column represents the actors that were the target of this communication initiated by the sender. The time column consists of unique exact time points of the communication initiated. The actors in the sender and receiver column are indicated with numbers. A total of 16 unique actors are present in the dataset. The dataset is fully observed, no data preparation is needed.

The Apollo 13 dataset consists of REH data. REH data is distinct from panel data in the sense that the ties in REH data are short lasting, no unobserved tie changes and occur in exact moments in time. In combination with the rapid increase of available REH data and the potential to greatly contribute to the understanding of dynamic social networks, REH data is valuable for understanding social networks. The Apollo 13 dataset thus gives an ideal illustration for the use of multiple imputation on missing values in REH data.

1. **Method**

To assess the effectiveness of MI on handling missing values in REH data, the criteria for evaluation must be determined. The missing values are induced in the complete dataset (amputation) and the missingness mechanism is defined. Subsequently MI is performed, generating multiple complete datasets, each of which is analyzed using REM. The results of these analyses are combined. Additionally, the REM is applied to the fully complete dataset. The accuracy of estimates of the MI and its impact on the REM estimates is examined. Below, each step will be discussed in more detail.

**3.1 Missingness data generation**

The generation of missing values in a complete dataset is called amputation. The Apollo 13 dataset is a fully observed dataset, to evaluate MI on this dataset missing values have to be induced. The package MICE (Multivariate Imputation via Chained Equations) is used in R (bron R, Van Buuren & Groothuis-Oudshoorn, 2011). This implements a method to handle missing data. The method is based on Fully Conditional Specification (FCS), which imputes each incomplete variable by a separate model. MICE works by iteratively imputing missing values using regression models based on the observed data (Van Buuren & Groothuis-Oudshoorn, 2011).

Based on “Strategies for simulated missingness” (Vink, 2022), missingness was simulated with the ampute() function (Schouten et al., 2018) from the package MICE. Model-based finite populations was used, a single finite observed set is taken as the comparative truth. The sampling variance from the evaluations of the imputation performance can thus be eliminated (Vink and Buuren, 2014). The noise induced by the sampling mechanism will not be taken into account, as this is not the topic of interest of this research. The proportion of missing was set at 50%/20%. Missingness was simulated a hundred times, to ensure the validity of the results. The missing data generation process was conducted two times: once assuming MCAR as the missingness mechanism and once assuming right-tailed MAR. All conditions are the same for MCAR and MAR. Verdiep verder in de keuze voor 1500 observaties en leg uit……………… The missing data pattern considered for amputation are all possible combinations of missingness involving sender (actor 1) and receiver (actor 2).

**3.2 Multiple imputation**

The variables sender, receiver and time are used as predictors for the missing values. In the past, it was thought that using a lower number of imputations, around 3-5 imputations, was sufficient to obtain excellent results (Schafer & Olsen, 1998). More recent research by Van Buuren (2018) and Graham et al. (2007) show that a higher number of imputations than previously recommended, would lead to better outcomes. However, considering the computational efficiency, the number of imputations is set to five. The number of imputations refers to the number of datasets generated, each containing missing values generated based on the number of iterations. The number of iterations is also set to five, as inferential validity is often achieved after five to ten iterations (Oberman et al., 2021). Proper convergence is achieved in 5 to 20 iterations (Van Buuren, 2018). The number of iterations refers to the amount of iterations through each variable to estimate new values for missing data.

A custom version of the predictive mean matching (pmm) method, pmm.conditional, was used as the method for imputation (tilburg bron). This method works by first predicting missing values using a regression model based on the observed or other values. Then it imputes values from a the observed dataset that are the closest to the predicted values in the regression model for the missing value. Predictive mean matching avoids creating loops in the data, which made it possible to use time as a predictor and prevents that actors are initiating interactions with themselves. The imputed value for the corresponding receiver or sender cannot be the same, this will be ensured using the pmm.conditional method. The generation of MCAR missingness can be seen as the situation in which all weight values are zero. Since the probability of being missing with MAR missingness by definition depends on the value of the observed variables, the weights of the variables that will be amputed is set to zero (Schouten et al., 2018).

MICE implements an iterative Markov Chain Mote Carlo (MCMC) algorithm, which refers to simulations by sampling many random values from a posterior distribution (Hammersley & Handscomb, 1964). The convergence and plausibility of the imputation is checked. The convergence is checked by the convergence plot for MAR and MCAR simulations. The convergence plot show the change in imputation estimates as the number of imputations increases and can thus tell whether the algorithm has stabilized or further iterations are needed for reliable imputations. Plausibility is checked by the density plots for MAR and MCAR simulations. The density plots show the distribution of the imputed values and the skewness of the imputed data distribution. The quantile-quantile and cumulative distribution plot also show the distribution of the imputed values (Nguyen et al., 2017). The Kolmogorov-Smirnov test is used to compare the distribution of the imputed values to a theoretical distribution in the quantile-quantile plot. For the cumulative distribution plot the distribution of the imputed values is compared with a fitted distribution.

**3.3 Data analysis**

The function remify() (Arena, 2003) and remstats() (Meijerink-Bosman et al., 2023) are used to calculate the statistics on the fully observed dataset. Three endogenous and one exogenous statistic is included into the model. The endogenous statistics used in this study are according to the statistics used by Kamalabad et al. (2023): reciprocity, indegree sender and outdegree receiver. The exogenous statistics is …

Reciprocity

DEFINITION RECIPROCITY

Indegree sender

DEFINITION

Outdegree receiver

DEFINITION

LAST ONE

DEFINTION

The data is altered before the Cox proportional hazard (coxph) function. A risk set is created, which consists of every possible interaction at a given time combined with the status of whether that interaction took place in the observed data. The status column shows a 1 when interaction took place and a 0 when no interaction took place. This altered data is used in the coxph function from the package (Therneau, 2023), which make the REM analyses possible. REM is fitted and pooled with each simulation. Reciprocity, indegree sender, outdegree receiver and … are set as the survival predictors in the coxph() function. The analysis is run over a hundred simulations for the imputed datasets assuming MCAR and the imputed datasets assuming MAR. These results are pooled and averaged for MCAR and MAR. The imputation method is evaluated using the following measures: raw bias (RB), percent bias (PB) coverage rate (CR), confidence interval (CI) and average width (AW).

1. **Results and analysis**

The code for the results work. Only need to run it with the right statistics.

1. **Discussion and conclusion**