# Chain event graphs for missing data: exploring informative missing data, with examples from longitudinal studies

Jane Hutton*
The University of Warwick, Coventry, UK - J.L.Hutton@warwick.ac.uk

**Abstract**

Chain event graphs (CEGs) extend graphical models to address situations in which, after one variable takes a particular value, possible values of future variables differ from those following alternative values. This provides a framework for modelling discrete processes which exhibit strong asymmetric dependence structures, such as health status for smokers and non-smokers. These graphs are derived from probability trees by merging those vertices in the trees whose associated conditional probabilities are the same. It is possible to score each model efficiently and in closed form. Hence standard Bayesian selection methods can be used to search over a wide variety of models, each with its own explanatory narrative.
The expression 'missing at random' (MAR) was coined by Rubin in 1976, but the concepts of informative missingness compared to MAR are not simple to explain in studies with several or many variables which are only partially observed. In much research, knowledge of the design of data collection and the subject area will indicate that missing data is likely to be related to variables of interest. Problems caused by missingness can be especially acute in longitudinal data analyses when it is typical for substantial amounts of data about certain units in the sample to be missing at some of the observation times.
In order to use this background knowledge, we developed new classes of models in which the dependence of missingness, as well as independence conditional on observed values of variables can be displayed. The CEGs can capture data which are missing not at random but nevertheless exhibit context-specific symmetries. One of the advantages of this method is that the selected maximum a posteriori model and other closely scoring models can be easily read back to scientists in a graphically transparent way.
The efficacy of our methods are illustrated using a longitudinal study from birth to age 25 of children in New Zealand, and a simulation based on a study of weight loss.
**Keywords**: chain event graphs; graphical models; probability trees; missing data.

## 1. Introduction

A new class of graphical models, Chain Event Graphs (CEGs) (Smith 2008) provide a powerful framework for the study of categorical data deriving from discrete processes which have an associated probability tree. The CEG can be seen as a graphical model which generalises discrete Bayesian Networks by taking into account asymmetries within the tree structure representation of the problem. Straightforward Bayes Factor search methods have been shown to lead to promising CEG models, which not only score significantly higher than the closest maximum a posteriori (MAP) Bayesian Network but also provide a refined set of conclusions (Barclay et al, 2013). An extended graph, the ordinal CEG, provides a more refined graphical representation for the effect of covariates on a binary outcome variable.
The idea of 'missing at random' (MAR), coined by Rubin in 1976, is appealing: given the data which is observed, there is no further information to be gained from knowing which variables for each subject were not observed. However, the concepts of informative missingness, or 'missing not at random' (MNAR) compared to MAR are not simple to explain (Seaman et al, 2013). In many studies, knowledge of the design of data collection and the subject area will indicate that missing data is likely to be related to variables of interest. In order to use this background knowledge, we developed new models in which the dependence of missingness, as well as independence conditional on observed values of variables can be displayed (Barclay et al, 2014). A graphical representation can be given of various missing data mechanisms that are not fully random but nevertheless might exhibit various conditional independence relations associated with the underlying tree. We demonstrate how the CEG lets us trace back the path each individual takes in the tree and can explicitly distinguishes the missing category from the remaining categories within its structure. This addresses the

problem of the naive mis-estimation of conditional probability which can arise when missingness is treated as an additional category for each variable. The probability of that an individual variable is missing category might depend strongly on the values of other variables.
In this paper we demonstrate how we can further exploit the structure of the CEG to represent different missing data mechanisms and show how this framework helps to differentiate different hypotheses associated with processes that lead to missing not at random (MNAR) data structures. An ordering of the variables in the tree is chosen so that the resulting MNAR models can be estimated.

## 2. Motivating examples

The Christchurch Health and Development Study (CHDS) followed up 1265 children born in Christchurch, New Zealand, with health and social data collected at a range of ages. Here we consider the effect of birth family structures and drug use aged 16-18 on hospital admissions aged 21-25 years, no related to pregnancy. The family type at birth,with no missing data, was: both parents, single parent, adopted by a couple. At age 16-18, subjects were asked about drug abuse, coded as none or user. Hospital admission was coded as no admission or at least one admission. Both drug abuse and admissions had missing data.
The second example will be based on a study which compared two methods of weight loss, with participants weighed at entry and then each month, with those who had lost 8% of initial weight by four months being eligible for a further study.

## 3. Chain Event Graphs

A CEG is defined using the following example based on the CHDS, with $Y_1$ = Birth family, $Y_2$ = Drug abuse and $Y_3$ = Admissions. The event tree is given in Figure 1; edges with the same colour have the same conditional probability.
We say that two non-leaf vertices, $v_i$ and $v_j$, of the tree, $T$, are in the same **stage**, $u$, if and only if the topology of florets $F(v)$ and $F(v^{'})$ the same, and there is a bijection between the florets such that the probabilities on corresponding edges are the same. The tree in Figure 1 has stages:

$$u_1 = \{v_1\}, u_2 = \{v_2, v_3, v_4\},$$

$$u_3 = \{v_5, v_7\}, u_4 = \{v_6, v_8, v_9\},$$

$$u_5 = \{v_{10}\}$$

**Positions** give a finer partition of the tree. Two vertices are in the same position, $w$, if their sub-trees have the same topology such that we have a bijection between the edges of the two sub trees and the conditional probabilities associated with the edges are the same. The tree in Figure 1 has positions:

$$w_1 = \{v_1\}, w_2 = \{v_2, v_3\},$$

$$w_3 = \{v_4\}, w_4 = \{v_5, v_7\},$$

$$w_5 = \{v_6, v_8, v_9\}, w_6 = \{v_{10}\}$$

$$w_\infty = \{l_1, l_2, l_3, \ldots, l_{10}, l_{11}, l_{12}\}$$

The CEG, $C(T)$, is the staged tree collapsed over its positions: the vertices are given by the positions of the tree. The leaf nodes are collected in a single position, $w_\infty$. Further, the edge set is defined as follows: Let $v(w)$ be a representative vertex for position $w$. Then there exists an edge from a position $w$ to a position $w^{'}$ in the CEG for every edge from the representative vertex $v(w)$ to any vertex $v \in w^{'}$. The stages of the tree are represented in the CEG by connecting two positions that are in the same stage by an undirected dotted line (Smith 2008); Figure 2 gives a CEG for CHDS variables.
Vertex $w_5$ represents participants who at birth had both or adoptive parents, and who used drugs at age 16-18, as well as those with a single parent at birth who did not use drugs at age 16-18; these are the paths to the same probability of admission to hospital. The sink node $w_\infty$ combines all categories of the admission variable.
An ordinal CEG can be defined if the leaf nodes represent a binary outcome, as there will be a unique ordering of the probabilities of the zero, or 'success' category of the outcome. The definition of an ordinal CEG is:

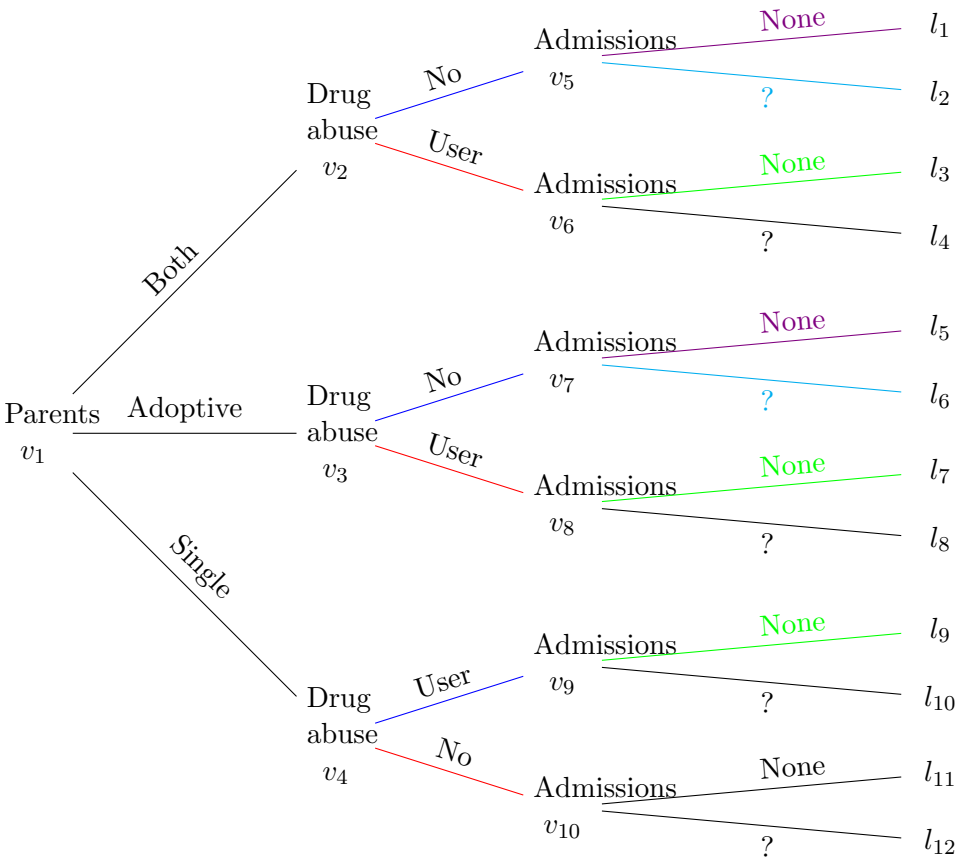


Figure 1: Example of a tree, $T$, on three variables, with coloured stages

Let $T$ be a tree on $p$ variables with a binary outcome variable $Y_p$ represented by the leaf nodes in the tree. We say that a CEG, $C(T)$, is an ordinal CEG with respect to $Y_p$ when the positions in each vertex subset associated with a variable $Y_i$, $V_{Y_i}$, are vertically aligned in descending order with respect to the predictive probability $P(Y_p = 0|D, C(T))$.

**4. Ordinal CEGs for missing data**

When data are MAR, the probabilities for the outcome associated with the missing category are a weighted mean of the probabilities for observed categories. Hence, Figure 3 is an example of an MAR CEG.

If the data were missing completely at random, the CEG in Figure 3 would have $w_2, w_3, w_4$ in same position. If the fitted ordinal CEG has conditional probabilities for a missing category which lie outside the corresponding probabilities for the observed categories, we can conclude that the data are MNAR; Figure 4 is an example.

In this case, the conditional probability of no hospital admissions for those with missing data on drugs is lower than for either observed category.

It is possible for the conditional probabilities of only some sub-trees of a vertex to show a MAR pattern. In this case, we can say that data are MAR conditional on the variable associated with that vertex. For example, the CEG in Figure 5 shows that, drug use is MAR for children in two parent families, but MNAR for children with a single parent.

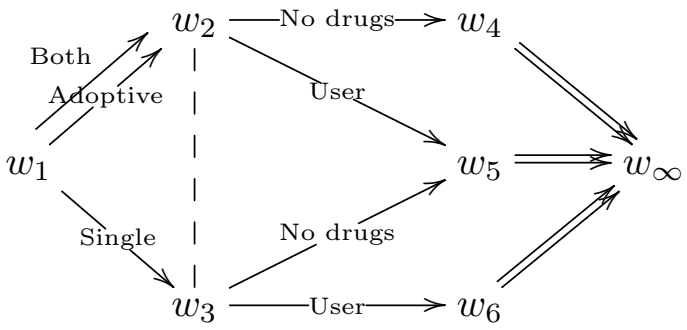


Figure 2: Possible CEG for CHDS

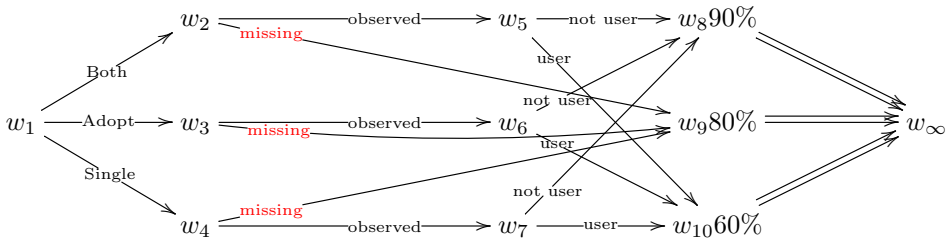


Figure 3: An ordinal CEG when data are MAR

**5. Example: Missing data tree and CEG for CHDS**

The observed percentages with no admissions for the CHDS are shown on the tree in Figure 6.

Calculating the scores for the tree, with uniform prior distributions on root to leaf paths leads to the conclusions that admissions for children from two parent families are not related to drug use, but for single parent families, drug use is associated fewer young adults avoiding admission. Missing data on drug is highly informative.

A full ordinal CEG structure may be complicated with many positions in each vertex subset. A reduced ordinal CEG focuses on the positions in the final subset, $Y_p$. Paths to these positions are re-expressed as new variables. The example of repeated measurements on weight over time will be used to illustrate this.

**6. Conclusions**

Ordinal chain event graphs provide conclusions directly, and can represent missing data structures explicitly. There are various possibilities for informative prior distributions, such as using population rates for hospital admissions in the CHDS example. The variables should be entered on the tree in a time or other logical ordering. As the number of covariates increases, the trees obviously expand rapidly. One approach is to consider a few variables, find the CEG, and use it to define new variables so that the number of categories, and hence vertices, is reduced. Further variables can then be added. Ordinal CEGs provide a framework for defining new variables. Sparsity can be awkward with trees, as with large dimensional contingency tables.

**References**

LM Barclay, JL Hutton, and JQ Smith. (2013) Refining a Bayesian Network using a Chain Event Graph. *Int. J. Approx. Reason.*, 54:13001309.

LM Barclay, JL Hutton, and JQ Smith. (2014) Chain event graphs for informed missingness. *Bayesian Analysis*, 9:53–76.

S Seaman, J Galati, D Jackson, and J Carlin. (2013) What is meant by missing at random? *Statist. Sci.*, 28:257–268.

JQS Smith. (2008) Conditional independence and chain event graphs. *Artificial Intelligence*,172:42–68.

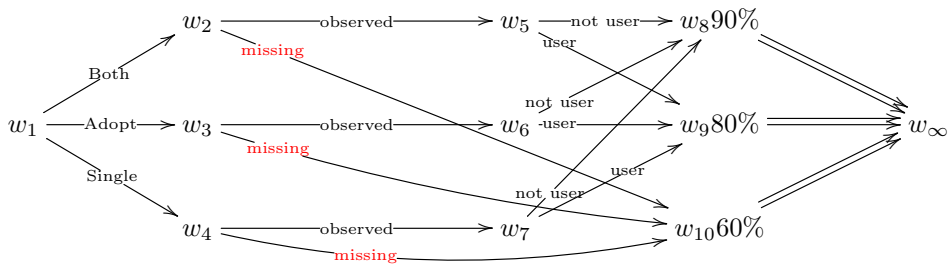


Figure 4: An ordinal CEG when data are MNAR

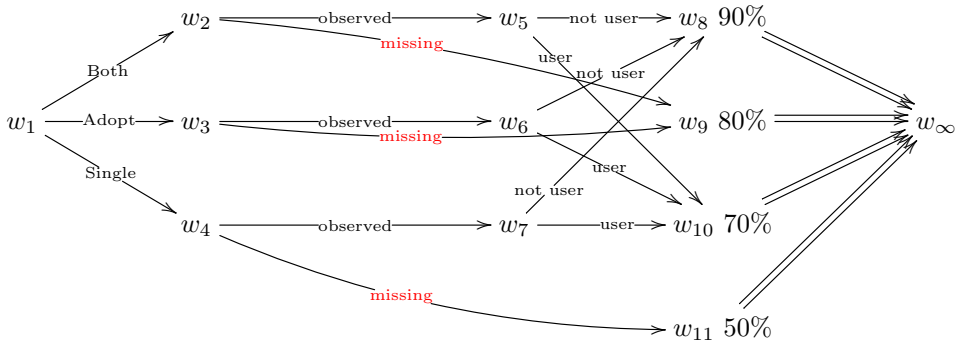


Figure 5: An ordinal CEG when data are MAR given two-parent family.

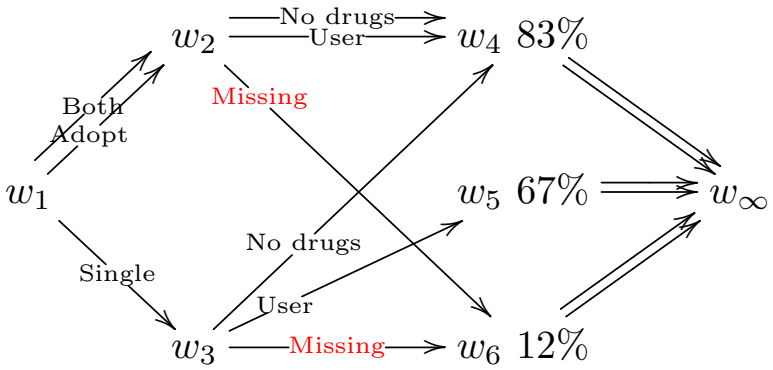


Figure 6: Missing data tree for CHDS