Semantics of Causal Loop Diagrams
Draft

Sólrún Halla Einarsdóttir

December 15, 2016

A Causal Loop Diagram (CLD) is a directed graph used to display causal relationships between variables. The nodes represent the variables and the edges represent qualitative causal relationships, which can be positive or negative.

We are interested in writing a DSL for CLDs in Haskell and propagating signs through CLDs using constraint programming, and would like to define a semantic that aids us in that.

We denote a positive causal relationship between A and B by $A \xrightarrow{+} B$ and a negative one by $A \xrightarrow{-} B$. Then $A \xrightarrow{+} B$ means that an increase in A causes an increase in B, while a decrease in A causes a decrease in B. On the other hand, $A \xrightarrow{-} B$ means that a decrease in A causes an increase in B, while an increase in A causes a decrease in B. We denote the sign of the edge from A to B by s_{AB}, so $s_{AB} = +$ if $A \xrightarrow{+} B$ and $s_{AB} = -$ if $A \xrightarrow{-} B$.

1   QPN approach

A qualitative probabilistic network (QPN) is defined as a directed acyclic graph $G = (V, E)$ where the vertices, V, correspond to variables and the edges, E to qualitative probabilistic influences. These influences can be positive (+), negative (-), or ambiguous (?).

The meaning of signs on edges is defined according to first order stochastic dominance, that is $s_{ab} = +$ implies that for all a_1, a_2 where $a_1 \geq a_2$, we must have

$$P(b \leq b_0 | a_1, x) \leq P(b \leq b_0 | a_2, x)$$

for all possible values b_0 of b and all consistent contexts x. The definition of $s_{ab} = -$ is the same but with $a_1 \leq a_2$.

In simpler terms, $s_{ab} = +$ means that greater values of a mean greater values of b are more likely, and $s_{ab} = -$ means that greater values of a mean smaller values of b are more likely.

These influences are symmetric, that is, $s_{xy} = s_{yx}$. Due to this symmetry it is possible to propagate an observed increase or decrease of one variable

around the graph and find whether other variables are then likely to have increased or decreased.

An overview of the subject can be found here: `http://www.staff.science.uu.nl/~renoo101/Prof/Research/qpns.html`.

This definition is broad enough to apply to many different systems and be applicable to various real world situations.

We found some issues with QPNs that lead us to explore other approaches. First of all, since QPNs were originally defined for acyclic graphs and the theory on them relies on acyclicity, they may not be the best fit to describe CLDs, in which loops are an important feature. Inference on QPNs containing loops is difficult to implement and can lead to ambiguous results.

Second, the formal semantics of inference on QPNs is difficult to formalize since it relies heavily on not-so-simple probability theory.

Additionally, as QPNs are defined solely based on qualitative relationships there is no obvious way to expand them to also describe quantitative relationships.

Lastly, since all inference in QPNs is probabilistic it leads to results that may not be as meaningful or concrete as we would like, such as "there is a heightened probability that x has increased", rather than "x has increased". For instance, a variable may decrease even though the cumulative probabilistic influence on it is positive.

2 Difference equation approach

Inspired by a system of tanks with water flowing from one to another, and in search of semantics that might also be extended to quantitative reasoning, we came up with the following approach.

We consider the value nodes of the graph to be functions of the same variable, such as a time variable t.

If we have a graph with two nodes, x and y, and one edge from x to y, then $s_{xy} = +$ implies that

$$\frac{\partial y}{\partial t} = F(x(t)),$$

where F is a monotone increasing function (monotone decreasing for negative causality, $s_{xy} = -$).

If the node y has multiple parent nodes $x_1, \ldots, x_n$, then $\frac{\partial y}{\partial t}$ depends on all the parent nodes.

In general we can then describe the causal relationship from x to y as

$$\frac{\partial \left(\frac{\partial y(t)}{\partial t} \right)}{\partial x(t)} = f(x(t)),$$

where f has a prime function F such that F is monotone increasing if $s_{xy} = +$ and monotone decreasing if $s_{xy} = -$.

This is somewhat more nuanced than $CLDs$ as they are described above, where $s_{xy} = +$ implies that an increase in x leads to an increase in y, and a decrease in x to a decrease in y. Here we may have some threshold value x_0 for which $F(x_0) = 0$, above which x causes an increase in y, but an increase in x causes a faster rate of increase in y and a decrease in x causes a slowed rate of increase in y, and vice versa.

Note that though $F(x)$ is monotone increasing, it may not be strictly increasing, so we could for instance have $F(x) = 0$ for all $x < C$ for some threshold value C.

If the y has parent nodes $x_1, \ldots, x_n$, then we have something like

$$\frac{\partial y}{\partial t} = \sum_{i=1}^{n} F_i(x_i),$$

where F_i is monotone increasing if $s_{x_i y} = +$ but monotone decreasing if $s_{x_i y} = -$.

In a discrete time system we consider $\Delta(x_t) = x_t - x_{t-1}$ instead of $\frac{\partial x}{\partial t}$, and write $\Delta(x_t) = F(y_{t-1})$ instead of $\frac{\partial x}{\partial t} = F(y(t))$. In simple cases we may only consider one time step with two values of t, t_{start} and t_{end}.

Here we explore how this approach relates to qualitative reasoning, but it could be extended to quantitative reasoning by solving appropriate differential equations.

2.1 Simple qualitative model

We consider a qualitative discrete time system where all values of node variables are either +, -, 0, or ? (ambiguous).

For simplicity we can consider the case where all initial values are set to zero and $F_i(0) = 0$ for all edges. If F is monotone increasing and G monotone decreasing we then write $F(+) = +$, $F(-) = -$, $G(+) = -$, and $G(-) = +$.

This is convenient for qualitative reasoning since then we are only concerned with increases and decreases rather than numerical values. The final values of the variables then tell us whether there was a net increase or decrease for each variable.

Consider a graph with three nodes, a node z and its two parent nodes x and y, $x \xrightarrow{s_{xz}} z$ and $y \xrightarrow{s_{yz}} z$. Then we have

$$\Delta(z_t) = F_{xz}(x_{t-1}) \oplus F_{yz}(y_{t-1}),$$

where F_{xz} and F_{yz} are monotone increasing or decreasing in accordance with s_{xz} and s_{yz}. Then we can say the total effect on z is $s_{xz} \oplus s_{yz}$.

3

Consider a graph with three nodes a, b and c and two edges, $a \xrightarrow{s_{ab}} b$ and $b \xrightarrow{s_{bc}} c$. Then we have

$$
\begin{aligned}
\Delta(c_t) &= F_{bc}(b_{t-1}) \\
&= F_{bc}(b_{t-2}) \oplus \Delta(b_{t-1}) \\
&= F_{bc}(b_{t-2} \oplus F_{ab}(a_{t-2}))
\end{aligned}
$$

If $b_{t-2} = 0$, for instance if $t - 2$ is the initial time value, then we have

$$
\Delta(c_t) = F_{bc} \circ F_{ab}(a_{t-2}),
$$

and the effect of a on c is $s_{ab} \otimes s_{bc}$.

3 Next steps

We will attempt to model CLDs within the `GenericLibrary`.

We will also formalize the semantics of `GenericLibrary`, perhaps using the difference equation approach described above.