

# First-order causal knowledge

Sean Stromsten (Sean\_Stromsten@Brown.edu)

Department of Cognitive and Linguistic Sciences; Box 1978

Brown University

Providence, RI 02912 USA

## Abstract

Recent advances in normative frameworks for causal reasoning have inspired psychologists to model human causal inference using these normative frameworks. These models are, however, vulnerable to the same criticisms as connectionist and other probabilistic models. A proposed response to these criticisms is the notion of first-order causal rules, which are a simple generalization of dependencies in causal Bayes nets, adding universally quantified logical variables.

## Psychology adopts normative theories of causality

Psychologists have long noted that when people encode correlations, they tend to do so in terms of *reasons why* elements should covary (for a review, see Murphy & Medin, 1985). One important species of such reasons is *causation*. Causation, however, is not merely a psychological phenomenon. Causes and effects, regarded as full-fledged citizens of the world, have gained in scientific respectability of late, because of new work on normative theories for representing, reasoning with, and learning about causality (Spirtes, Glymour, & Scheines, 1993; Pearl, 2000).

The most influential recent formalism for causal knowledge is causal Bayes nets (hereafter, CBNs). CBNs are Bayes nets wherein all arrows go from causes to their direct effects (for an introduction to Bayes nets, see, for instance, Heckerman (1996)). Some properties of CBNs suggest that it is not a brute fact that people encode causes rather than correlations, but that there are also good reasons for them to do so:

- Causally-ordered graphs tend to have the fewest arrows—that is, to capture the greatest number of conditional independencies—and therefore require the fewest parameters and the least computation.
- Causal relationships are more likely than merely probabilistic ones to be preserved across changes in the world. For instance, the probability of a symptom given a disease is quite stable, even though the prevalence of the disease might change. While it may often be the quantity of greater

interest, the probability of the disease given the symptom is sensitive to changes in the prevalence of the disease, and so is less stable.

- Causal graphs support queries about what is likely to happen when a random variable's value is *set* rather than observed. To reason with variables that have been set ('interventions'), is to 'observe' the chosen values for the set variables, and sever the arcs to them from their normal causes. This ability to isolate variables can simplify learning.

Inspired by this work, and in the spirit of 'rational' psychology (Anderson, 1990), some psychologists have proposed models, based on this framework, of human causal inference (for instance, Gopnik, Sobel, Shultz, & Glymour, 2001; Lagnado & Sloman, 2002; Tenenbaum & Griffiths, 2001).

## Causal graphical models are probabilistic models

A bewildering variety of knowledge representation formats have been proposed, which can be broadly classified as 'logical' or 'probabilistic'. In the former class I include first-order logic (and the restricted subsets supported by logic programming languages), frames, scripts, and schemata. In the latter, I include both explicit probabilistic models, including Bayes nets, and connectionist networks. Causal Bayes nets are the latest kind of probabilistic knowledge representation to pique the interest of cognitive scientists.

Partisans of either kind of representation are fond of noting what the other kind can't do. Logic (by which I mean propositional or first-order logic) fails to capture the strong, but not exceptionless, regularities of the world, so cannot combine multiple sources of weak evidence (for instance, expectations and perceptual cues), reason between possible causes of an event, or reverse a (probable) conclusion on receipt of further evidence.

Probabilistic models, on the other hand, cannot represent an infinite set of propositions and their dependencies with finite means, the way logic can. In probabilistic models, propositions are *atomic*—that is, they have no compositional semantics. A node in

a network may represent the probability that the cat chased the dog, but there are no parts of this representation corresponding to ‘the cat’, ‘the dog’, and ‘chased’ that could be reused to represent ‘the dog chased the cat’ or rules for combining those parts in two different ways that would distinguish the two propositions. As pointed out by Fodor and Pylyshyn (1988), any model of our knowledge as a network of atomic propositions fails to account for systematicities in both the propositions we can entertain and the inferences we make<sup>1</sup>.

From a practical standpoint, large, fixed-structure networks have other drawbacks. For one, very little of a network may be computationally relevant to a particular query, making inference unnecessarily costly. Among the random variables that are, strictly speaking, relevant to a query, many may be so weakly connected to the query as to be practically irrelevant. Also, when parent/child or cause/effect relations of the same *type* have independent parameters, estimates of those parameters from data will have unnecessarily high variance, or posteriors over those parameters will be unnecessarily diffuse.

Causal Bayes nets are as vulnerable to these criticisms as any other probabilistic representation. The current flowering of theoretical and empirical research on causal representation and reasoning makes it a fitting moment to promote a format for causal knowledge which both supports uncertain reasoning, including those forms of reasoning special to causal models, and has combinatorial semantics (and therefore enforces systematicity). What follows is a sketch of such a format.

### Causal rules

In response to the question of how humans reason about causes, one leading Bayes net researcher has speculated that “fragmented structures of causal organizations are constantly being assembled on the fly, as needed, from a stock of functional building blocks” (Pearl, 1999, p. 74).

This strategy, sometimes called ‘knowledge-based model construction’ (after Breese, 1992; hereafter called, ‘KBMC’), has been pursued by several AI researchers in the 1990s (Goldman & Charniak, 1993; Glesner & Koller, 1995; Ngo & Haddawy, 1996; Mahoney & Laskey, 1998). While these systems have been designed with an eye toward performance, I will argue that, modulo some resource restrictions, KBMC has promise as a model of human reasoning.

<sup>1</sup>If you can think that Bob ate the cake, and that Mary ate the cookies, you can think that Mary ate the cake. A network model, however, might represent the first two propositions but not the third. If you infer from the fact that Bob loves Mary that Bob will be nice to Mary, then (all other things being equal) you would make the same inference about John with regard to Nancy. Again, nothing about probabilistic network formalisms demands that these inferential links have the same strength, or even direction.

The ‘functional building blocks’ in a KBMC scheme can be called ‘causal rules’, because of their resemblance to logical rules. Causal rules are built from *classes* of random variables. A class of random variables defines a random variable for every tuple of objects satisfying its constraints. For instance, we could define a binary random variable class *Bites(x)* which is defined for any *x* for which the predicate *Dog(x)* is true.

Consider this probabilistic causal rule, in which arrows are from causes to effect:

$$\begin{array}{ccc} \textit{Wealth}(x) & \textit{Health}(y) & \textit{LoveFor}(x, y) \\ & \searrow & \downarrow & \swarrow \\ & \textit{BorrowsFrom}(y, x) & \end{array} \quad (1)$$

This rule says that whether *y* borrows money from *x* depends on how much money *x* has, whether *y* is in poor health (and presumably has trouble making money, or needs extra for medical expenses), and how much *x* loves *y*. The rule covers all cases where *x* and *y* are both people, and they know each other. As with a logical rule, the meaning and coherence of a causal rule depends on the sharing of logical variables across parts; the rule applies only when the *x* in *Wealth(x)* and the *x* in *LoveFor(x, y)* are bound to the same person.

An associated conditional probability distribution specifies the exact numerical form of the dependency<sup>2</sup>.

Several other rules (with similar constraints) will be used below to work through an example of reasoning with such rules:

$$\begin{array}{ccc} \textit{Wealth}(x) & \textit{Health}(x) & \sum_z \textit{LoveFor}(z, x) \\ & \searrow & \downarrow & \swarrow \\ & \textit{Happiness}(x) & \end{array} \quad (2)$$

$$\begin{array}{c} \textit{LoveFor}(x, y) \\ \downarrow \\ \sum_z \textit{LoveFor}(z, y) \end{array} \quad (3)$$

$$\begin{array}{ccc} \textit{TenderHeartedness}(x) & & \textit{Loveability}(y) \\ & \searrow & \swarrow \\ & \textit{LoveFor}(x, y) & \end{array} \quad (4)$$

(These rules are for illustration only; I don’t mean to suggest either (1) that they constitute a reasonable theory of anything or (2) that the concepts over which we reason correspond closely to words in English.)

<sup>2</sup>It is possible for more than one rule to have the same effect. In this case, some method must be agreed on for combining the influence of these multiple causes in the CPD. We pass over this wrinkle, for the moment.

To be coherent, a set of rules must generate only legal Bayes nets—that is, it must not generate any cycles. This can be done either, as above, by making sure the random variable classes themselves are orderable, or by adding a time index to each random variable class, and making sure the time of effects is some positive offset from the time of each of its causes.

To make use of the causal rules, we need some individuals, some facts about them, and a question we want answered. The rules give us our inventory of random variable classes, with the set of individuals then specifying all the particular random variables in those classes. In fact, a set of rules and some individuals (and some satisfied constraints) defines a (possibly very large, or even infinite) Bayes net which would be sufficient to answer any query about the random variables defined over those individuals. Representing this Bayes net implicitly, however, has several desirable properties: (1) it yields a probabilistic knowledge base with just the right representational and inferential systematicities, and systematicity is maintained if new individuals are encountered; (2) only small pieces of it need be represented explicitly to answer any particular question.

Here are some facts, assumed in the example inference below:

Health(Mary) = 'good'.		Knows(John,Fido).
LoveFor(Mary,Fred) = 'medium'.		Knows(John,Mary).
Happiness(John) = 'high'.		Knows(Fred,Mary).
BorrowsFrom(John,Mary) = 'yes'.		Knows(Fred,John).
		Knows(Ted,Phil).
Person(Mary).	Person(Phil).	Knows(Sarah,Sue).
Person(John).	Person(Bob).	Knows(Sue,Bob).
Person(Fred).	Person(Ted).	Knows(Sarah,Ted).
Person(Sarah).	Person(Phil).	
Person(Sue).	Dog(Fido).	

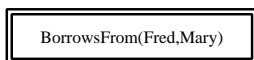
where 'Knows()' is defined to be symmetric. Note that these facts are divided into two kinds: assertions of values for some random values, and assertions of the truth of some constraints. The system I am describing here treats these two kinds of knowledge differently: constraints not asserted are assumed to be false. I will return to this later.

The Bayes net defined by the above rules, individuals, and constraints has about 80 nodes—not large, by modern standards, but, as we will see, quite a bit larger than necessary for answering a typical query.

## Reasoning with causal rules

To answer the question “how probable is it that Fred borrows money from Mary?”, we proceed as follows:

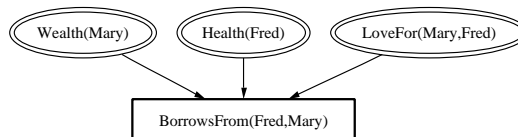
(1) Start with a network consisting only of the query.



The rectangular border of the node is to indicate that it is a query, and the double border to indicate that it has not yet been expanded.

(2) On each successive expansion step, match the the random variable classes in the causal rules against the nodes in the network. Binding the logical variables in a causal rule with the objects in a matched random variable yields candidate causes and effects to add to the network. When all rules have been matched against a node, and all candidate causes and effects to add have been found, then that node is said to be expanded.

For instance, because Fred and Mary are both people, and because they know each other, rule 1 (and only rule 1) matches the query, with  $y$  bound to Fred and  $x$  bound to Mary. This yields three candidates for addition to the network:



Now the original node has a single border to indicate that it has been expanded, while the newly proposed nodes have double borders.

## Checking relevance

Mere matching is not sufficient to ensure the relevance of these nodes. We might, for instance, have no facts about Fred or Mary that might bear on the query, in which case any network constructed amounts to an elaborate approximation of the marginal distribution of the query. A node proposed is irrelevant, and should not be added, unless it is on an *active path* from some fact to the query. A (non-looping) path is active unless some node on that path is *blocked*. A node on a path is blocked if either (a) one or both arrows are pointing away from the node, and the value of the node is known, or (b) both arrows point into the node, and neither the node's value nor the values of any of its descendents (effects) is known.

There are intuitions behind these rules: if *Camping* causes *PoisonIvy*, which, in turn, causes *Itching*, and supposing also (unrealistically) that this is the only causal path by which camping causes itching, then finding out whether someone went camping affects your guess as to whether he is itching, and vice-versa, but only by way of affecting your estimate of whether *PoisonIvy* is true. If you know whether he touched poison ivy, then *Itching* and *Camping* are independent—that is, a known value for *PoisonIvy* blocks the inferential

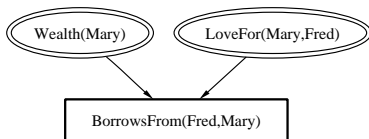
path from *Camping* to *Itching*. When the node in question is the common cause of two effects, the reasoning is similar: if *Smoking* is the common cause of both *StainedFingers* and *YellowTeeth*, then finding out the value of *StainedFingers* will change our guess about its potential cause, *Smoking*, which will, in turn, change our guess about the other effect of smoking, *YellowTeeth*. But if we know whether someone smokes, this inferential path is blocked.

The common effect case is different. If *Rain* and *Sprinkler* are both causes of *WetGrass*, then finding out that it rained increases our belief that the ground is wet, but this does not, in turn, lead us to increase our belief that the sprinkler has been on. That inferential path is blocked by *not* knowing the value of *WetGrass*. If we know the value of *WetGrass*, then the path from *Rain* to *Sprinkler* through *WetGrass* is *unblocked*; finding out that one of the value of *WetGrass* is ‘yes’ causes raises our belief in both causes, but, given that the grass is wet, finding out that one of these causes is present, and might therefore be the explanation for the wet grass, decreases the probability of the other one. If some effect of *WetGrass*, say, *WetShoes*, were observed, instead, the competition between *Rain* and *Sprinkler* as explanations would be much the same.

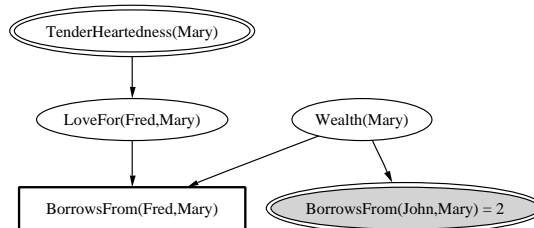
Determining the relevance of a node being considered for addition to the network requires possibly costly information: we need to know all the causes, both direct and indirect, of every fact, and of every candidate cause proposed for addition to the network. This cost may be controlled in several ways:

- Facts about what (indirectly) causes what can be generalized and cached, in the same form as the original causal rules.
- We can be conservative and just add a node if we can’t prove (in reasonable time) that it is irrelevant. This increases the size of the network, but doesn’t affect the answers we get. Put another way, overestimating the set of causal ancestors of a variable may be acceptable.
- We can limit the depth of the backtracking, betting that very long paths will bear only weakly on the query.

Returning to the example, we can determine now that one of these proposed nodes is on no active path to a fact, and remove it.



We then mark the query node as expanded, and try matching rules against the other nodes. *Wealth(Mary)* matches one of the causes in rule 1, so any individual satisfying the constraints of the rule—a person who knows Mary—defines a candidate effect of *Wealth(Mary)*. John is such a person. The proposed node *BorrowsFrom(John,Mary)* is not merely relevant; its value is known. The other unexpanded node above, *LoveFor(Fred,Mary)*, matches the effect of rule 4. Of the two proposed causes, only *TenderHeartedness(Mary)* survives the relevance test. After these steps, the network has five nodes, and has finally connected the query to some evidence.



The shading of *BorrowsFrom(John,Mary)* is to indicate that its value is known.

Several stopping criteria are reasonable. The usual practice in KBMC has been to continue construction until no more relevant candidates can be added—build the complete query-relevant network. Continuing expansion in the example above until there are no unexpanded nodes yields a network of thirteen nodes (shown in Figure 1), considerably fewer than are in the complete Bayes net defined by the rule set, individuals, and satisfied constraints.

It may be reasonable to stop earlier, however. If the query is input to a decision, it makes sense to stop expanding the network when the bounds on the query probabilities leave no ambiguity in that decision. Such bounds can be obtained, in the absence of any more clever scheme, by instantiating the set of unexpanded nodes to all possible sets of values. If the set of relevant random variables is infinite, or a decision is required in a hurry, construction must stop at some point, even if tighter bounds are desirable.

**Why divide the facts into two kinds?** One problematic aspect of the knowledge base, as formulated, is the division of knowledge into the hard constraints under which random variable and causal mechanism classes are defined over tuples of objects, and the random variable classes themselves. There is no reason, in principle, why the constraints themselves can’t be binary random variable classes, but there is an apparent price to pay in the complexity of

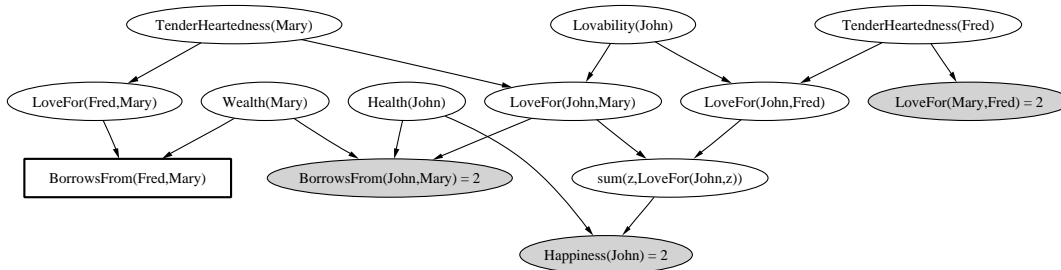


Figure 1: The complete query relevant network for the example discussed.

reasoning. If the constraints themselves are treated as random variables, answering a query means generating not one network, but a tree of networks. This may, however, accurately reflect the branching structure of some reasoning problems.

### What limitations distinguish people from KBMC programs?

A program for stringing together causes and effects, like a program to string together logical proofs, can easily build long inferential chains that seem quite inhuman. If humans reason about causes by stringing together causal networks on the fly, they probably so under some severe limitations. Some limitations worth investigating are

- the length of chains considered in both determination of relevance and in inference.
- the number of objects that can be reasoned over at once. To reason about many objects, we seem to need to aggregate them into composite objects. Interestingly, representing complex objects in terms of simpler ones seems to require similar representational resources to those advocated here for representing causal knowledge.
- the number of nodes that can be reasoned over at once.
- the precision of conditional probability estimates.

In explicit reasoning, at least, people also seem to prefer to say ‘don’t know’ when there is no nearby (in graphical terms), strong evidence for or against a proposition, or to take action to obtain such evidence, rather than making fine distinctions among answers far from certainty.

What I am advocating here is not a theory of any particular task or ability, but a style of representation that could be useful in theory building. Cognitive theories proliferate when researchers have languages in which to formulate them. New languages

that combine some of the expressiveness of logic with probabilistic and causal reasoning could, I believe, stimulate creative cognitive theory building.

### Acknowledgments

For helpful discussion, I thank Amy Hoff, David Lagnado, Mark Johnson, Steven Sloman, David Sobel, and Joshua Tenenbaum. This work was supported by NSF IGERT grant 9870676 at Brown University.

### References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Breese, J. S. (1992). Construction of belief and decision networks. *Computational Intelligence*, 8(4), 624-647.
- Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28, 3-71.
- Glesner, S. & Koller, D. (1995) Constructing flexible dynamic belief networks from first-order probabilistic knowledge bases. In *Froidevaux, C., & Kohlas., J., Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Springer-Verlag.
- Gopnik, A., Sobel, D. M., Shultz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37, 620-629.
- Goldman, R. P. & Charniak, E. (1993). A language for construction of belief networks. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 15(3), 196-208.
- Heckerman, D. (1998). A tutorial on learning in Bayesian networks. In Jordan, M. I., ed., *Learning*

- in graphical models*, Cambridge, MIT.
- Lagnado, D. & Sloman, S. A. (2002). Learning causal structure. *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society* (pp. ?-?). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Mahoney, S. M. & Laskey, K. (1998). Constructing situation-specific Bayesian Networks. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*.
- Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.
- Ngo, L. & Haddawy, P. (1996). Answering queries from context-sensitive probabilistic knowledge bases. *Theoretical Computer Science*.
- Pearl, J. (1999). Bayesian networks. In Wilson, R. A., & Keil, F. C., eds., *The MIT encyclopedia of cognitive science*, 72-74.
- Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.
- Spirtes, P., Glymour, C., & Scheines, R. (1993) *Causation, prediction, and search*. New York: Springer-Verlag.
- Tenenbaum, J. B. & Griffiths, T. L. (2001). Structure learning in human causal induction. *Advances in Neural Information Processing*, 13.