Artificial Intelligence 238 (2016) 154–165

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

A note on the complexity of the causal ordering problem

Bernardo Gonçalves [a,*], Fabio Porto [b]

[a] IBM Research, São Paulo, Brazil
[b] National Laboratory for Scientific Computing (LNCC), Petrópolis, Brazil

ARTICLE INFO

Article history:
Received 5 September 2015
Received in revised form 14 June 2016
Accepted 17 June 2016
Available online 18 June 2016

Keywords:
Causal ordering
Causal reasoning
Structural equations
Hypothesis management

ABSTRACT

In this note we provide a concise report on the complexity of the causal ordering problem, originally introduced by Simon to reason about causal dependencies implicit in systems of mathematical equations. We show that Simon's classical algorithm to infer causal ordering is NP-Hard—an intractability previously guessed but never proven. We present then a detailed account based on Nayak's suggested algorithmic solution (the best available), which is dominated by computing transitive closure—bounded in time by $O(|\mathcal{V}| \cdot |\mathcal{S}|)$, where $\mathcal{S}(\mathcal{E}, \mathcal{V})$ is the input system structure composed of a set $\mathcal{E}$ of equations over a set $\mathcal{V}$ of variables with number of variable appearances (density) $|\mathcal{S}|$. We also comment on the potential of causal ordering for emerging applications in large-scale hypothesis management and analytics.

1. Introduction

The causal ordering problem has long been introduced by Simon as a technique to infer the causal dependencies implicit in a deterministic mathematical model [1]. For instance, let $f_1(x_1)$ and $f_2(x_1, x_2)$ be two equations defined over variables x_1, x_2. Then the causal ordering problem is to infer all existing causal dependencies, in this case the only one is (x_1, x_2), read 'x_2 causally depends on x_1.' It is obtained by first matching each equation to a variable that appears in it, e.g., $f_2 \mapsto x_2$. Intuitively, this means that f_2 is to be assigned to compute the value of x_2—using the value of x_1, which establishes a direct causal dependency between these two variables. Indirect dependencies may then arise and can be computed, which is specially useful when the system of equations is very large.

Causal ordering inference can then support users with uncertainty management, say, towards the discovery of what is wrong with a model for enabling efficient and effective modeling intervention. If multiple values of x_1 are admissible for a modeler, then as a user of the causal ordering machinery she has support to track their influence on the values of x_2. One major application for that is probabilistic database design [2].

Back in the 1950s, Simon was motivated by studies in econometrics and not very concerned with the algorithmic aspects of the Causal Ordering Problem (COP). Yet his vision on COP and its relevance turned out to be influential in the artificial intelligence literature. In a more recent study, Dash and Druzdzel revisit and motivate it in light of modern applications [3]. They show that Simon's original algorithm, henceforth the Causal Ordering Algorithm (COA), is correct in the sense that any valid causal ordering that can be extracted from a self-contained (well-posed) system of equations must be compatible with the one that is output by COA [3]. Their aim has also been (sic.) to validate decades of research that has shown the causal ordering to provide a powerful tool for operating on models. In addition to the result on the correctness of COA, their note also provides a convenient survey of related work that connects to Simon's early vision on causal reasoning.

* Corresponding author.
E-mail addresses: begoncalves@acm.org (B. Gonçalves), fporto@lncc.br (F. Porto).

http://dx.doi.org/10.1016/j.artint.2016.06.004
0004-3702/

However, Simon's formulation of COP into COA—originally in [1], and reproduced in [3], turns out to be intractable. There is a need to distinguish Simon's COA from COP itself. The former still seems to be the main entry point to the latter in the specialized literature. In fact, there is a lack of a review on the computational properties of COA—and as we show in this note, it tries to address an NP-Hard problem as one of its steps. The interested reader who needs an efficient algorithmic approach to address COP in a real, large-scale application can only scarcely find some comments spread through Nayak [4, pp. 287–291], and then Iwasaki and Simon [5, p. 149] and Pearl [6, p. 226] both pointing to the former. Regarding Simon's COA itself, the classical approach to COP, it is only Nayak who suggests in words that (sic.) '[it] is a worst-case exponential time algorithm' [7, p. 37]. We discuss this ambiguity that exists in the most up-to-date literature shortly in §1.2.

COP is significant also in view of emerging applications in large-scale hypothesis management and analytics [2]. The modeling of physical and socio-economical systems as a set of mathematical equations is a traditional approach in science and engineering and a very large bulk of models exist which are ever more available in machine-readable format. Simon's early vision on the automatic extraction of the "causal mechanisms" implicit in (large-scale) models for the sake of informed intervention finds nowadays new applications in the context of open simulation laboratories [8], large-scale model management [9] and online, shared model repositories [10–12].

In this paper we review the causal ordering problem (§2). Our core contributions are (§3) to originally show that COA aims at addressing an NP-Hard problem, confirming Nayak's earlier intuition; and then (§4) to organize into a concise yet complete note his hints to solve COP in polynomial time.

1.1. Informal preliminaries

Given a system of mathematical equations involving a set of variables, the *causal ordering problem* consists in detecting the hidden asymmetry between variables. As an intermediate step towards it, one needs to establish a one-to-one mapping between equations and variables [1].

For instance, Einstein's famous equation $E = mc^2$ states the equivalence of mass and energy, summarizing (in its scalar version) a theory that can be imposed two different asymmetries for different applications. Say, given a fixed amount of mass $m = m_0$ (and recalling that c is a constant), find the particle's relativistic rest energy E; or rather, given the particle's rest energy, find its mass or potential for nuclear fission. That is, the causal ordering depends on what variables are set as input and which ones are "influenced" by them. Suppose there is uncertainty, say, a user considers two values to set the mass, $m = m_0$ or $m = m_0'$. Then this uncertainty will flow through the causal ordering and affect all variables that are dependent on it (energy E).

For really large systems, having structures say in the order of one million equations [13], the causal ordering problem is critical to provide more specific accountability on the accuracy of the system—viz., what specific variables and subsystems account for possibly inaccurate outcomes. This is key for managing and tracking the uncertainty of alternative modeling variations systematically [8,13].

1.2. Related work

COA. Dash and Druzdzel [3] provide a high-level description of how Simon's COA [1] proceeds to discover the causal dependencies implicit in a structure. It returns a 'partial' causal mapping: from partitions on the set of equations to same-cardinality partitions on the set of variables—a 'total' causal mapping would instead map every equation to exactly one variable.

They show then that any valid total causal mapping produced over a structure must be consistent with COA's partial causal mapping. Nonetheless, no observation at all is made regarding COA's computational properties in [3], leaving in the most up-to-date literature an impression that Simon's COA is the way to go for COP. In this note we show that Simon's COA tries to address an NP-Hard problem in one of its steps, and then clearly recommend Nayak's efficient approach to COP as a fix to COA.

COP. Inspired by Serrano and Gossard's work on constraint modeling and reasoning [14], Nayak describes an approach that is provably efficient to process the causal ordering: extract any valid total causal mapping and then compute the transitive closure of the direct causal dependencies, viz, the causal ordering. Nayak's is a provably correct approach to COP, as all valid 'total' causal mappings must have the same causal ordering.

In this note we arrange those insights into a concise yet detailed recipe that can be followed straightforwardly to solve COP efficiently.

2. The causal ordering problem

We start with some preliminaries on notation and basic concepts to eventually state COP formally.

For an equation $f(x_1, x_2, \ldots, x_\ell) = 0$, we will write $Vars(f) \triangleq \{x_1, x_2, \ldots, x_\ell\}$ to denote the set of variables that appear in it.

Definition 1. A **structure** is a pair $\mathcal{S}(\mathcal{E}, \mathcal{V})$, where $\mathcal{E}$ is a set of equations over a set $\mathcal{V}$ of variables, $\mathcal{V} \triangleq \bigcup_{f \in \mathcal{E}} Vars(f)$, such that: