Doctoral Dissertations 2014-current

Causal Discovery for Relational Domains: Representation, Reasoning, and Learning

Marc Maier, *University of Massachusetts - Amherst*

Date of Award
Summer 6-10-2014

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

First Advisor
David Jensen

Second Advisor
Andrew Barto

Third Advisor
Hanna Wallach

Fourth Advisor
Andrew Papachristos

Keywords
causal discovery, relational data, machine learning, graphical models

Subject Categories
Artificial Intelligence and Robotics

Abstract
Many domains are currently experiencing the growing trend to record and analyze massive, observational data sets with increasing complexity. A commonly made claim is that these data sets hold potential to transform their corresponding domains by providing previously unknown or

unexpected explanations and enabling informed decision-making. However, only knowledge of the underlying causal generative process, as opposed to knowledge of associational patterns, can support such tasks.

Most methods for traditional causal discovery—the development of algorithms that learn causal structure from observational data—are restricted to representations that require limiting assumptions on the form of the data. Causal discovery has almost exclusively been applied to directed graphical models of propositional data that assume a single type of entity with independence among instances. However, most real-world domains are characterized by systems that involve complex interactions among multiple types of entities. Many state-of-the-art methods in statistics and machine learning that address such complex systems focus on learning associational models, and they are oftentimes mistakenly interpreted as causal. The intersection between causal discovery and machine learning in complex systems is small.

The primary objective of this thesis is to extend causal discovery to such complex systems. Specifically, I formalize a *relational* representation and model that can express the causal and probabilistic dependencies among the attributes of interacting, heterogeneous entities. I show that the traditional method for reasoning about statistical independence from model structure fails to accurately derive conditional independence facts from relational models. I introduce a new theory—*relational d-separation*—and a novel, lifted representation—the *abstract ground graph*—that supports a sound, complete, and computationally efficient method for algorithmically deriving conditional independencies from probabilistic models of relational data. The abstract ground graph representation also presents causal implications that enable the detection of causal direction for bivariate relational dependencies without parametric assumptions. I leverage these implications and the theoretical framework of relational *d*-separation to develop a sound and complete algorithm—the *relational causal discovery* (RCD) algorithm—that learns causal structure from relational data.