Learning
Causal Models of Behavior
in Networks

David Jensen

Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst

Supported by:

MIT Lincoln Laboratory
Networks
Multimode networks
...with attributes
Models of networks with attributes
Associational models
Causal models
The Financial Industry Regulatory Authority (FINRA) regulates virtually every securities firm doing business with the US public. They register member firms, write rules to govern their behavior, examine them for compliance and discipline those that fail to comply.
FINRA data

1.2 Million brokers

300,000 branches

16,000 firms

400,000 disclosures
Broker fraud

<0.1%
Relational probability trees

(Neville, Jensen, Friedland, and Hay 2003)
Relational probability trees

(Neville, Jensen, Friedland, and Hay 2003)
Results — Objective accuracy
Results — Aiding examiners

(Neville et al. 2005)
Relational dependency networks

(Neville & Jensen 2004, 2006)
Statistical relational learning

- Early work in:
  - Probabilistic Inductive Logic Programming
  - Object-oriented Bayesian Networks (1997)
- Now a well-developed suite of techniques
- Good general reference: Getoor & Taskar 2007
Questions we answered...

- Which reps are likely to commit fraud in the next 12 months?
- What attributes of reps are predictive of fraud risk?
- What predictive dependencies hold among attributes of reps, firms, branches, and disclosures?
...Questions FINRA is really asking

- What changes will current events cause in the securities industry over the next 36 months? (a “weather report”)
- What is causing those events to happen? (an “intelligent tutor”)
- What would be the effect of alternative approaches to enforcement and education if they are enacted by FINRA? (a “flight simulator”)

...
Experts interpret models causally

It’s what they don’t want to do
Assertion 1
More research should focus on representing, learning, and managing *causal knowledge* about systems of interacting entities.
What is causality?

“The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. ... Causation implies that by varying one factor, I can make another vary.”

– Cook & Campbell (1979)
Why discover causality?

• A frequent goal of data analysis projects is **actionable** knowledge.

• Statistical association alone is insufficient to distinguish among different causal models.

• Each causal model implies different actions, if we wish to influence the value of B.
Existing methods for causal discovery

• A small but highly regarded literature exists on algorithms for causal discovery in observational data (Pearl 2000/2009; Spirtes, Glymour & Scheines 1993/2001)

• However, existing methods assume non-relational dependence
How does it work?

• Basic strategy
  (1) Use observational data to infer conditional independence;
  (2) Use inferred independence to identify constraints on the space of possible causal models; and
  (3) Identify when all models in the space imply the same causal dependence.

• Example — The PC Algorithm (Spirtes & Glymour 1991)
Assertion 2

Relational data could enable a new generation of methods and algorithms for discovering causal knowledge.
New opportunities for automation

• Causal reasoning is greatly aided by representing and reasoning about relations

• Recent developments make this possible

• *Widespread use of relational databases* — Datasets with large and complex relational schemas are increasingly available

• *Development of relational models* — New methods from relational learning and social network analysis
Why should relations help?

- New variables that could be common causes or effects.
- New opportunities to represent and adjust for instance dependence.
- New opportunities to condition on latent variables by blocking on relational structure.
Strategy 1: Relational PC

- Automatically formulate and search for conditional independence in relational data and identify how that constrains the space of possible causal models.

- Identify and use new opportunities for conditional independence and edge orientation.

(Maier, Taylor, Oktay, & Jensen 2010; Rattigan & Jensen 2010)
Evaluating RPC

Precision = \frac{|\text{Correct}|}{|\text{Found}|}

Recall = \frac{|\text{Found}|}{|\text{True}|}
Experimental results
Experimental results

![Graph showing performance vs. sample size for skel. precision, skel. recall, comp. precision, and comp. recall. The graph indicates that as the sample size increases, the performance metrics generally decrease, with skel. precision and comp. precision showing more significant drops compared to skel. recall and comp. recall.](image-url)
Strategy 2: Quasi-experimental design

- Social scientists, economists, and medical researchers regularly infer causality from observational data using manual methods called quasi-experimental designs (QEDs).

- Relational data sets contain much of the information necessary to perform this reasoning automatically.

- Initial results indicate that automated reasoning with first-order logic can identify some QEDs automatically.

(Jensen, Fast, Taylor, and Maier 2008)
The opportunity

Manual Social Science
Largely uninterested in automated search of large hypothesis spaces

Knowledge Discovery in Data
Largely uninterested in learning causality, and thus uninterested in designs.

Relational, Temporal, and Spatial Representations

Causality

Automation

Causal Discovery
Has not exploited the recent advances in knowledge representation and reasoning
“...I see no greater impediment to scientific progress than the prevailing practice of focusing all of our mathematical resources on probabilistic and statistical inferences while leaving causal considerations to the mercy of intuition and good judgment.”

- Judea Pearl (2000)
Contact:

jensen@cs.umass.edu
kdl.cs.umass.edu
/papers, /proximity

Thanks to:

Brian Taylor, Andrew Fast, Lisa Friedland, Henry Goldberg, Michael Hay, John Komoroske, Marc Maier, Jennifer Neville, Matthew Rattigan, Ted Senator, and Agustin Schapira