Building Undirected Influence Ontologies Using Pairwise Similarity Functions

Tamlin Love, Ritesh Ajoodha

Introduction

An ontology is a set of concepts or variables and the relations between them [1].
Influence ontologies encode influence reltations, such as causal influence between random variables.

How do we recover structure?

We model ontologies as graphs whose vertices encode concepts/variables and whose edges encode relations.
We use correlation metrics (Pearson and Spearman for continuous data, Cramer's V for categorical data) to detect influence in the observations.
We present algorithms 1 and 2 to recover the ontology structures from observations.
Algorithm 1 weights edges by the similarity function and preserves edges whose weights are above a threshold parameter *t*, which controls the density of the reconstruction.
Algorithm 2 assumes a sparse, tree structure by finding the maximum weighted spanning tree over the complete graph, based on [2].

How do we evaluate structure?

We present a modification to the minimum graph edit distance, which measures how many operations are needed to transform one graph into another [3]. Our modified scaled GED score ranges from 0 to 1, where 0 indicates a perfect reconstruction and 1 indicates the worst possible reconstruction.

Experiment

We randomly generated 200 Bayesian networks over varying number of variables and densities. We then sampled 20,000 observations from each. Applying each of our algorithms, we attempted to reconstruct our ground truth structures.

Results

For sparse structures, the MWST approach performs best. For dense structures, the threshold approach performs best for low *t*. Overall, sparse structures can be recovered with much greater fidelity than dense structures.

Example Application: CHILD

We test the effectiveness of our algorithms in reconstructing a real-world Bayesian network: the CHILD network, used to diagnose "blue baby syndrome" in infants [4]. The best threshold reconstruction (for *t=0.4715*) achieves a GED score of 14 (*0.07368* scaled) and the best MWST reconstruction achieves a GED score of 8 (*0.042105* scaled).

Conclusion

While these methods cannot replace traditional Bayesian network structure-learning techniques, they are useful as computationally cheap data exploration tools and in knowledge discovery over structures which cannot be modelled as Bayesian networks.

Can we recover the **undirected** skeleton of an **influence ontology** structure?

We present two algorithms that achieve this using **pairwise similarity functions**, and a modified metric to evaluate the reconstructed structures.

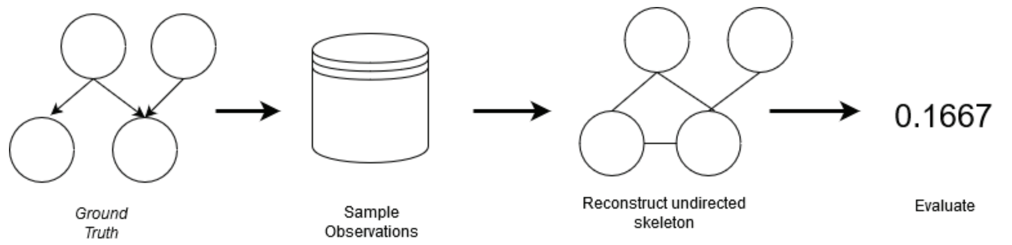
Algorithm 1 reconstructs **dense structures** best, while algorithm 2 reconstructs **sparse structures** best.

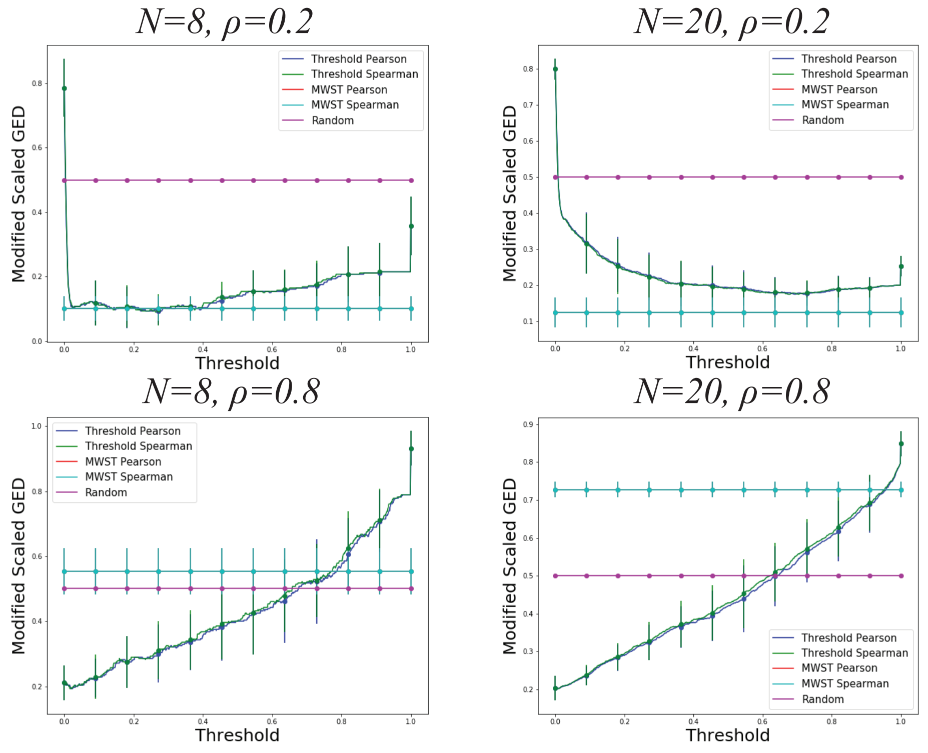
Algorithm 1 Build Influence Ontology - Threshold Approach

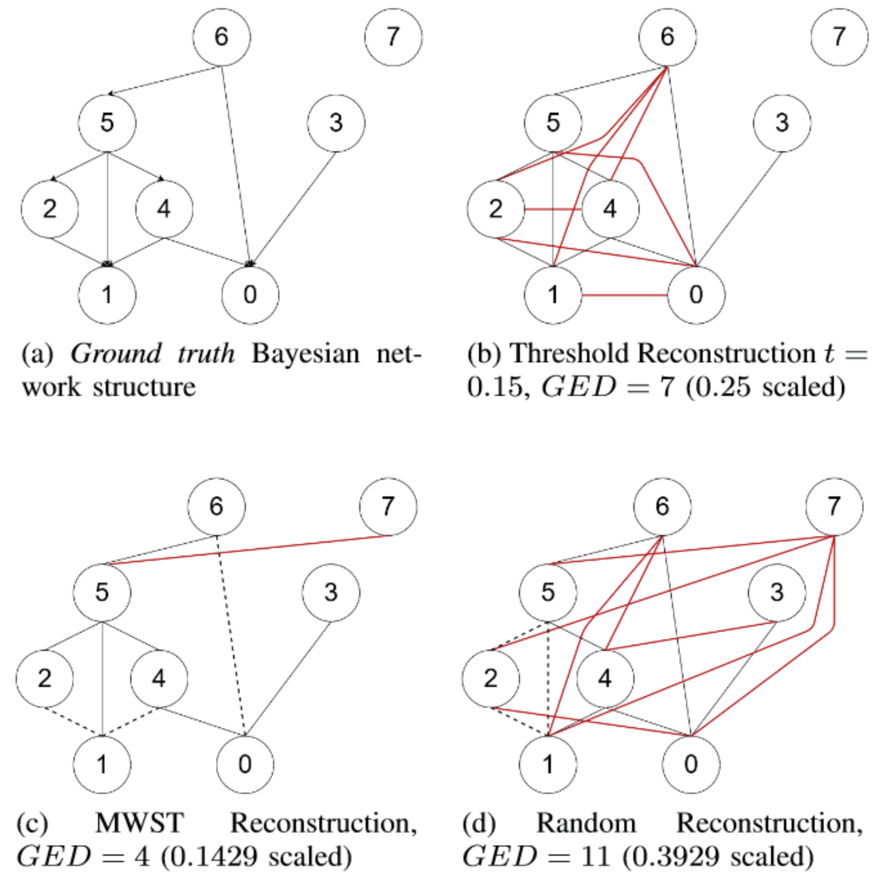
```
procedure BUILD_GRAPH($\mathcal{X}$, $\mathcal{D}$, $similarity$, $t$)
    $\mathcal{G} \leftarrow (V = \mathcal{X}, E = \emptyset)$
    $N \leftarrow Dim(\mathcal{G})$
    $S_{ij} \leftarrow similarity(\mathcal{D}, i, j)$
    for $i \in [0, ..., N-1]$ do
        for $j \in [0, ..., N-1]$ do
            if $i \neq j$ and $|S_{ij}| \geq t$ then
                $addEdge(\mathcal{G}, i, j)$
    return $\mathcal{G}$
```

Algorithm 2 Build Influence Ontology - MWST Approach

```
procedure BUILD_GRAPH($\mathcal{X}$, $\mathcal{D}$, $similarity$)
    $\mathcal{G} \leftarrow (V = \mathcal{X}, E = \emptyset)$
    $N \leftarrow Dim(\mathcal{G})$
    $S_{ij} \leftarrow similarity(\mathcal{D}, i, j)$
    for $i \in [0, ..., N-1]$ do
        for $j \in [0, ..., N-1]$ do
            if $i \neq j$ then
                $addEdge(\mathcal{G}, i, j)$ with $weight = |S_{ij}|$
    $\mathcal{T} \leftarrow Kruskal\_Get\_MWST(\mathcal{G})$
    return $\mathcal{T}$
```





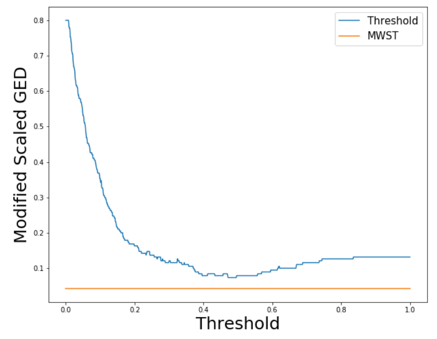


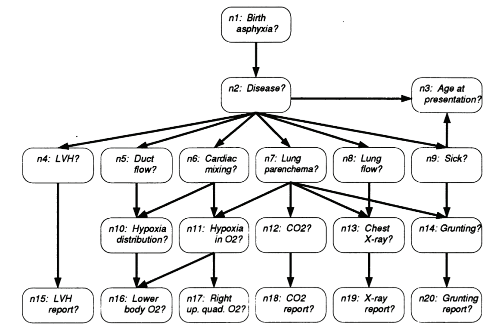
(a) *Ground truth* Bayesian network structure

(b) Threshold Reconstruction $t = 0.15$, $GED = 7$ (0.25 scaled)

(c) MWST Reconstruction, $GED = 4$ (0.1429 scaled)

(d) Random Reconstruction, $GED = 11$ (0.3929 scaled)





References

[1] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International Journal Human-Computer Studies*, vol. 43. pp. 907-928, 1993.
[2] X.-w. Chen, G. Anantha and X. Wang, "An effective structure learning method for constructing gene networks," *Bioinformatics*, vol. 22, no. 11, pp. 1367-1374, 2006.
[3] D. Justice and A. Hero, "A linear formulation of the graph edit distance for graph recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2005.
[4] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, R. G. Cowell, et al., "Bayesian analysis in expert systems," *Statistical science*, vol. 8, no. 3, pp. 219-247, 1993.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG