Ancestor Relations in the Presence of Unobserved Variables

Pekka Parviainen and Mikko Koivisto

Helsinki Institute for Information Technology HIIT
Department of Computer Science
University of Helsinki

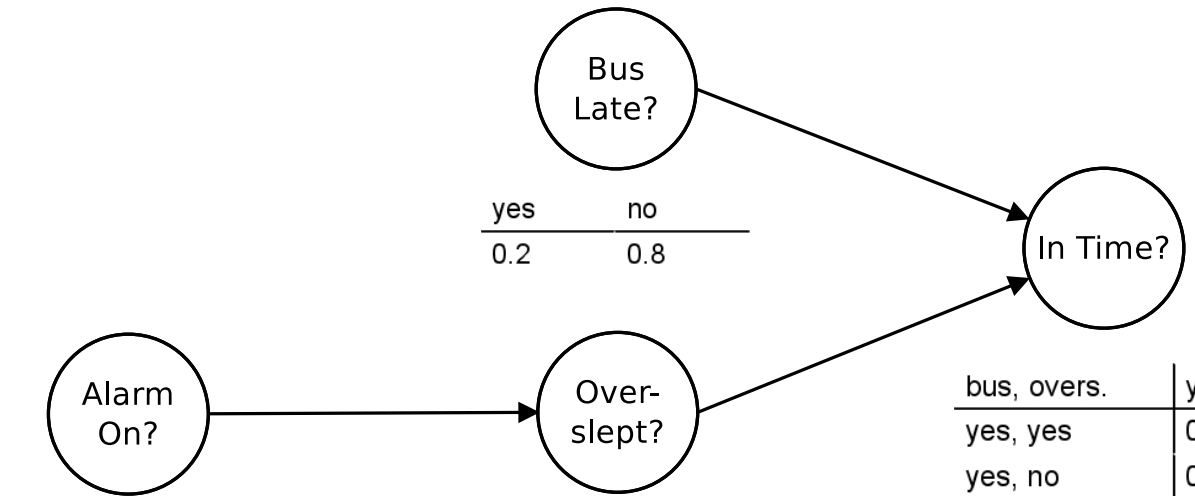
ECML PKDD
6.9.2011

Outline

- What are ancestor relations and why should anybody care about them?
- How can ancestor relations be learned?
- Are ancestor relations useful in practice?

Bayesian networks

- Representations of joint probability distributions
- Consist of:
  - The structure is a directed acyclic graph (DAG) that represents conditional independencies between variables.
  - The local conditional probability distributions that are specified by parameters.



yes	no
0.2	0.8

yes	no
0.9	0.1

alarm	yes	no
yes	0.1	0.9
no	0.9	0.1

bus, overs.	yes	no
yes, yes	0.1	0.9
yes, no	0.2	0.8
no, yes	0.3	0.7
no, no	0.9	0.1

Bayesian networks

- Compact, flexible and interpretable
- Sometimes arcs are interpreted as cause-effect pairs

Structure Discovery

- Construct a best-fit DAG from observational data.
- Challenges:
  - The set of conditional independencies can be represented by a number of different DAGs (Markov equivalence class).
  - There may be unobserved variables.
  - Computational complexity.

Approaches

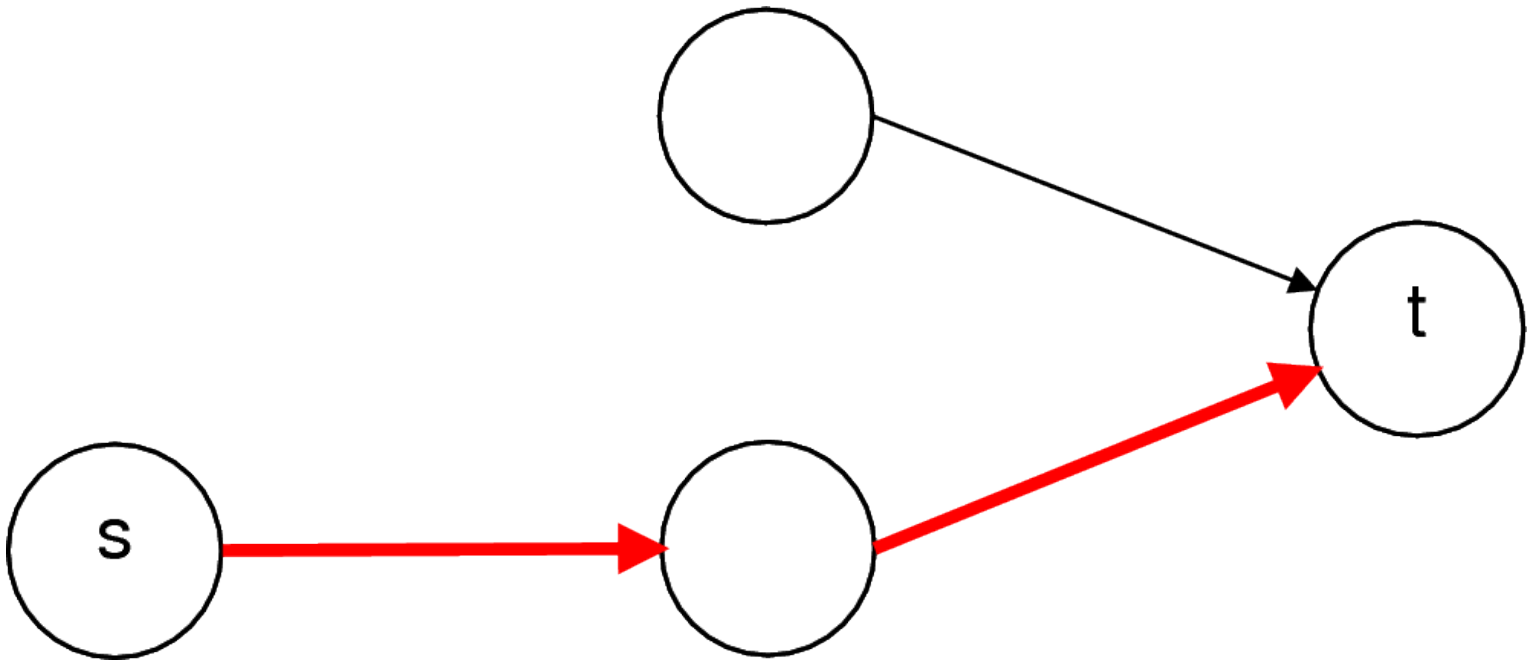
- Constraint-based
  - Test conditional independencies between variables.
  - Theoretically sound treatment of unobserved variables.
- Score-based
  - Assign each DAG a score based on how well it fits to data.
  - Flexible, enables incorporating prior information.
  - Hard to handle unobserved variables in a computationally efficient manner.

Structural features

- There may be several almost equally good DAGs (or Markov equivalence classes) and the best-fit DAG may be highly unlikely.
- Therefore, instead of learning a best-fit DAG, it may be useful report posterior probabilities of some *structural features* of interest, e.g., arcs.
- Every DAG has a posterior probability, the posterior probability of a structural feature is the sum over the posterior probabilities of all DAGs that have the feature in question. This is called *(full) Bayesian averaging*.

Ancestor relations

Node s is an ancestor of node t, denoted by $s \rightsquigarrow t$, if there is a directed path from s to t.



Ancestor relations

- Ancestor relations can unveil causal information.
- Can ancestor relations be learned in a computationally efficient manner?
- Can ancestor relations be learned reliably if there are some unobserved variables at work?
- Does learning ancestor relations yield more information than learning arcs?

Algorithm

- Compute $p(s \rightsquigarrow t|D)$, where D is the data.
- (Full) Bayesian averaging
- Our algorithm computes exact posterior probabilities.
- Based on dynamic programming

Assumptions

- Modular likelihood score, i.e.,

$$p(D|A) = \prod_{v \in N} p(D_v | D_{A_v}, A_v),$$

  where A is the (arc set of a) DAG and A_v are the parents of v.

- Order-modular structure prior, i.e.,

$$p(A) = \sum_{L} p(A, L),$$

  where L is a linear order and $p(A, L) = \prod_{v \in N} \rho_v(L_v) q_v(A_v)$.

Dynamic programming - outline

- Goal: compute $p(s \rightsquigarrow t|D)$.
- For every node set $S \subseteq N$ and $T \subseteq S$, compute $g_s(S, T)$, the contribution of the DAGs on S that have a directed path from s to every $u \in T$ and not to any other node.
- $p(s \rightsquigarrow t, D) = \sum_{T:t \in T} g_s(N, T)$.

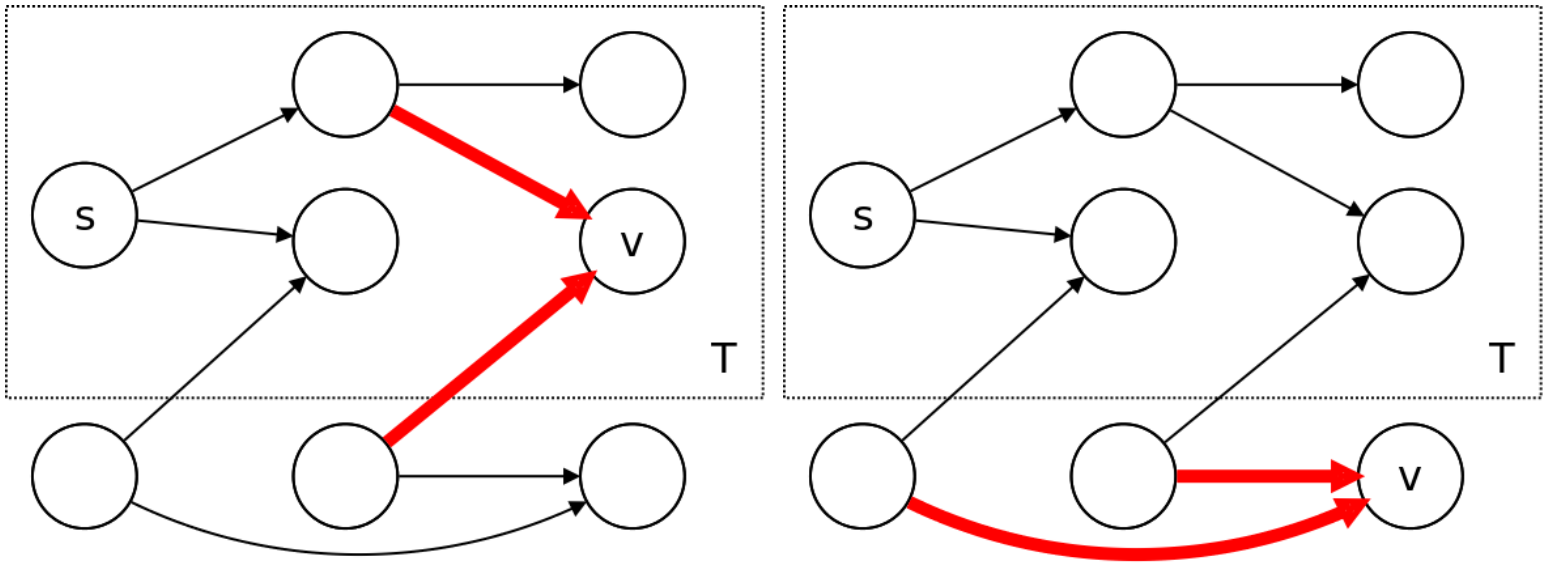
Dynamic programming - outline

- How to compute $g_s(S,T)$?
-

$$g_s(S,T) = \sum_{v \in S} g_s(S \setminus \{v\}, T \setminus \{v\}) \rho_v(S \setminus \{v\}) \bar{\beta}_v(S,T),$$

where $\bar{\beta}_v(S,T)$ is the sum over all possible parent sets of v given that there is a directed path from s exactly to the nodes in T.

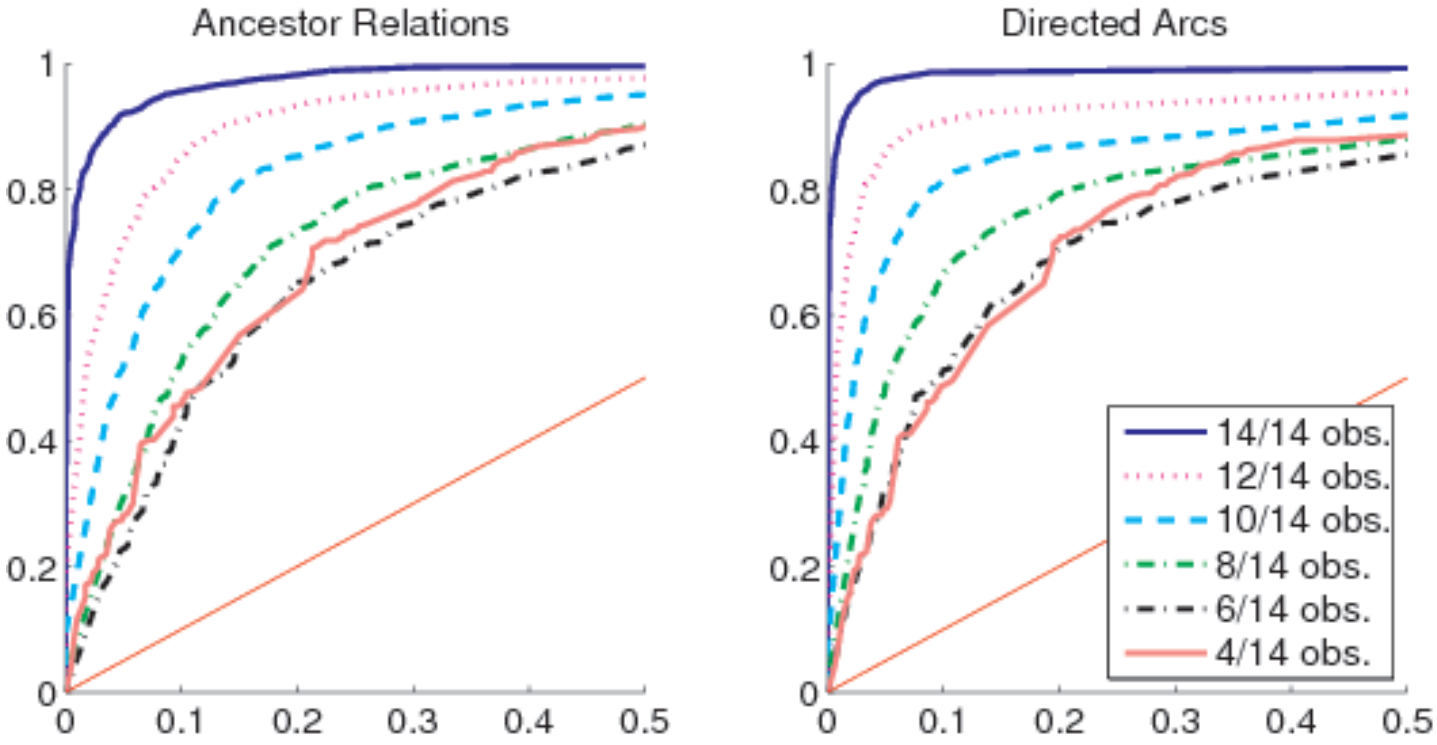
Dynamic programming - outline



Time and space complexity

- $O(n3^n)$ time and $O(3^n)$ space for any s and t.
- $O(n^2 3^n)$ time and $O(3^n)$ space for all pairs s and t.

Learning power



n = 10,000

Full vs. partial Bayesian averaging

		Predicted Ancestor Relations			
m	ℓ	both	partial	full	none
100	0	13.6	1.1	1.8	165.5
100	4	5.3	0.3	0.5	84.0
500	0	30.5	0.5	1.3	149.7
500	4	12.7	0.5	0.6	76.3
2000	0	39.7	0.2	0.4	141.8
2000	4	18.6	0.2	0.4	70.8
10000	0	40.9	0.1	0.4	140.7
10000	4	21.6	0.1	0.2	68.0

Conclusions

- Bayesian learning of ancestor relations is computationally feasible (when the number of nodes is moderate).
- Ancestor relations can be discovered with reasonable power even in the presence of unobserved variables.
- Partial Bayesian averaging, i.e., deducing the ancestor relations from arc probabilities seems to work almost as well as full Bayesian averaging.

Conclusions

- Bayesian learning of ancestor relations is computationally feasible (when the number of nodes is moderate).
- Ancestor relations can be discovered with reasonable power even in the presence of unobserved variables.
- Partial Bayesian averaging, i.e., deducing the ancestor relations from arc probabilities seems to work almost as well as full Bayesian averaging.

Thank you!