Segmenting multi-attribute sequences using Dynamic Bayesian Networks

Janne Toivola
with Robert Gwadera and Jaakko Hollmén

Laboratory of Computer and Information Science
Helsinki University of Technology
Finland
28.10.2007

Motivation

- Discovering **changing dependencies** in multi-attribute event sequences
- a.k.a. synchronized multi-stream sequences and multi-sequences
- I parallel sequences with T events

Applications

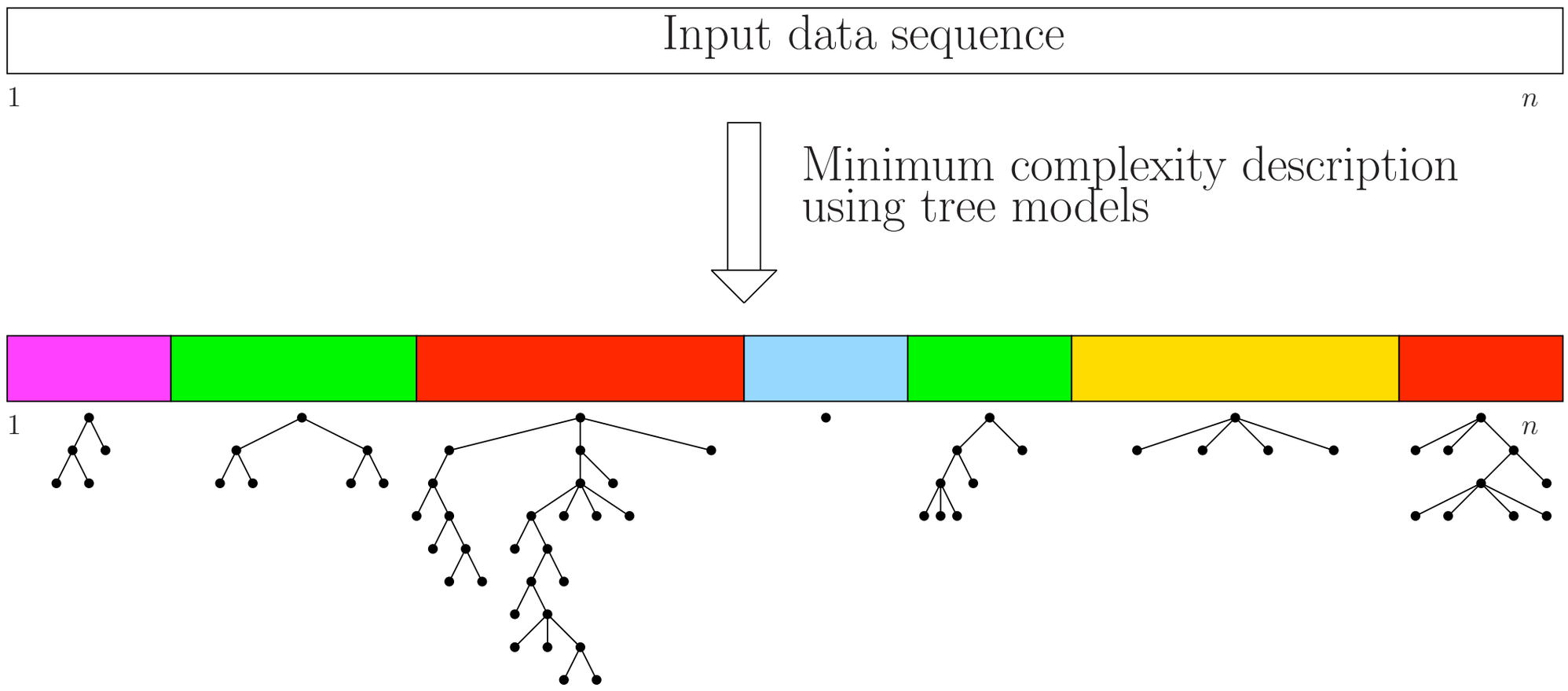
- Monitoring systems / sensor networks: events correspond to configuration of values from different sensors
- Molecular biology: gene expression levels etc. measured for neighboring genes in a chromosome simultaneously
- More..?

Previous work

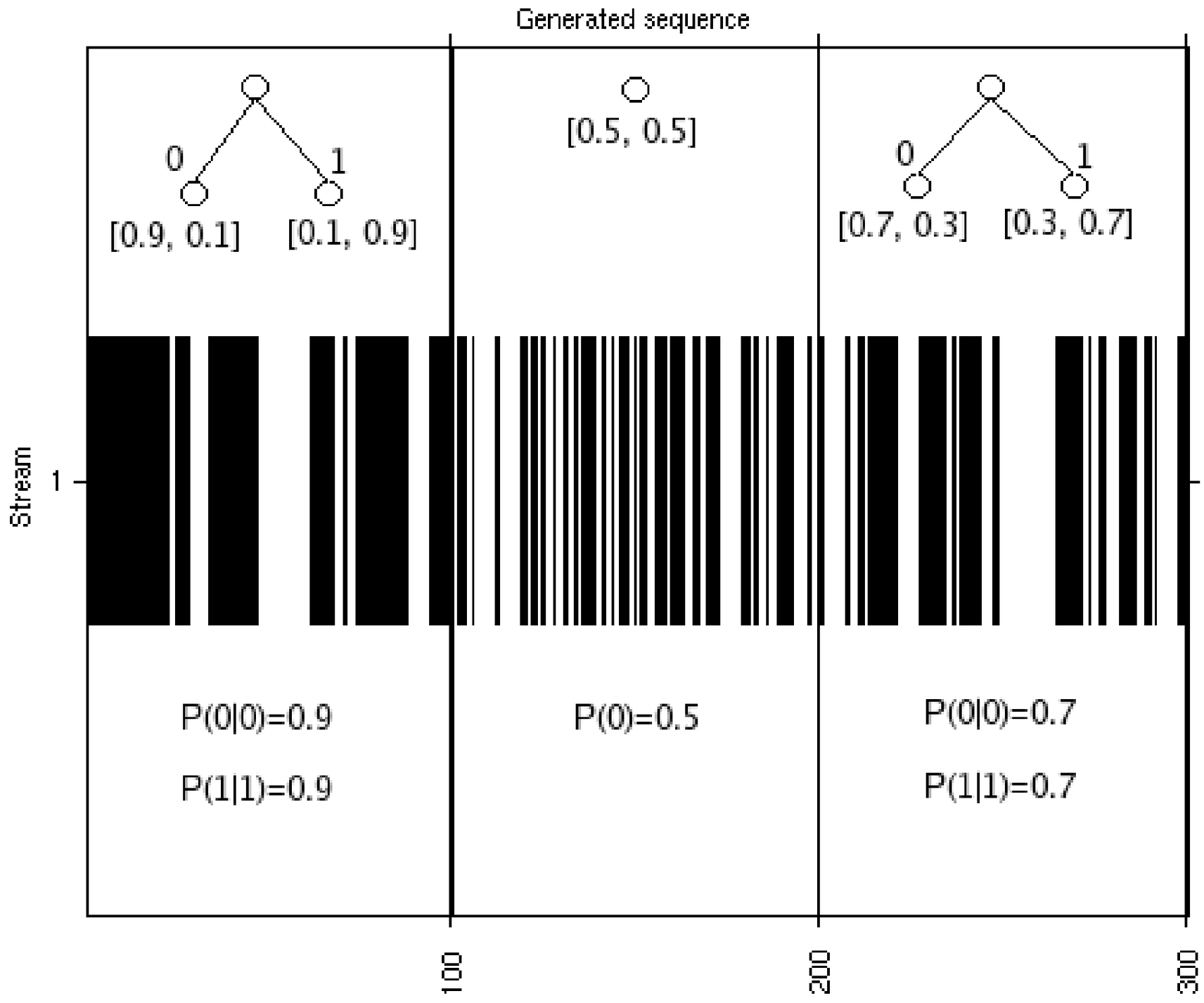
- Significant occurrences of complex inter and intra-stream patterns (frequent patterns or episodes etc.)
- Correlation between entire sequences
  - Boolean networks
  - Probabilistic relationships in Dynamic Bayesian Networks (DBN)
  - Do not model segments with different sets of dependencies (explicitly at least...)

Single sequence segmentation

- R.Gwadera, A.Gionis, H.Mannila, *Optimal Segmentation using Tree Models*, (ICDM-06): a single-sequence segmentation algorithm using probabilistic tree models

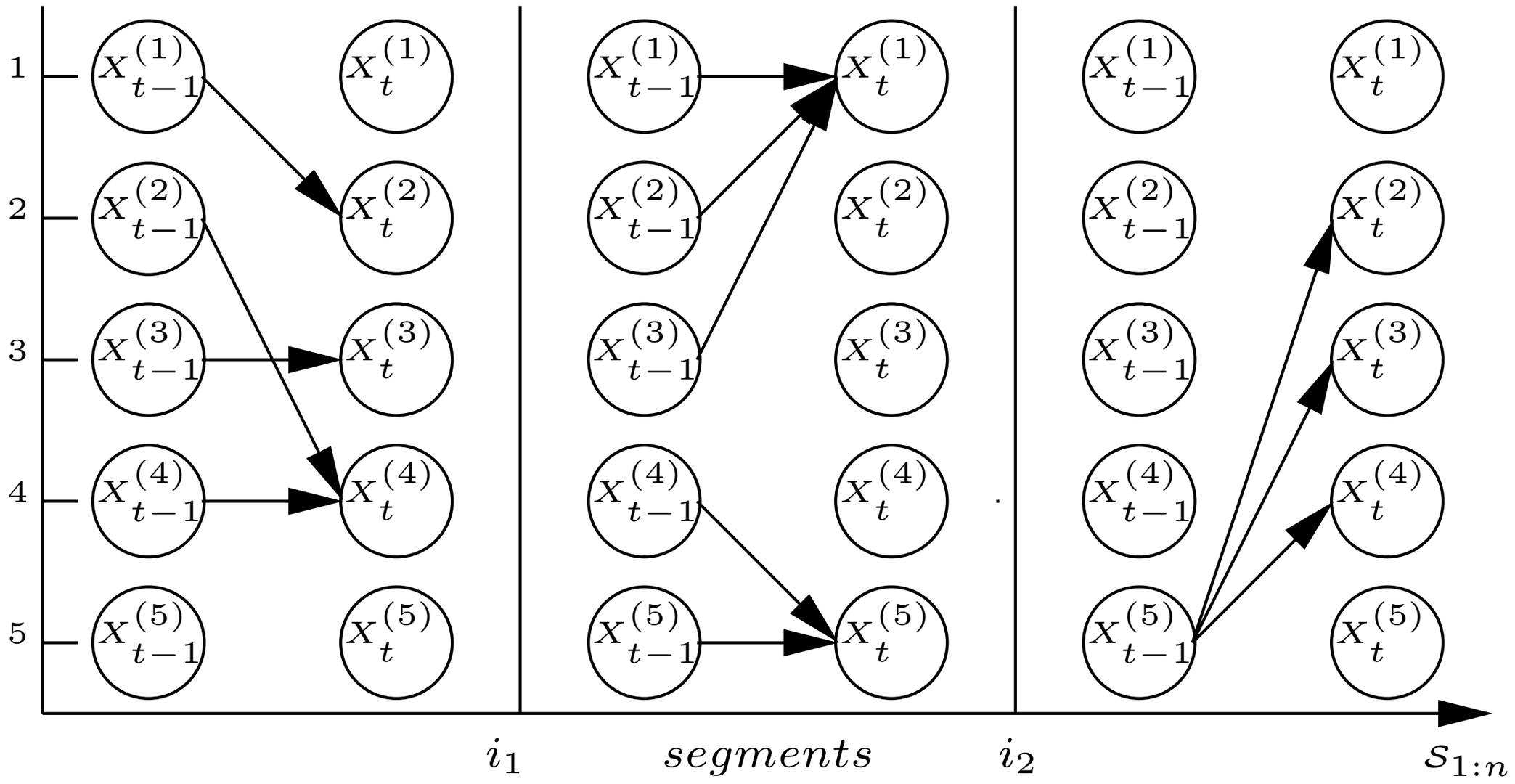


Example of tree models



DBNs instead of trees

- An example of 5-attribute sequence



Segmentation + DBNs

- Given: a finite multi-sequence $\mathcal{S}_{1:n}$, maximum number of segments K, DBN structure space $\mathcal{G} = \{G_1, ..., G_{|\mathcal{G}|}\}$, cost function for a segment $cost(\mathcal{S}_{a:b})$, and border insertion penalty B
- Find partition points $\mathbf{i} = [i_1, ..., i_k]$ by minimizing

$$\sum_{j=0}^{k} cost(\mathcal{S}_{i_j:i_{j+1}}) + k \cdot I \cdot B$$

Finding DBN structure

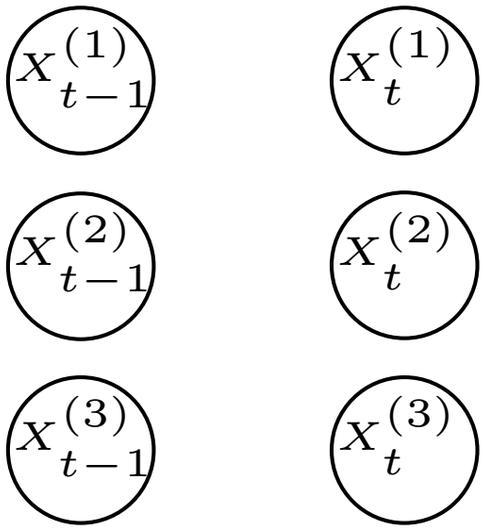
- BIC for Bayesian networks:

$$BIC_G(\mathcal{S}_{1:n}) = -\log_2(\mathcal{L}(\mathcal{S}_{1:n}|G,\hat{\Theta})) + \sum_{i=1}^{I} \frac{|\mathcal{A}|^{d_i}(|\mathcal{A}|-1)}{2}\log_2(n)$$

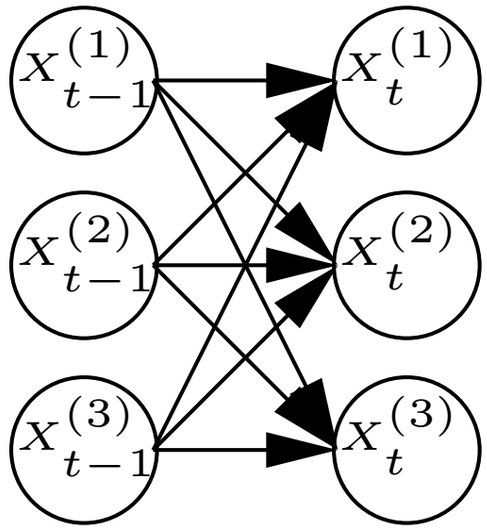
$$\hat{G}_{BIC}(\mathcal{S}_{1:n}) = \min_{G\in\mathcal{G}}(BIC_G(\mathcal{S}_{1:n}))$$

Choices

- Our structure space



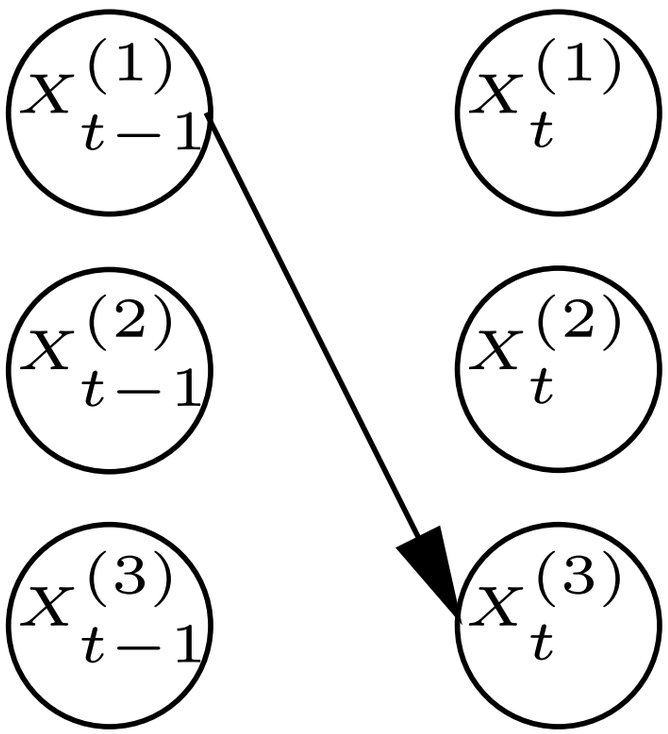
Empty



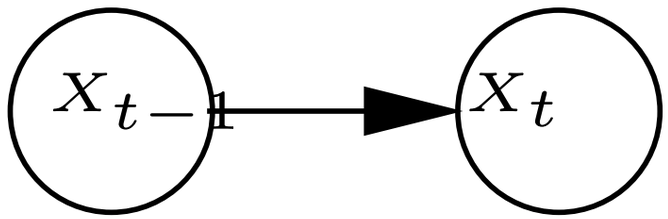
Maximal connections

A bad choice

- Conversion from a DBN to a Markov chain to use a single-stream segmentation



DBN



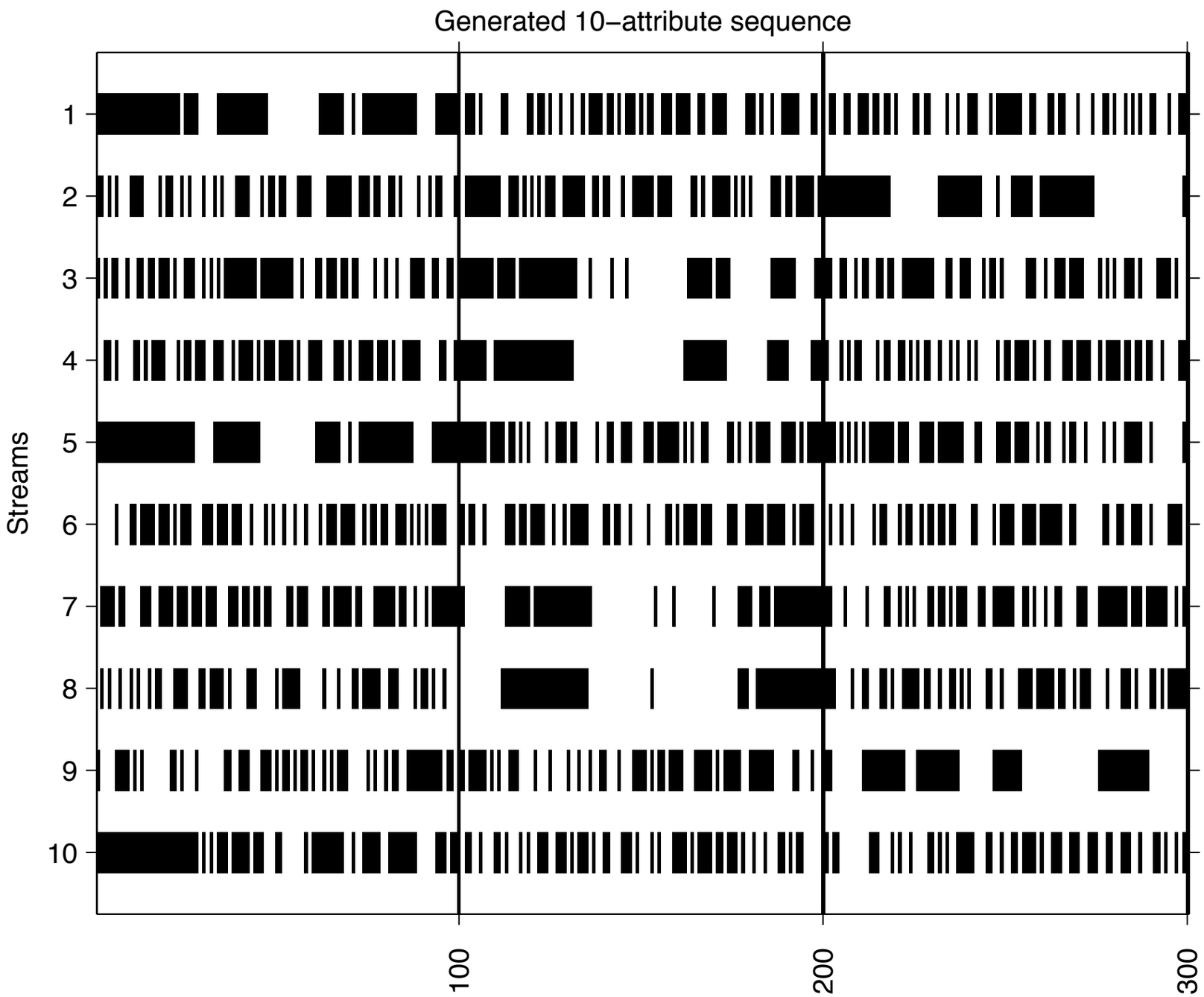
Markov chain

Why is it bad?

- Combined alphabet is exponential in the number of streams
- There is a risk of under-segmentation: few dependencies "get lost in noise"
- Difficult to restrict the set of parents for individual nodes

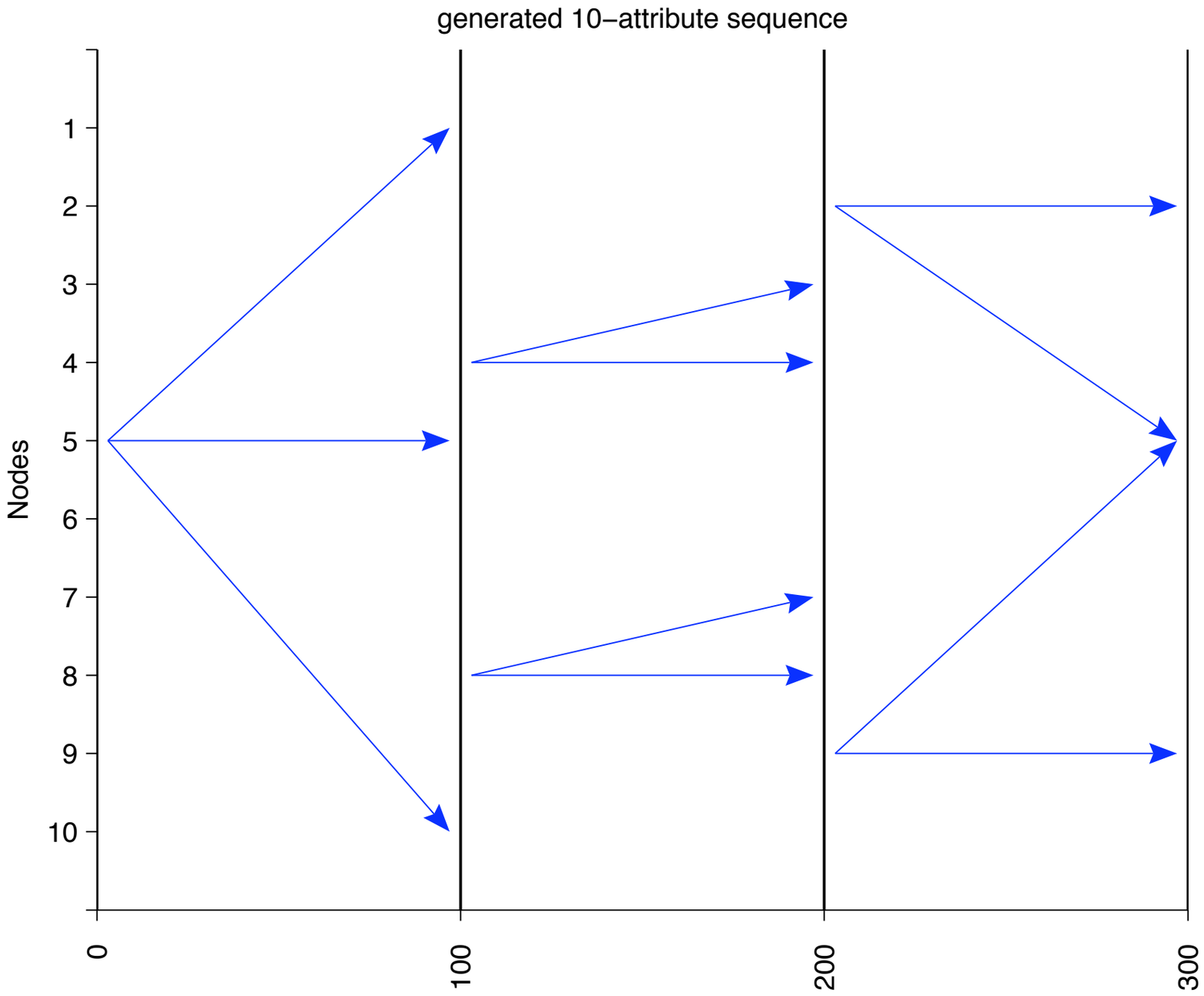
Artificial data

- A 10-attribute sequence with 3 segments



Result of segmentation

- The correct models and segments found!



Conclusions etc.

- Segmentation and probabilistic models
- On generated data, the algorithm finds the correct segmentation and corresponding DBNs
- Doesn't handle missing data well
- Complexity grows easily with the models
- Other potential applications?