

# Sequence modelling for analyzing student interaction

Christian Hansen, Casper Hansen, Stephen Alstrup & Christina Lioma  
chrh@di.ku.dk

## Takeaway: Simple mixture of Markov chain models allow for human interpretable interaction models

### Unsupervised clustering of log interaction data

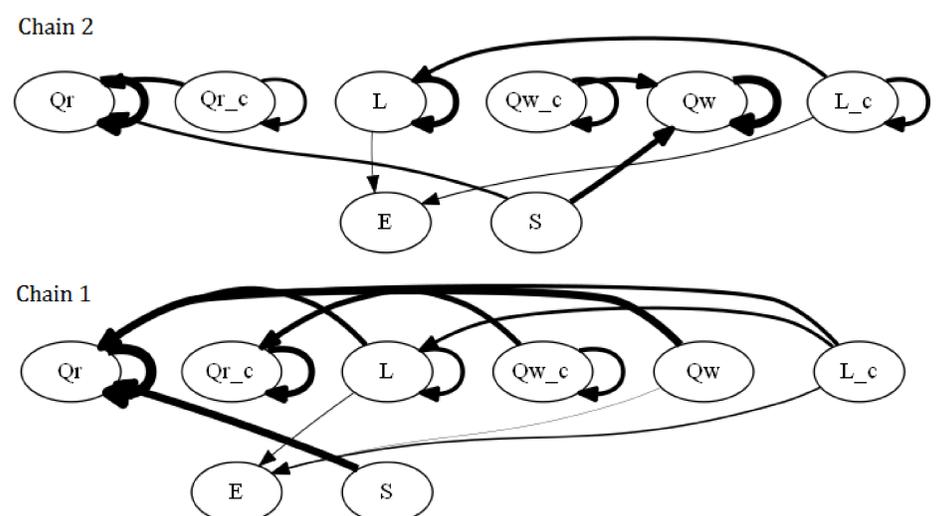
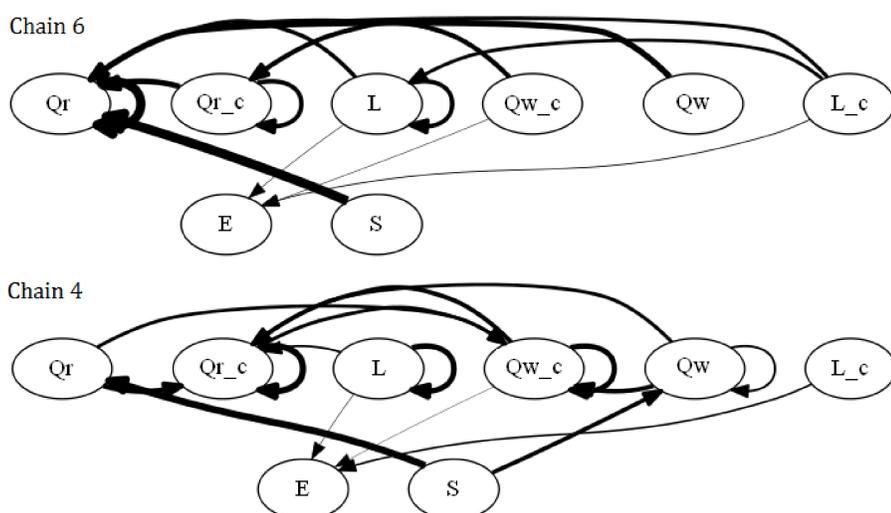
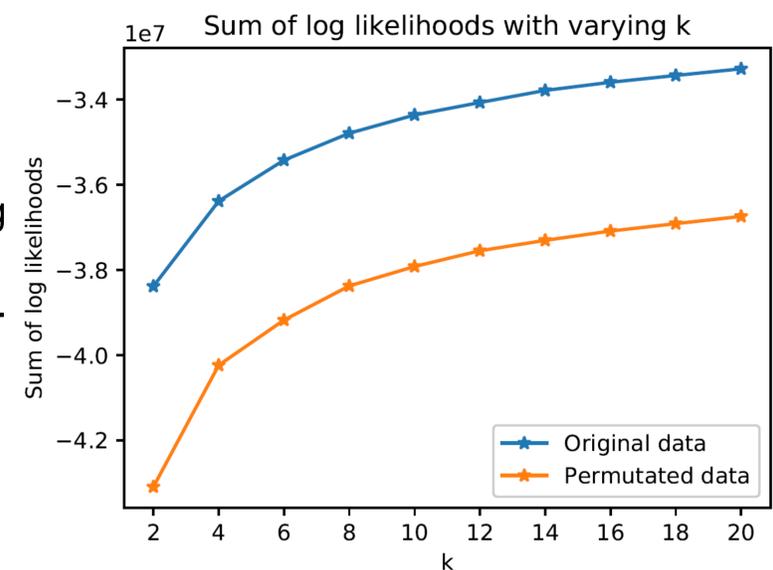
- Edulab, the largest provider of online mathematics in Denmark, produces a lot of log interaction data. The data have previously been mostly unused.
- Wish to develop method for unsupervised clustering, such that unwanted behavior in the system can be identified, without defining beforehand the unwanted behavior.

Number of sequences	1.08M
Number of actions	37.5M
Number of lessons	1.35M
Number of correctly answered questions	27.44M
Number of wrongly answered questions	8.71M

- Log data generated by students in age 12-14 from august 2016 to February 2017 is used.

### Experiment

- Convert student interaction into a state space, consisting of watching video lessons, answering questions right, and answering questions wrong. The state depend on if there is a switch of topic of the material.
- Convert all log data to the state space representation and use K-means like clustering optimizing MLE with random initial chains to cluster the log data
  - **Qr** is a right question, **Qw** is a wrong question, and **L** is lesson,
  - **\_c** corresponds to a change of topic
  - **S** and **E** correspond to start and end of session
- Run on original and permutated dataset.



### Findings

- The resulting chains are illustrated, with thicker lines corresponding to larger transitional probabilities.
- Resulting chains are humanly interpretable, and insights can be derived directly.
- Very binary behavior was observed (chain 2), for close to 12% of the sequences.
- Focused study in single subject (chain 6) and switching between topics (chain 4) is also observed
- Chain 1 is the most frequent chain

### Conclusions and Future work

- The current system allows for unproductive use of lessons, since multiple lessons are often watched subsequently
- Negative educational loops (always be right/wrong) are not prevented.
- Future work will be considering time dependent models and a larger state space allowed by using more of the logged data.