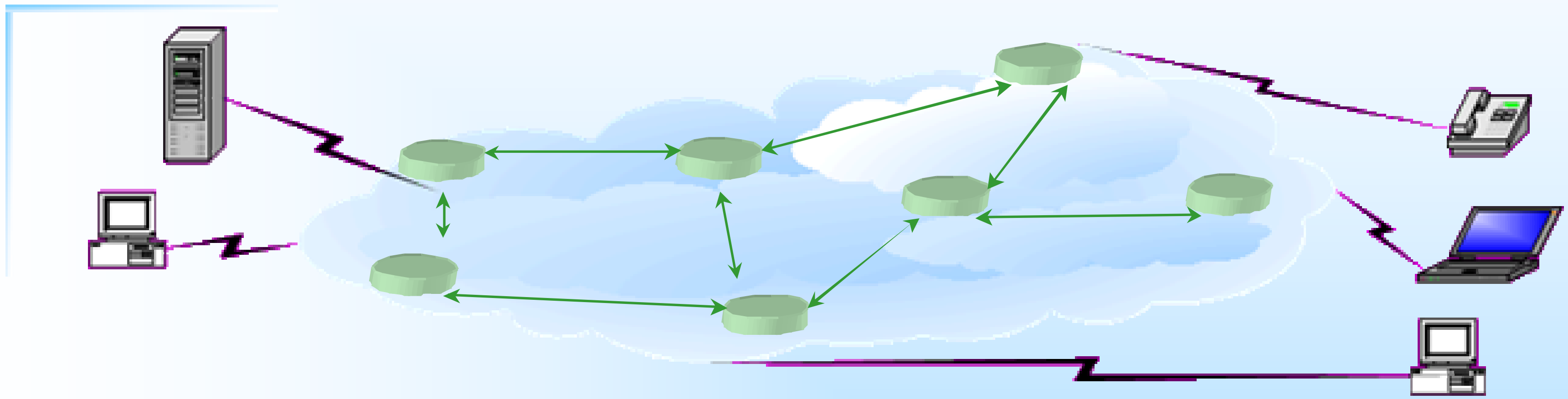
Modelling of Real Network Traffic by Phase-Type distribution

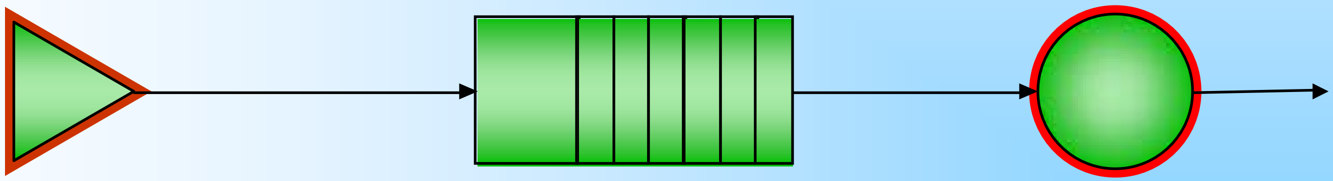
Andriy Panchenko
Dresden University of Technology

Outline

1. Problem overview
2. Fitting algorithm
3. Experiments

Traffic





Need of a traffic model

simple but adequate
applicable to further performance analysis

Failure of Poisson process

- Interarrivals are not exponentially distributed
- Heavy tailed distributions

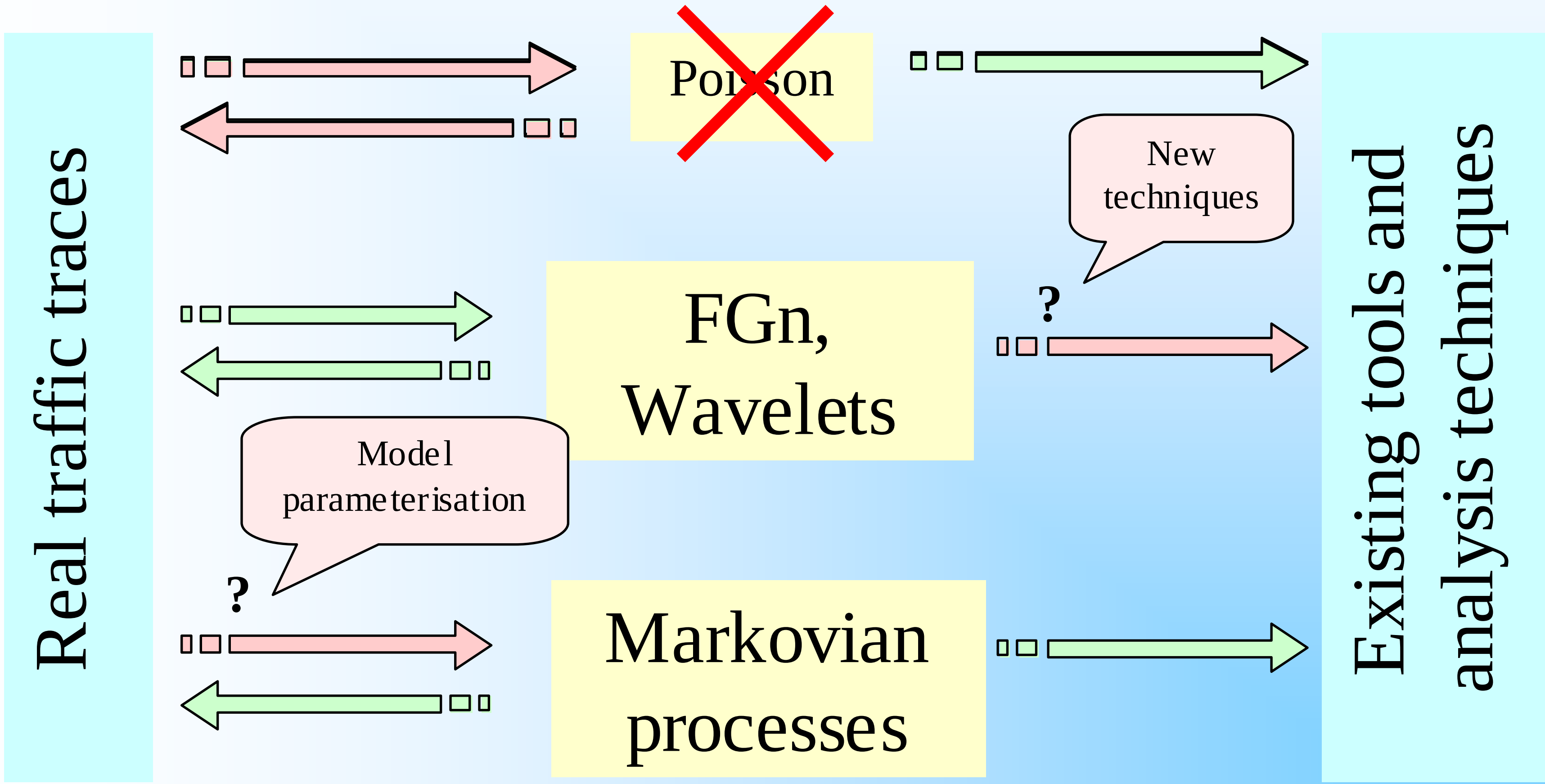
  - Presence of huge values that are generated with small probability but have a great impact on the lower moments of pdf.

Traffic modeling goals

- Goal: to develop a traffic model
  - simple but adequate
  - applicable to further analysis (possibly by means of existing tools and techniques)

- Approaches
  - Separate modeling of each transmitting process and their further aggregation
    (simulation)
  - Statistical description of traffic processes
    (analytical model)
  - …

Modeling compromise

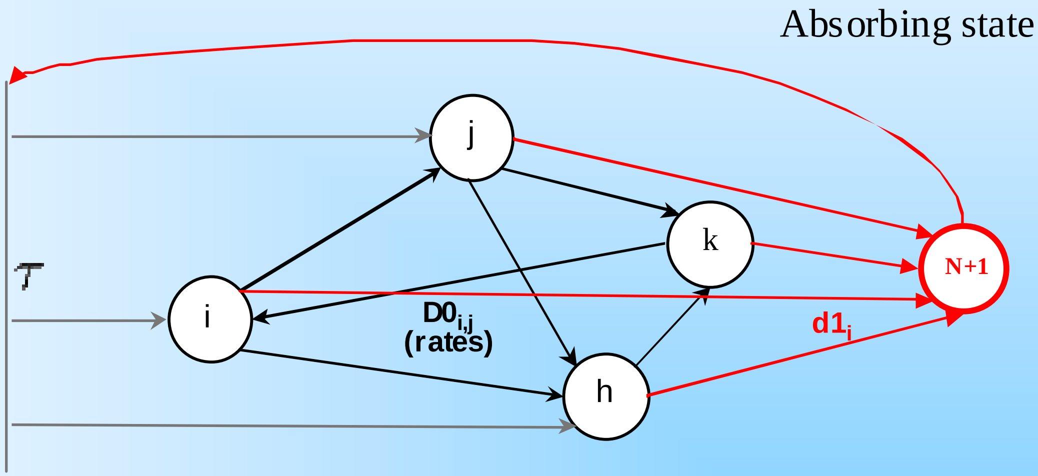


Phase-Type (PH) distribution

- CTMC with N states described by a generator matrix:
  - D0 includes local transitions rates and negative diagonal elements
  - d1 includes transition rates accompanied by an arrival

$$Q = \begin{bmatrix} \mathbf{D}_0 & \mathbf{d}_1 \\ 0 & 0 \end{bmatrix}$$

$$d_1 = -D_0 e^T$$



- π is the stationary distribution (π Q=0)
- τ the distribution after an arrival

$< \tau, \mathbf{D}_0 >$ complete definition of PH-distribution

Properties of PH-distribution

- distribution function

$$F(x) = 1 - \tau \exp(D_0 x) \cdot e^T$$

- probability density function

$$f(x) = F'(x) = \tau \exp(D_0 x) \cdot d_1$$

- k-th moment

$$\mathbf{E}[X^k] = k!\, \tau(-D_0^{-1})^k e^T$$

Randomisation technique

- computing of CTMC transient measures $\exp(D_0 \cdot t_i)$

$$\exp(A) = \sum_{l \geq 0} (1/l!) A^l$$

- Randomisation

$$\exp(D_0 \cdot t_i) = \sum_{k=0}^{\infty} \beta(k, \alpha t_i) \cdot (P_0)^k$$

where

$$P_0 = \frac{D_0}{\alpha} + I \qquad p_1 = \frac{d_1}{\alpha}$$

- Randomisation rate $\alpha \geq \max_{1 \leq j \leq n} |D_0[j, j]|$
- Poisson probability $\beta(k, \alpha t_i)$

General fitting task

- Adjust the free parameters of PH-distribution $< \tau, \mathbf{D}_0 >$
- Maximize the density function over all values t_i

$$L(t;\tau,D_0) = \prod_{i=1}^{m} \tau exp(D_0 t_i) d_1$$

EM-algorithm (Expectation-Maximization)

- t – sample; O – model parameters
- EM iterative algorithm for maximum likelihood parameter estimation if data are incomplete or hidden.
- Problem: find maximum likelihood estimate O_{ML}

$$O_{ML} = \arg\max_{O \in \Omega} L(t; O)$$

- Principle: sequential mapping $O^{(r)} \to O^{(r+1)}$ such that $\boxed{L(O^{(r+1)}) \geq L(O^{(r)})}$
  - E-step: evaluation of conditional expectation
  - M-step: maximization
    - Differentials w.r.t. model parameters *O*
    - Generalized EM $Q(O^{(r+1)}, O^{(r)}) > Q(O^{(r)}, O^{(r)})$

M&S

Exploited features

- Aggregated trace (approx.)
- Fixed initial distribution
- Reduced Likelihood $L\left(D_0^{(s)};t\right)$
- Exploit the randomization technique

$$P_0 = \frac{D_0}{\alpha} + I \qquad p_1 = \frac{d_1}{\alpha}$$

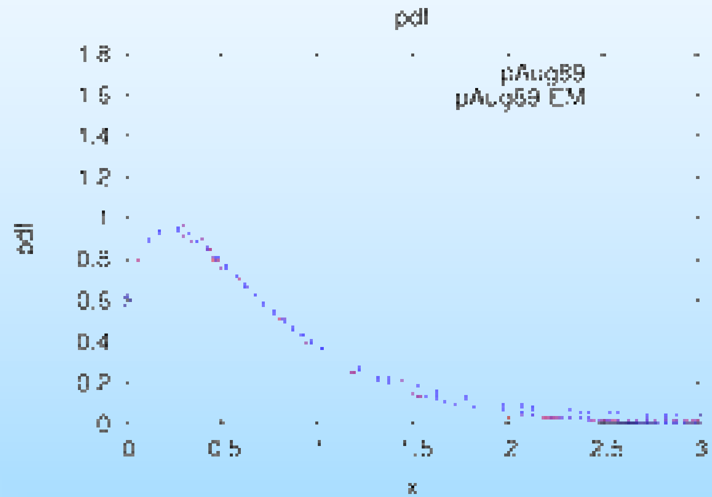
$$P_0 e^T + p_1 = e^T$$

Empirical pdf

- Represent trace *t* as set of tuples $\langle \hat{t}_i, g(\hat{t}_i) \rangle$

$$t_j \in T_i \ , \ t_j \in (\Delta_{i-1}, \Delta_i]$$

- 1: Identical intervals
  - Not enough intervals
  - Heavy tails

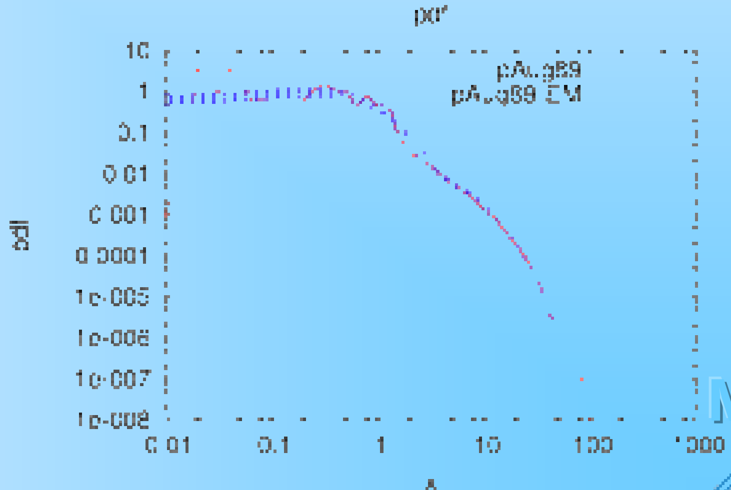


- 2: Log-scaled intervals

$$l_{min} = \lfloor \log_{10}(\min(t)) \rfloor$$

$$l_{max} = \lceil \log_{10}(\max(t)) \rceil$$

$$(10^l, 10^{l+1}]$$



Steps of our EM-algorithm

1. start with some PH-distribution
2. compute the likelihood of the PH-distribution according to the each element in the trace
3. compute the new transition probabilities in the matrices for the given likelihood
4. check for convergence and stop or repeat steps 2-4.

M&S

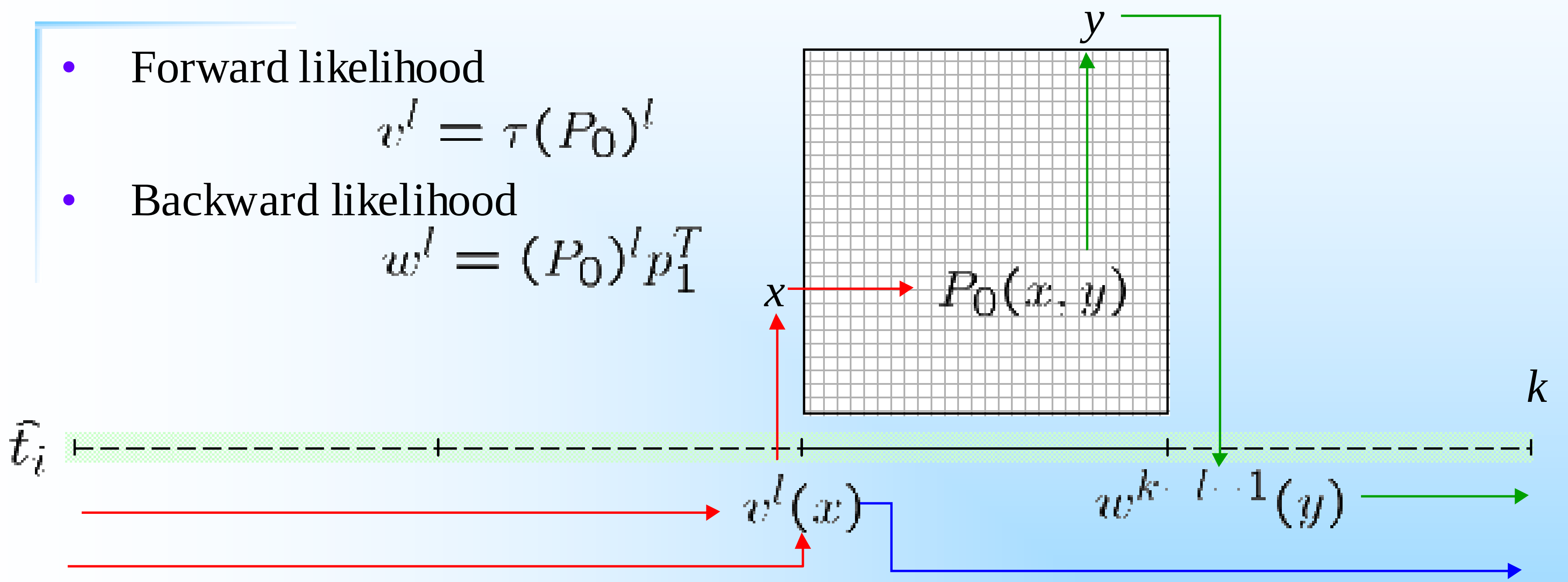
EM-steps

- Forward likelihood

$$v^l = \tau (P_0)^l$$

- Backward likelihood

$$w^l = (P_0)^l p_1^T$$



$$X_0^{(i)}(x,y) = \sum_{k=l_i}^{r_i} \beta(k,\alpha\hat{t}_i) \cdot \sum_{l=0}^{k-1} \underline{v^l(x)} P_0(x,y) \underline{w^{k-l-1}(y)}$$

$$x_1^{(i)}(x) = \sum_{k=l_i}^{r_i} \beta(k,\alpha\hat{t}_i) \cdot \underline{v^k(x)} \underline{p_1(x)}$$

EM-steps

- Collect weighted values of likelihood

$$Y_0 = \sum_{i=1}^{R} g(\hat{t_i}) \cdot X_0^{(i)}$$

$$y_1 = \sum_{i=1}^{R} g(\hat{t_i}) \cdot x_1^{(i)}$$

- Normalize

$$\hat{Y_0}(x, y) = \frac{Y_0(x,y)}{c_r Y_0 e^T + y_1(x)}$$

$$\hat{y_1} = e^T - \hat{Y_0} e^T$$

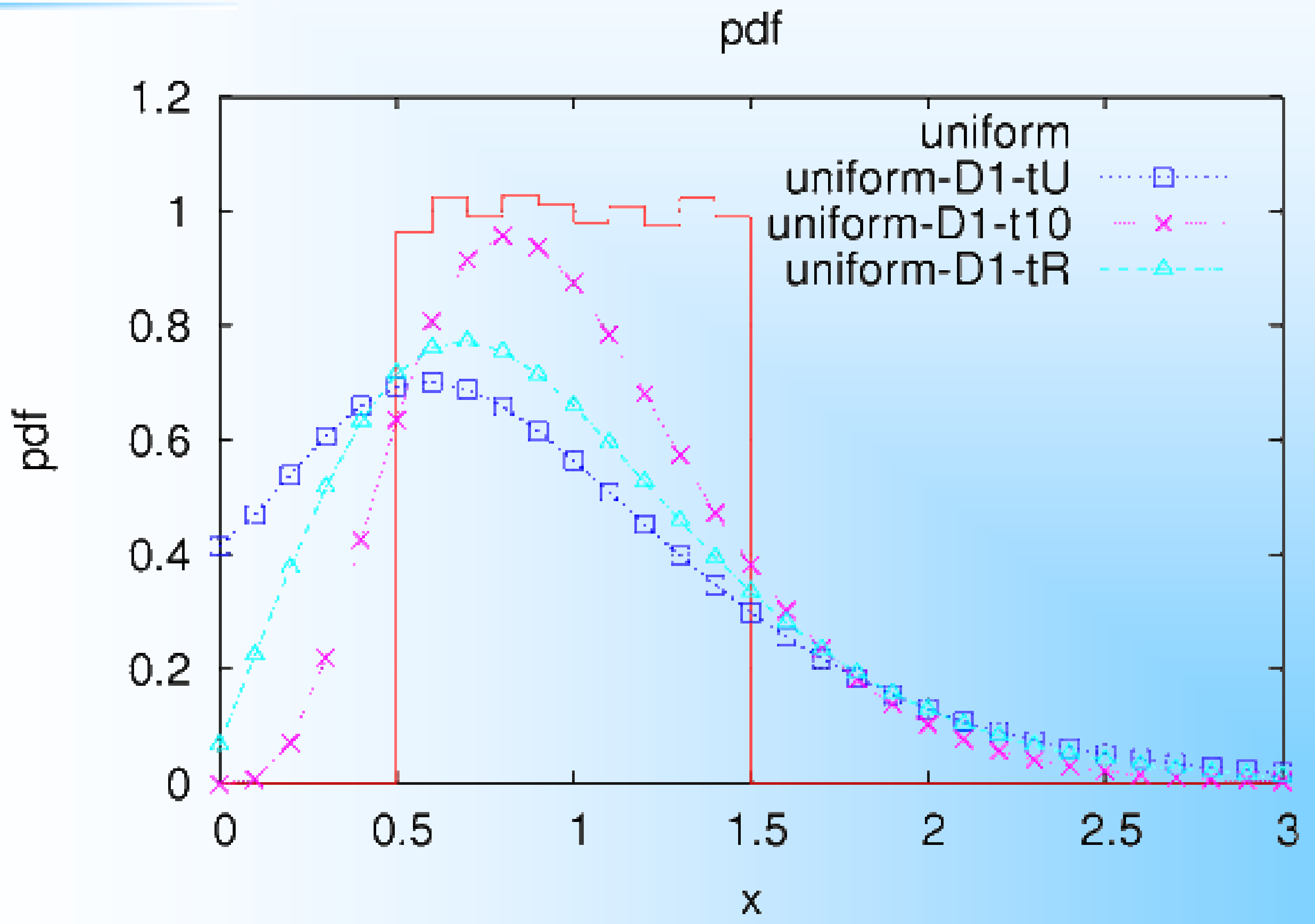
Algorithm framework

Input: Trace, Initial distribution, Randomization rate
Build aggregated trace
Encode matrices $P_0 = \frac{D_0}{\alpha} + I$

REPEAT
 Calculate forward / backward likelihoods
 FOR all intervals
  Calculate likelihood matrices $X_0^{(i)}, x_1^{(i)}$
 END for
 Collect and normalize likelihood matrices in Y_0
 $P_0 = Y_0$
UNTIL converges

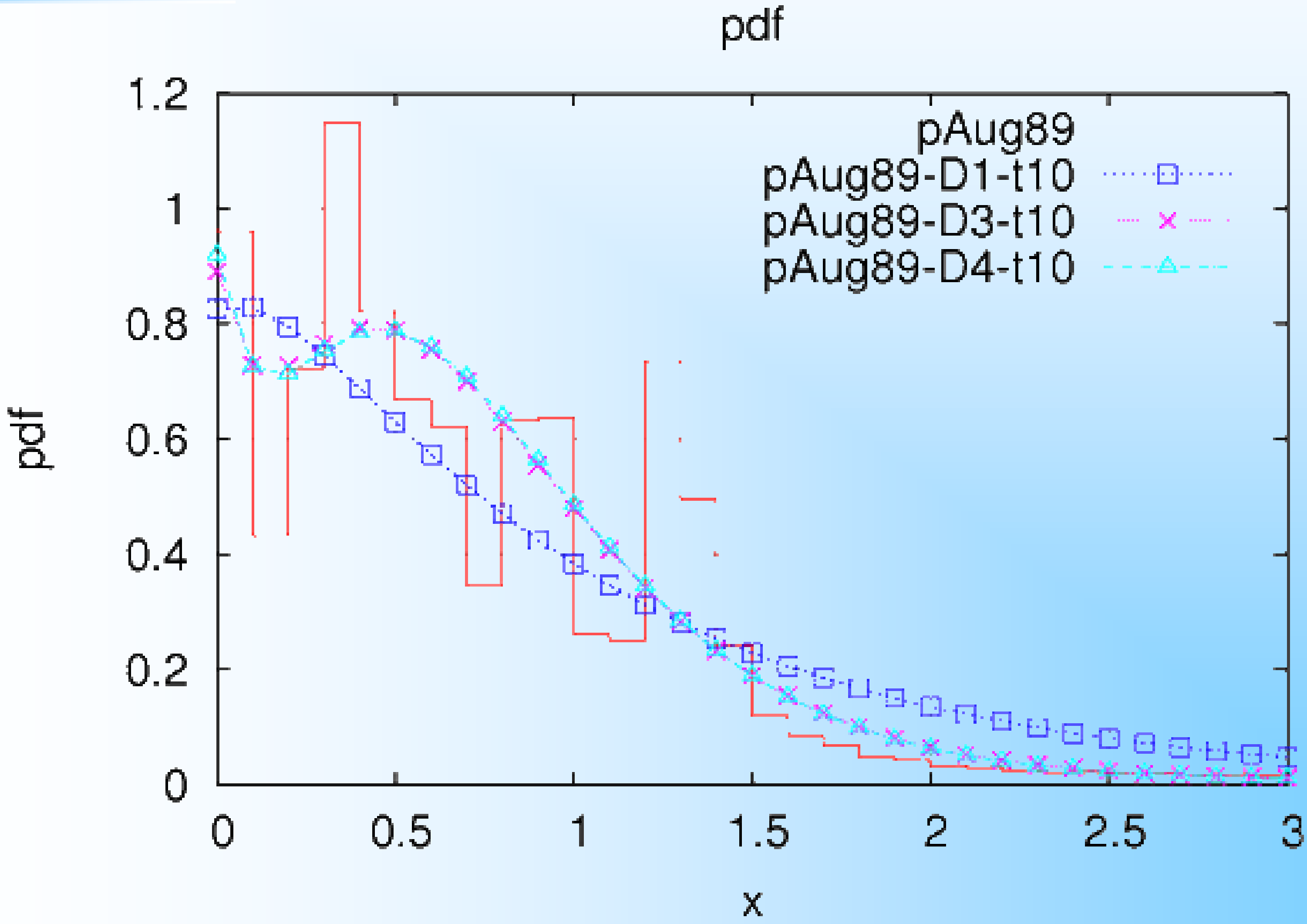
Decode matrices $D_0 = \alpha(P_0 + I)$

Influence of initial distribution



uniform
(1,0,…,0)
random

Influence of matrix form

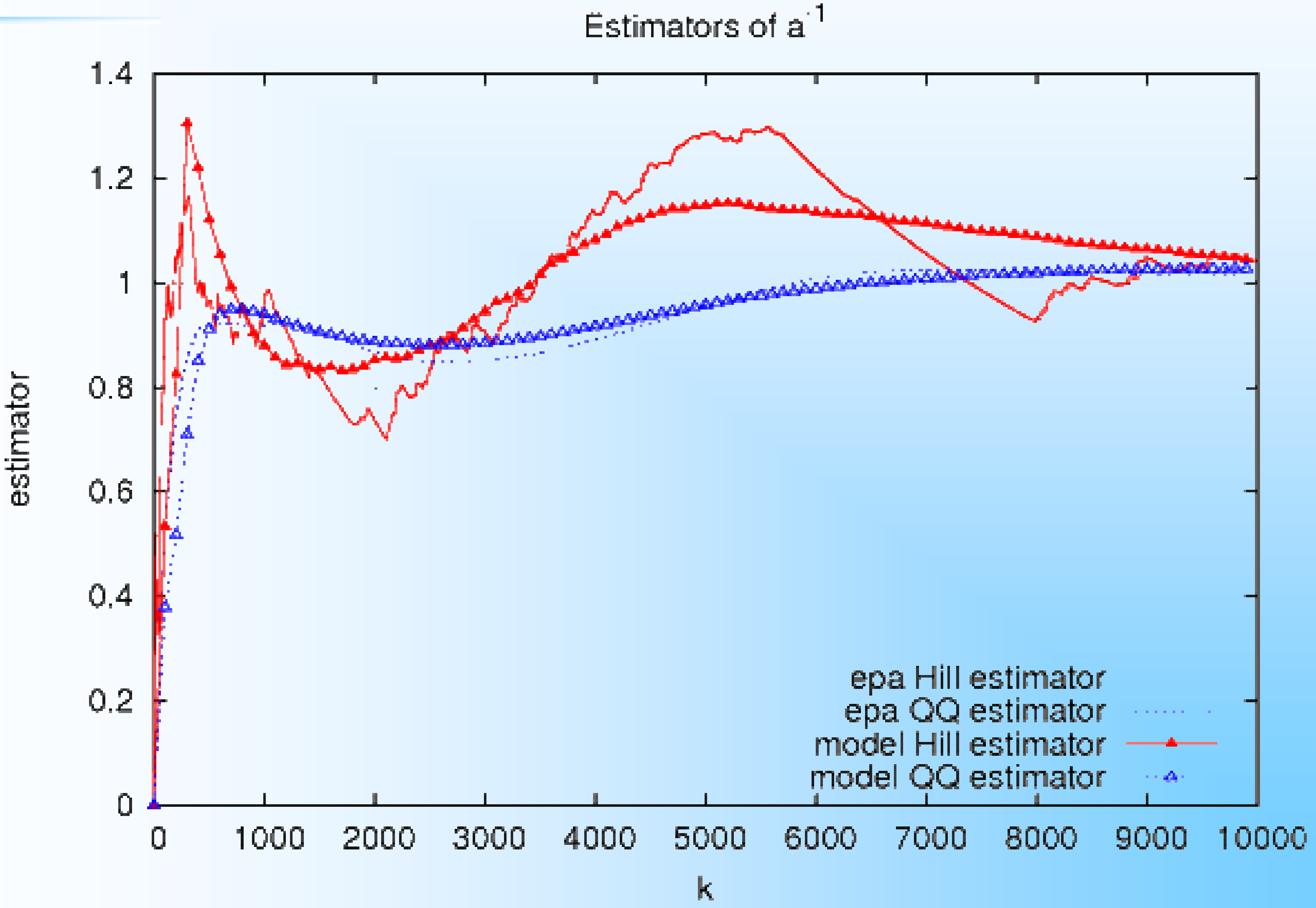


dense

\\

upper triang

Fitting of heavy-tailed distributions



Conclusion

- good capture of main part as well as tail behaviour of empirical distribution
- a larger number of phases yields a better likelihood value
- best approximation was achieved with
  - initial vector (1,0,…,0)
  - matrix form: upper triangular or with main and first upper diagonals
- h.t.d. properties are captured with relatively compact models with 5 states

M&S

Thank You!