Aalto University
School of Electrical
Engineering

Lecture 7:
Review of Markov Processes

Pasi Lassila
Department of Communications and Networking

Contents

- Markov processes theory recap
- Elementary queuing models for data networks
- Simulation of Markov processes

Markov process

- Consider a **continuous-time and discrete-state** stochastic process $X(t)$
  - with state space $S = \{0, 1, \ldots, N\}$ or $S = \{0, 1, \ldots\}$
- **Definition**: The process $X(t)$ is a **Markov process** if

$$
\begin{aligned}
&P\{X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1, \ldots, X(t_n) = x_n\} = \\
&P\{X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n\}
\end{aligned}
$$

for all n, $t_1 < \ldots < t_{n+1}$ and $x_1, \ldots, x_{n+1}$

- This is called the **Markov property**
  - Given the current state, the future of the process does not depend on its past (that is, *how* the process has evolved to the current state)
  - As regards the future of the process, the current state contains all the required information

Time-homogeneity, transition probabilities

- **Definition**: Markov process $X(t)$ is **time-homogeneous** if

$$\boxed{P\{X(t+h)=y \mid X(t)=x\} = P\{X(h)=y \mid X(0)=x\}}$$

for all $t,\ h \geq 0$ and $x, y \in S$

  – In other words, probabilities $P\{X(t+h)=y \mid X(t)=x\}$ are independent of t

  – further, the conditional probability depends only on the difference of times, h

State transition rates

- Consider a time-homogeneous Markov process $X(t)$
- The **state transition rates** q_{ij}, where $i, j \in S$, are defined as follows:

$$\boxed{q_{ij} := \lim_{h \downarrow 0} \tfrac{1}{h} P\{X(h) = j \mid X(0) = i\}}$$

  – Transition rate q_{ij} describes the rate of probability mass from state i to state j

- The initial distribution $P\{X(0) = i\}$, $i \in S$, and the state transition rates q_{ij} together determine the state probabilities $P\{X(t) = i\}$, $i \in S$, by the **Kolmogorov equations**
- Note that we will consider only time-homogeneous Markov processes

Dynamic behavior: Exponential holding times

- Assume that a Markov process is in state i
- During a short time interval $(t, t+h]$, the conditional probability that there is a transition from state i to state j is $q_{ij}h + o(h)$ (independently of the other time intervals)
- Let q_i denote the total transition rate out of state i, that is:

$$q_i := \sum_{j \neq i} q_{ij}$$

- Then, during a short time interval $(t, t+h]$, the conditional probability that there is a transition from state i to any other state is $q_i h + o(h)$ (independently of the other time intervals)
- This is clearly a memoryless property
- Thus, the holding time in (any) state i is exponentially distributed with intensity q_i

Dynamic behavior: State transition probabilities

- Let T_i denote the holding time in state i and T_{ij} denote the (potential) holding time in state i that ends to a transition to state j

$$T_i \sim \mathrm{Exp}(q_i), \quad T_{ij} \sim \mathrm{Exp}(q_{ij})$$

- T_i can be seen as the minimum of independent and exponentially distributed holding times T_{ij}

$$T_i = \min_{j \neq i} T_{ij}$$

- Let then p_{ij} denote the conditional probability that, when in state i, there is a transition from state i to state j (the **state transition probabilities**);

$$p_{ij} = P\{T_i = T_{ij}\} = \frac{q_{ij}}{q_i}$$

Transition rate matrix

- The state transition rates q_{ij} and q_i define the **transition rate matrix** Q

$$\boxed{Q := (q_{ij}; i, j \in S)}$$

where

$$\boxed{q_{ii} := -q_i = -\sum_{j \neq i} q_{ij}}$$

- Example: for $S=\{0,1,2\}$:

$$Q = \begin{pmatrix} -q_0 & q_{01} & q_{02} \\ q_{10} & -q_1 & q_{12} \\ q_{20} & q_{21} & -q_2 \end{pmatrix}$$

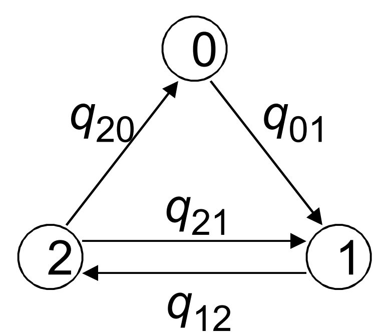
State transition diagram

- A time-homogeneous Markov process can be represented by a **state transition diagram**, which is a directed graph where
  - nodes correspond to states and
  - one-way links correspond to potential state transitions

$$\text{link from state } i \text{ to state } j \quad \Leftrightarrow \quad q_{ij} > 0$$

- Example: Markov process with three states, $S = \{0,1,2\}$

$$Q = \begin{pmatrix} -q_{01} & q_{01} & 0 \\ 0 & -q_{12} & q_{12} \\ q_{20} & q_{21} & -(q_{20} + q_{21}) \end{pmatrix}$$



Irreducibility

- **Definition**: There is a **path** from state i to state j $(i \to j)$ if there is a directed path from state i to state j in the state transition diagram.
  - In this case, starting from state i, the process visits state j with positive probability (sometimes in the future)
- **Definition**: States i and j **communicate** $(i \leftrightarrow j)$ if $i \to j$ and $j \to i$.
- **Definition**: Markov process is **irreducible** if all states $i \in S$ communicate with each other
  - Example: The Markov process presented in slide 9 is irreducible

Irreducible Markov processes and equilibrium distribution

- An irreducible Markov process $X(t)$ with a finite state space has always a unique equilibrium distribution π.
  - Can be solved from the global balance equations (GBE) for each state together with the normalization condition (N)

$$\forall i, \quad \sum_{j \neq i} \pi_i q_{ij} = \sum_{j \neq i} \pi_j q_{ji} \ (GBE) \quad , \quad \sum_i \pi_i = 1 \quad (N)$$

- The quilibrium distribution can be calculated numerically from

$$\pi = e \cdot (Q + E)^{-1}$$

  - where e is a vector of 1's and E is a matrix of 1's

Birth-death process

- Consider a continuous-time and discrete-state Markov process $X(t)$
  - with state space $S = \{0,1,\ldots,N\}$ or $S = \{0,1,\ldots\}$
- **Definition**: The process $X(t)$ is a **birth-death process** (BD) if state transitions are possible only between neighbouring states, that is:

$$|i-j|>1 \quad \Rightarrow \quad q_{ij} = 0$$

- In this case, we denote

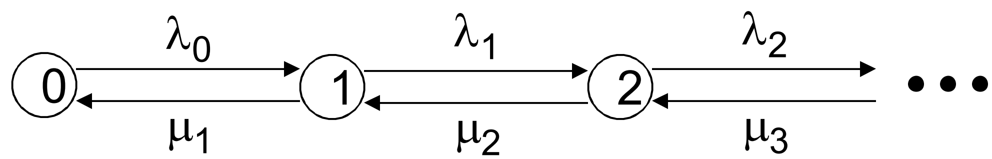
$$\mu_i := q_{i,i-1} \geq 0$$
$$\lambda_i := q_{i,i+1} \geq 0$$

  - In particular, we define $\mu_0 = 0$ and $\lambda_N = 0$ (if $N < \infty$)
  - the rates are called the death and birth rates, respectively.

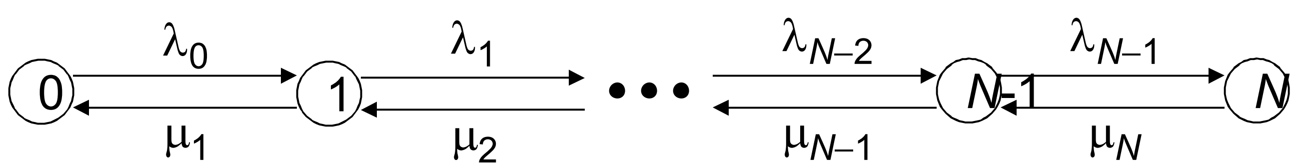
Irreducibility

- **Proposition**: A birth-death process is irreducible if and only if $\lambda_i > 0$ for all $i \in S \setminus \{N\}$ and $\mu_i > 0$ for all $i \in S \setminus \{0\}$

- State transition diagram of an infinite-state irreducible BD process:



- State transition diagram of a finite-state irreducible BD process:



B-D processes and their equilibrium distribution

- Consider an irreducible birth-death process $X(t)$
- Equilibrium distribution $\pi = (\pi_i \mid i \in S)$ (if it exists) given by LBEs
- Local balance equations (LBE):

$$\pi_i \lambda_i = \pi_{i+1} \mu_{i+1} \qquad \text{(LBE)}$$

- Thus we get the following recursive formula:

$$\pi_{i+1} = \frac{\lambda_i}{\mu_{i+1}} \pi_i \quad \Rightarrow \quad \pi_i = \pi_0 \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}$$

- Normalizing condition (N):

$$\sum_{i \in S} \pi_i = \pi_0 \sum_{i \in S} \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j} = 1 \qquad \text{(N)}$$

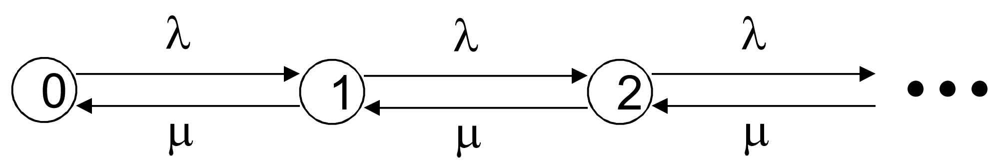
Contents

- Markov processes theory recap
- Elementary queuing models for data networks
- Simulation of Markov processes

BD-processes and Kendall’s notation

- Birth-death processes are an important sub-class of Markov processes because they represent elementary queueing models

- Example:
  - Assume birth rate λ and death rate μ (both independent of state)



  - Corresponds to a system where customers arrive at constant rate λ and they are served in FIFO order by a server with constant service rate μ
  - In Kendall’s notation this is the M/M/1 queueing model
    - Poisson arrivals (M), memoryless = exponential service times (M) and 1 server

A/B/n/p/k [Kendall (1953)]

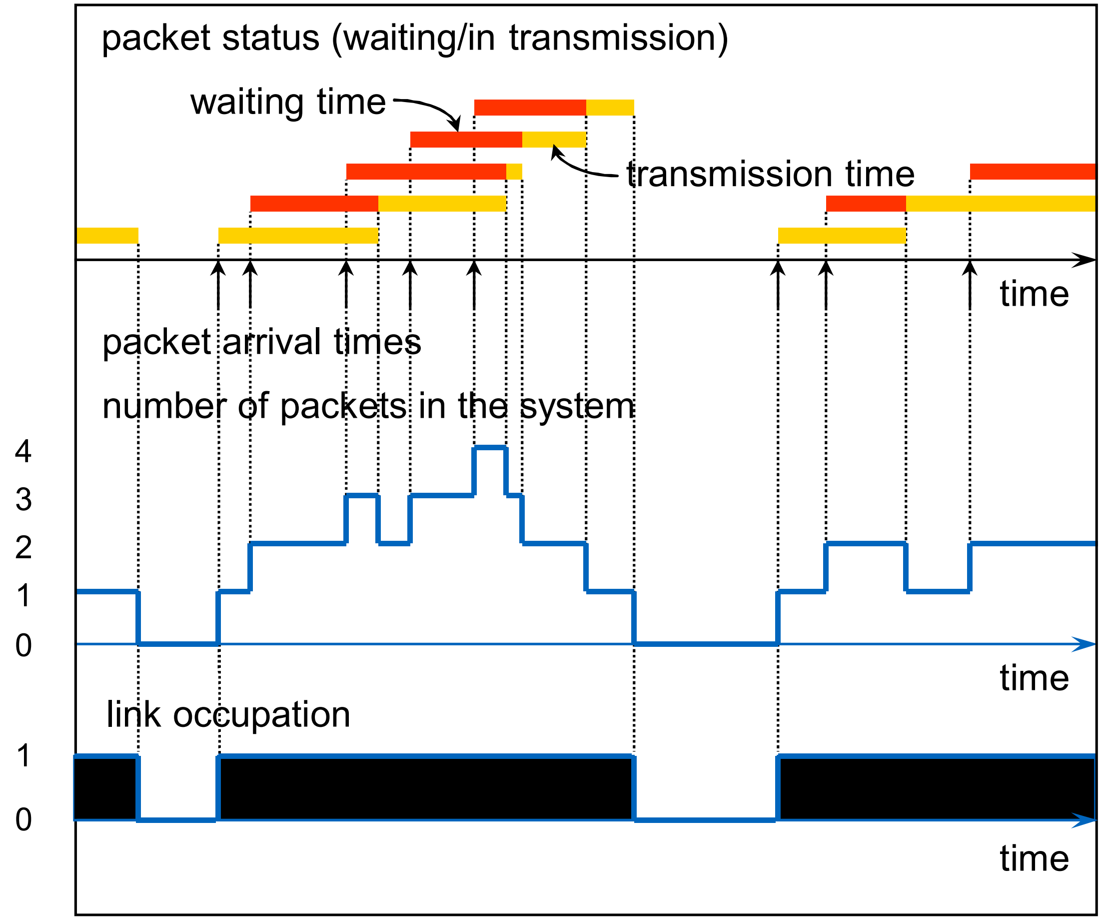
- A refers to the **arrival process**.
  Assumption: IID interarrival times.
  Interarrival time distribution:
  - M = exponential (memoryless)
  - D = deterministic
  - G = general
- B refers to **service times**.
  Assumption: IID service times.
  Service time distribution:
  - M = exponential (memoryless)
  - D = deterministic
  - G = general
- n = nr of (parallel) servers
- p = nr of system places
  = nr of servers + waiting places
- k = size of customer population
- Default values (usually omitted):
  - $p = \infty, k = \infty$
- Examples:
  - M/M/1
  - M/D/1
  - M/G/1
  - G/G/1
  - M/M/n
  - M/M/n/$n+m$
  - M/M/∞ (Poisson model)
  - M/M/n/n (Erlang model)
  - M/M/k/k/k (Binomial model)
  - M/M/n/n/k (Engset model, $n < k$)

IID = independently and identically distributed

Packet level model for data traffic (1)

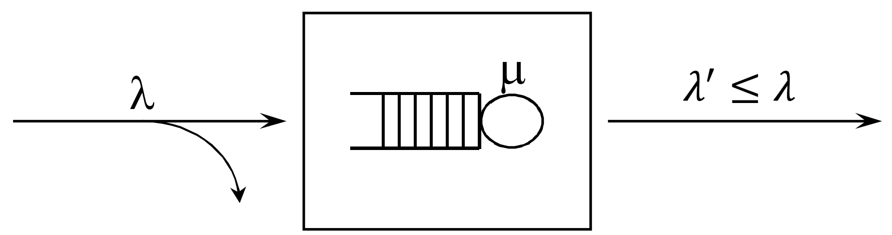
- Consider a single link in a data network (such as, IP network)
- Data traffic consists of **packets**
  - packets compete with each other for the processing and transmission resources (statistical multiplexing)
  - packet characterisation: **length** (in data units)
- Modelling of offered traffic:
  - **packet arrival process** (at which moments new packets arrive)
  - **packet length distribution** (how long they are)
- Link model: a **single server queueing system**
  - the service rate μ depends on the **link capacity** and the **average packet length**
  - when the link is busy, new packets are buffered, if possible, otherwise they are lost
  - Packets are served in FIFO manner

Packet level traffic process



Packet level model for data traffic (2)

- The link is modelled as a **queueing system** with a single server and (in)finite buffer
  - customer = packet
    - λ = packet arrival rate (packets per time unit)
    - L = average packet length (bits/bytes)
  - server = link, waiting places = finite buffer
    - C = link speed (bits per time unit)
  - service time = packet transmission time
    - $1/\mu = L/C$ = average packet transmission time (time units)



Traffic load

- The strength of the offered traffic is described by the traffic load ρ
- By definition, the **traffic load** ρ is the ratio between the arrival rate λ and the service rate $\mu = C/L$:

$$\boxed{\rho = \frac{\lambda}{\mu} = \frac{\lambda L}{C}}$$

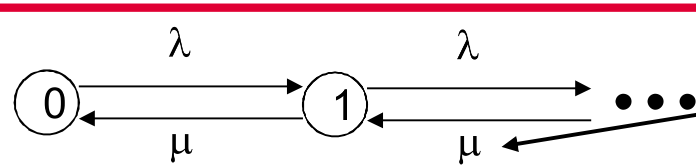
  - The traffic load is a **dimensionless** quantity
  - By Little's formula, it tells the **utilization factor** of the server, which is the probability that the server is busy, if buffer is assumed infinite

Performance model (1)

- System capacity
  - C = link speed in kbps
- Traffic load
  - λ = packet arrival rate in pps (considered here as a variable)
  - L = average packet length in kbits
- Quality of service (from the users' point of view)
  - $E[D]$ = mean delay (from arrival to departure)
- We can model this as an M/M/1 queue!

Performance model (2)

- The **M/M/1 queueing system**:
  - packets arrive according to a **Poisson process** (with rate λ)
  - packet lengths are i.i.d. according to the **exponential distribution** with mean L, server processes packets at rate C
  - Thus, service times are exponential with mean $1/\mu = L/C$
  - queuing discipline is **FIFO** , with 1 server and infinite queue size
- This is just a birth death process



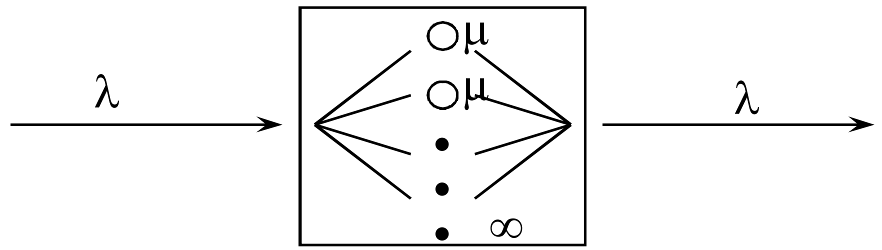
$\mu = C/L$

  - Mean delay $E[D]$ is (due to Little)

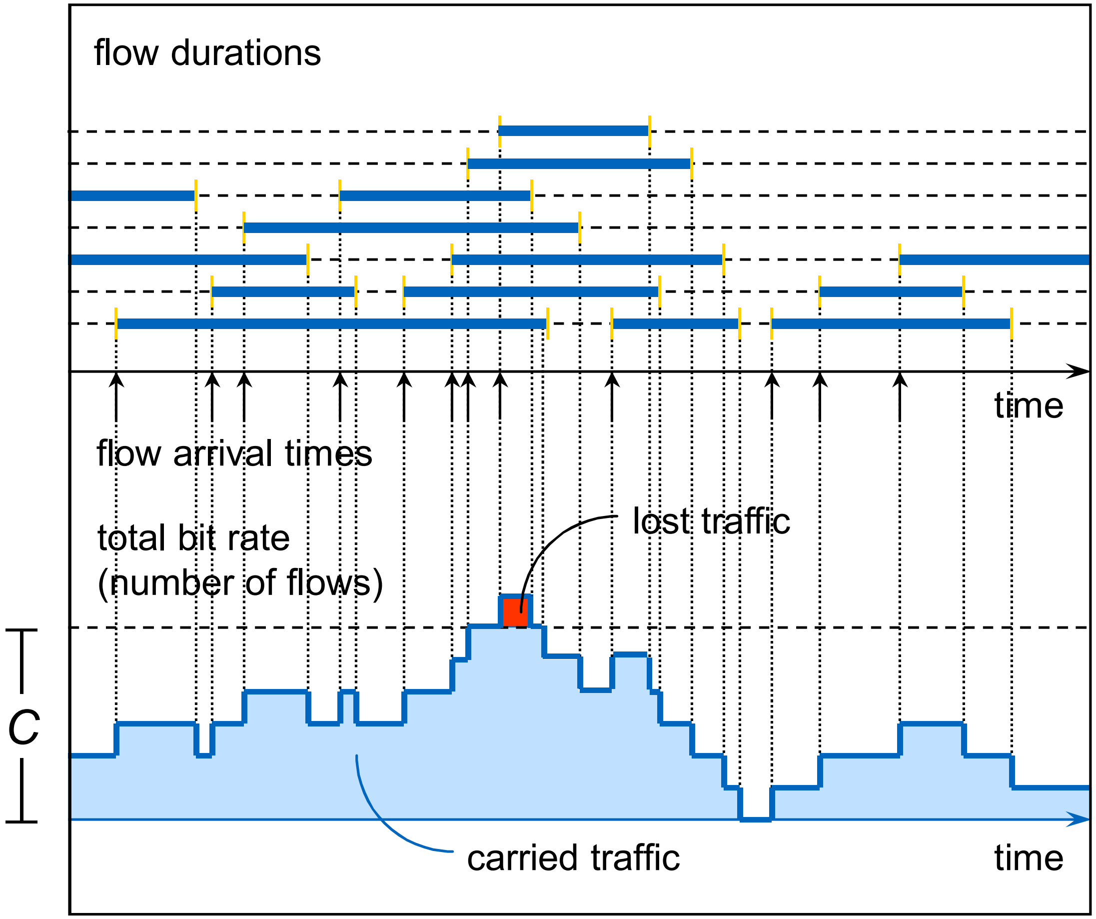
$$E[D] = \frac{E[X]}{\lambda} = \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1}{\mu-\lambda}$$

Flow level model for streaming CBR traffic (1)

- Consider a link between two routers
  - traffic consists of UDP flows carrying CBR traffic (like VoIP)
- Link model: an **infinite system**
  - customer = UDP flow = CBR bit stream
    - λ = flow arrival rate (flows per time unit)
  - service time = flow duration
    - $h = 1/\mu$ = average flow duration (time units)
- **Bufferless** flow level model:
  - when the total transmission rate of the flows exceeds the link capacity, bits are lost (uniformly from all flows)



Traffic process



Offered traffic

- Let r denote the bit rate of any flow
- The volume of offered traffic is described by average total bit rate R
  - By Little's formula, the average number of flows is

$$\boxed{a = \lambda h}$$

  - This may be called **traffic intensity** (cf. slide 6)
  - It follows that

$$\boxed{R = ar = \lambda hr}$$

Loss ratio

- Let N denote the number of flows in the system
- When the total transmission rate Nr exceeds the link capacity C, bits are lost with rate

$$Nr - C$$

- The average loss rate is thus

$$E[(Nr - C)^+] = E[\max\{Nr - C, 0\}]$$

- By definition, the **loss ratio** p_{loss} gives the **ratio between the traffic lost and the traffic offered**:

$$p_{\text{loss}} = \frac{E[(Nr - C)^+]}{E[Nr]} = \frac{1}{ar} E[(Nr - C)^+]$$

Performance model (1)

- System capacity
  - $C = nr$ = link speed in kbps

- Traffic load
  - $R = ar$ = offered traffic in kbps
  - r = bit rate of a flow in kbps.
  - h = average duration of a flow

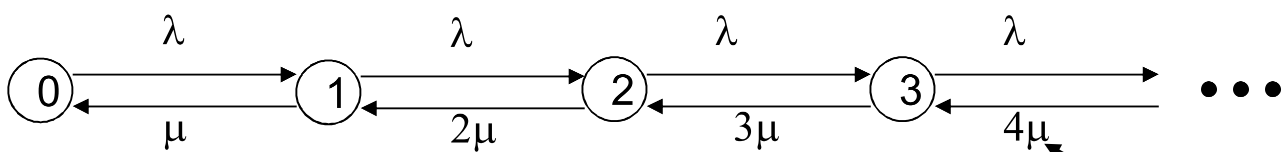
- Quality of service (from the users' point of view)
  - p_{loss} = loss ratio

- We can model this using the M/M/∞ model!

Performance model (2)

- Assume an **M/M/∞ infinite system**:
  - flows arrive according to a **Poisson process** (with rate λ)
  - flow durations are independent and identically distributed according to **exponential distribution** with mean h

- Again, this is just a BD-process!



$$\mu = 1/h$$

  - But to estimate the performance, one must record the amount of lost traffic (see earlier slide)

$$p_{loss} = \frac{1}{ar} E[(Nr - C)^+] = \frac{e^{-a}}{ar} \sum_{n=C/r+1}^{\infty} \frac{a^n}{n!}(nr - C)$$

Max number of flows with no loss!

Contents

- Markov processes theory recap
- Elementary queuing models for data networks
- Simulation of Markov processes

Simulation of a Markov process

- Given current state i, one simply needs to generate the time the process stays in state i and what is the next state j

- Basically 2 ways to implement
  - The methods follow directly from the dynamic behavior of the Markov process as described on slides 7-8
  - One can consider the next transition following from
    - Method 1. a minimum of exponentially distributed r.v.'s or
    - Method 2. time until next event and then using the branching probabilities

- Consider a finite state Markov process $X(t)$ (does not have to be irreducible) with transition rate matrix Q and state space S=$1,\ldots,N$

Method 1

- Aim: Simulate process $X(t)$ with initial state x_0 for K transitions

- Initialize: state $x{=}x_0$, time $t = 0$ and transition counter $\mathrm{step}{=}0$
- Stopping condition: If $\mathrm{step} < K$, then
  - Draw a sample $t_j(x)$ of times to next possible events in state x for all $j{=}1,\ldots,N$, i.e., each $t_j(x) \sim \mathrm{Exp}(q_{xj})$
  - The holding time (time to next transition) in state x, is given by $\min\,(t_1(x),\ldots,\, t_N(x))$
  - Time $t = t + \min\,(t_1(x),\ldots,\, t_N(x))$
  - Next state x where the process moves is $x = \arg\min\,(t_1(x),\ldots,\, t_N(x))$
  - Increase step counter: $\mathrm{step}{=}\mathrm{step}{+}1$

- Note: there is no statistics collection here!

Method 2

- Aim: Simulate process $X(t)$ with initial state x_0 for K transitions

- Initialize: state $x{=}x_0$ and transition counter $\mathrm{step}{=}0$
- Stopping condition: If $\mathrm{step} < K$, then
  – Holding time: draw a sample $t(x)$ of time to next transition, i.e., $t(x) \sim \mathrm{Exp}(q_x)$ (recall q_x is the sum of transition rates out from state x)
  – Time $t = t + t(x)$
  – Next state y is selected from the discrete distribution so that with probability q_{xy} / q_x the process moves to state y
  – Increase step counter: $\mathrm{step}{=}\mathrm{step}{+}1$

- Note: there is no statistics collection here!

Simulation of a Markov chain

- Markov chain is the discrete time counter part of the Markov process
  - That is, in addition to the state being discrete, also time is discrete
  - Can be used to model systems where time is slotted (e.g., cellular systems)

- Characterized by matrix P, where each element p_{ij} gives the probability to move from state i to state j in the next transition

- Simulation then just corresponds to simulating these ”jumps” from one time step to the next