Markov Chains

Sarah Filippi
Department of Statistics
http://www.stats.ox.ac.uk/~filippi
TA: Patrick Gemmel

With grateful acknowledgements to Prof. Yee Whye Teh's slides from 2013–14.

My use of statistics

- I work on a range of topics in computational statistics focused on understanding biological processes and their relation to disease.
- Using Bayesian statistics and stochastic models.
- I did a PhD in reinforcement learning, a field at the intersection of statistics, machine learning and artificial intelligence.

Schedule

- 09:30-10:30 Lecture: Introduction to Markov chains
- 10:30-12:00 Practical
- 12:00-13:00 Lunch
- 13:00-14:00 Lecture: Continuous Markov Chain
- 14:00-15:30 Practical

Practicals

- Some mathematical derivations.
- Some programming in:
  - R
  - MATLAB
- Probably not possible to do all practicals; pick and choose.

- Package available at

http://www.stats.ox.ac.uk/~filippi/teaching.html

Markov Chains

Andrey Andreyevich Markov
1856-1922

Sequential Processes

- Sequence of random variables $X_0, X_1, X_2, X_3, \ldots$
- Not iid (independently and identically distributed).
- Examples:
  - X_i = Rain or shine on day i.
  - X_i = Nucleotide base at position i.
  - X_i = State of system at time i.
- Joint probability can be factorized using Bayes' Theorem:

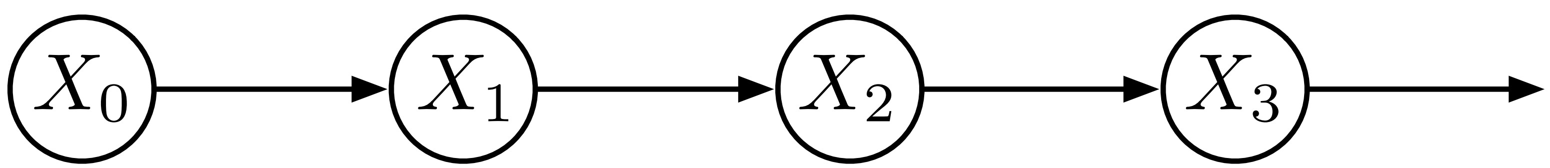
$$
\begin{aligned}
&\mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2 \ldots) \\
=&\mathbb{P}(X_0 = x_0)\mathbb{P}(X_1 = x_1 | X_0 = x_0)\mathbb{P}(X_2 = x_2 | X_0 = x_0, X_1 = x_1) \cdots
\end{aligned}
$$

Markov Assumption

- Markov Assumption: each X_i **only** depends on the previous X_{i-1}.

$$
\begin{aligned}
&\mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2 \ldots) \\
=&\mathbb{P}(X_0 = x_0)\mathbb{P}(X_1 = x_1 | X_0 = x_0)\mathbb{P}(X_2 = x_2 | X_0 = x_0, X_1 = x_1) \cdots \\
=&\mathbb{P}(X_0 = x_0)\mathbb{P}(X_1 = x_1 | X_0 = x_0)\mathbb{P}(X_2 = x_2 | X_1 = x_1) \cdots
\end{aligned}
$$

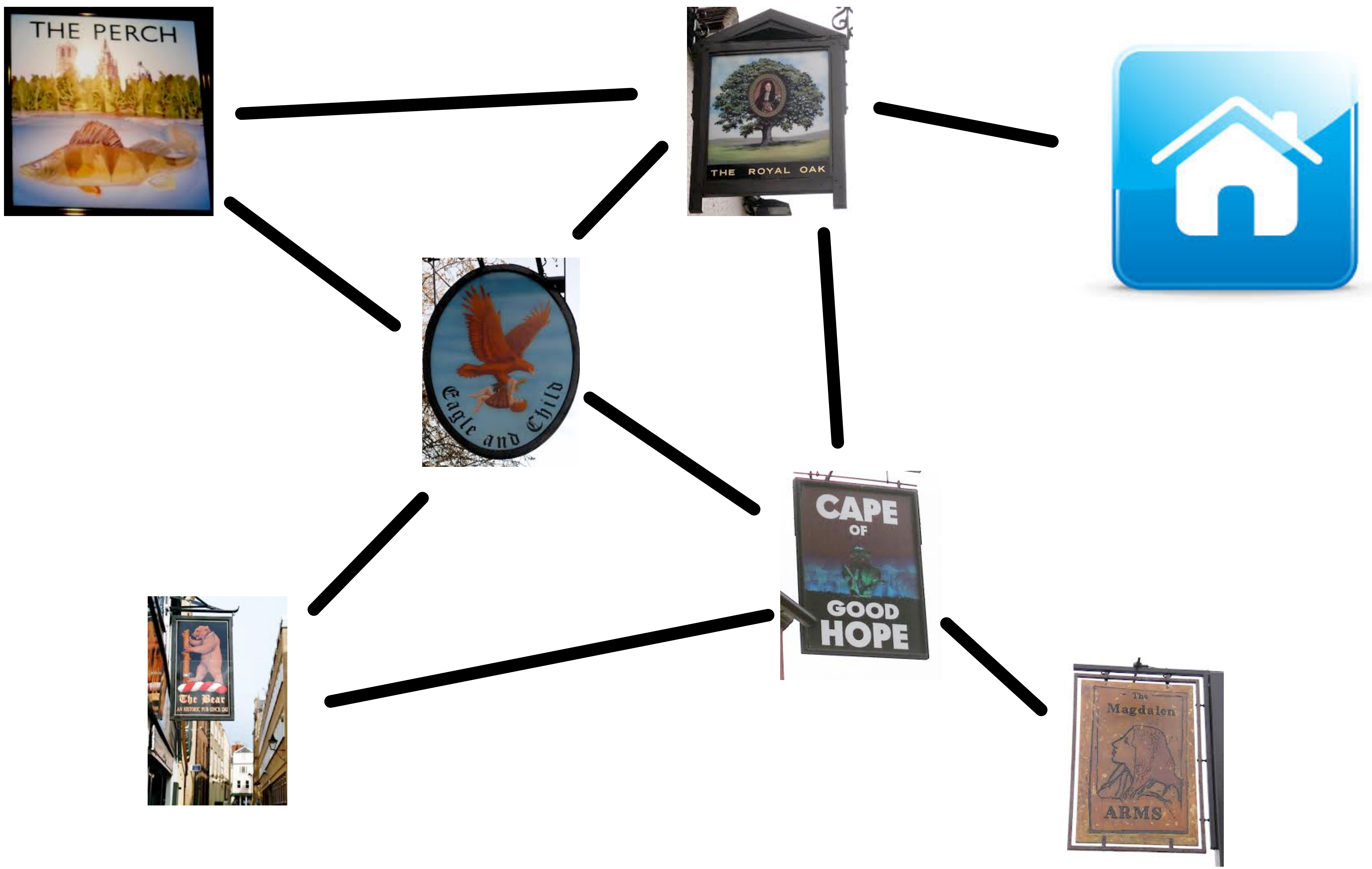
- Future is independent of the past, given the present.
- Process “has no memory”.



- Higher order Markov chains:

$$
\begin{aligned}
&\mathbb{P}(X_t = x_t | X_0 = x_0, \ldots, X_{t-1} = x_{t-1}) \\
=&\mathbb{P}(X_t = x_t | X_{t-k} = x_{t-k}, \ldots, X_{t-1} = x_{t-1})
\end{aligned}
$$

Random Pub Crawl



Jukes-Cantor DNA Evolution

GCTCATGCCG

|

GCTAATGCCG

|

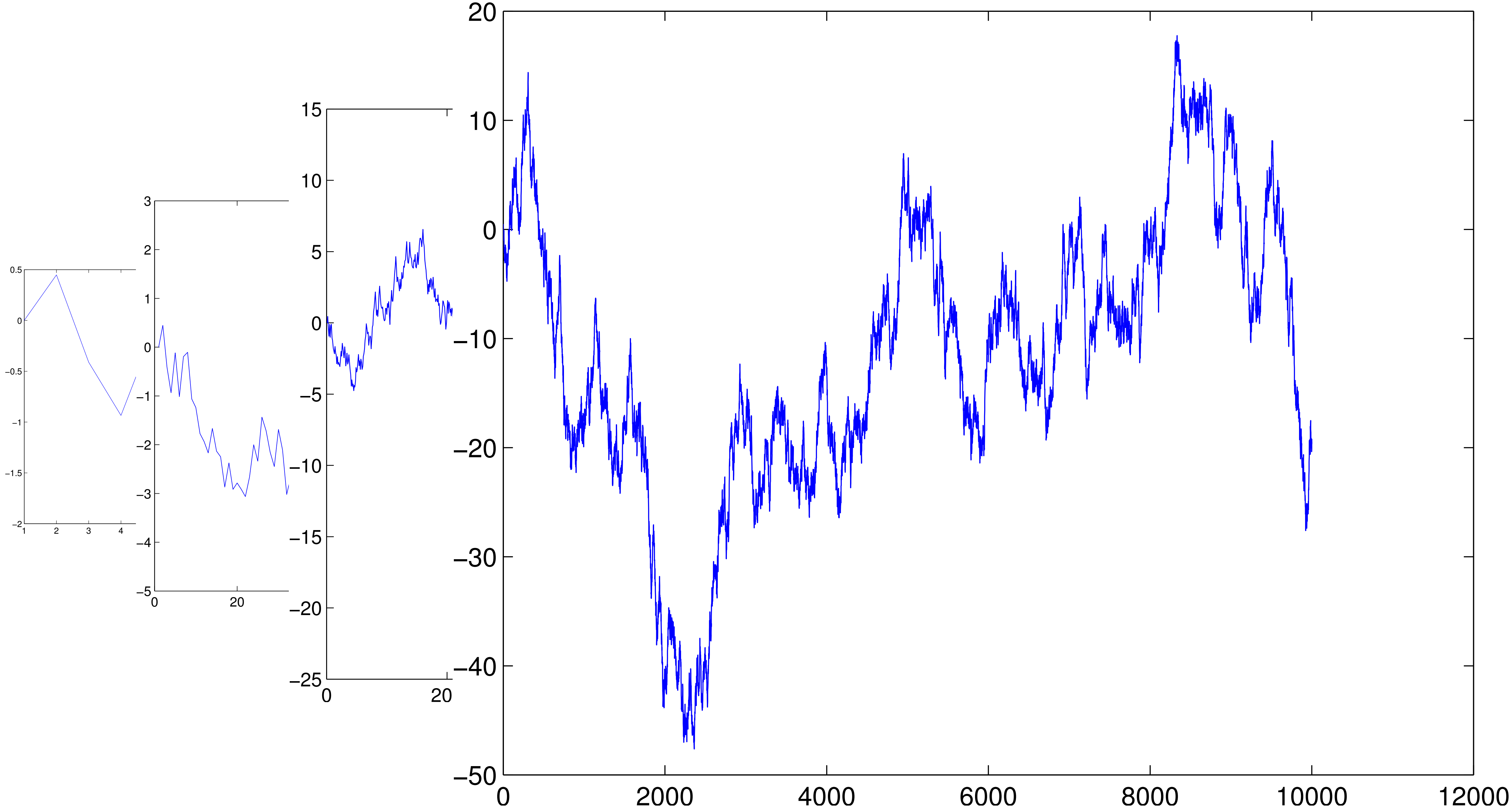
GCTATTGGCG

	$\to A$	$\to G$	$\to C$	$\to T$
A	$1-3\epsilon$	ϵ	ϵ	ϵ
G	ϵ	$1-3\epsilon$	ϵ	ϵ
C	ϵ	ϵ	$1-3\epsilon$	ϵ
T	ϵ	ϵ	ϵ	$1-3\epsilon$

- Mutation process operates independently at each position.
- Small total probability 3ε of a mutation happening at each generation.

Random Walk on **Z**

- Start at 0.
- Move up or down with probability 1/2.



Parameterization

$$X_0 \longrightarrow X_1 \longrightarrow X_2 \longrightarrow X_3 \longrightarrow$$

- Initial distribution:

$$\mathbb{P}(X_0 = i) = \lambda_i$$

- Transition probability matrix:

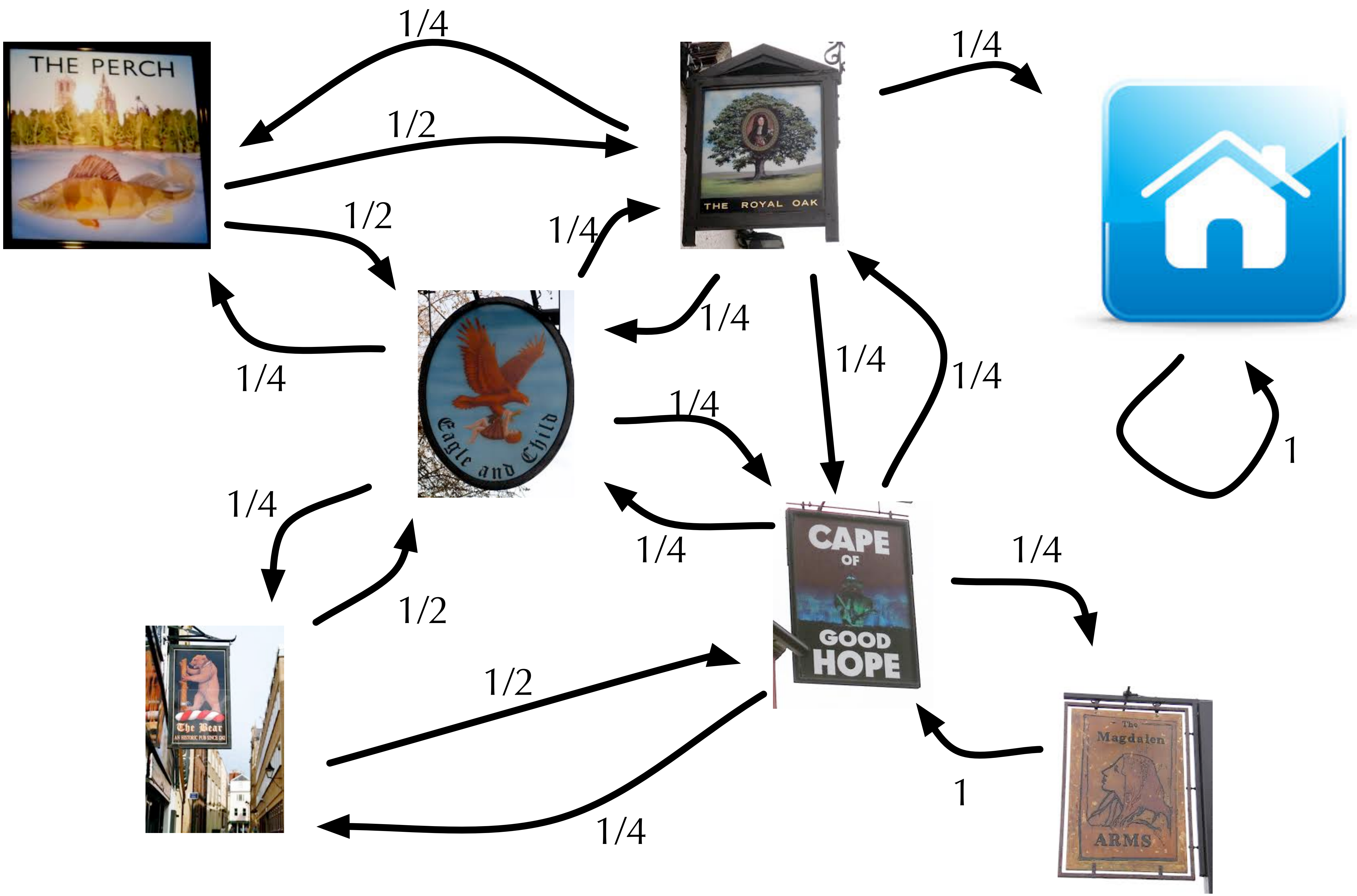
$$\mathbb{P}(X_t = j | X_{t-1} = i) = T_{ij}$$

  - Homogeneous Markov chains (transition probabilities do not depend on the time step)
  - Inhomogeneous Markov chains - transitions do depend on time step.

Transition probability matrix

- Transition probability matrix T has to have:
  - non-negative entries
  - rows that sum to 1
- Any such matrix is a transition probability matrix.

State Transition Diagrams



Simulating Random Pub Crawl (*)

- Write a programme to simulate from the random pub crawl. (From the "home" state allow probability 1/2 of going back to the Royal Oak).
- Starting from the Home state, run your programme 1000 times, each time simulating a Markov chain of length 100.
- Each simulation should be a random sequence of values $(s_1, s_2, s_3 \ldots, s_{100})$ where each s_i is a pub.
- Collect statistics of the number of times each state is visited at each time step $t = 1 \ldots 100$.
- How do the statistics defer if you started at Magdelen Arms?
- Does the distribution over states visited at step t converge for large t?
- Approximately how long does it take for the chain to "forget" whether it started at Home or at Magdelen Arms?

Useful Properties of Markov Chains

Chapman-Kolmogorov Equations

- We can calculate multi-step transition probabilities recursively:

$$\begin{aligned}\mathbb{P}(X_{t+2} = j | X_t = i) = &\sum_k \mathbb{P}(X_{t+2} = j | X_{t+1} = k)\mathbb{P}(X_{t+1} = k | X_t = i) \\ = &\sum_k T_{ik}T_{kj} \\ = &(T^2)_{ij}\end{aligned}$$

- Similarly:

$$\begin{aligned}P_{ij}^{(m)} := &\mathbb{P}(X_{t+m} = j | X_t = i) \\ = &\sum_k \mathbb{P}(X_{t+m} = j | X_{t+1} = k)\mathbb{P}(X_{t+1} = k | X_t = i) \\ = &\sum_k P_{ik}^{(m-1)} T_{kj} \\ = &(T^m)_{ij}\end{aligned}$$

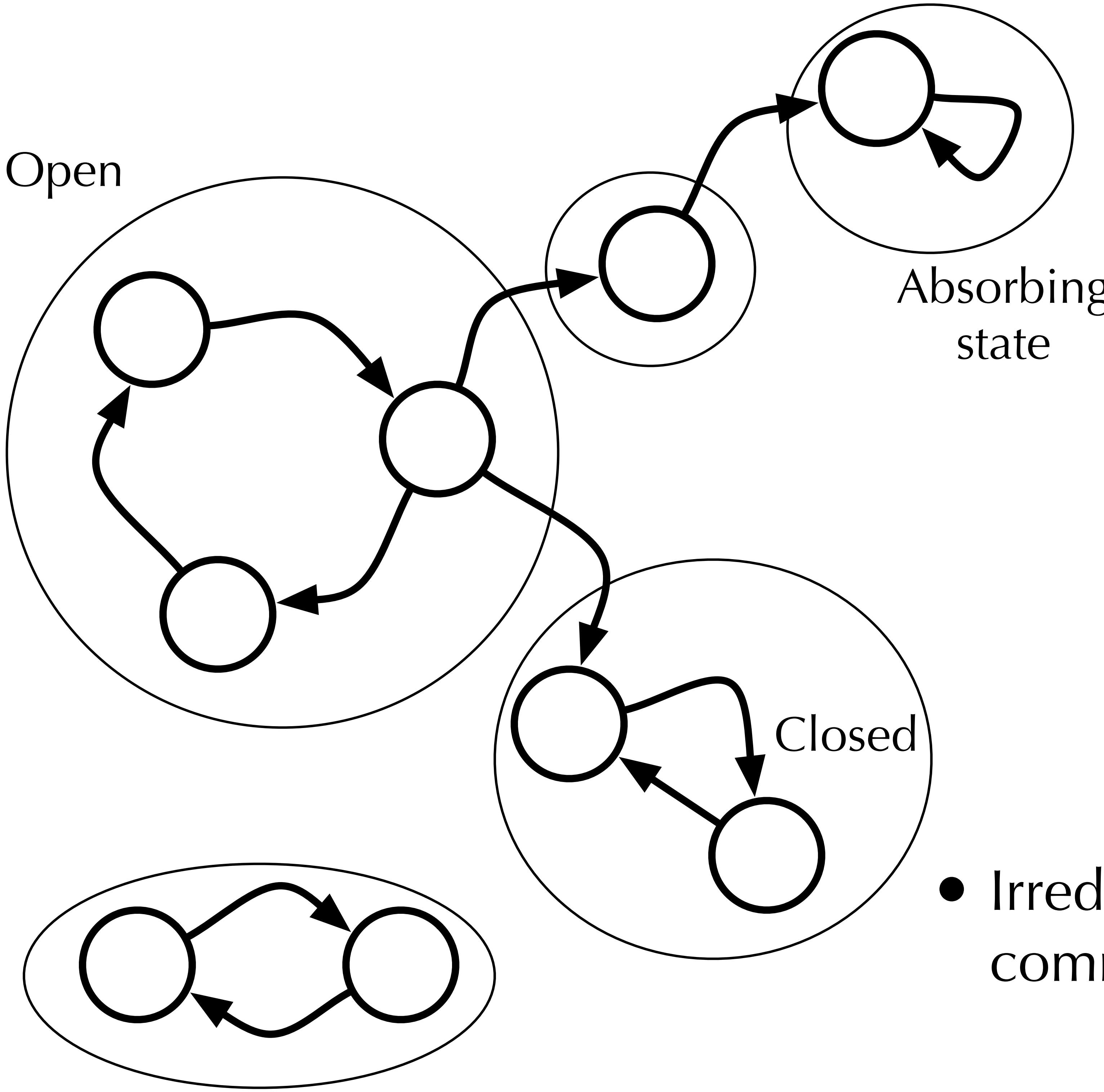
Marginal Distributions

- Similarly we can calculate the marginal probabilities of each X_i recursively:

$$
\begin{aligned}
P_i^{(t)} :=& \mathbb{P}(X_t = i) \\
=& \sum_k \mathbb{P}(X_{t-1} = k)\mathbb{P}(X_t = i | X_{t-1} = k) \\
=& \sum_k P_k^{(t-1)} T_{ki} \\
=& (\lambda T^t)_i
\end{aligned}
$$

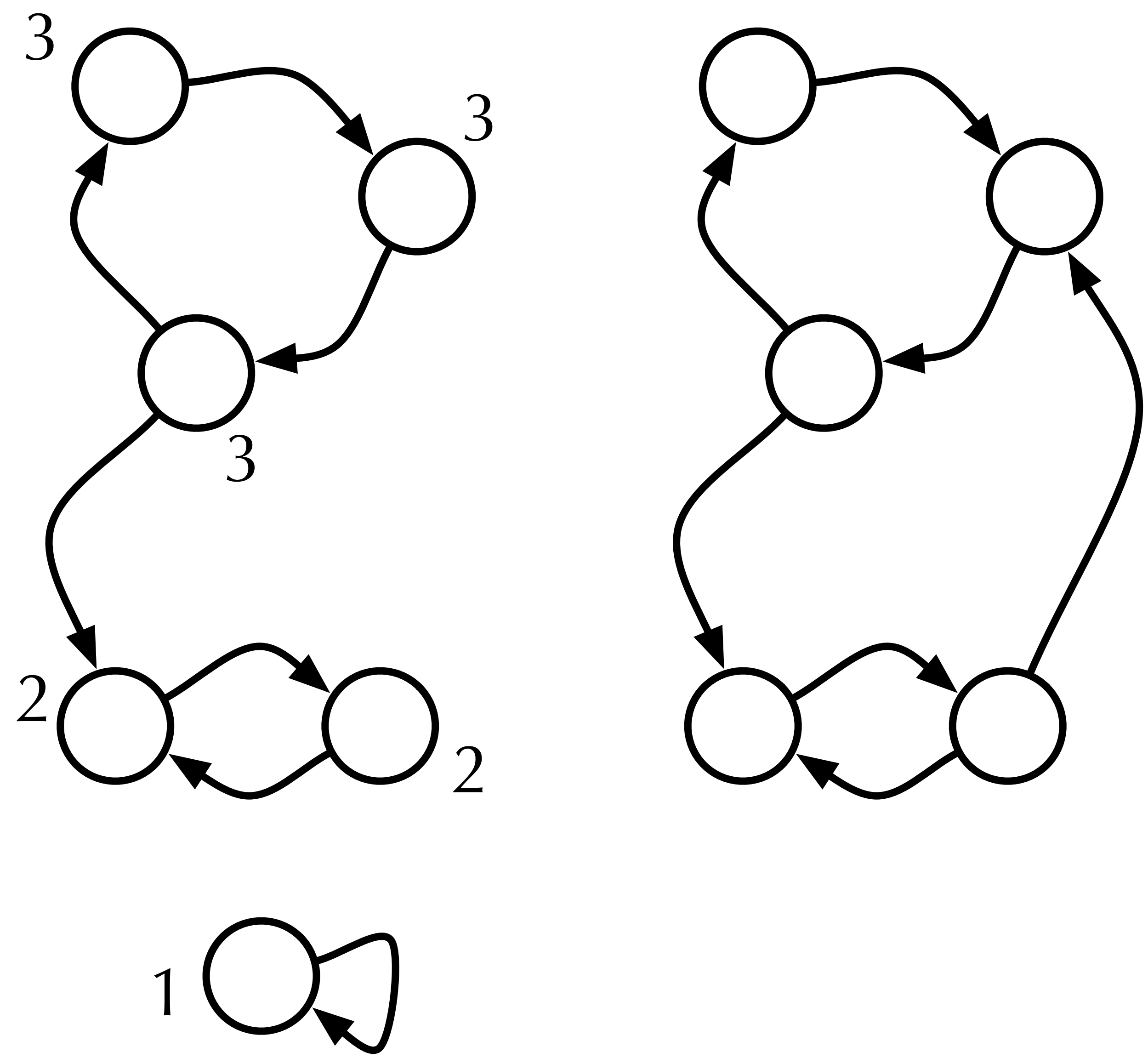
- where we take λ to be a row vector describing the initial probability distributions.

Communicating Classes



- Irreducible: One communicating class.

Periodicity



- Period of i:

$$gcd\{n : \mathbb{P}(\text{returning to } i \text{ from } i \text{ in } n \text{ steps}) > 0\}$$

- If a chain is irreducible, then all states have the same period.
- If the period is 1, then we say the chain is aperiodic.

Recurrence and Transience

- If we start at state *i*, what is the chance that we will return to *i*?
- Two possibilities:

  Case 1:

$$\mathbb{P}(\exists t > 0 : X_t = i | X_0 = i) = p < 1$$

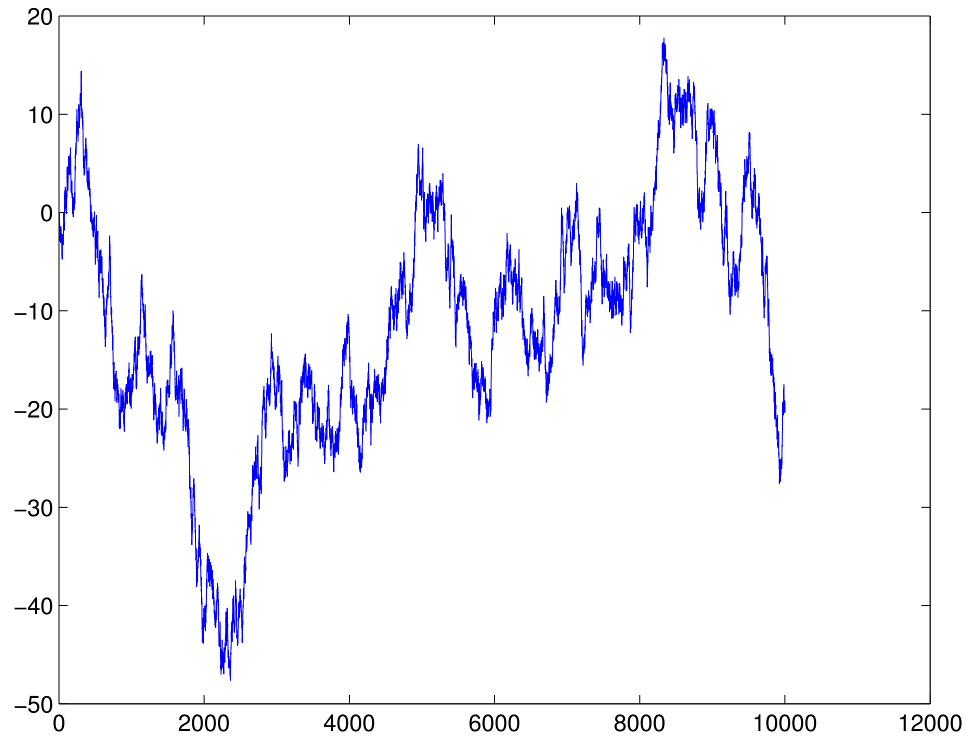
  - Total number of times we will encounter *i* in all future will be finite.
  - State *i* is **transient.**

  Case 2:

$$\mathbb{P}(\exists t > 0 : X_t = i | X_0 = i) = 1$$

  - We will return to *i* infinitely often.
  - State *i* is **recurrent**.
- A state *i* is recurrent if and only if $\sum_t P_{ii}^{(t)} = \infty$

Random Walk on **Z**



- Start at 0.
- Move up or down with probability 1/2.
- X_{2t} is the sum of 2t iid {+1,-1} variables.
- It equals 0 if there are exactly t +1's, and t -1's.
- This probability is

$$P_{00}^{(2t)} = \frac{(2t)!}{t!t!}\left(\frac{1}{2}\right)^{2t} \approx \frac{1}{\sqrt{\pi}\sqrt{t}}$$

  using Stirling's Formula:

$$n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$$

- This sums to infinity over t, so chain is recurrent.

Positive Recurrence and Null Recurrence

- Recurrence:
  - Chain will revisit a state infinitely often.
  - From state i we will return to i after a (random) finite number of steps.
- But the expected number of steps can be infinite!
  - This is called null recurrence.
- If expected number of steps is finite, this is called positive recurrent.
- Example: random walk on $\mathbf{Z}$.

UNIVERSITY OF OXFORD

Communicating Classes

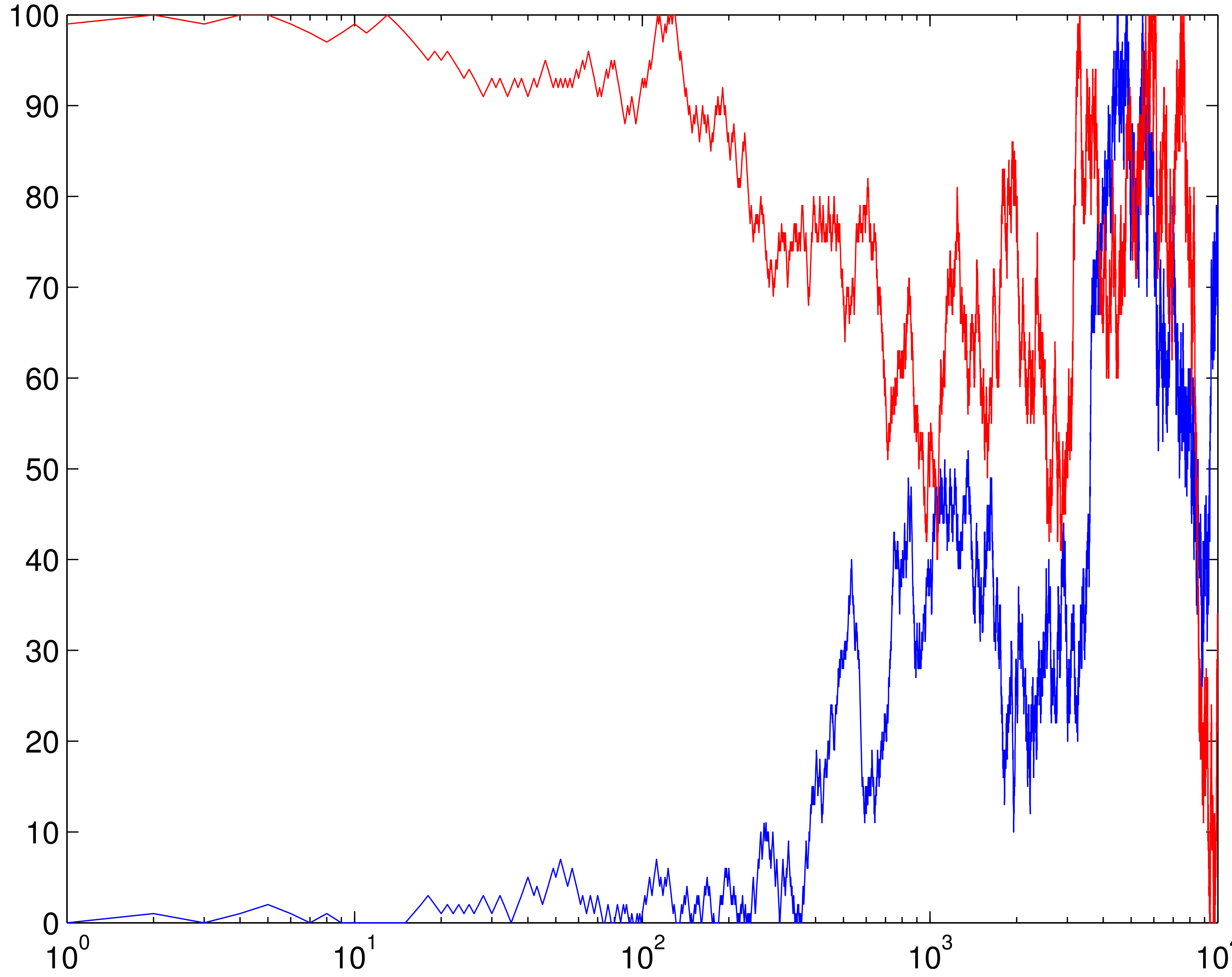
- Find the communicating classes and determine whether each class is open or closed, and the periodicity of the closed classes.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/4 & 0 & 3/4 & 0 \\ 0 & 0 & 1/3 & 0 & 2/3 \\ 1/4 & 1/2 & 0 & 1/4 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{pmatrix}$$

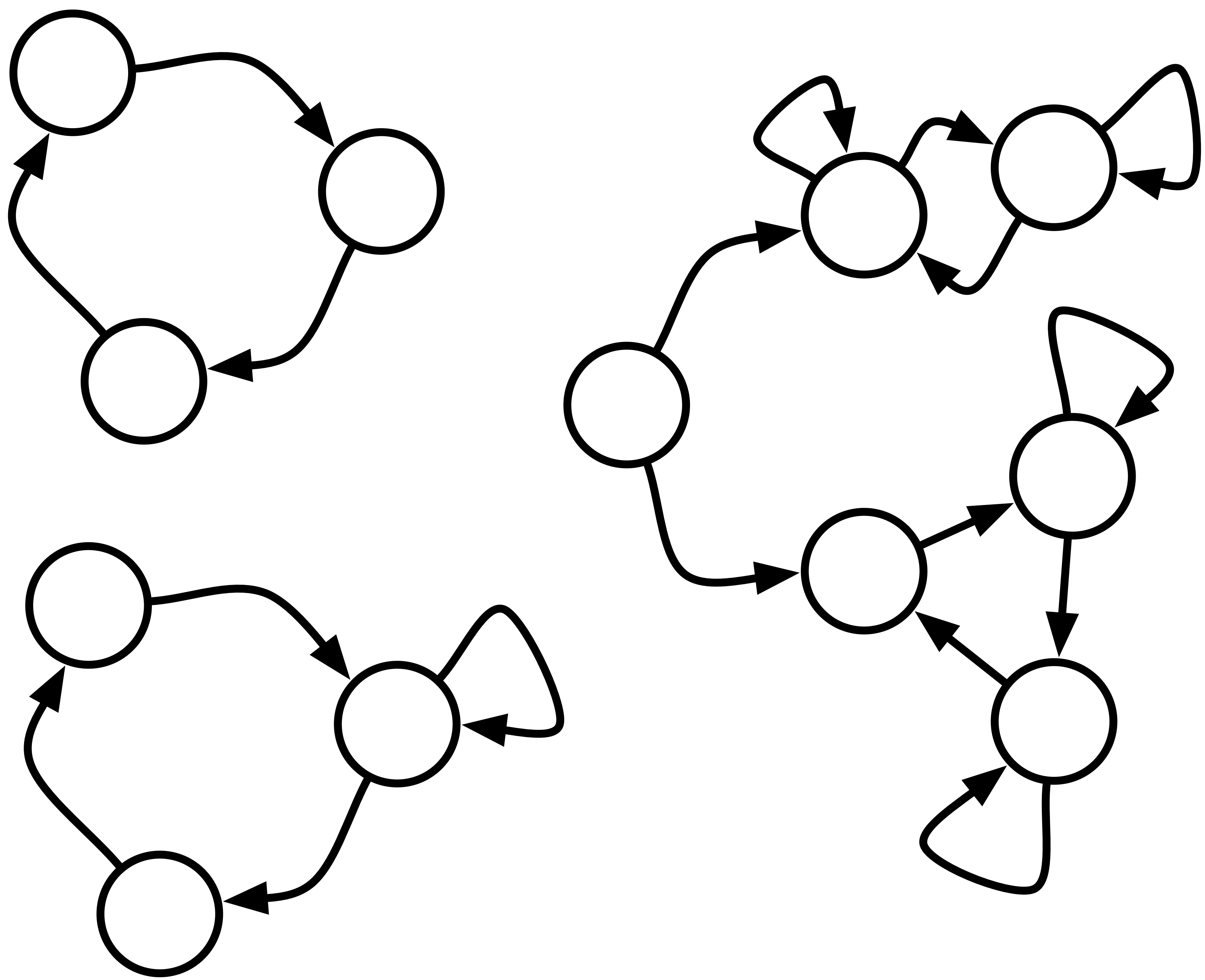
Convergence of Markov Chains

Do Markov Chains Forget?

- A Markov chain on {0,1,2,...,100}.
- At each step: move up or down by 1 at random, except at boundaries.
- Start at 0 and at 100.



Do Markov Chains Forget?



Stationary Distribution

- If a Markov chain "forgets" then for any two initial distributions/probability vectors λ and γ,

$$\lambda T^n \approx \gamma T^n \quad \text{for large } n$$

- In particular, there is a distribution/probability vector π such that

$$\lambda T^n \to \pi \quad \text{as } n \to \infty$$

- Taking $\lambda = \pi$, we see that

$$\pi T = \pi$$

- Such a distribution is called a stationary or equilibrium distribution.
    - When do Markov chains have stationary distributions?
    - When are stationary distributions unique?

Convergence Theorems

- A positive recurrent Markov chain T has a stationary distribution.
- If T is irreducible and has a stationary distribution, then it is unique and

$$\pi_i = \frac{1}{m_i}$$

  where m_i is the mean return time of state i.
- If T is irreducible, aperiodic and has stationary distribution π then

$$\mathbb{P}(X_n = i) \to \pi_i \quad \text{as } n \to \infty$$

- (Ergodic Theorem): If T is irreducible with stationary distribution π then

$$\frac{\#\{t \leq n : X_t = i\}}{n} \to \pi_i \quad \text{as } n \to \infty$$

Stationarity and Reversibility

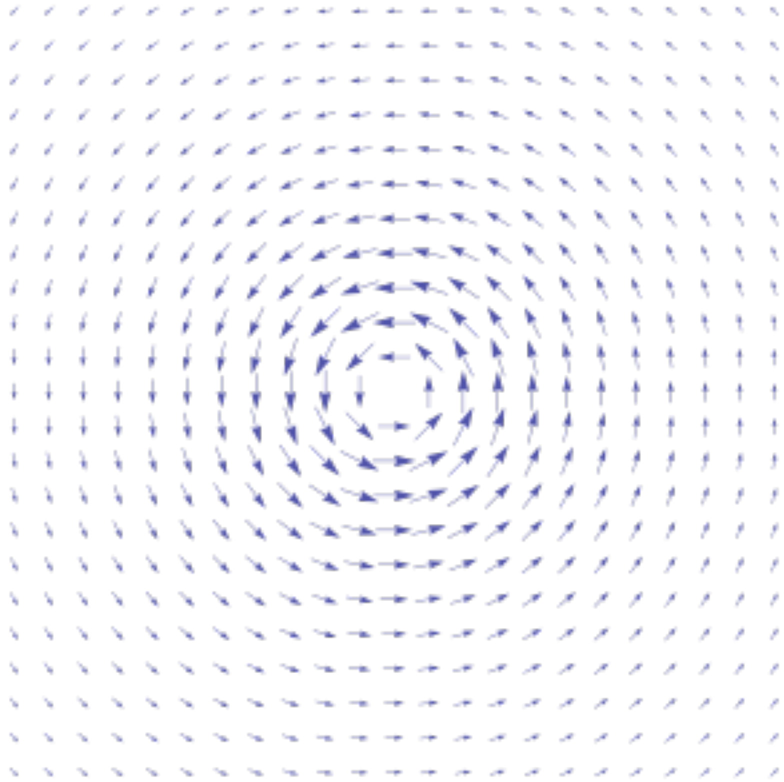
- Global balance: at a stationary distribution, the flow of probability mass into and out of each state has to be balanced:

$$\sum_{i=1}^{K} \pi_i T_{ij} = \pi_j = \sum_{k=1}^{K} \pi_j T_{jk}$$

- Detailed balance: the flow of probability mass between each pair of states is balanced:

$$\pi_i T_{ij} = \pi_j T_{ji}$$

- A Markov chain satisfying detailed balance is called reversible. Reversing the dynamics leads to the same chain.
- Detailed balance can be used to check that a distribution is the stationary distribution of a irreducible, periodic, reversible Markov chain.



Eigenvalue Decomposition

- The stationary distribution is a left eigenvector of T, with eigenvalue 1.

$$\pi T = \pi$$

- All eigenvalues of T have length ≤ 1. (Some eigenvalues can be complex valued).
- If there is another eigenvector with eigenvalue 1, then stationary distribution is not unique.

Random Walk

- Show that a random walk on a connected graph is reversible, and has stationary distribution π with π_i proportional to deg(i), the number of edges connected to i.
- What is the probability the drinker is at home at Monday 9am if he started the pub crawl on Friday?

Stationary Distributions

- Solve for the (possibly not unique) stationary distribution(s) of the following Markov chains.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/4 & 0 & 3/4 & 0 \\ 0 & 0 & 1/3 & 0 & 2/3 \\ 1/4 & 1/2 & 0 & 1/4 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \end{pmatrix}$$

Estimating Markov Chains

Maximum Likelihood Estimation

- Observe a sequence $x_0, x_1, x_2, x_3, \ldots x_t$.
- Likelihood of the sequence under the Markov chain model is:

$$\mathcal{L}(\lambda, T) = \lambda_{x_0} \prod_{s=1}^{t} T_{x_{s-1} x_s} = \lambda_{x_0} \prod_{i=1}^{K} \prod_{j=1}^{K} T_{ij}^{N_{ij}}$$

where N_{ij} is the number of observed transitions i -> j.

- We can solve for the maximum likelihood estimator:

$$T_{ij} = \frac{N_{ij}}{\sum_{k=1}^{K} N_{ik}}$$

UNIVERSITY OF OXFORD

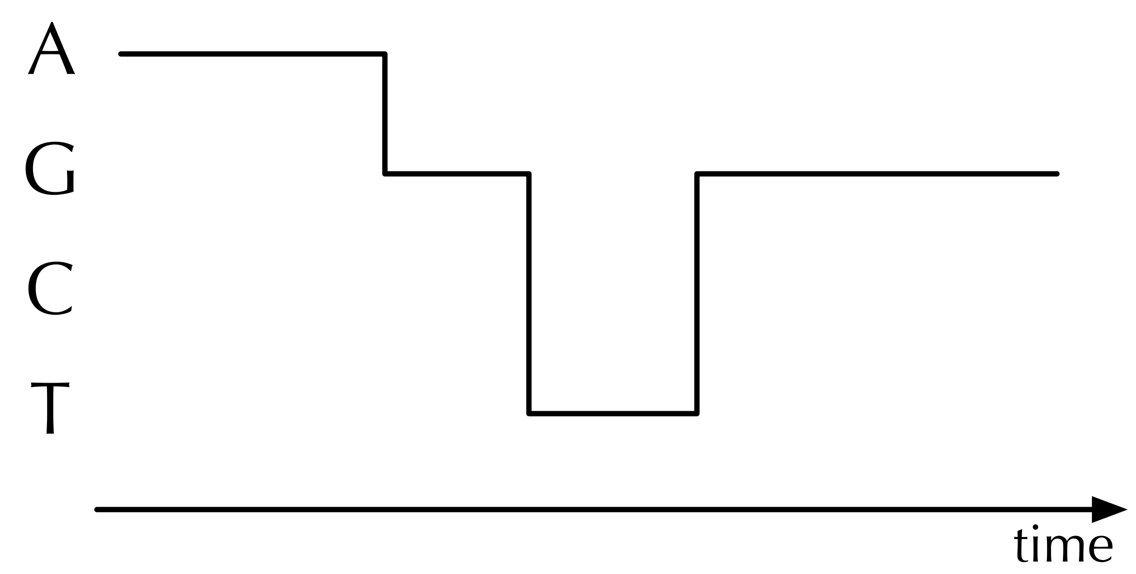
Markov Model of English Text (*)

- Download a large piece of English text, say "War and Peace" from Project Gutenberg.
- We will model the text as a sequence of characters.
- Write a programme to compute the ML estimate for the transition probability matrix.
- You can use the file markov_text.R or markov_text.m to help convert from text to the sequence of states needed. There are $K = 96$ states and the two functions are text2states and states2text.
- Generate a string of length 200 using your ML estimate.
- Does it look sensible?

Continuous-Time Markov Chains

Jukes-Cantor DNA Evolution

	$\to A$	$\to G$	$\to C$	$\to T$
A	$1-3\epsilon$	ϵ	ϵ	ϵ
G	ϵ	$1-3\epsilon$	ϵ	ϵ
C	ϵ	ϵ	$1-3\epsilon$	ϵ
T	ϵ	ϵ	ϵ	$1-3\epsilon$



- Probability of mutation is O(ε) per generation.
- mutations will appear at rate of once every O(1/ε) generations.
- Measuring time in units of 1/ε leads to a continuous-time Markov chain.
- In each time step of length ε, total probability of a mutation is 3ε.

$$P = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix} = I + \epsilon \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}$$

Continuous-time Markov Chains

- A collection of random variables $(X_t)_{t\geq 0}$.
- An initial distribution λ and a transition rate matrix R.
- Suppose $X_t = i$. Then in the next ε time,

$$\mathbb{P}(X_{t+\epsilon} = j | X_t = i) = I_{ij} + \epsilon R_{ij}$$

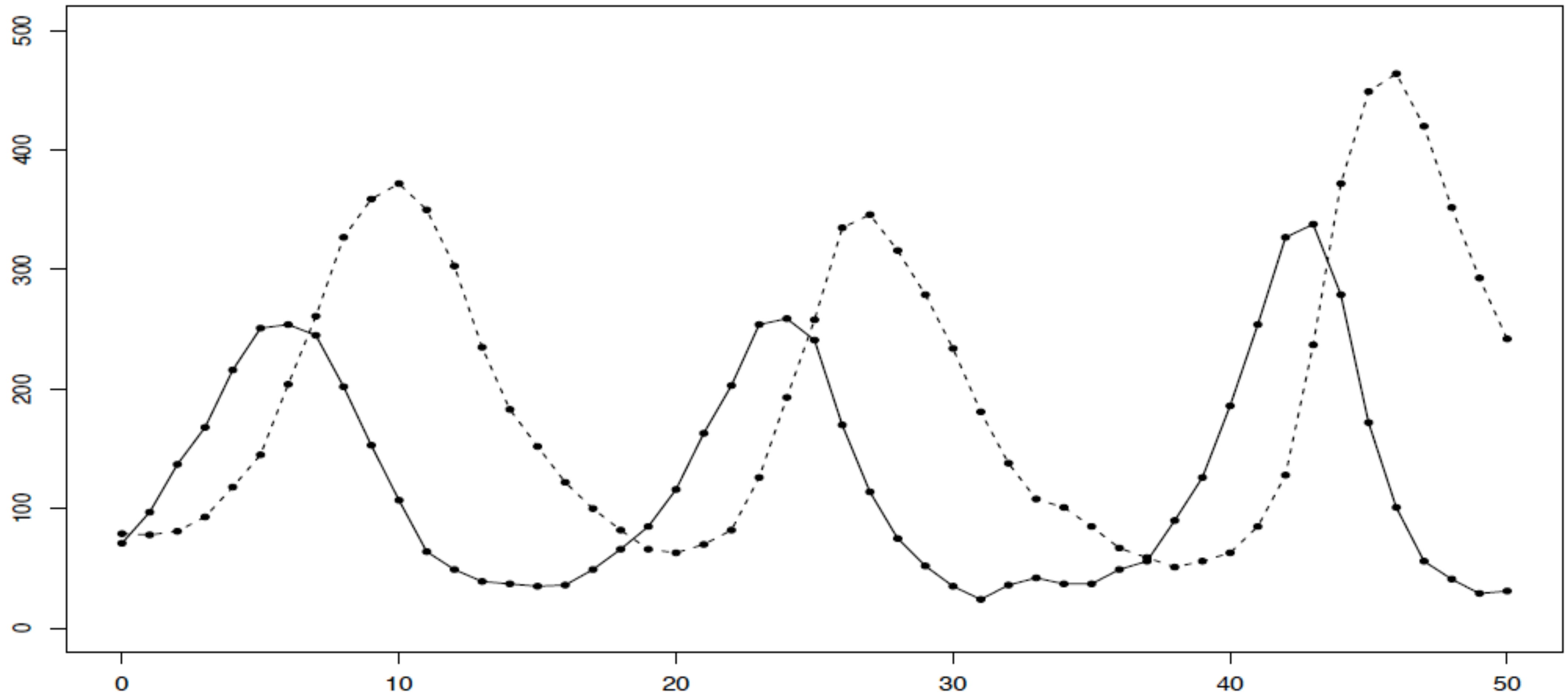
$$\begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}$$

  - Rows of R sum to 0.
  - Off-diagonal entries are non-negative.
  - On-diagonal entries are negative of sum of off-diagonal ones.

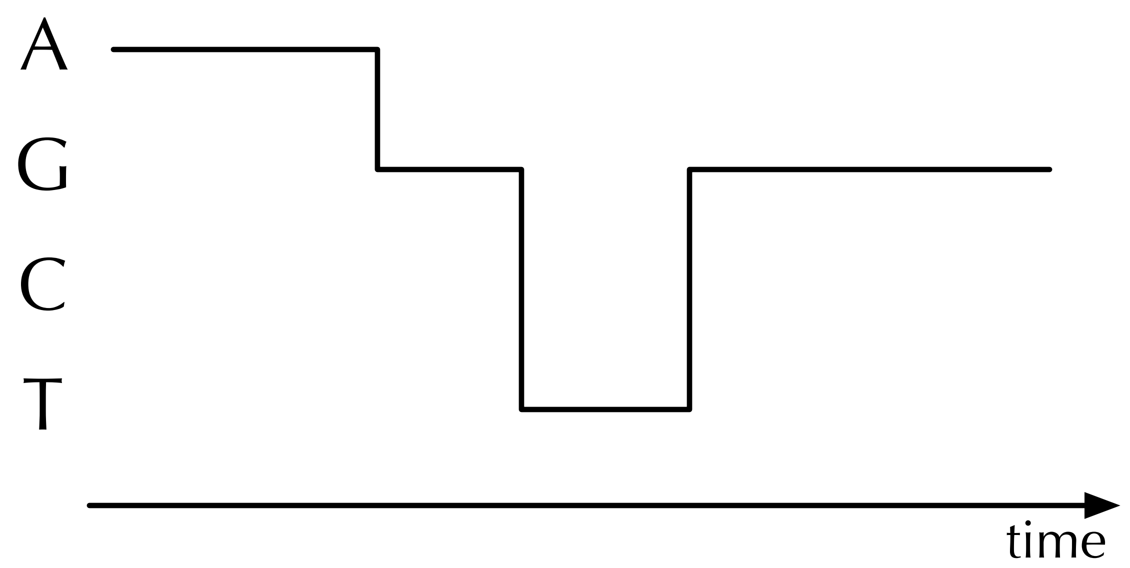
Lotka-Volterra Process (Predator-Prey)

- A continuous-time Markov chain over $\mathbf{N}^2$, number of predators and preys in an ecosystem.

$$R(\{x,y\} \to \{x+1,y\}) = \alpha x \qquad R(\{x,y\} \to \{x-1,y\}) = \beta xy$$
$$R(\{x,y\} \to \{x,y+1\}) = \delta xy \qquad R(\{x,y\} \to \{x,y-1\}) = \gamma y$$



Gillespie's Algorithm



- Start by sampling X_0 from initial distribution λ.
- When in state i, wait in state for an amount of time distributed as

$$\mathrm{Exp}(|R_{ii}|)$$

- At end of waiting time, transition to a different state $j \neq i$ with probability

$$\frac{(R_{i1}, \ldots, R_{ij-1}, 0, R_{ij+1}, \ldots, R_{iK})}{|R_{ii}|}$$

Chapman-Kolmogorov Equations

- Denote $P(t)$ the transition probability matrix over time interval t.
- Transition probabilities can be computed using matrix exponentiation:

$$\begin{aligned}
P(t)_{ij} :=& \mathbb{P}(X_t = j | X_0 = i) \\
=& \mathbb{P}(X_{tn\frac{1}{n}} = j | X_0 = i) \\
\approx& ((I + \tfrac{1}{n}R)^{tn})_{ij} \to \exp(tR)_{ij}
\end{aligned}$$

- Composition of transition probability matrices:

$$P(t+s) = \exp((t+s)R) = \exp(tR)\exp(sR) = P(t)P(s)$$

- Forward/backward equations:

$$\begin{aligned}
P(t+\epsilon) &= P(t)(I + \epsilon R) \\
\frac{\partial P(t)}{\partial t} &\approx \frac{P(t+\epsilon) - P(t)}{\epsilon} \to P(t)R = RP(t)
\end{aligned}$$

Convergence to Stationary Distribution

- Suppose we have an
  - irreducible,
  - aperiodic and
  - positive recurrent

  continuous-time Markov chain with rate matrix *R*.
- Then it has a unique stationary distribution π which it converges to:

$$\mathbb{P}(X_t = i) \to \pi_i \qquad \text{as } t \to \infty$$

$$\frac{1}{T}\int_0^T \mathbf{1}(X_t = i) \to \pi_i \qquad \text{as } t \to \infty$$

Reversibility and Detailed Balance

- If a Markov chain with rate matrix *R* has reached its stationary distribution π, then flow of probability mass into and out of states is balanced.
- Global balance:

$$\begin{aligned} \sum_{i=1}^{K} \pi_i R_{ij} = &\quad 0 &= \sum_{k=1}^{K} \pi_j R_{jk} \\ \sum_{i \neq j} \pi_i R_{ij} = &\quad \pi_j |R_{jj}| &= \sum_{k \neq j} \pi_j R_{jk} \end{aligned}$$

- Detailed balance for reversible chains:

$$\pi_i R_{ij} = \pi_j R_{ji}$$

Kimura 80 Model

- Rate matrix:

$$\begin{array}{c|cccc} & \to A & \to G & \to C & \to T \\ \hline A & -\kappa-2 & \kappa & 1 & 1 \\ G & \kappa & -\kappa-2 & 1 & 1 \\ C & 1 & 1 & -\kappa-2 & \kappa \\ T & 1 & 1 & \kappa & -\kappa-2 \end{array}$$

- Distinguish between transitions A↔G (purine) and C↔T (pyrimidine) and transversions.

- Practical: show that the stationary distribution of K80 model is uniform over {A,G,C,T}.

Felsenstein 81 Model

- Rate matrix:

	$\to A$	$\to G$	$\to C$	$\to T$
A	$-\pi_G - \pi_C - \pi_T$	π_G	π_C	π_T
G	π_A	$-\pi_A - \pi_C - \pi_T$	π_C	π_T
C	π_A	π_G	$-\pi_A - \pi_G - \pi_T$	π_T
T	π_A	π_G	π_C	$-\pi_A - \pi_G \pi_C$

- Practical: Find the stationary distribution of the F81 model.

Predator-Prey Model (*)

- Use Gillespie's algorithm to simulate from the predator-prey model:

$$R(\{x,y\} \to \{x+1,y\}) = \alpha x \qquad R(\{x,y\} \to \{x-1,y\}) = \beta xy$$
$$R(\{x,y\} \to \{x,y+1\}) = \delta xy \qquad R(\{x,y\} \to \{x,y-1\}) = \gamma y$$

- You can represent the continuous-time trajectory using a sequence of pairs (t_0,s_0), (t_1,s_1),... (t_A,s_A), where
  - $t_0 = 0$, s_0 is the initial state at time 0,
  - each subsequent pair (t_a,s_a) is the next state and the time the chain jumps to the state.
- How is the dynamics of the model affected by the parameters α, β, γ, δ?

Summary

- Definition of Markov Chains
- Properties of Markov Chains
    - Communication classes
    - Recurrence and Transience
    - Periodicity
- Convergence of Markov Chains : Stationary Distribution
- Estimation of Markov Chains
- Continuous Time Markov Chains