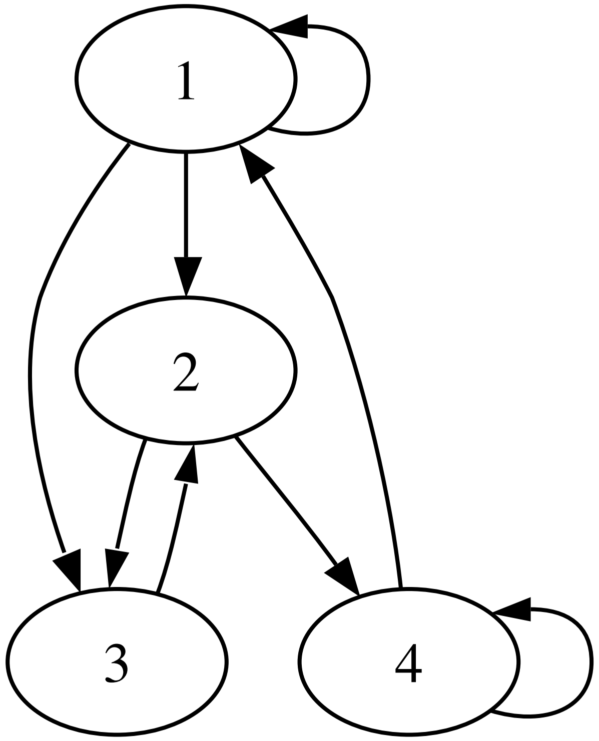
CS109/Stat121/AC209/E-109 Data Science

Markov Chain Monte Carlo

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu

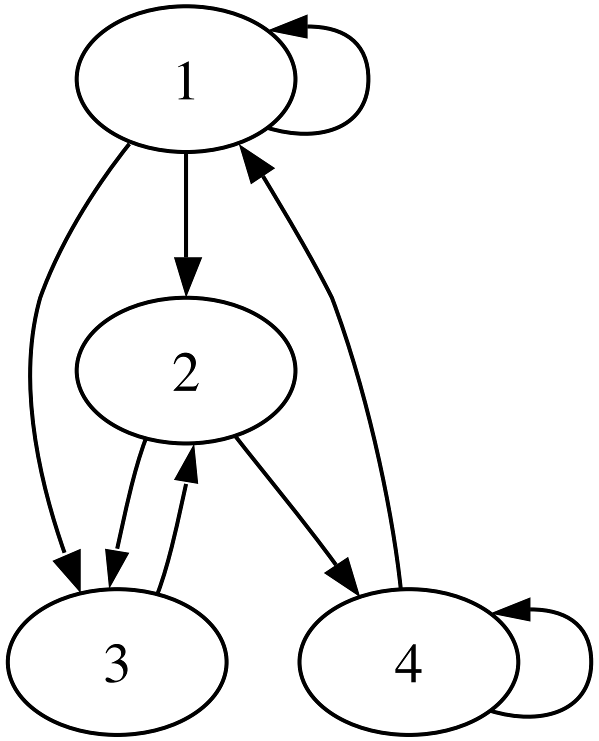


This Week

- HW3 due next Thursday (Oct 17) at 11:59 pm – start now!
- Friday lab **10-11:30 am** in MD G115

What is a Markov Chain?

Chain with 4 states:

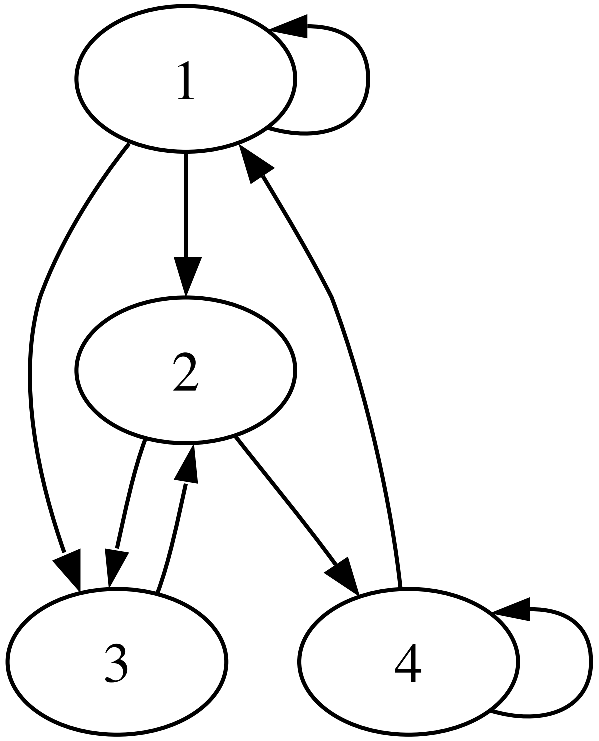


Transition matrix, if at each stage an arrow is followed uniformly at random:

$$Q = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 \end{pmatrix}$$

Definition of Markov Chain

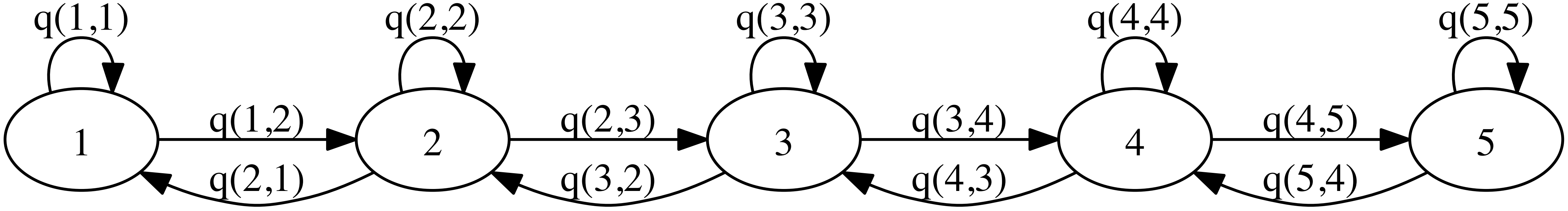
Chain with 4 states:



Definition 1. A sequence of random variables $X_0, X_1, X_2, \ldots$ taking values in the *state space* $\{1, \ldots, M\}$ is called a *Markov chain* if there is an M by M matrix $Q = (q_{ij})$ such that for any $n \geq 0$,

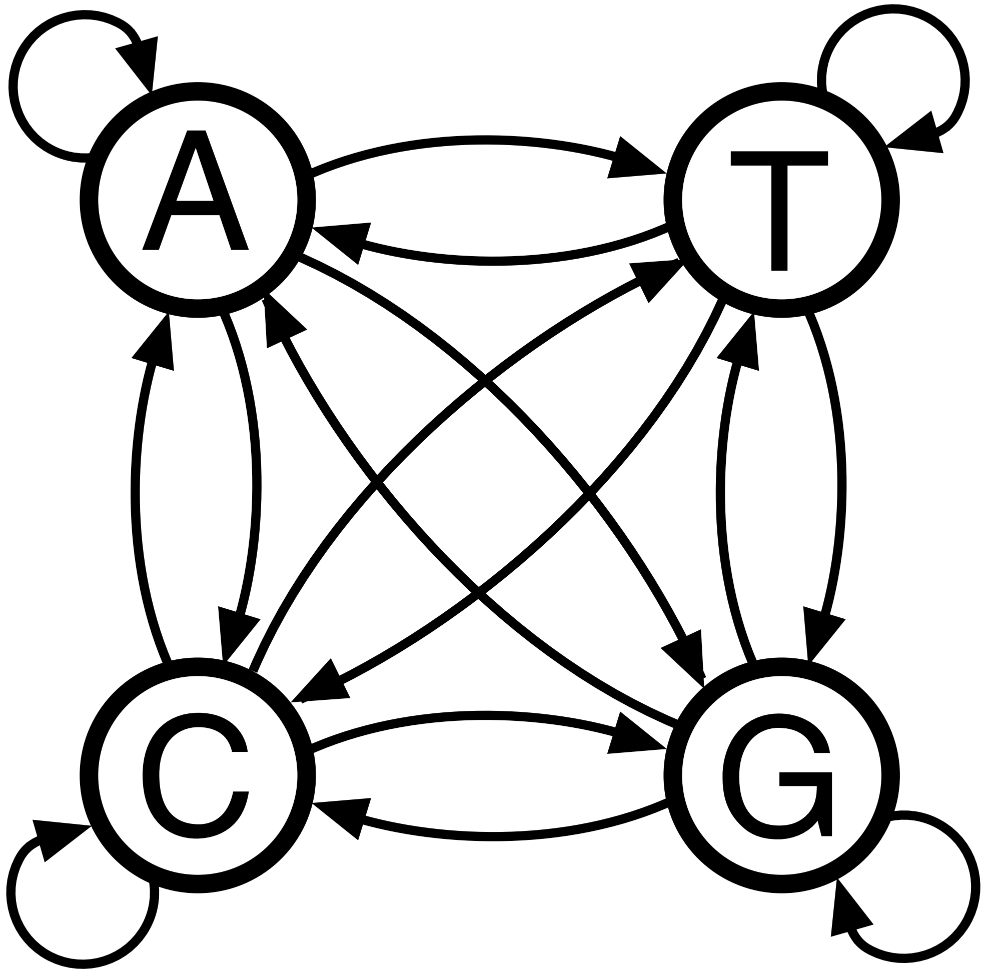
$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i) = q_{ij}.$$

Example: Birth-Death Chain



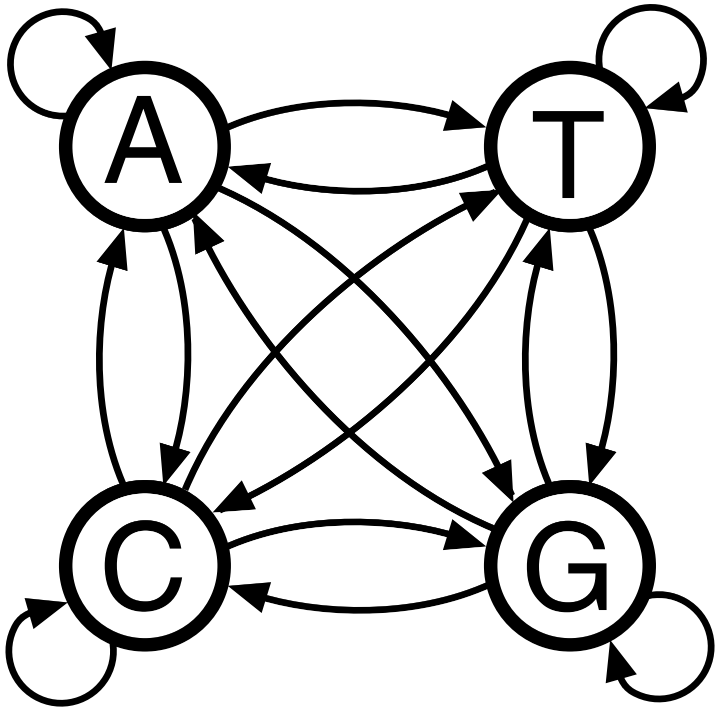
From j, can only go to j-1 or j+1, or stay at j (at boundaries, only 2 of these are possible).

Application: DNA Sequence Analysis, CpG Islands



source: Durbin et al, *Biological Sequence Analysis*

Application: DNA Sequence Analysis, CpG Islands



In C-G dinucleotides, the C often mutates to a T due to methylation. In a “CpG island” the methylation is suppressed.

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

—	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

source: Durbin et al

Application: DNA Sequence Analysis, CpG Islands

Now can use *likelihood ratios* to decide whether a given sequence was from a CpG island or the rest of the "ocean"!

Log-likelihood ratio: $\log \frac{P(x|\text{model}\,+)}{P(x|\text{model}\,-)}$

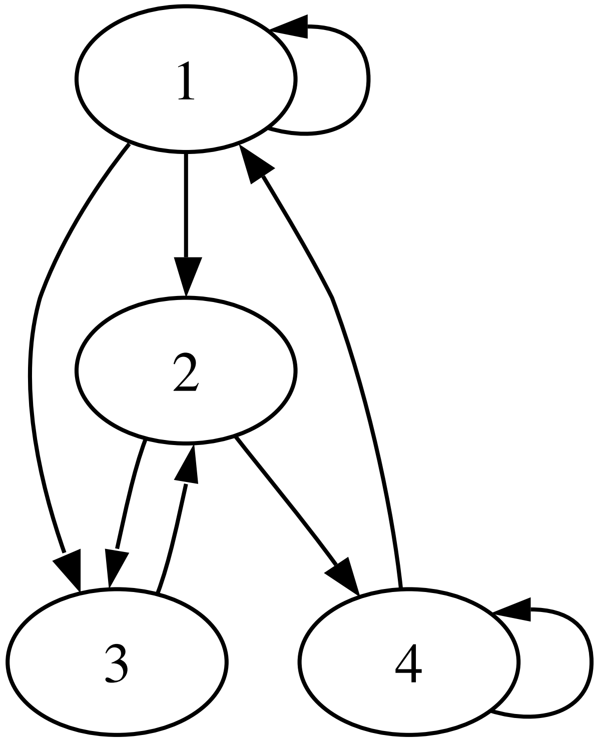
+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

source: Durbin et al

Google PageRank (Page-Brin)

Imagine someone randomly surfing the web...



How "important" a page is depends not only on how many other pages link to it, but also on how important *those* pages are!

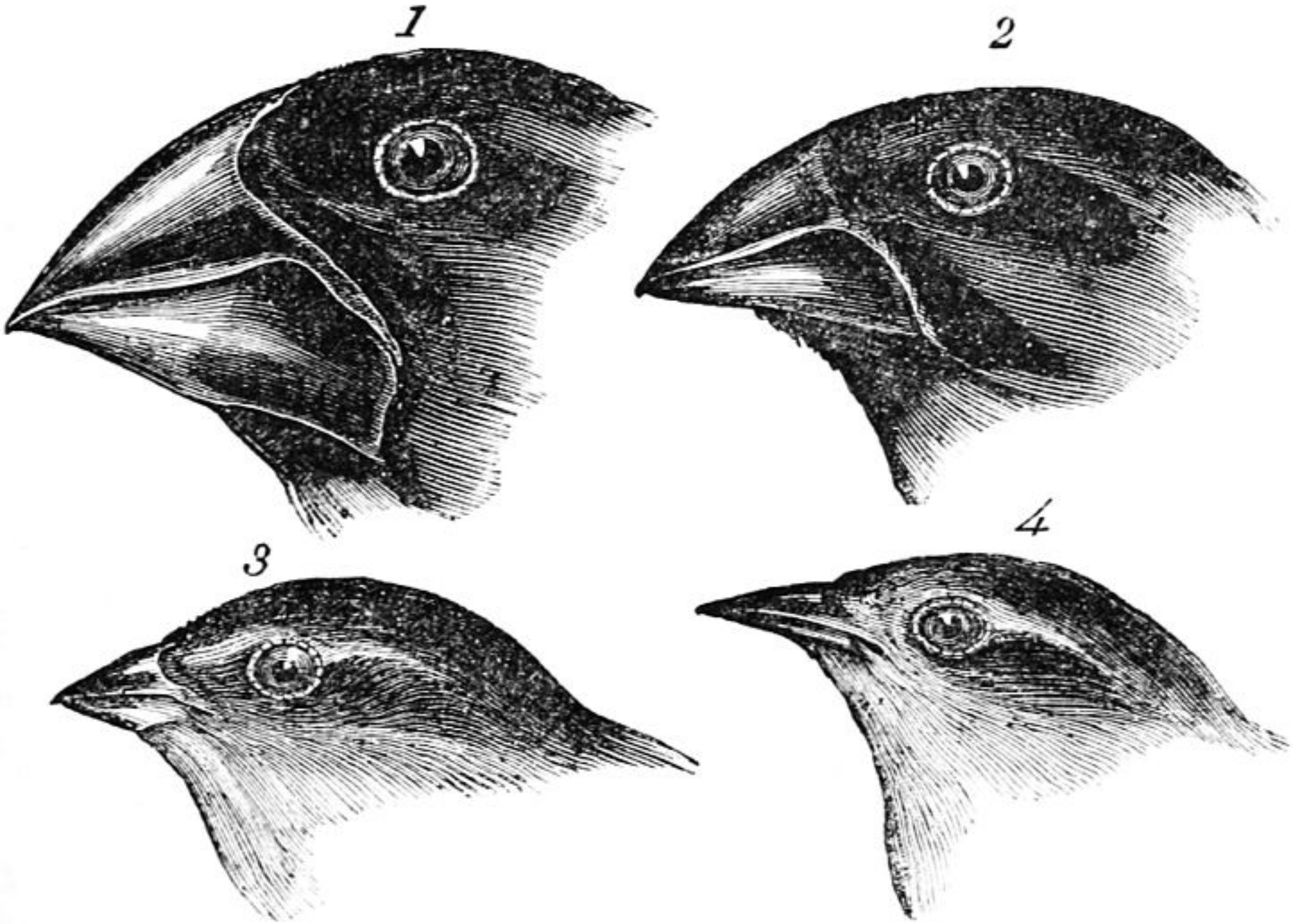
Markov Chain Monte Carlo (MCMC)

Key idea: simulate complicated distributions and approximate hard-to-compute averages by designing and running a Markov chain!

The chain is constructed so that its *stationary distribution* (the distribution it converges to in the long run) is the desired distribution.

But why isn’t that at least as hard as the original problem?

Darwin’s Finches



1. Geospiza magnirostris.
3. Geospiza parvula.

2. Geospiza fortis.
4. Certhidea olivacea.

http://en.wikipedia.org/wiki/Darwin's_finches

Darwin’s Finches

“Seeing this gradation and diversity of structure in one small, intimately related group of birds, one might really fancy that from an original paucity of birds in this archipelago, one species had been taken and modified for different ends.” – Charles Darwin

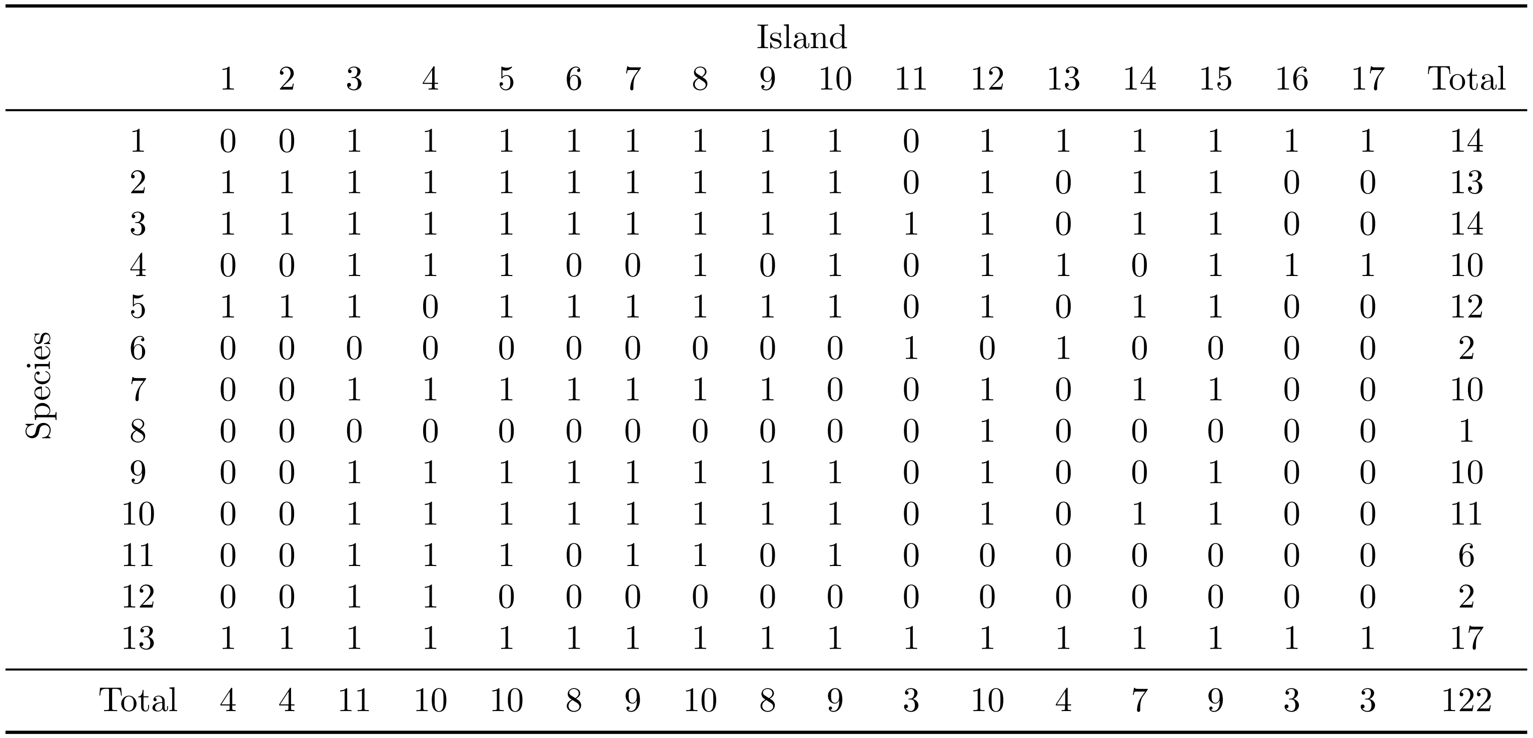
Darwin's Finches and Binary Tables

		Island																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
Species	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	14
	2	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	13
	3	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	14
	4	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1	10
	5	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0	12
	6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
	7	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0	10
	8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	9	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	10
	10	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	11
	11	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	6
	12	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
	Total	4	4	11	10	10	8	9	10	8	9	3	10	4	7	9	3	3	122

data from Sanderson (2000)

Jared Diamond defined a *checkerboard* as a pair of species that never co-occur on an island. Here there are 10 checkerboards out of 78 possible. Is that a lot or a little?

Darwin’s Finches: A Monte Carlo Algorithm



		Island																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
Species	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	14
	2	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	13
	3	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	14
	4	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1	10
	5	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0	12
	6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
	7	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0	10
	8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
	9	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	10
	10	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	11
	11	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	6
	12	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
	13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
	Total	4	4	11	10	10	8	9	10	8	9	3	10	4	7	9	3	3	122

One “move” of a Markov chain that preserves row and column sums:

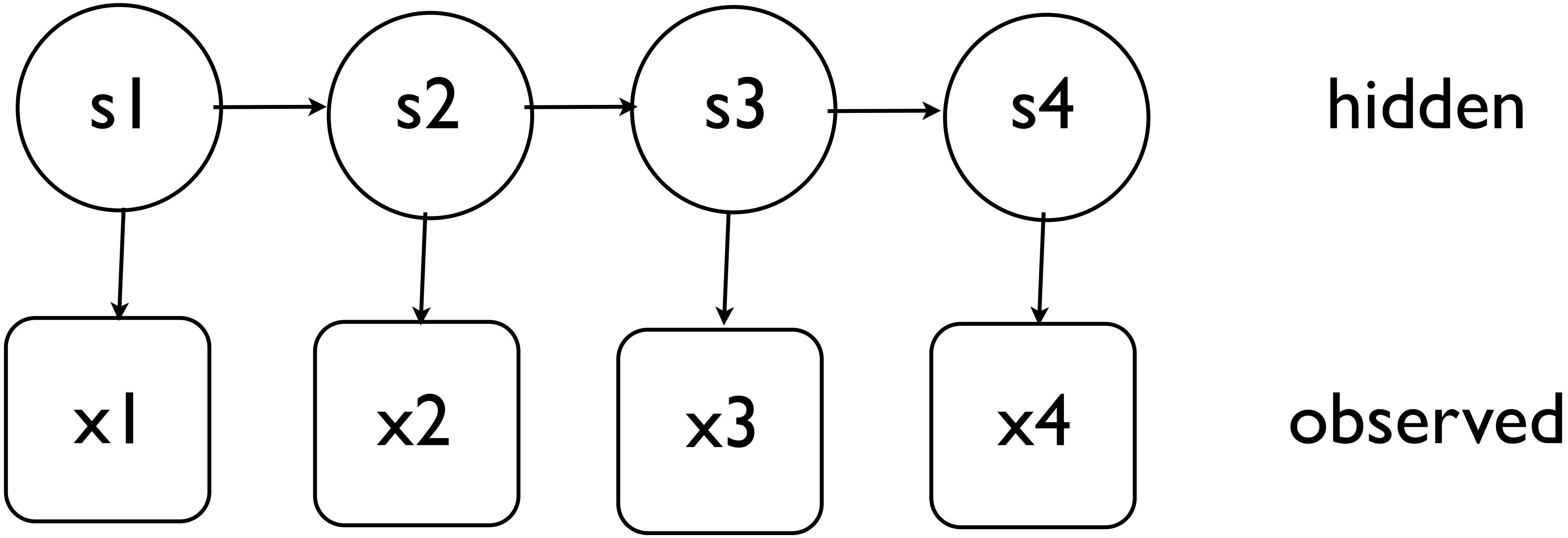
1. Pick 2 random rows, 2 random columns.
2. If submatrix is $\begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$ or $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

then swap between them;
else stay at the current state.

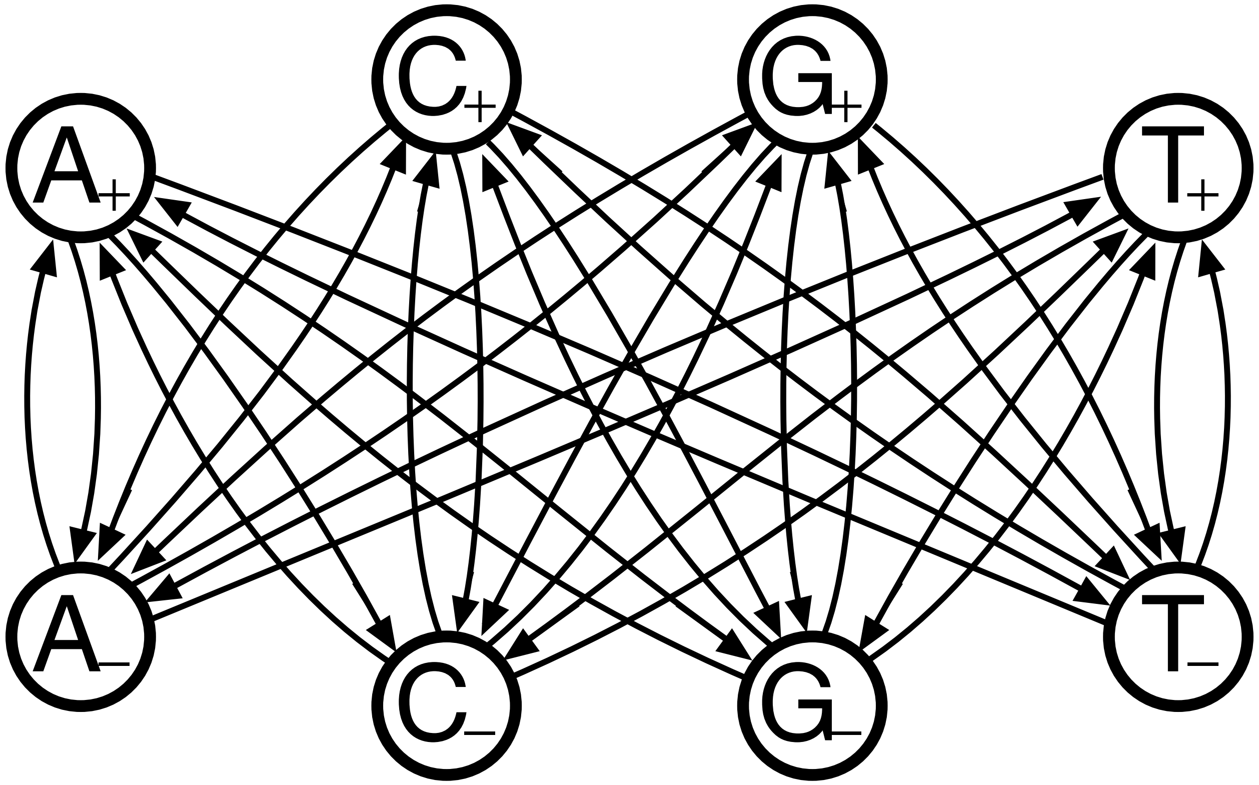
Hidden Markov Models (HMM)

Many applications in biology (e.g., identifying special regions in a long sequence, sequence alignment for two sequences), in speech recognition, and elsewhere.

Assume there is an underlying Markov chain running but that we can't directly observe the sequence of states. Instead, we observe "emissions" released just after each step in the chain.



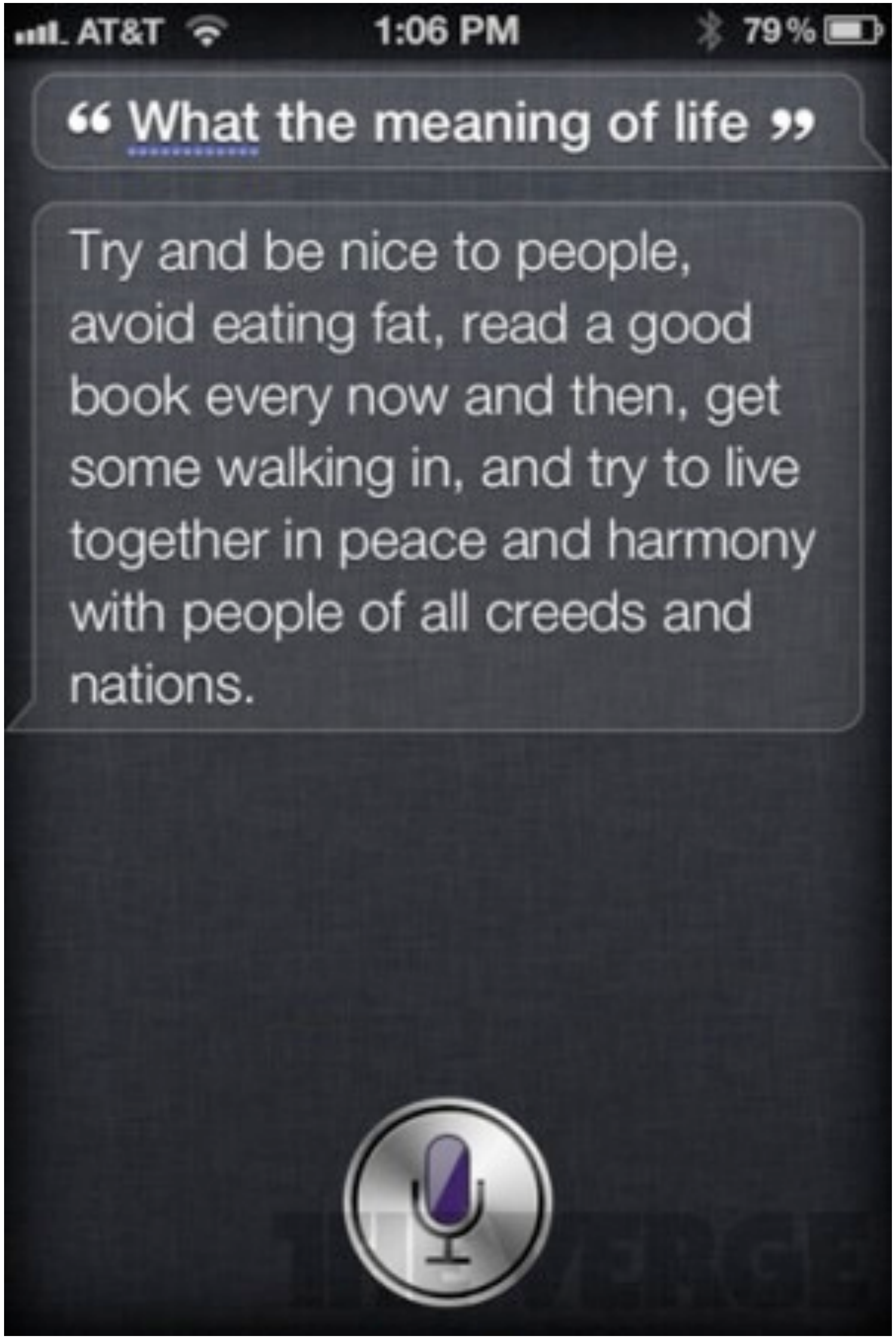
HMM Application: CpG Islands



source: Durbin et al

The challenging part is that we can only observe the sequence of A's, C's, T's, G's, not whether the state was an island (+) or non-island (-).

HMM Application: Speech Recognition

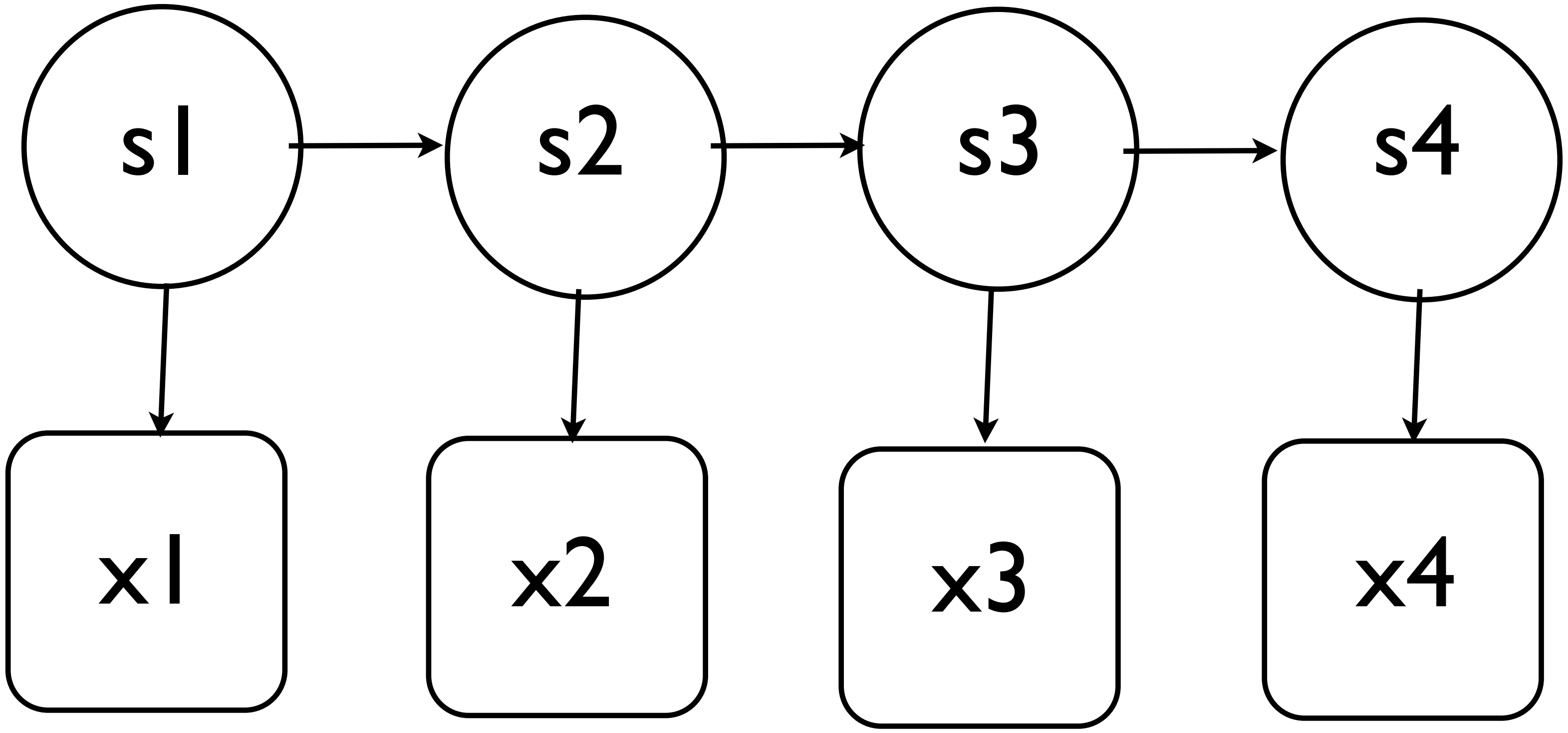


“What the meaning of life”

Try and be nice to people, avoid eating fat, read a good book every now and then, get some walking in, and try to live together in peace and harmony with people of all creeds and nations.

Three Fundamental Questions for HMMs

1. Find p(x), the probability of the observed sequence.
2. Find the most likely state sequence s, given the data x.
3. Estimate the model parameters (transition and emission probabilities), given the data x.



Three Fundamental Questions for HMMs

1. Find p(x), the probability of the observed sequence.
2. Find the most likely state sequence s, given the data x.
3. Estimate the model parameters (transition and emission probabilities), given the data x.

Methods:

1. Forward algorithm, backward algorithm (dynamic programming, recursive)
2. Viterbi algorithm (dynamic programming, recursive)
3. Baum-Welch algorithm (a form of the EM algorithm [Dempster-Laird-Rubin])

Finding the probability p(x)

Naive method: $$p(x) = \sum_s p(x, s) = \sum_s p(x|s)p(s)$$

Each term is easy to compute (assuming the transition and emission probabilities are known).

But note how many terms there are... For 10 possible states and 100 observations, the number of terms is 10^{100} – intractable even on the fastest supercomputer!

Finding the probability p(x): forward algorithm

Let $f_n(k) = p(x_1, \ldots, x_n, s_n = k)$

and compute these recursively!

$$f_{n+1}(l) = e_l(x_{n+1}) \sum_k f_n(k) q_{kl}$$

where the e's are emission probabilities and the q's are transition probabilities.

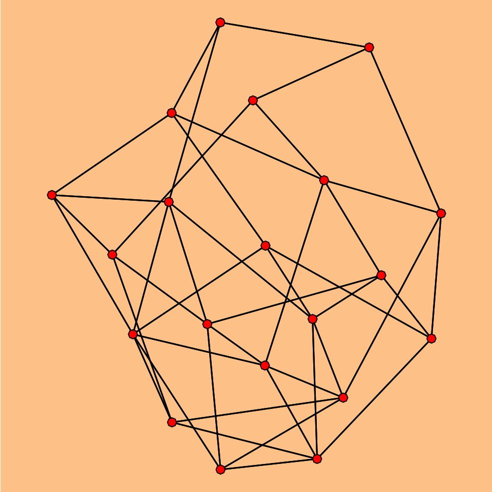
Then sum over k to get p(x).

Approximate number of multiplication operations needed for 10 possible states and 100 observations:

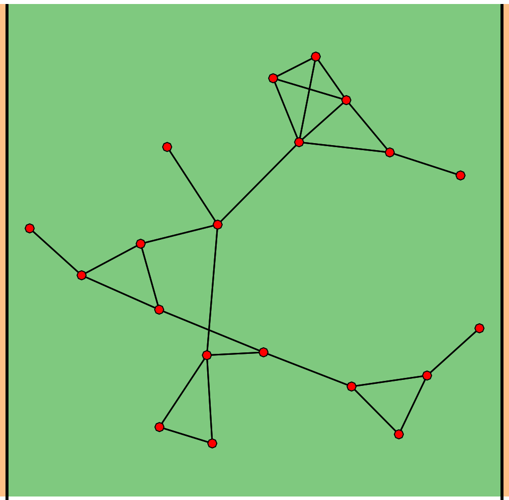
$$\text{naive method: } 2 \times 10^{102}$$

$$\text{forward method: } 2 \times 10^{4}$$

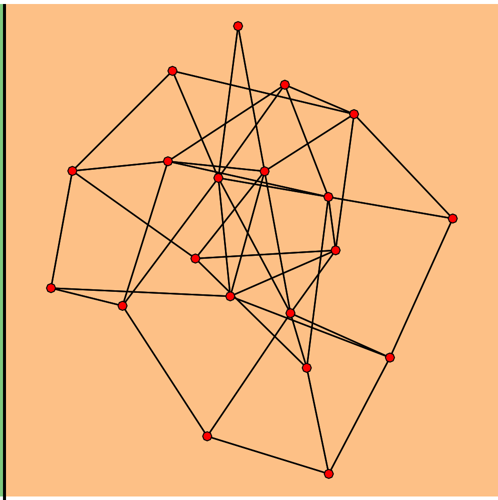
Social Networks: Geometric Features



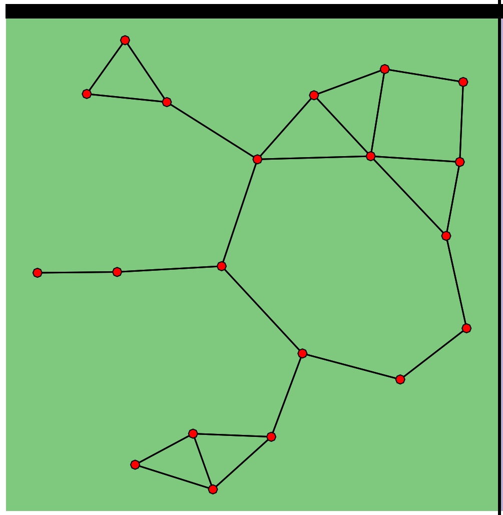
53 edges,
1 triangle,
148 two-stars



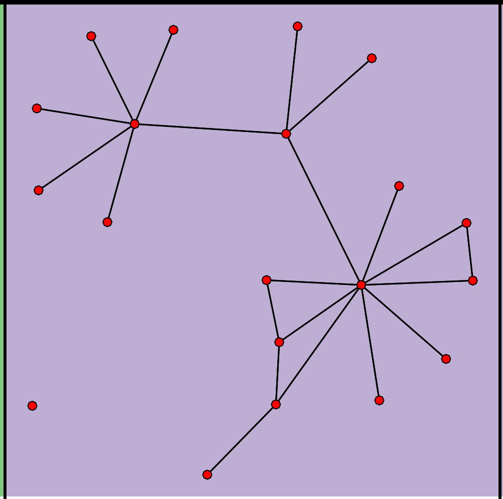
30 edges,
8 triangles,
58 two-stars



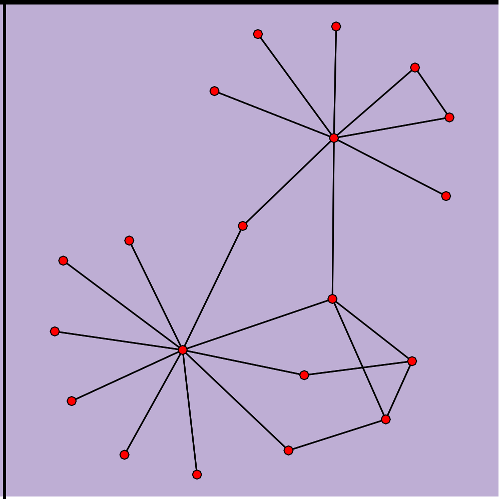
53 edges,
2 triangles,
158 two-stars



27 edges,
6 triangles,
53 two-stars

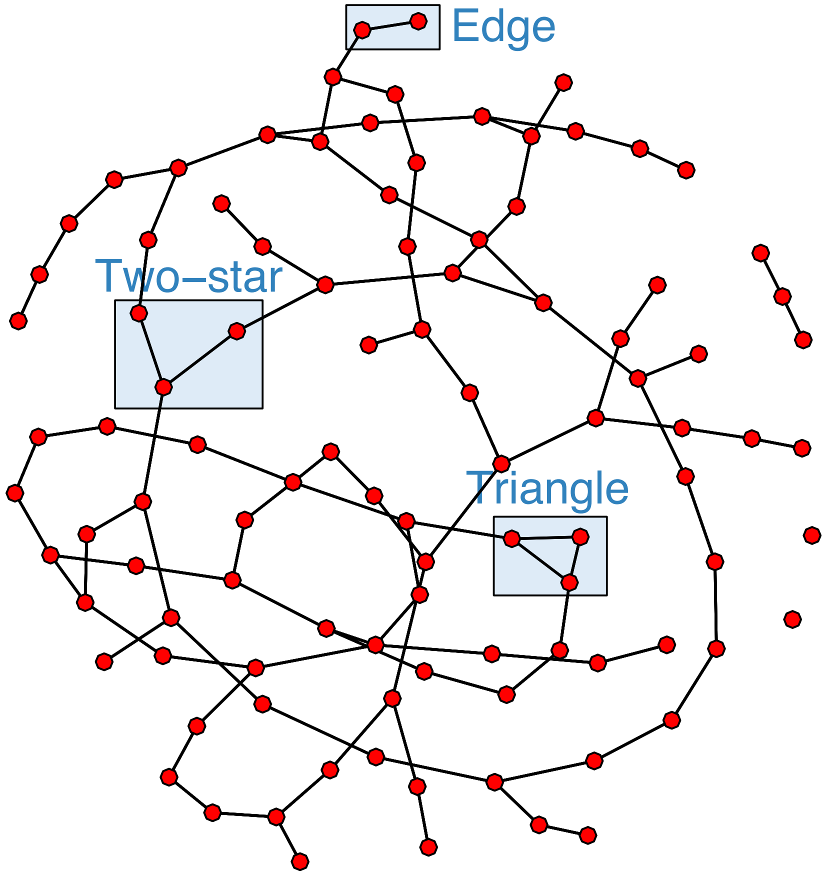


21 edges,
3 triangles,
66 two-stars



27 edges,
2 triangles,
90 two-stars

- Define three features:
  - Edge ($\binom{n}{2}$ possible)
  - Triangle ($\binom{n}{3}$ possible)
  - Two-star ($3\binom{n}{3}$ possible)
- Can overlap; e.g., triangle contains three two-stars



Exponential Random Graph Model

Idea: Make edge, triangle, two-star totals be sufficient statistics in an exponential family.

$$\begin{aligned} p_{\boldsymbol{\beta}}(G) \;\propto\; \exp(&\beta_{\text{edges}} \cdot (\#\text{ edges}) \\ &+ \beta_{\text{triangles}} \cdot (\#\text{ triangles}) \\ &+ \beta_{\text{two-stars}} \cdot (\#\text{ two-stars})) \end{aligned}$$

(Number of nodes presumed fixed.) More generally,

$$p_{\boldsymbol{\beta}}(G) \;\propto\; \exp(\boldsymbol{\beta}'\boldsymbol{x}(G))$$

To get $=$ instead of $\propto$, need normalizing constant:

$$p_{\boldsymbol{\beta}}(G) \;=\; \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}(G))}{c(\boldsymbol{\beta})}$$

Normalizing constant $c(\boldsymbol{\beta})$ is unknown!

For 20 nodes, sum involves 10^{57} terms....

MCMC for Generating Random Networks

Pick a random pair of nodes, and toggle whether there is an edge there.

This gives a uniform stationary distribution.

Pick a random pair of nodes, and "try" to toggle whether there is an edge there.

$$\frac{p_{\boldsymbol{\beta}_0}(G')}{p_{\boldsymbol{\beta}_0}(G)} = \frac{\exp(\boldsymbol{\beta}_0'\boldsymbol{x}(G'))/c(\boldsymbol{\beta}_0)}{\exp(\boldsymbol{\beta}_0'\boldsymbol{x}(G))/c(\boldsymbol{\beta}_0)} = \exp\big(\boldsymbol{\beta}_0'(\boldsymbol{x}(G') - \boldsymbol{x}(G))\big)$$

Flip a coin with this probability of Heads (or one if this exceeds one), and accept the toggle if Heads.
This gives the desired stationary distribution on networks!

Gibbs Sampler

Explore space by updating one coordinate at a time.

2D parameter space version:

Draw new θ_1 from conditional distribution of $\theta_1|\theta_2$

Draw new θ_2 from conditional distribution of $\theta_2|\theta_1$

Repeat

Gibbs Sampler

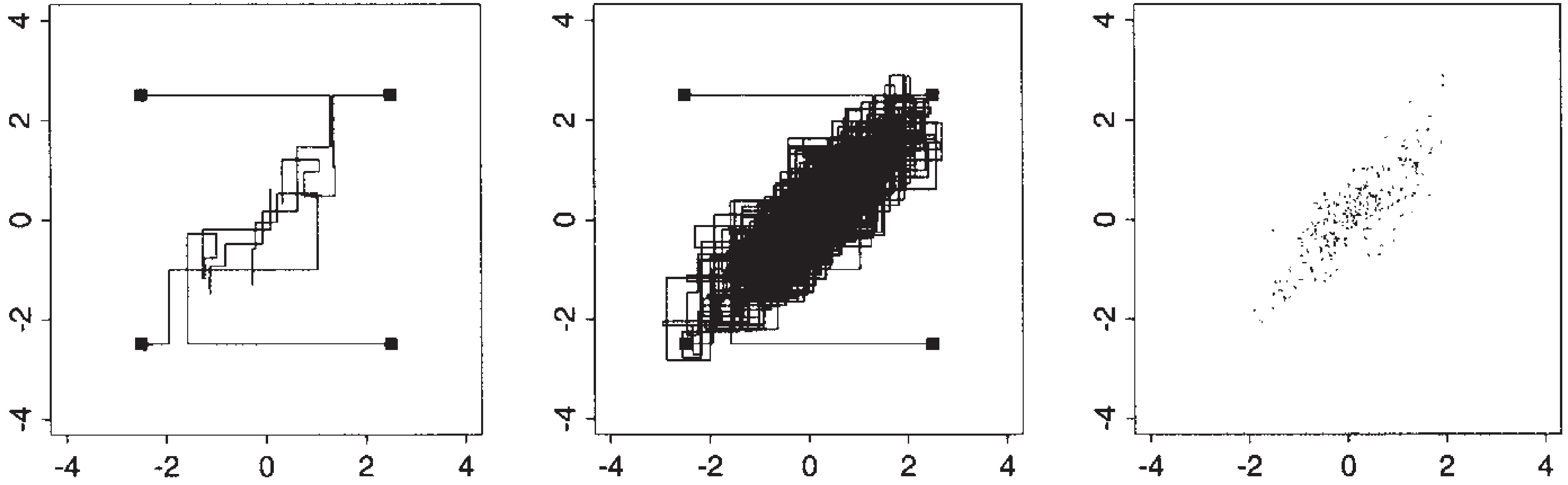


Figure 11.3 *Four independent sequences of the Gibbs sampler for a bivariate normal distribution with correlation ρ=0.8, with overdispersed starting points indicated by solid squares. (a) First 10 iterations, showing the component-by-component updating of the Gibbs iterations. (b) After 500 iterations, the sequences have reached approximate convergence. Figure (c) shows the iterates from the second halves of the sequences.*

Gelman et al, *Bayesian Data Analysis*

Metropolis-Hastings Algorithm

Modify a Markov chain on a state space of interest to obtain a new chain with *any* desired stationary distribution!

1. If $X_n = i$, propose a new state j using the transition probabilities p_{ij} of the original Markov chain.

2. Compute an *acceptance probability*,
$$a_{ij} = \min\left(\frac{s_j p_{ji}}{s_i p_{ij}}, 1\right).$$

3. Flip a coin that lands Heads with probability a_{ij}, independently of the Markov chain.

4. If the coin lands Heads, accept the proposal and set $X_{n+1} = j$. Otherwise, stay in state i; set $X_{n+1} = i$.

Metropolis-Hastings Algorithm

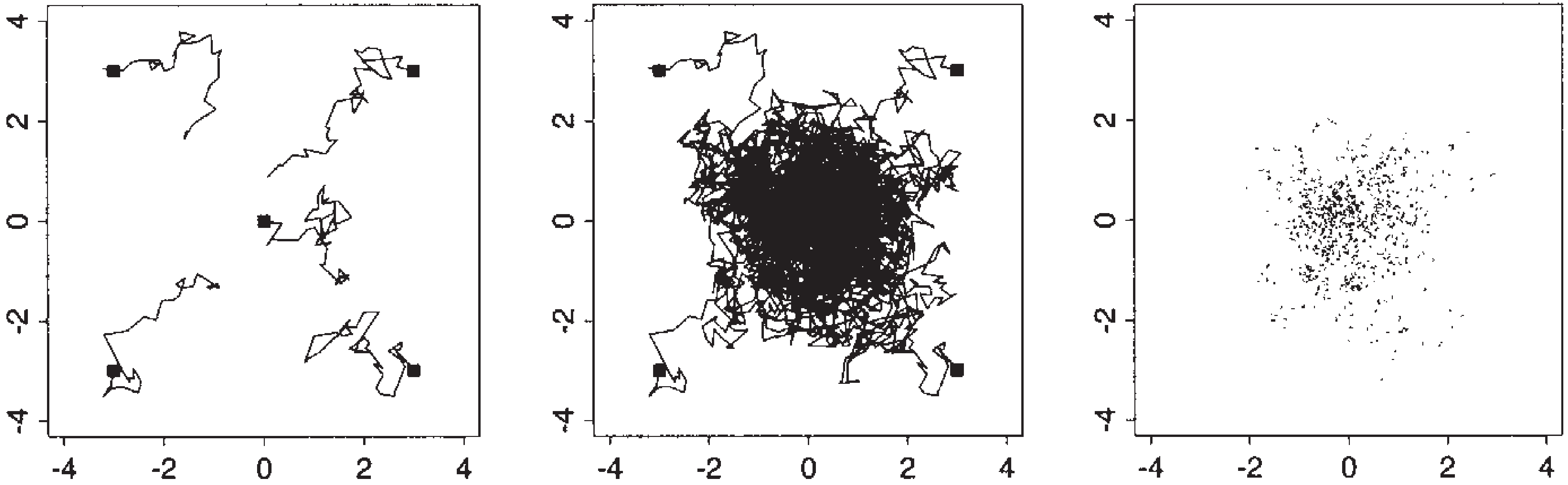


Figure 11.2 *Five independent sequences of a Markov chain simulation for the bivariate unit normal distribution, with overdispersed starting points indicated by solid squares. (a) After 50 iterations, the sequences are still far from convergence. (b) After 1000 iterations, the sequences are nearer to convergence. Figure (c) shows the iterates from the second halves of the sequences. The points in Figure (c) have been jittered so that steps in which the random walks stood still are not hidden. The simulation is a Metropolis algorithm described in the example on page 290.*

Gelman et al, *Bayesian Data Analysis*

Regression

Example: student performances in school

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

But what about differences between schools?

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{2i} + \beta_3 I_{3i} + \cdots + \beta_m I_{mi} + \epsilon_i$$

Ugly...

$$y_i = \alpha_{j[i]} + x_i \beta_1 + \epsilon_i$$

But then what about school level covariates? What about prior information?

Multilevel (Hierarchical) Models

$$y_i = \alpha_{j[i]} + x_i\beta_1 + \epsilon_i$$
$$\alpha_j \sim \mathcal{N}(\alpha_0, \sigma_0^2)$$

$$y_i = \alpha_{j[i]} + x_i\beta_1 + \epsilon_i$$
$$\alpha_j = \gamma_0 + \gamma_1 z_j + \delta_j$$
$$\delta_j \sim \mathcal{N}(0, \sigma_1^2)$$

Then use MCMC to study the joint posterior (density of all the parameters, given the data)