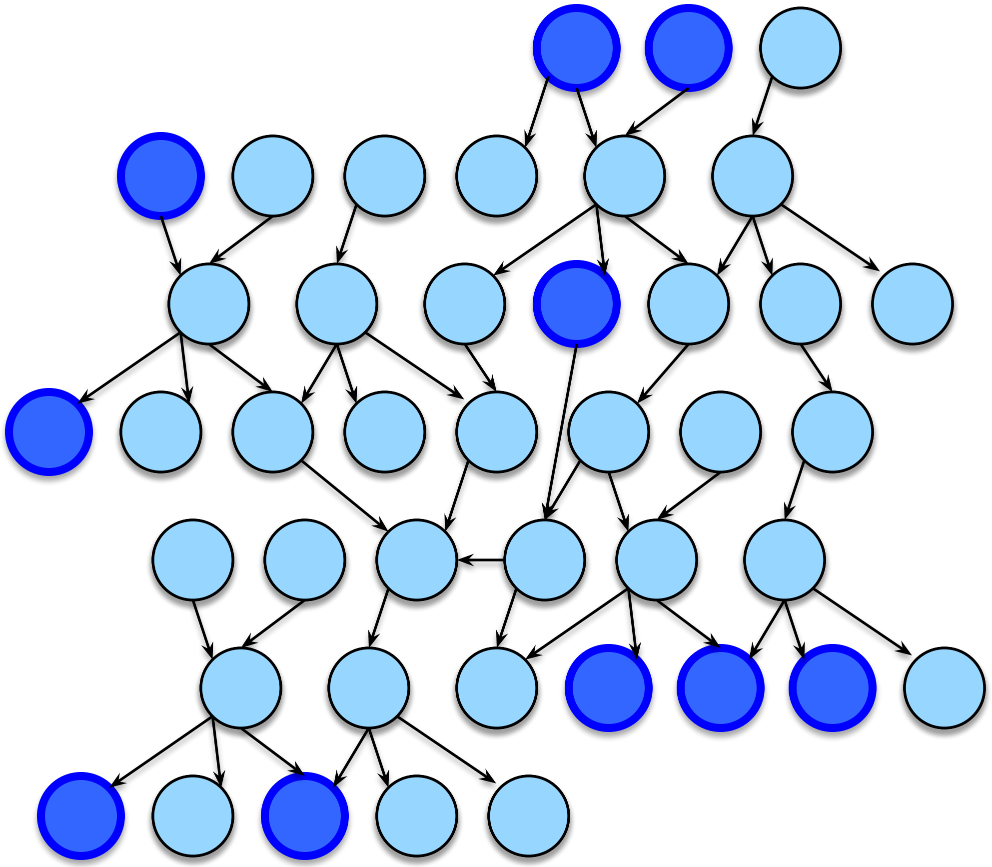
Markov Chain Monte Carlo

- MCMC (Markov chain Monte Carlo) is a family of randomized algorithms for approximating some quantity of interest over a very large state space
  - Markov chain = a sequence of randomly chosen states (“random walk”), where each state is chosen conditioned on the previous state
  - Monte Carlo = an algorithm (usually based on sampling) that has some probability of producing an incorrect answer
- MCMC = wander around for a bit, average what you see

Gibbs sampling

- A particular kind of MCMC
  - States are complete assignments to all variables
    - (Cf local search: closely related to simulated annealing!)
  - Evidence variables remain fixed, other variables change
  - To generate the next state, pick a variable and sample a value for it conditioned on all the other variables: $X_i' \sim P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n)$
    - Will tend to move towards states of higher probability, but can go down too
    - In a Bayes net, $P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n) = P(X_i \mid markov\_blanket(X_i))$
- Theorem: Gibbs sampling is consistent*
- Provided all Gibbs distributions are bounded away from 0 and 1 and variable selection is fair

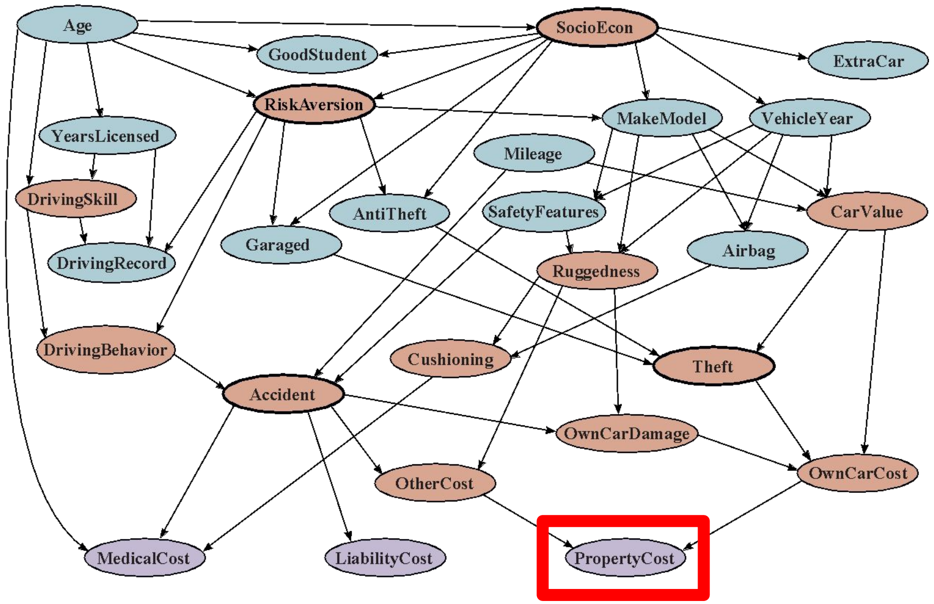
Advantages of MCMC

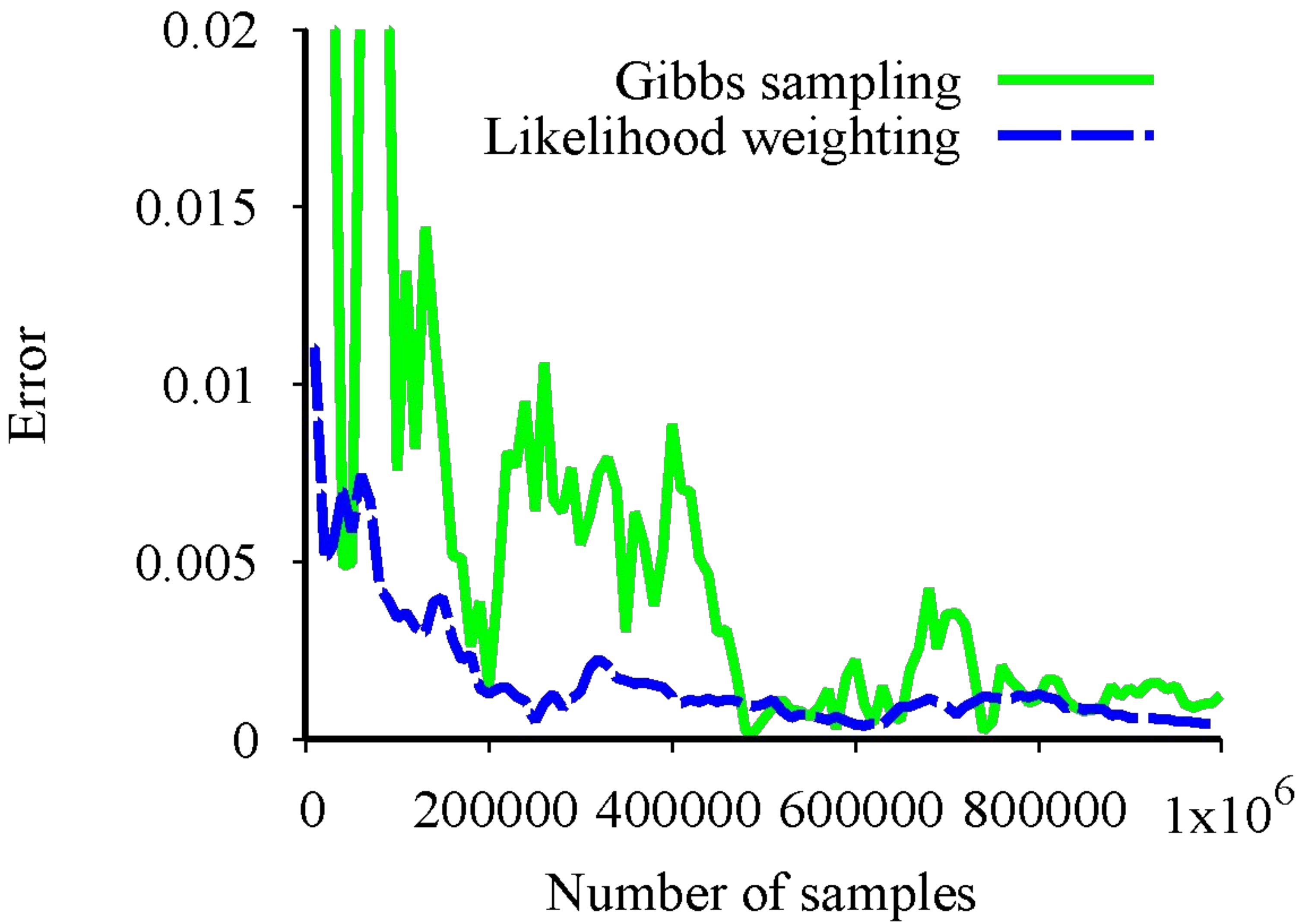


Samples soon begin to reflect all the evidence in the network

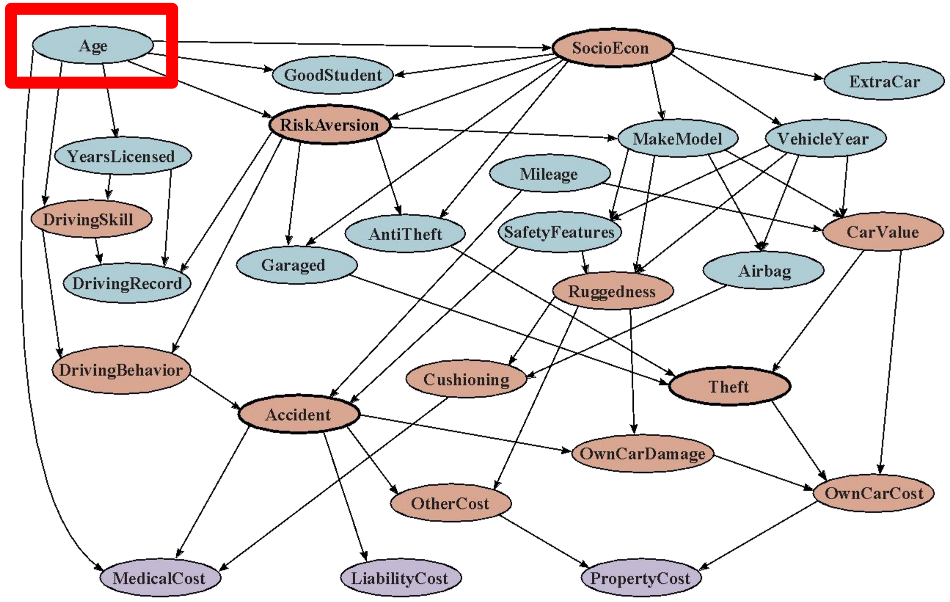
Eventually they are being drawn from the true posterior!

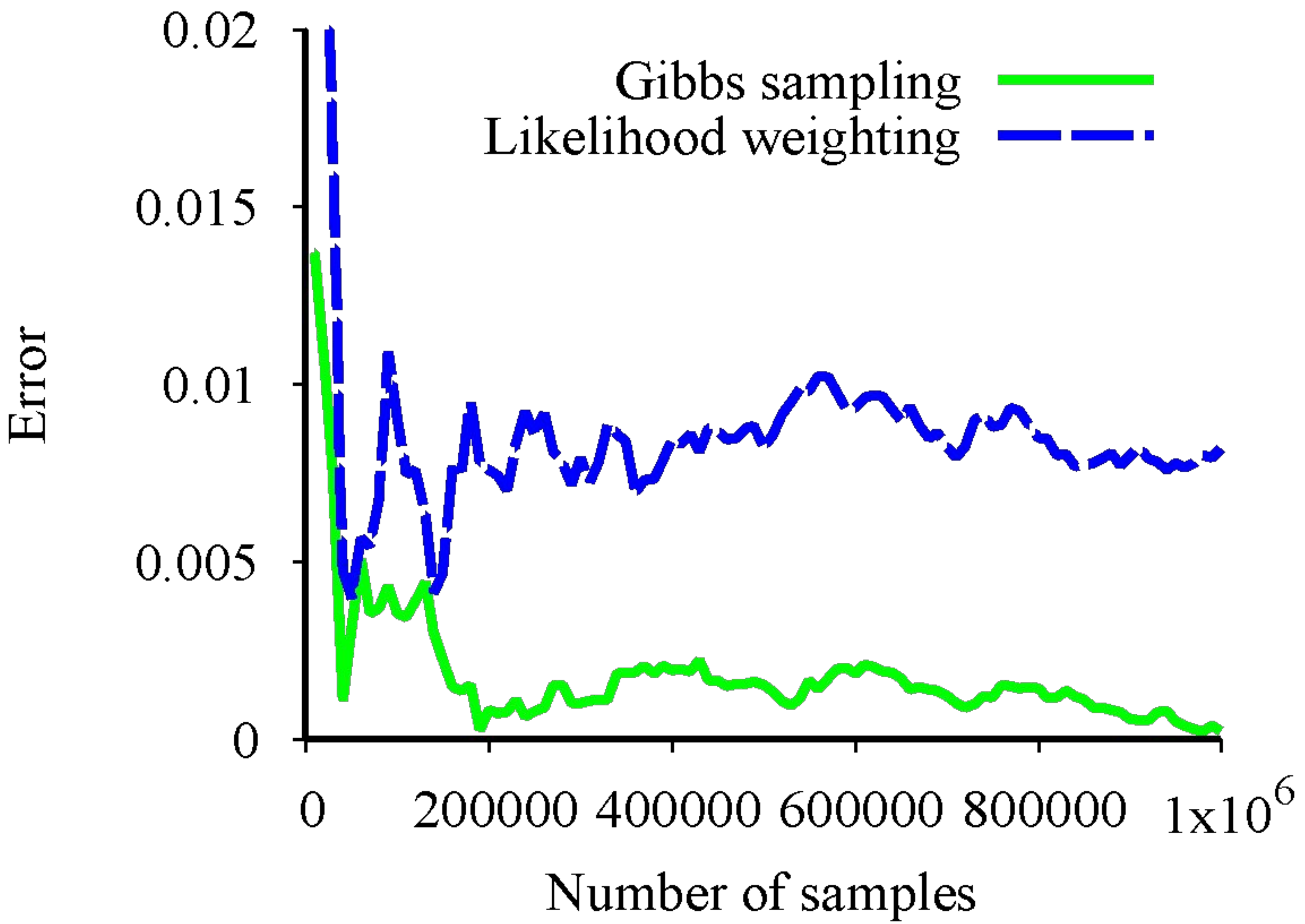
Car Insurance: *P*(*PropertyCost* | ***e***)





Car Insurance: *P*(*PropertyCost* | ***e***)



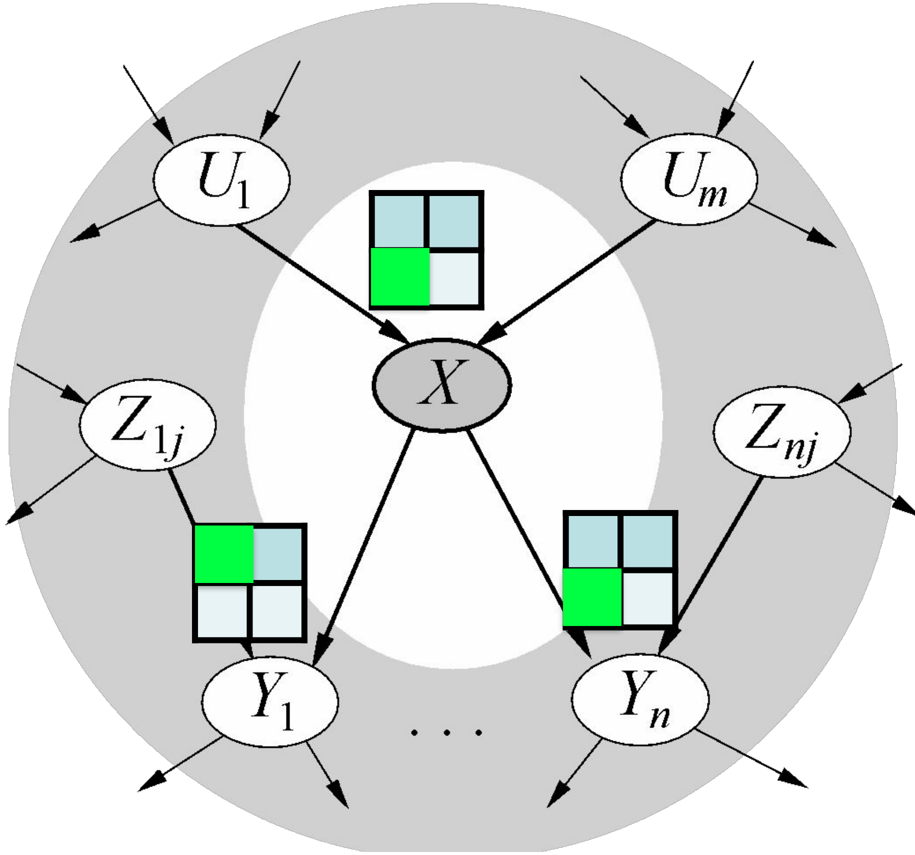


Gibbs sampling algorithm

- Repeat many times
  - Sample a non-evidence variable X_i from

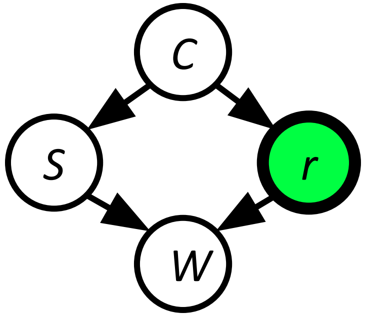
$P(X_i \mid x_1,..,x_{i-1},x_{i+1},..,x_n) = P(X_i \mid markov\_blanket(X_i))$

$= \alpha \; P(X_i \mid parents\,(X_i)) \; \prod_j P(y_j \mid parents(Y_j))$

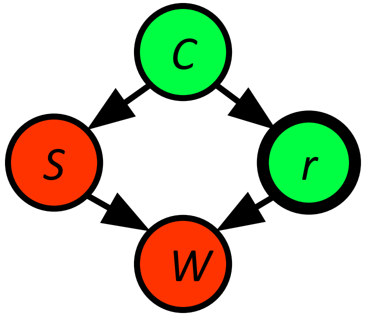


Gibbs Sampling Example: $P(S \mid r)$

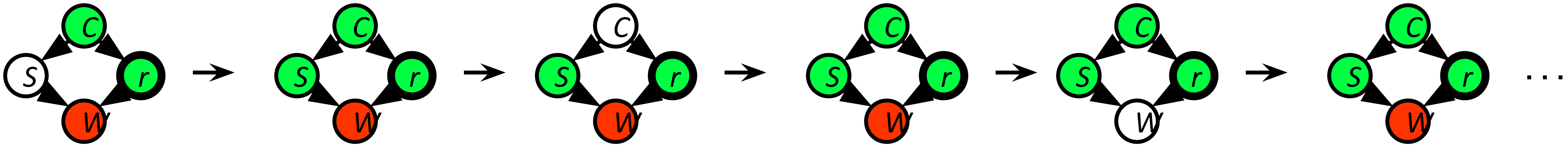
- Step 1: Fix evidence
  - *R* = true



- Step 2: Initialize other variables
  - Randomly

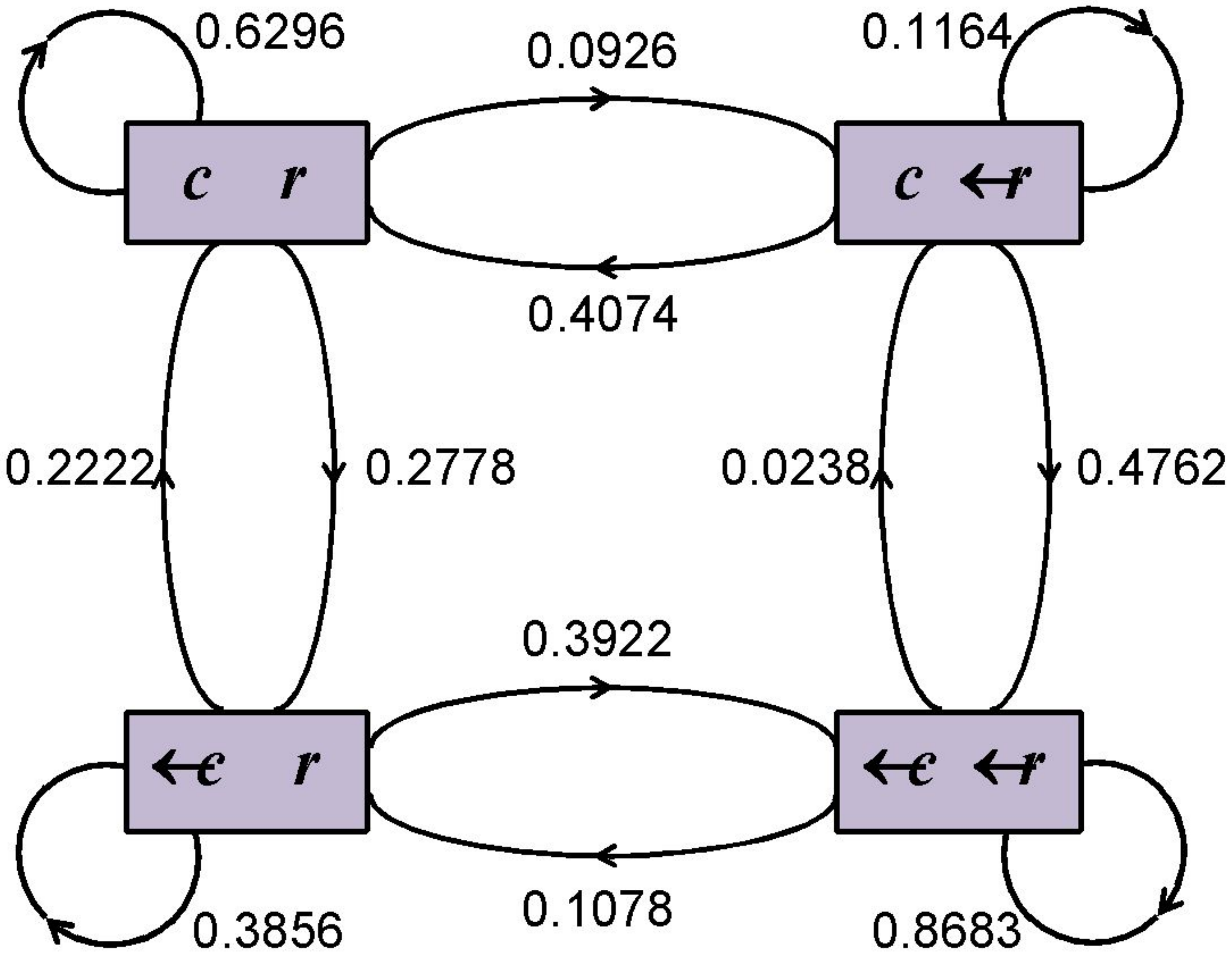


- Step 3: Repeat
  - Choose a non-evidence variable *X*
  - Resample *X* from *P*(*X* | *markov_blanket*(*X*))



Sample $S \sim P(S \mid c, r, \neg w)$ Sample $C \sim P(C \mid s, r)$ Sample $W \sim P(W \mid s, r)$

Markov chain given *s*, *w*



Gibbs sampling and MCMC in practice

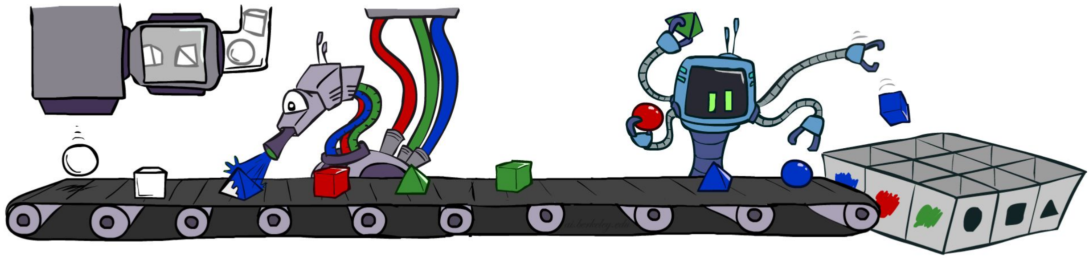
- The most commonly used method for large Bayes nets
  - See, e.g., BUGS, JAGS, STAN, infer.net, BLOG, etc.
- Can be *compiled* to run very fast
  - Eliminate all data structure references, just multiply and sample
  - ~100 million samples per second on a laptop
- Can run asynchronously in parallel (one processor per variable)
- Many cognitive scientists suggest the brain runs on MCMC

Consistency of Gibbs (see AIMA 13.4.2 for details)

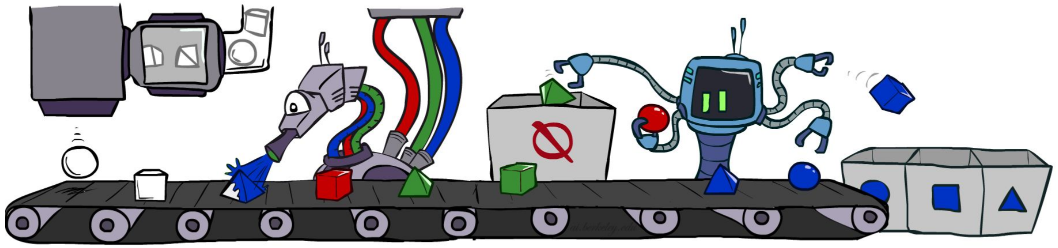
- Suppose we run it for a long time and predict the probability of reaching any given state at time t: $\pi_t(x_1,\ldots,x_n)$ or $\pi_t(\underline{\mathbf{x}})$
- Each Gibbs sampling step (pick a variable, resample its value) applied to a state $\underline{\mathbf{x}}$ has a probability $k(\underline{\mathbf{x}}' \mid \underline{\mathbf{x}})$ of reaching a next state $\underline{\mathbf{x}}'$
- So $\pi_{t+1}(\underline{\mathbf{x}}') = \sum_{\underline{\mathbf{x}}} k(\underline{\mathbf{x}}' \mid \underline{\mathbf{x}})\, \pi_t(\underline{\mathbf{x}})$ or, in matrix/vector form $\pi_{t+1} = \mathbf{K}\pi_t$
- When the process is in equilibrium $\pi_{t+1} = \pi_t = \pi$ so $\mathbf{K}\pi = \pi$
- This has a unique* solution $\pi = P(x_1,\ldots,x_n \mid e_1,\ldots,e_k)$
    - * Markov chain must be ***ergodic***, i.e., completely connected and aperiodic
    - Satisfied if all probabilities are bounded away from 0 and 1
- So for large enough t the next sample will be drawn from the true posterior
    - "Large enough" depends on CPTs in the Bayes net; takes *longer* if nearly deterministic

Bayes Net Sampling Summary

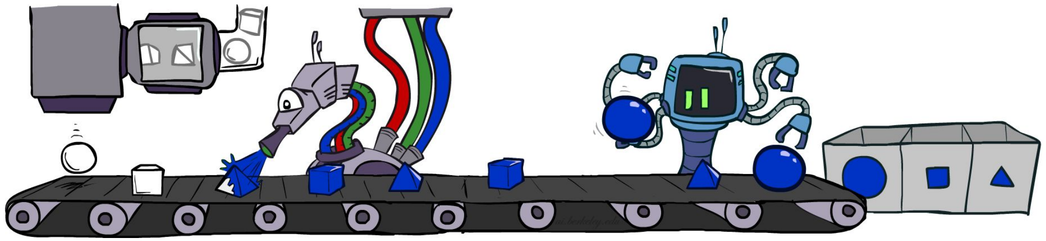
- Prior Sampling P :
  - Generate complete samples from $P(x_1,\ldots,x_n)$



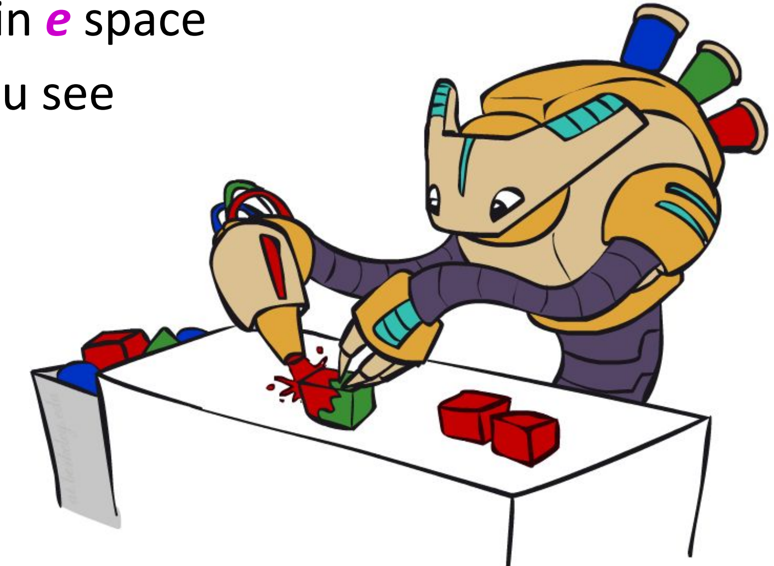
- Rejection Sampling $P(Q \mid \mathbf{e})$:
  - Reject samples that don't match $\mathbf{e}$



- Likelihood Weighting $P(Q \mid \mathbf{e})$:
  - Weight samples by how well they predict $\mathbf{e}$

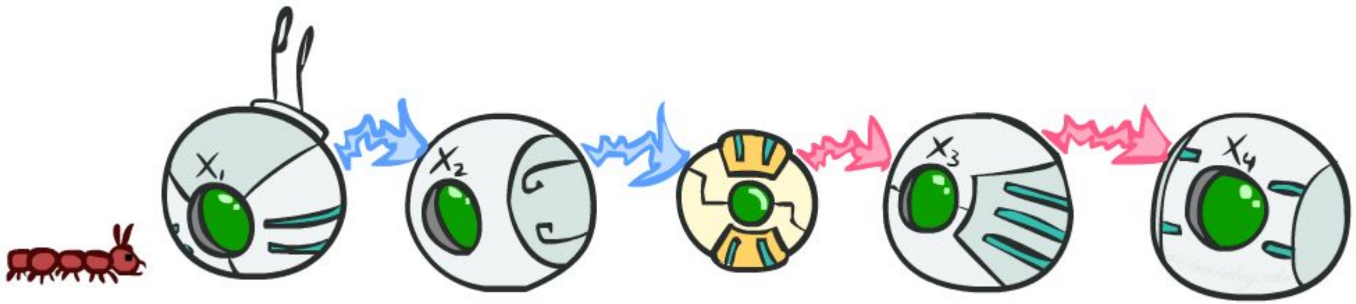


- Gibbs sampling $P(Q \mid \mathbf{e})$:
  - Wander around in $\mathbf{e}$ space
  - Average what you see



CS 188: Artificial Intelligence

Markov Models



Instructors: Stuart Russell and Peyrin Kao

University of California, Berkeley

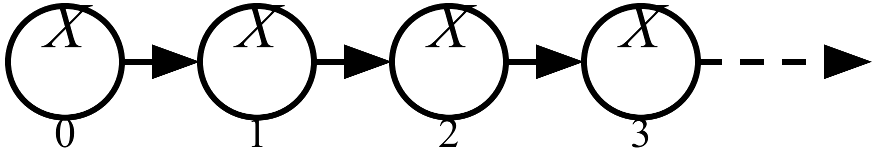
Uncertainty and Time

- Often, we want to reason about a ***sequence*** of observations where the state of the underlying system is ***changing***
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring
  - Global climate

- Need to introduce time into our models

Markov Models (aka Markov chain/process)

- Value of X at a given time is called the ***state*** (usually discrete, finite)



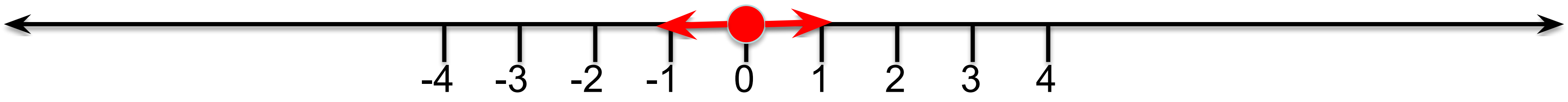
$P(X_0)$ $P(X_t \mid X_{t-1})$

- The ***transition model*** $P(X_t \mid X_{t-1})$ specifies how the state evolves over time
- ***Stationarity*** assumption: transition probabilities are the same at all times
- ***Markov*** assumption: "future is independent of the past given the present"
  - X_{t+1} is independent of $X_0, \ldots, X_{t-1}$ given X_t
  - This is a ***first-order*** Markov model (a kth-order model allows dependencies on k earlier steps)
- Joint distribution $P(X_0, \ldots, X_T) = P(X_0) \prod_t P(X_t \mid X_{t-1})$

Quiz: are Markov models a special case of Bayes nets?

- Yes and no!
- Yes:
  - Directed acyclic graph, joint = product of conditionals
- No:
  - Infinitely many variables (unless we truncate)
  - Repetition of transition model not part of standard Bayes net syntax

Example: Random walk in one dimension



- State: location on the unbounded integer line
- Initial probability: starts at 0
- Transition model: $P(X_t = k \mid X_{t-1} = k \pm 1) = 0.5$
- Applications: particle motion in crystals, stock prices, gambling, genetics, etc.
- Questions:
  - How far does it get as a function of t?
    - Expected distance is $O(\sqrt{t})$
  - Does it get back to 0 or can it go off for ever and not come back?
    - In 1D and 2D, returns w.p. 1; in 3D, returns w.p. 0.34053733

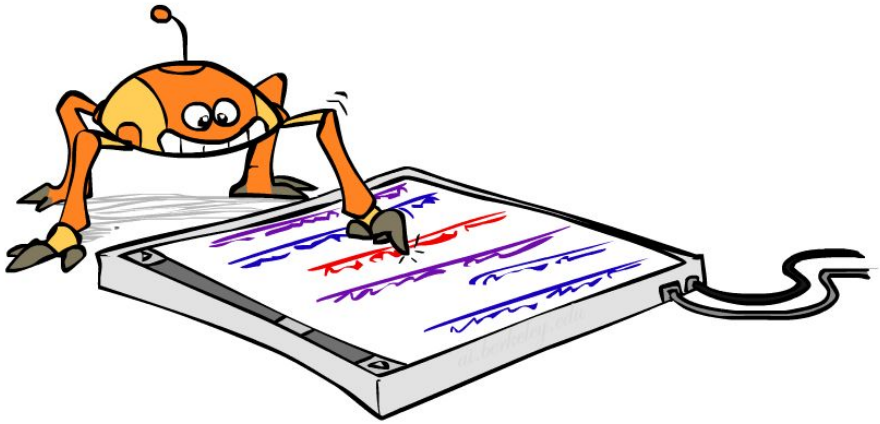
Example: n-gram models

We call ourselves *Homo sapiens*—man the wise—because our **intelligence** is so important to us.
For thousands of years, we have tried to understand *how we think*; that is, how a mere handful of matter can perceive, understand, predict, and manipulate a world far larger and more complicated than itself. ….

- State: word at position t in text (can also build letter n-grams)
- Transition model (probabilities come from empirical frequencies):
  - Unigram (zero-order): $P(Word_t = i)$
    - "logical are as are confusion a may right tries agent goal the was . . ."
  - Bigram (first-order): $P(Word_t = i \mid Word_{t-1} = j)$
    - "systems are very similar computational approach would be represented . . ."
  - Trigram (second-order): $P(Word_t = i \mid Word_{t-1} = j, Word_{t-2} = k)$
    - "planning and scheduling are integrated the success of naive bayes model is . . ."
- Applications: text classification, spam detection, author identification, language classification, speech recognition

Example: Web browsing

- State: URL visited at step *t*
- Transition model:
  - With probability *p*, choose an outgoing link at random
  - With probability (1-*p*), choose an arbitrary new page
- Question: What is the ***stationary distribution*** over pages?
  - I.e., if the process runs forever, what fraction of time does it spend in any given page?
- Application: Google page rank



Example: Weather

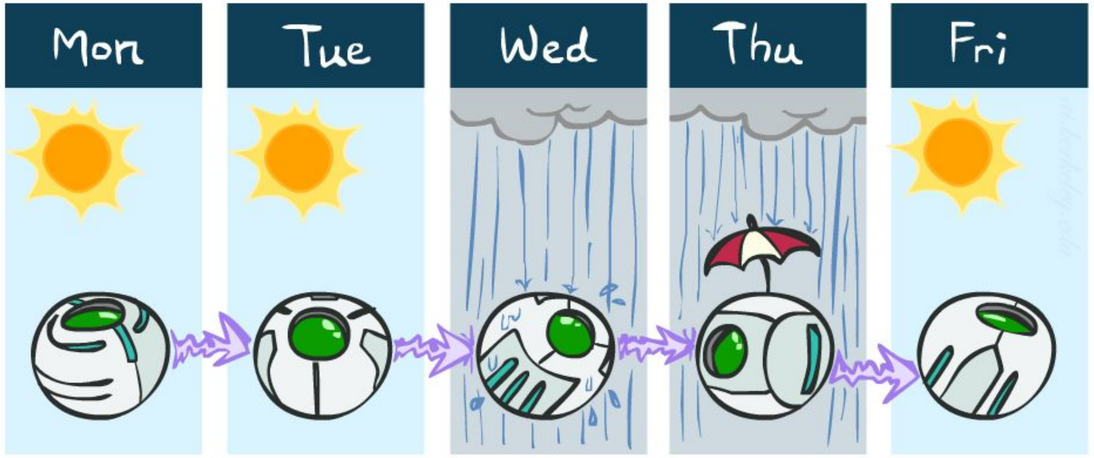
- States {rain, sun}

- Initial distribution $P(X_0)$

$P(X_0)$	
sun	rain
0.5	0.5

- Transition model $P(X_t \mid X_{t-1})$

X_{t-1}	$P(X_t\|X_{t-1})$	
	sun	rain
sun	0.9	0.1
rain	0.3	0.7



Two new ways of representing the same CPT