Topic Models 1

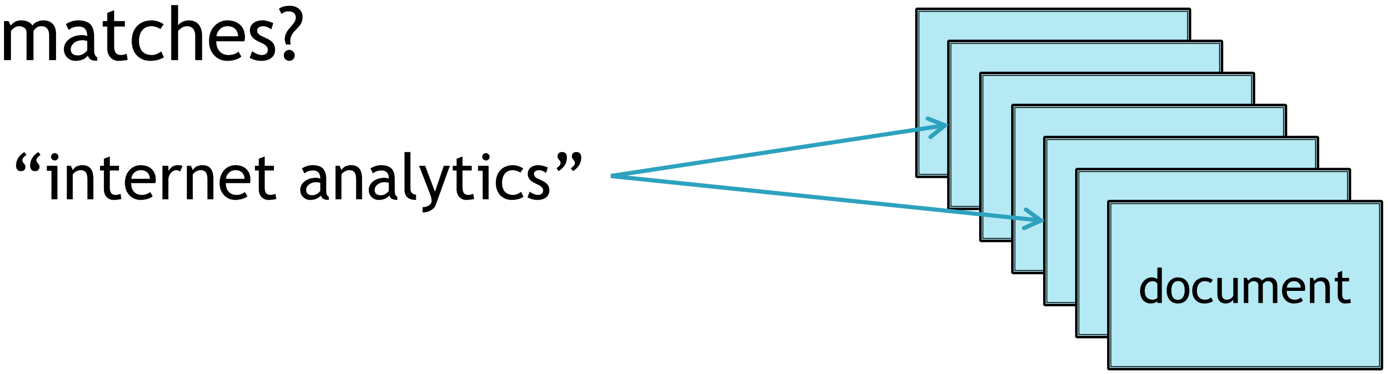
Internet Analytics (COM-308)

Prof. Matthias Grossglauser
School of Computer and Communication Sciences

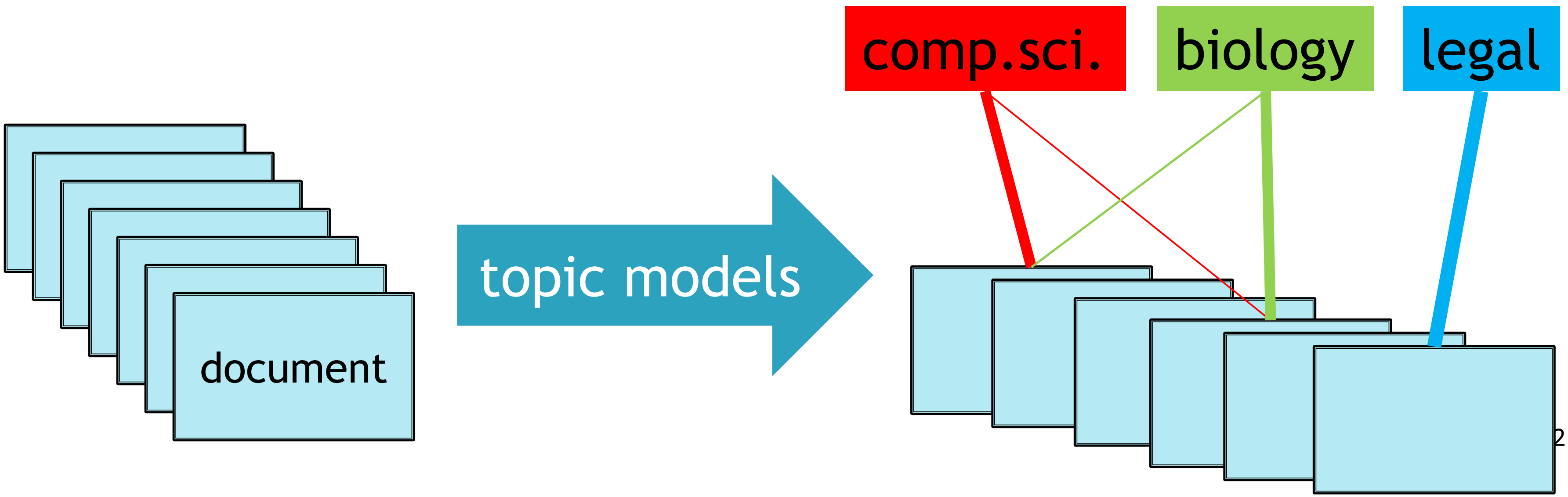
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Overview

- Previously: Given a query, how to find best matches?



- Today: Without a query, how to describe a corpus?

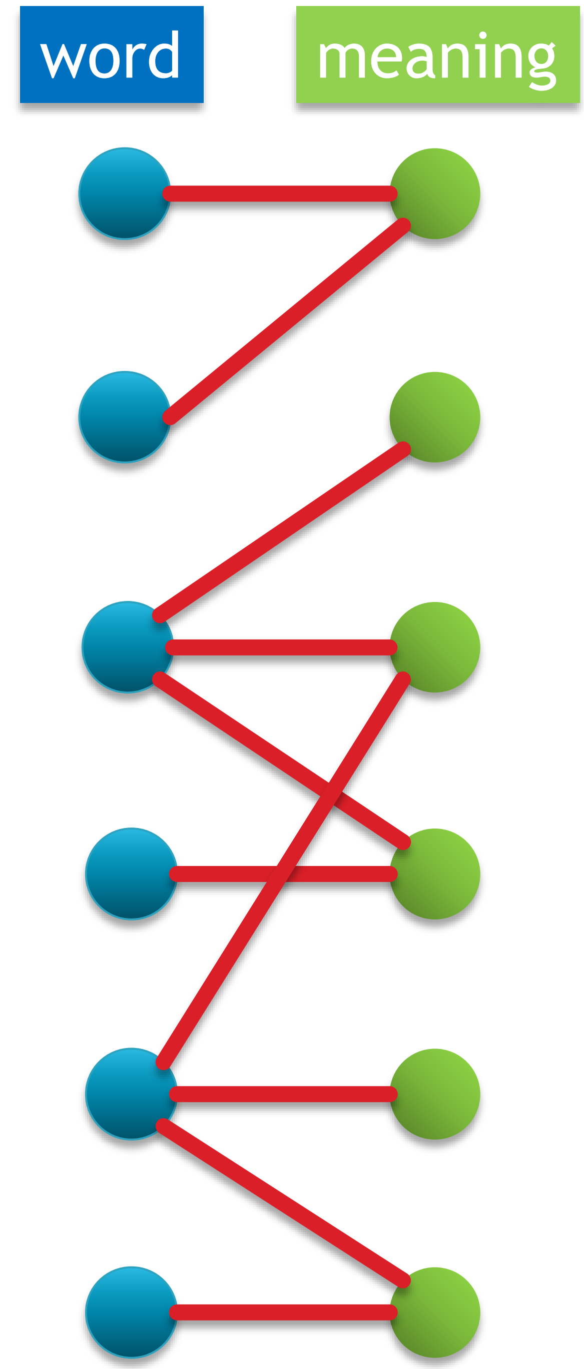


Topic models

- Document classification
- Supervised: training set with known classes
  - Generalization of binary classification (spam/not spam)
- Unsupervised: need to identify sensible topic classes by comparing documents
- Assumptions:
  - Number of words per document >> 1
  - Number of topics << number of documents
- Examples:
  - News articles: topics = {countries, business, politics, celebrity, …}
  - Scientific literature: {physics, mathematics, engineering, chemistry, life sciences,…}

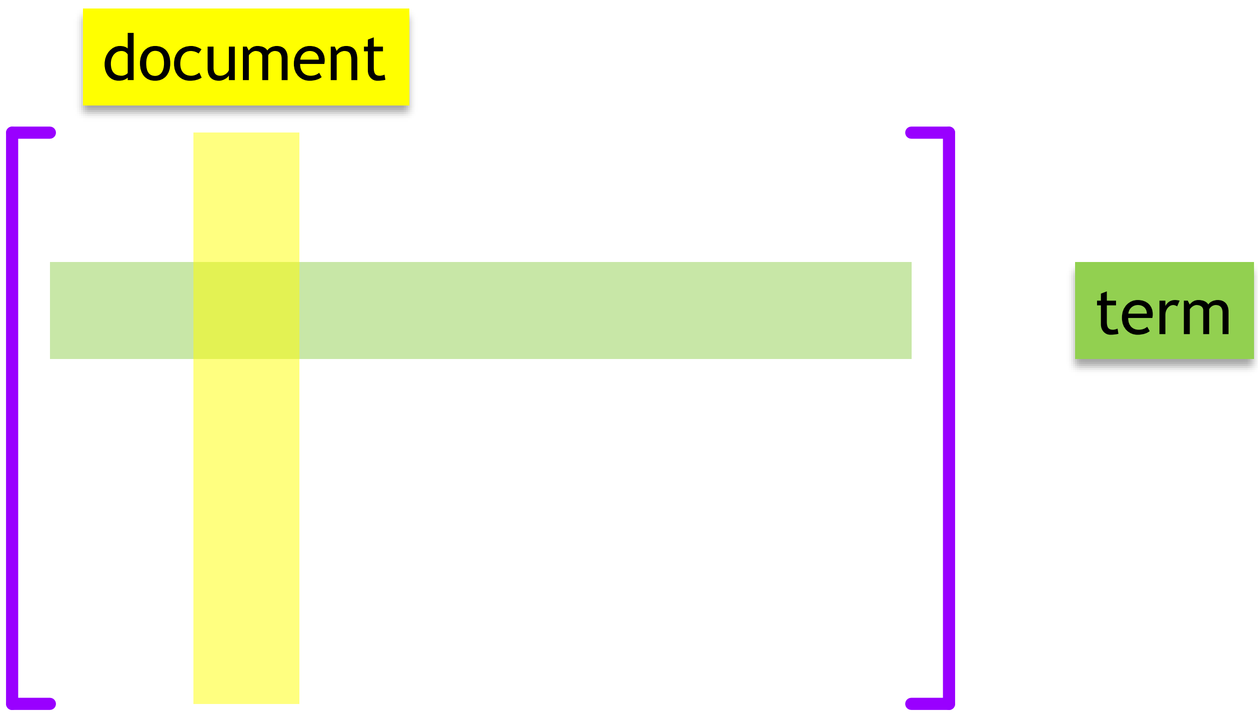
Synonymy and polysemy

- Synonymy:
  - Different words with the same meaning
  - “car” and “automobile”
- Polysemy/homonymy:
  - One word with different meanings
  - “jaguar”: animal, brand of car
- Topic models:
  - We see the words of docs, but we want to classify the meanings of docs
  - Ambiguity of individual words – but many words per doc helps!



Approach 1: Latent Semantic Indexing (LSI)

- Synonymous: Latent Semantic Analysis (LSA)
- Starting point: TF-IDF matrix of corpus
- $X =$



- Remember: high TF-IDF means “term that is rare overall, but prominent in this doc”

SVD of TF-IDF matrix

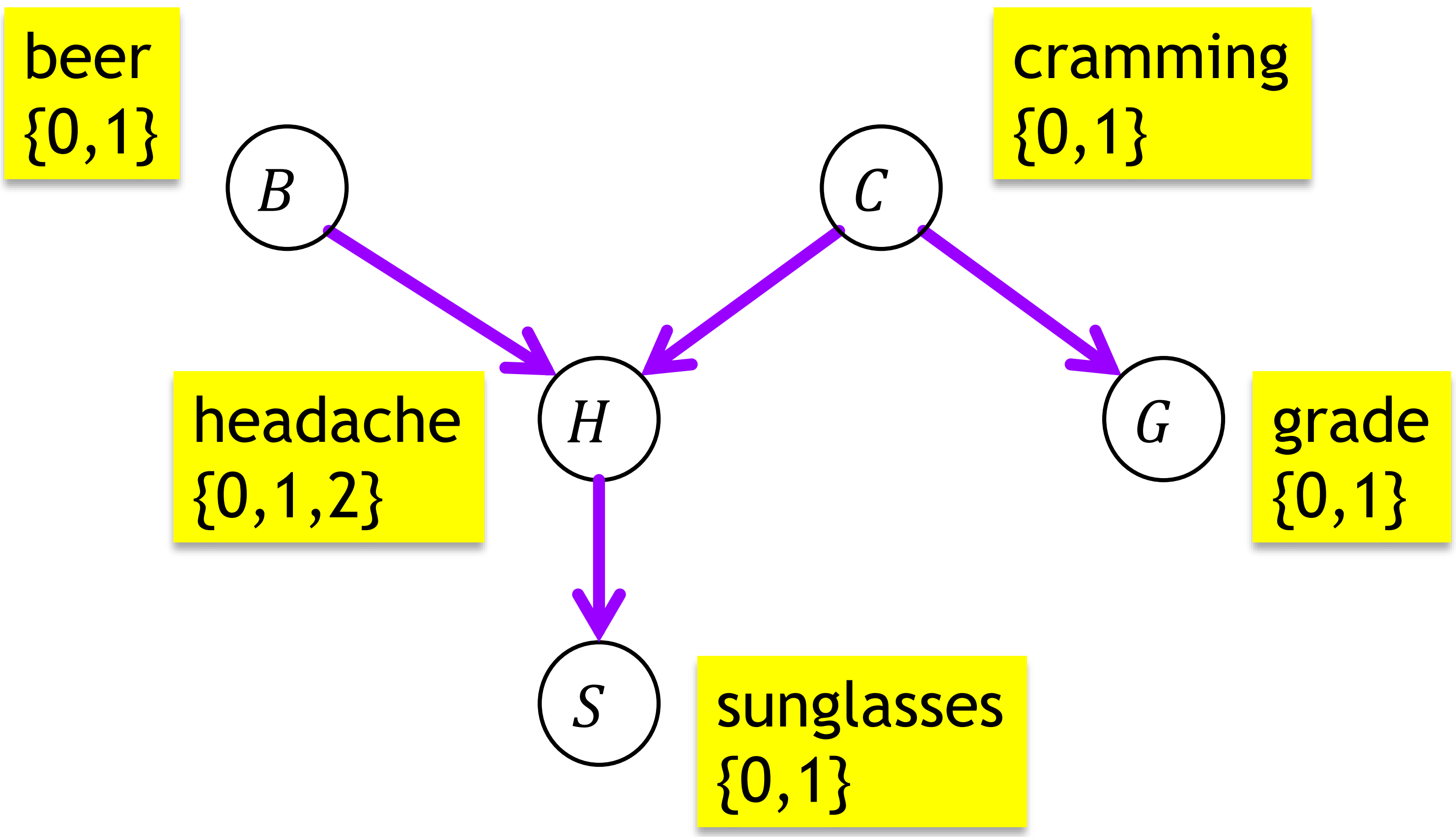
- Latent factors: “topics”
- Typically 100-300
- Should bunch together synonyms
- Should separate homonyms
- Critique:
  - Heuristic, no clean statistical foundation
  - Sometimes difficult to interpret results
  - Modern approaches based on probabilistic models:
    - better performance
    - better interpretability
    - generative

Gentle introduction to graphical models

- Modeling a multivariate distribution
- Example: insights from an expert:
  - "Drinking too much beer can result in headaches"
  - "Studying too much can cause headaches as well"
  - "To get a good grade, one must study"
  - "Wearing sunglasses tempers the pain of a headache"
- How to translate this into a probabilistic model?
  - Random variables
  - Dependencies?
    - Option: define/learn full joint distribution → many parameters, memory-intensive, hard to learn
    - Option: encode «causal structure» into model

Bayesian Network

- Edges = “direct” influence



Bayesian Network

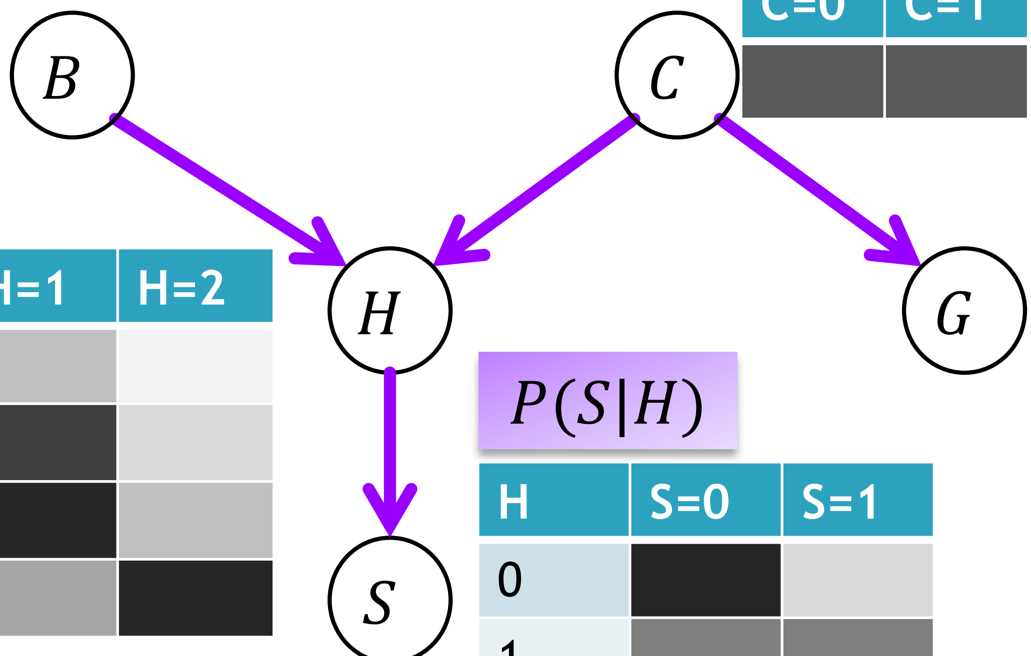
- One conditional distribution per node → full joint distribution

$P(B)$

B=0	B=1

$P(C)$

C=0	C=1



$P(H|B,C)$

(B,C)	H=0	H=1	H=2
(0,0)			
(0,1)			
(1,0)			
(1,1)			

$P(G|C)$

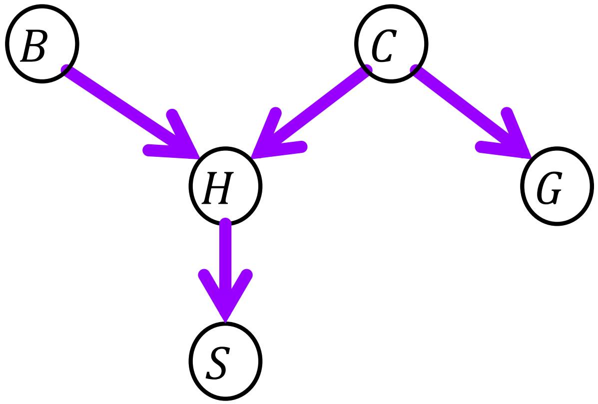
C	G=0	G=1
0		
1		

$P(S|H)$

H	S=0	S=1
0		
1		
2		

Full joint dist.: 47 params
BN: 1+1+8+3+2=15 params

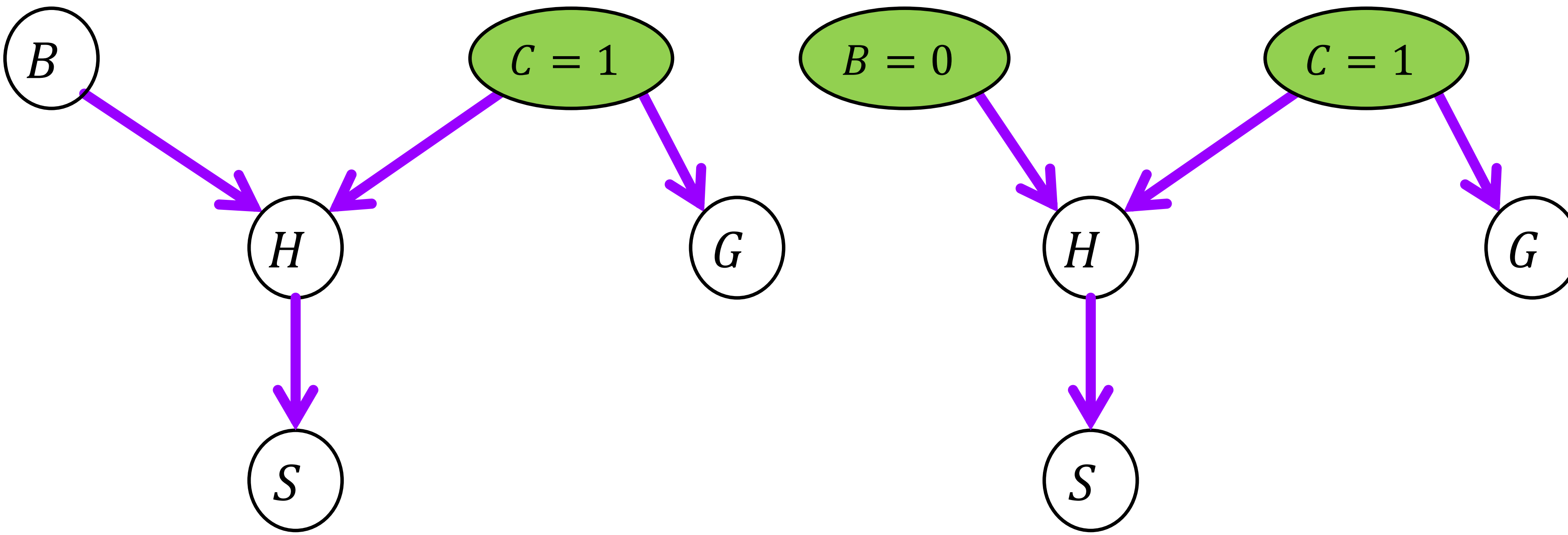
Joint distribution from CPDs



- Joint distribution from chain rule
- $P(B, C, H, G, S) =$
- $= P(C, H, G, S|B)P(B) =$
- $= P(H, G, S|B, C)P(C)P(B) =$
- $= P(H, S|B, C)P(G|B, C)P(C)P(B) =$
- $= P(S|B, C, H)P(H|B, C)P(G|C)P(C)P(B) =$
- $= P(S|H)P(H|B, C)P(G|C)P(C)P(B)$
- Joint distribution = product of all individual per-node factors
  - With the joint distribution, everything else follows: all marginal and conditional distributions we could want

Types of reasoning

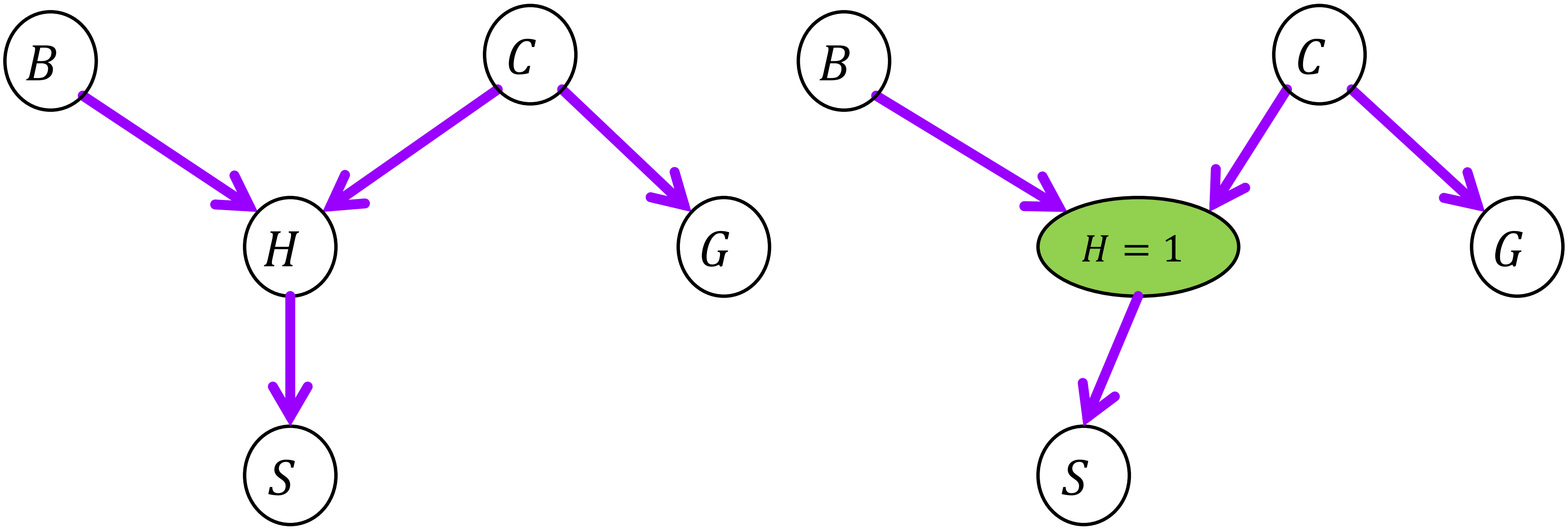
- Causal reasoning / prediction: downstream flow of influence



$$P(S=1|C=1) > P(S=1|B=0,C=1)$$

Types of reasoning

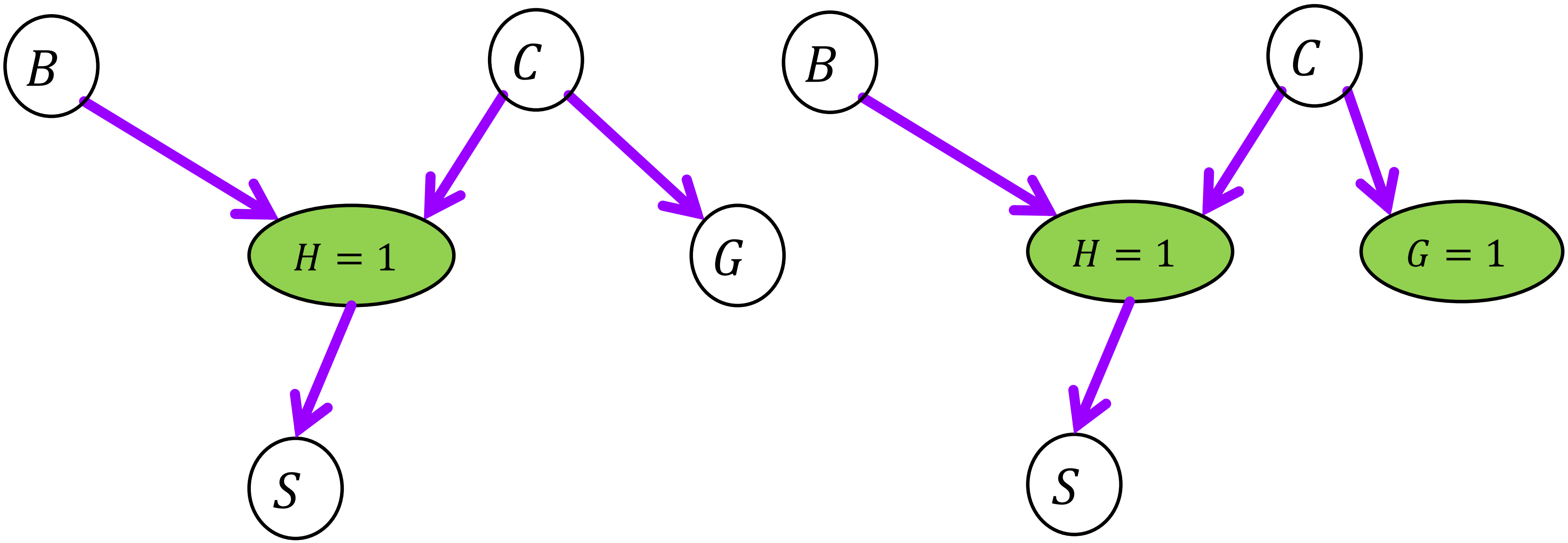
- Evidential reasoning / explanation: upstream flow of influence



$$P(B = 1|H = 1) > P(B = 1)$$
$$P(C = 1|H = 1) > P(C = 1)$$

Types of reasoning

- Intercausal reasoning: combination of upstream/downstream

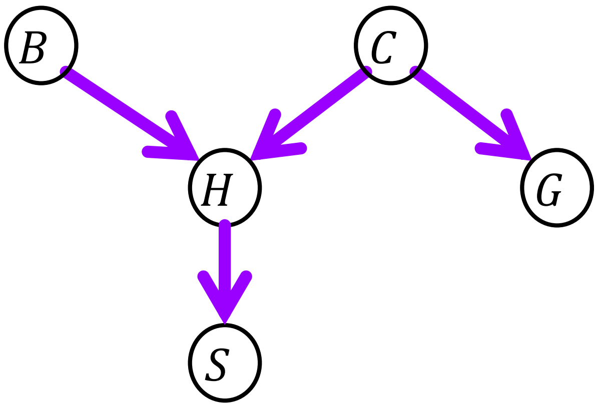


$$P(B=1|H=1) > P(B=1|H=1,G=1)$$

Explaining away: the "good grade" explains the "headache", making the possible cause "beer" less likely

Basic independencies in BNs

- Example: "the wearing of sunglasses depends only on the presence and strength of a headache"
  - Formally: $(S \perp B, C, G | H)$



- Also:
  - $(G \perp B, H, S | C)$
  - $(B \perp C)$
  - $(H \perp G | B, C)$
  - $(B \perp C, G)$
- How about $(H \perp S, G | B, C)$?
  - No! Intuition: suppose we know $B = 0$ and $C = 1$; then the guess for H changes according to $S = 0,1$

Basic conditional independencies in BNs

- Bayesian Network: directed acyclic graph (DAG) G
- Def: $Pa(X_i)$=parents of X_i in G
- Def: $ND(X_i)$=non-descendents of X_i in G
- Property: G has the following local independence properties:
  - For each X_i:

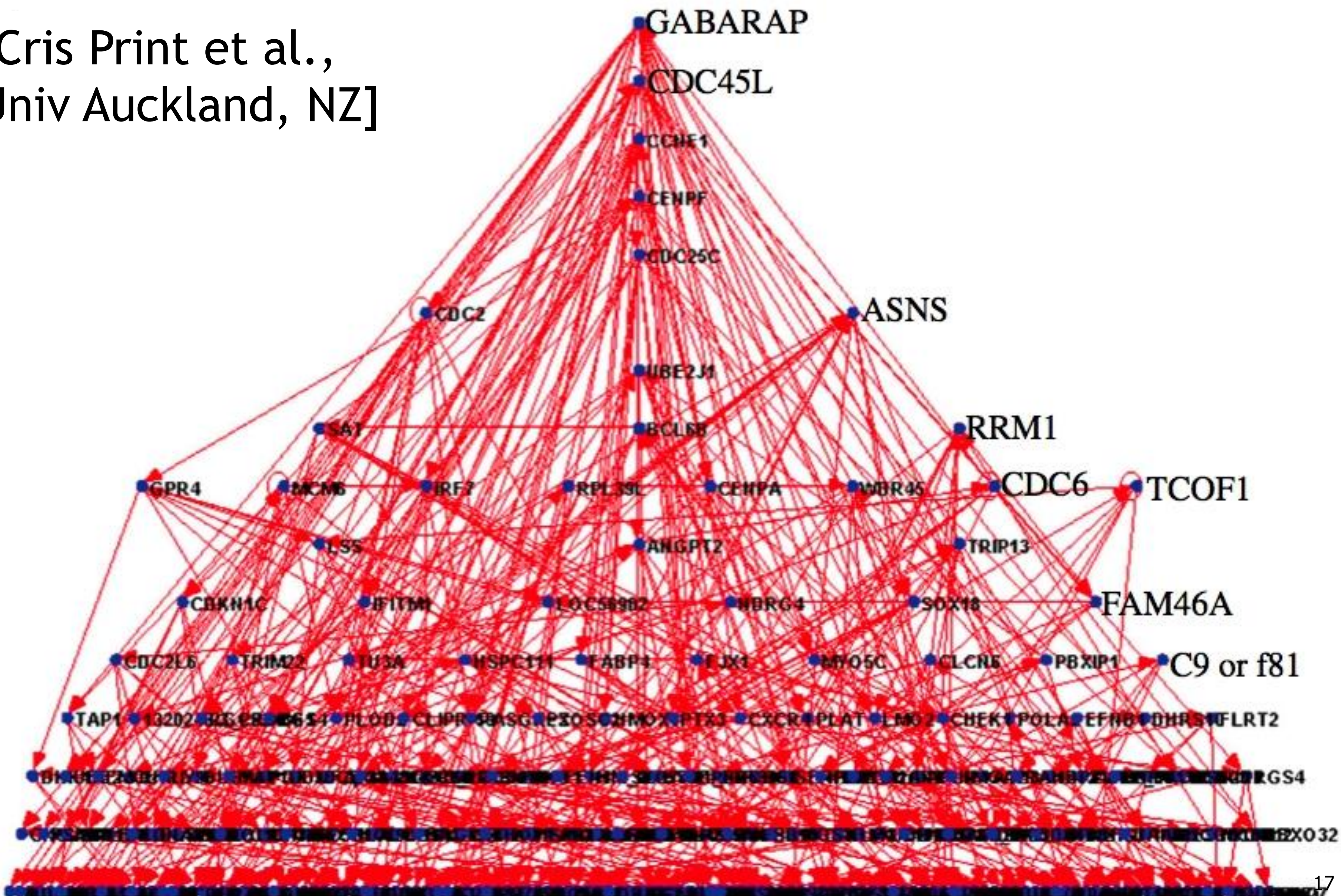
$$(X_i \perp ND(X_i) \mid Pa(X_i))$$

Bayesian Networks: recap

- Defines a multivariate probability distribution
- Models direct causal influences
  - This comes from expert knowledge, underlying mechanisms, data about the problem,…
- In practice: as sparse as possible
- Conditional independence properties as graph (path) properties
- Inference:
  - Observe some variables (observables)
  - Obtain conditional distribution of some other variables of interest → estimate
  - Some variables we do not care (latent)

Example: gene network

[Cris Print et al., Univ Auckland, NZ]



Computational challenge in large models

- Suppose G large; a few variables $Y \subset X$ are observed, $Z = X \backslash Y$ are not observed
- Want to estimate $P(Z_{573}|Y)$, where Z_{573} is e.g. one of many diseases in a medical diagnostic system
- Need to compute $P(Z_{573}|Y) =$

$$\sum_{Z_1, Z_2, \ldots, Z_{572}, Z_{574}, \ldots} P(Z_1, Z_2, \ldots, Z_{572}, Z_{573}, Z_{574}, \ldots |Y)$$

  - Very costly to marginalize out all other latent variables
- Inference methods:
  - Exact
  - Markov Chain Monte Carlo (MCMC)
  - Variational inference

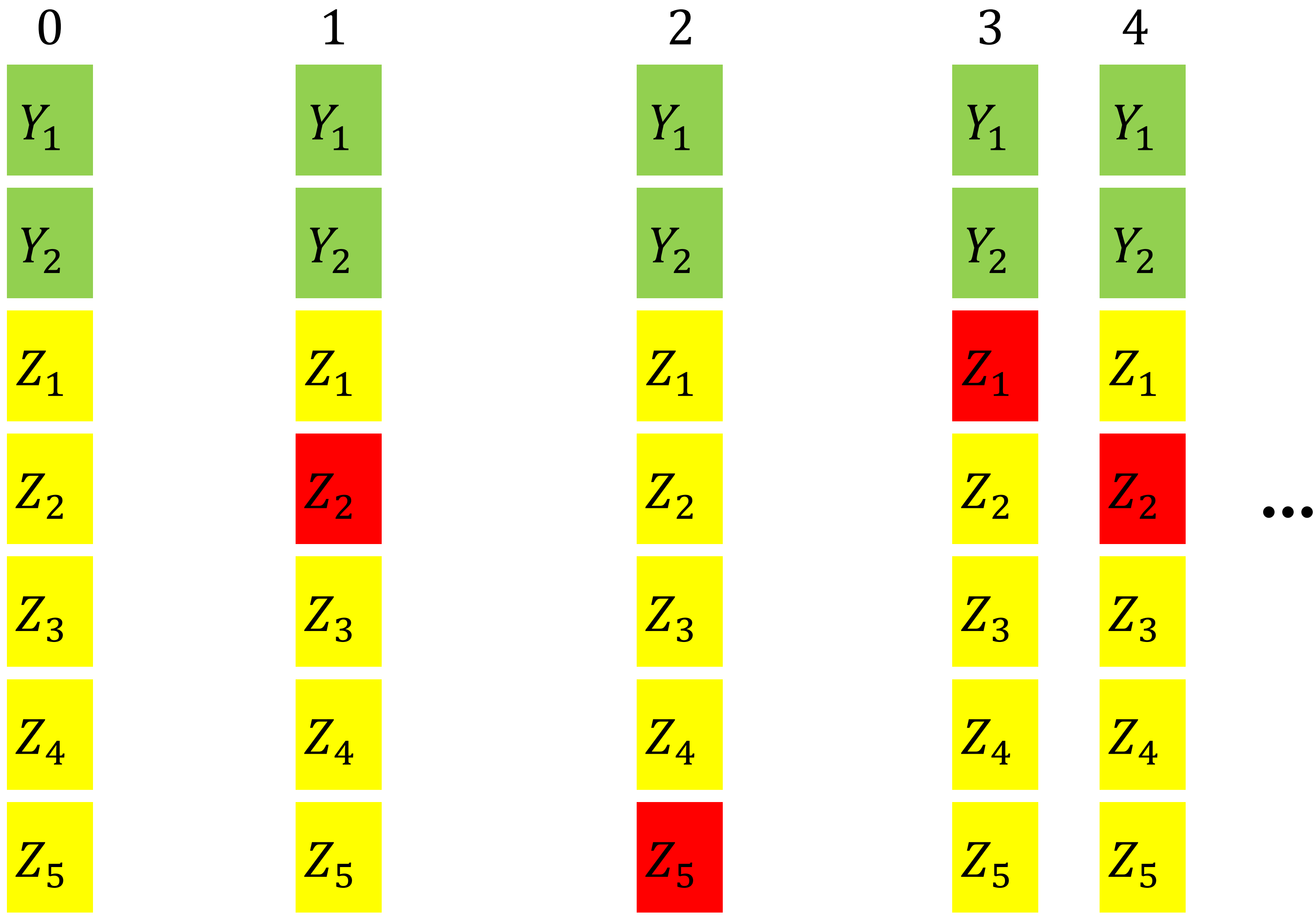
Inference: MCMC

- Probabilistic model:
  - Joint distribution $P(x)$ over $X = (X_1, X_2, \dots, X_n) = (Z, Y)$
  - $Y = (Y_1, \dots, Y_a)$: observed variables
  - $Z = (Z_1, \dots, Z_b)$: unobserved/latent variables
- Goal:
  - Obtain samples from $P(Z|Y = y)$

Gibbs sampling

- Markov chain Q:
  - State of Q is a variable assignment Z
  - Pick K uniformly from $\{1, \ldots, b\}$ (or cycle through)
  - Sample Z_K from $P(Z_K | Z_1, Z_2, \ldots, Z_{K-1}, Z_{K+1}, \ldots, Z_b, Y = y)$
  - Repeat
- Possible transition in Q:
  - Def: $z' \sim_k z$ if $z' = (z_1, z_2, \ldots, z_{K-1}, *, z_{K+1}, \ldots, z_b)$,i.e., equal to z except at position k
  - Transition $z \rightarrow z'$ only possible for $z' \sim_k z$ for some k
- Transition matrix of $Q(z, z') =$
- $= \begin{cases} \dfrac{P(Z = z' | Y = y)}{b \sum_{z'' \sim_k z} P(Z = z'' | Y = y)} & z' \sim_k z \\ 0 & otherwise \end{cases}$

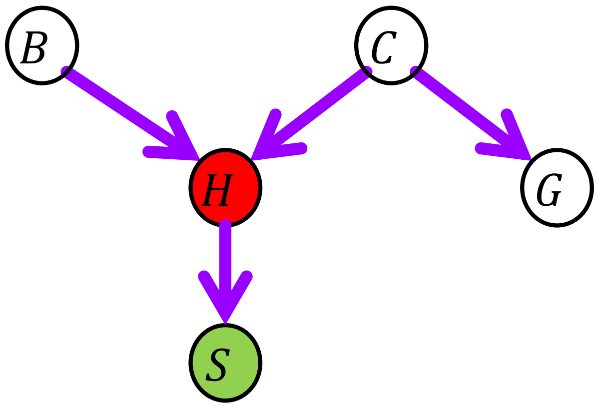
Gibbs sampling: illustration



$Z(1) \sim_2 Z(0)$ $\quad$ $Z(2) \sim_5 Z(1)$

Gibbs sampling for BNs: example

- Resampling variable H conditional on S
- $P(H|B,C,G,S) =$
- $= \frac{P(H,B,C,G,S)}{P(B,C,G,S)} =$
- $= \frac{P(H,B,C,G,S)}{\sum_H P(H,B,C,G,S)} =$
- $= \frac{P(B)P(C)P(H|B,C)P(G|C)P(S|H)}{\sum_{H'} P(B)P(C)P(H'|B,C)P(G|C)P(S|H')} =$
- $= \frac{P(H|B,C)P(S|H)}{\sum_{H'} P(H'|B,C)P(S|H')}$



Sampling from a variable only involves factors (CPDs) "touched" by this variable!

Gibbs sampling

- Claim:
  - Q is a reversible MC with stationary distribution
$$\pi(z) = P(Z = z|Y = y)$$
  - Interpretation: run the MC Q and collect large # of samples of $Z|Y = y$, then compute whatever statistic needed: mean, moments, confidence intervals, etc.
  - But: samples are correlated!
- Reminder:
  - An ergodic MC (irreducible, aperiodic, pos-recurrent) MC has a single stationary distribution π
  - Ergodic theorem: temporal averages → ensemble expectations
  - Reversible MC: if Q is ergodic and we can find a $\pi(.)$ such that for all z, z', $\pi(z)Q(z, z') = \pi(z')Q(z', z)$, then $\pi(.)$ is the stationary distribution

Gibbs sampling: proof

- Proof:
  - $\pi(z)Q(z,z') =$
  - $= P(Z=z|y)Q(z,z') =$
  - $= \frac{P(Z=z|y)P(Z=z'|y)}{b\sum_{z''\sim_k z} P(Z=z''|y)} =$
  - $= \frac{P(Z=z'|y)P(Z=z|y)}{b\sum_{z''\sim_k z'} P(Z=z''|y)} =$
  - $= P(Z=z'|y)Q(z',z) =$
  - $= \pi(z')Q(z',z)$

Note: z and z' only differ at position k; therefore,
$z''\sim_k z \Leftrightarrow z''\sim_k z'$

Detailed balance equations
→ global balance equations
→ $\pi(z)$ is stationary distrib. of MC Q

Bayesian Network: key ideas

- Two functions:
  - Compact representation for a set of conditional independence assumptions among RVs
  - A data structure to encode a joint distribution compactly through its factors
- Flexibility: model does not specify observables
- Example: 100 binary RVs
  - Full joint distribution: $2^{100} - 1 \sim 10^{30}$ values
  - All independent: 100 values, but very limiting
  - In practice, much closer to «everything independent» than to «full joint distribution»
    - Tradeoff: compact representation & efficient inference, but still capture main dependencies
- Next week: topic models using graphical models

References

- [D. Koller, N. Friedman: Probabilistic Graphical Models, MIT Press, 2009]
- [Ch. D. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval, Cambridge, 2008]
- [C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006]