CS 583: Probabilistic Graphical Models

Topic: Template-based representations

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Motivation

- We have so far assumed that the set of variables $\mathcal{X}$ is fixed
- In some domains, the size of $\mathcal{X}$ can depend on the input
- Examples
  - Text, video, social networks, …
- How can we specify the structure and the parameters for $\mathcal{X}$, where $\mathcal{X}$ depends on the input?

Template-based models

- We would like to specify a template model (both structure and parameters) where
  - The structure can be built automatically by using the template on the input
  - The parameters can be reused shared

Template-based model examples

- Temporal models
- Plate models
- Relational models

Temporal models

- The state of the world, represented by a set of variables, evolves over time
  - $\mathcal{X}^{(0)}, \mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(T)}$
- Each possible world is a trajectory: an assignment to $\mathcal{X}^{(0:T)}$
- Examples
  - Dynamic Bayesian networks
  - State observation models
  - Hidden Markov models
  - Kalman filters

Temporal models

$$P\left(\mathcal{X}^{(0:T)}\right) = P\left(\mathcal{X}^{(0)}\right)\prod_{t=0}^{T-1} P\left(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(0:t)}\right)$$

Nothing magic; just the chain rule.

Markov assumption

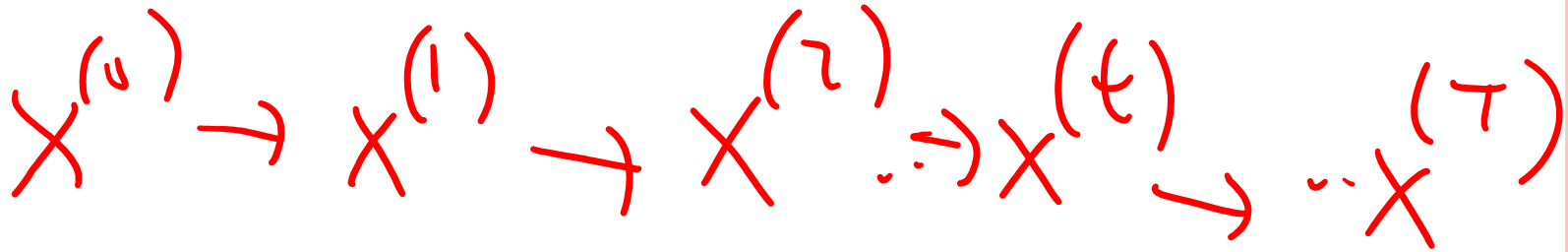
$$\mathcal{X}^{(t+1)} \perp \mathcal{X}^{(0:t-1)} \mid \mathcal{X}^{(t)}$$

Is this a reasonable assumption?

$X^{(0)}, X^{(1)}, X^{(2)}, \ldots X^{(t)}, \ldots X^{(T)}$

Temporal models

$$P\left(\mathcal{X}^{(0:T)}\right) = P\left(\mathcal{X}^{(0)}\right)\prod_{t=0}^{T-1} P\left(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(t)}\right)$$



STATIONARY

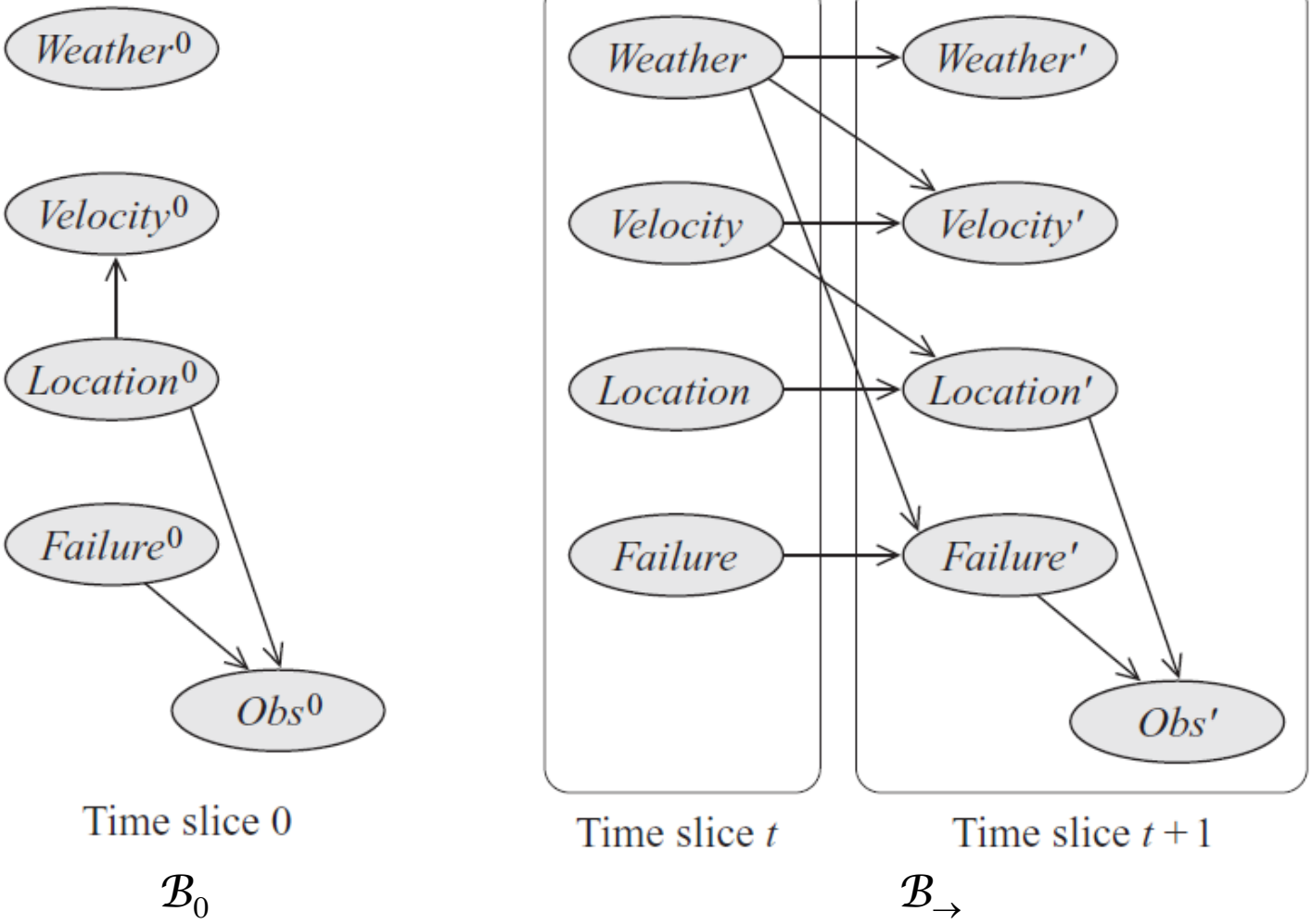
- We still need to specify the CPD for all variables in $\mathcal{X}^{(0:T)}$
- A Markovian dynamic system is *stationary* (*time-invariant*, *homogeneous*) if $P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(t)})$ is same for all t. Then, the process can be represented using a *transition model*:

$$P\left(\mathcal{X}^{(t+1)} = \xi' \mid \mathcal{X}^{(t)} = \xi\right) = P\left(\mathcal{X}' = \xi' \mid \mathcal{X} = \xi\right)$$

Dynamic bayesian networks

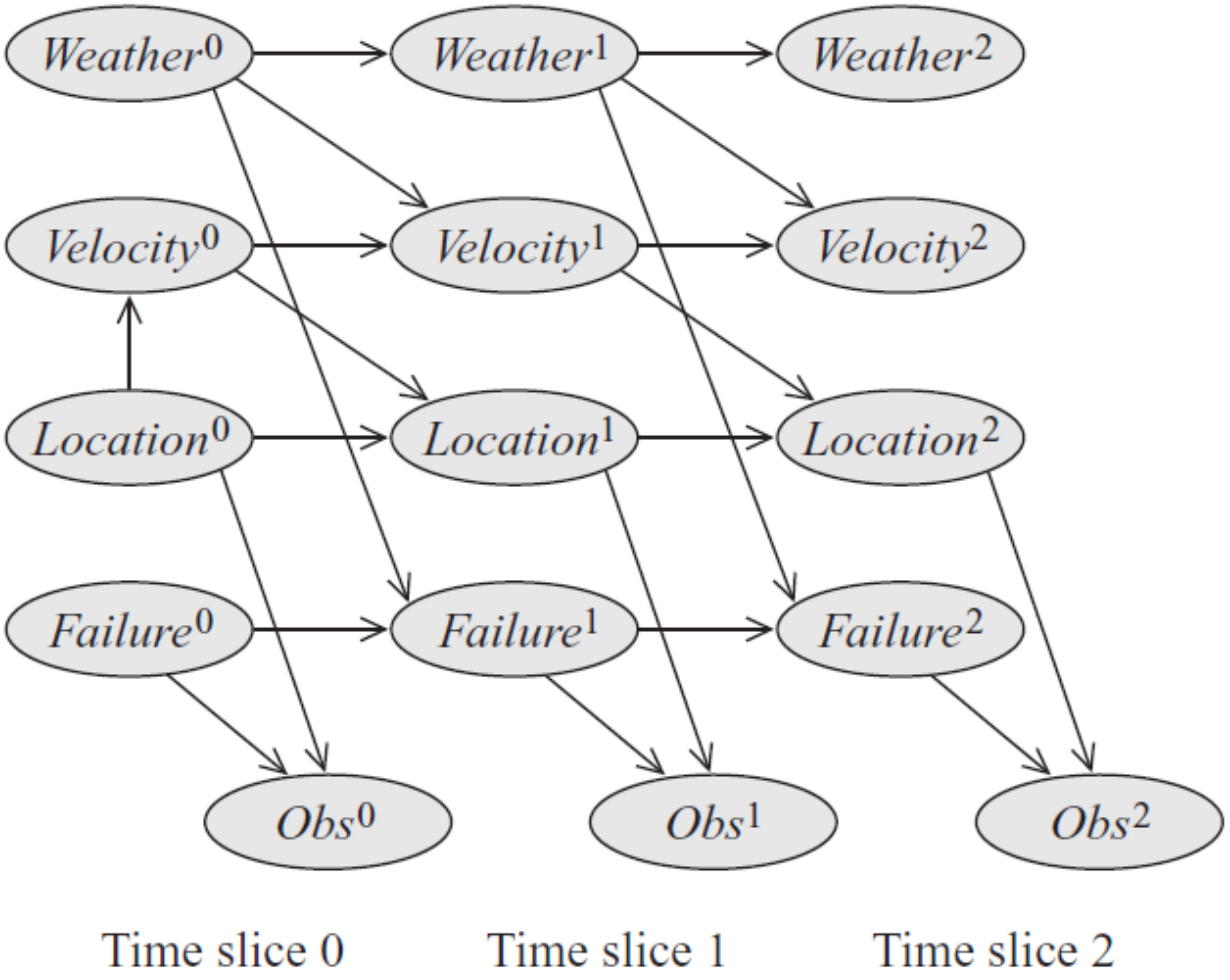
- Definition: A dynamic Bayesian network (DBN) is a pair $<\mathcal{B}_0, \mathcal{B}_{\rightarrow}>$, where $\mathcal{B}_0$ is a Bayesian network over $\mathcal{X}^{(0)}$, representing the initial distribution over states, and $\mathcal{B}_{\rightarrow}$ is a 2-time-slice Bayesian network for the process.

DBN EXAMPLE



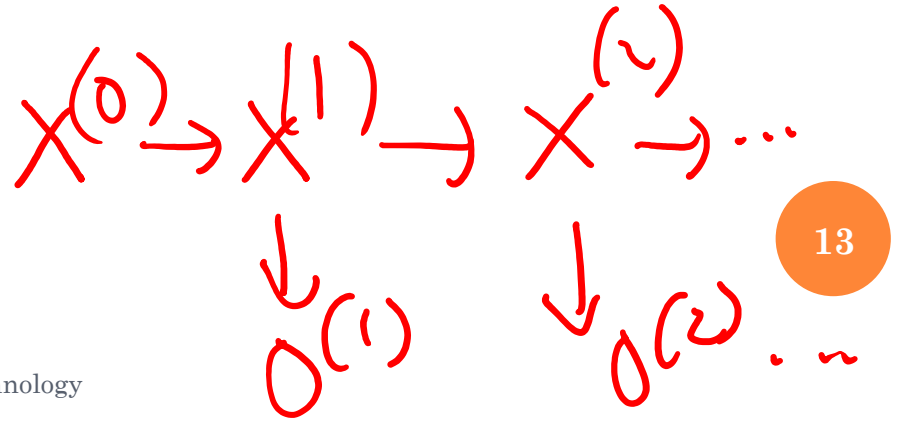
$\mathcal{B}_0$ $\mathcal{B}_{\rightarrow}$

DBN EXAMPLE


Weather0
Weather1
Weather2
Velocity0
Velocity1
Velocity2
Location0
Location1
Location2
Failure0
Failure1
Failure2
Obs0
Obs1
Obs2
Time slice 0
Time slice 1
Time slice 2

State observation models

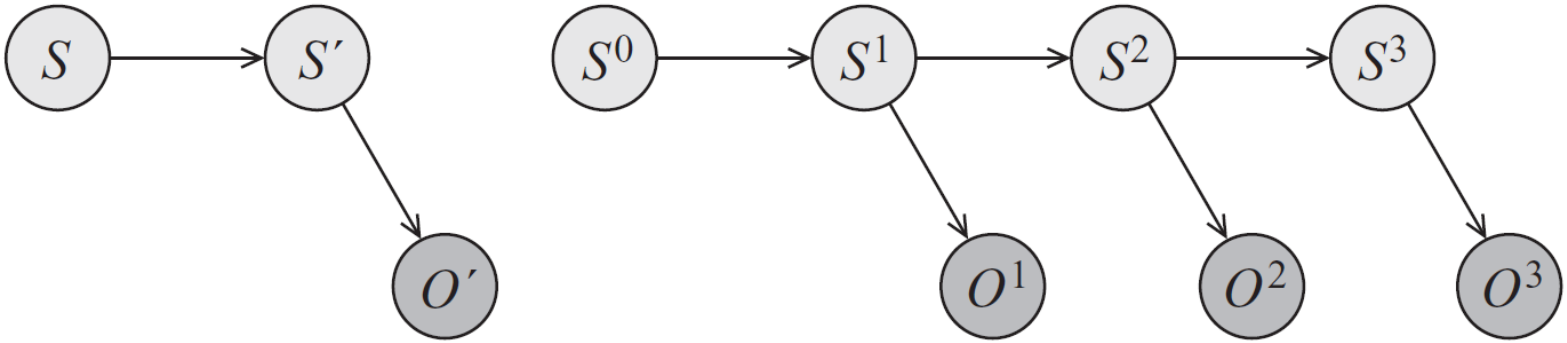
- System evolves on its own, with our observations occurring in a separate process
  - Our observations do not change the system
- Observations are obtained from a noisy sensor
- Two key assumptions
  - States are Markovian: $\boldsymbol{X}^{(t+1)} \perp \boldsymbol{X}^{(0:t-1)} \mid \boldsymbol{X}^{(t)}$
  - Observations at t, $\boldsymbol{O}^{(t)}$, are conditionally independent of the entire state sequence given the state at t, $\boldsymbol{X}^{(t)}$:
    - $\boldsymbol{O}^{(t)} \perp \boldsymbol{X}^{(0:(t-1))}, \boldsymbol{X}^{(t+1:\infty)} \mid \boldsymbol{X}^{(t)}$



HIDDEN MARKOV MODELS

$P(S^{t+1} \mid S^t)$

$P(S^0)$



(a) (b)

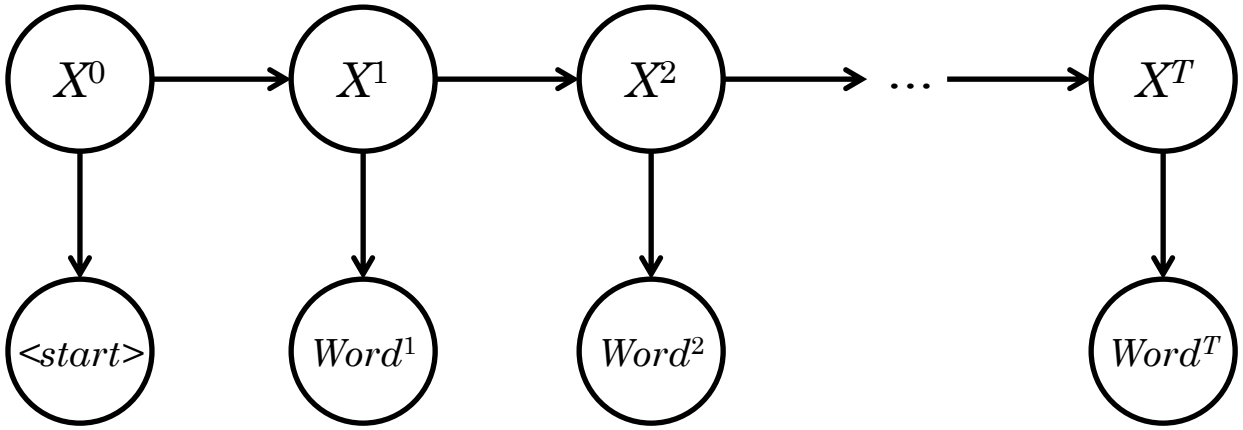
$P(O^t \mid S^t)$

Part-of-speech tagging

- Task: Given a sentence, identify, for each word, the part of speech
  - E.g., noun, verb, adjective, adverb
- Part-of-speech depends on the context
  - E.g., "can": modal or noun?
- Nearby words provide essential information
  - N + "can": more likely to be a modal
  - "The" + "can": more likely to be a noun

Part-of-speech tagging using HMMs

- States (***X***): Part-of-speech roles of the words
- Observations (***O***): The words themselves



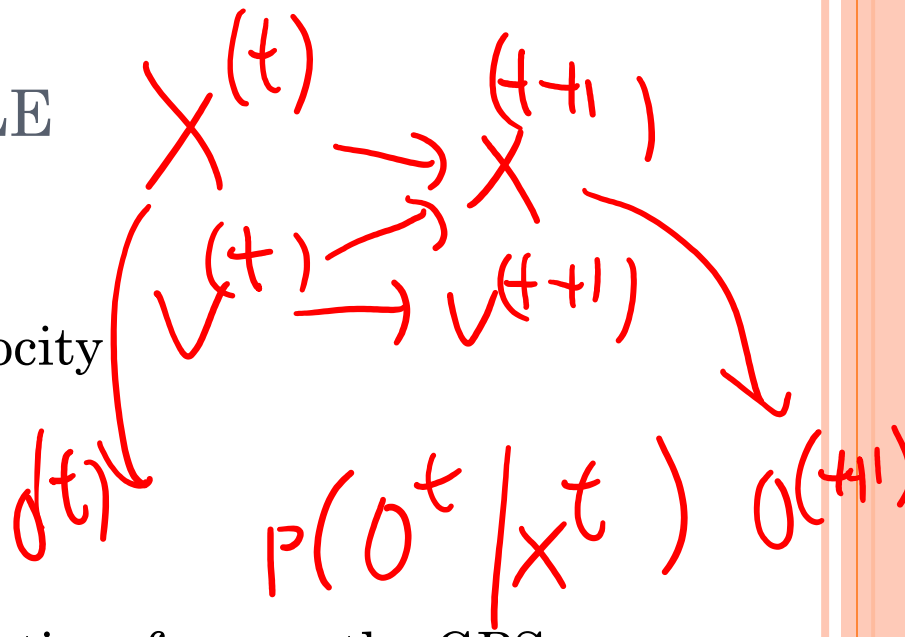
A Great Read

- "Speech and Language Processing" book by Jurafsky and Martin
- The full book is freely available at https://web.stanford.edu/~jurafsky/slp3/
- Chapter 8: Sequence Labeling for Parts of Speech and Named Entities
  - https://web.stanford.edu/~jurafsky/slp3/8.pdf
  - Introduces and discusses HMMs and CRFs

Kalman Filters

- A linear dynamical system
- Real-valued variables evolve linearly over time, with some Gaussian noise
- Can be considered a specific kind of dynamic Bayesian network where
  - All variables are continuous
  - All dependencies are linear Gaussian
- Often used for tracking systems, for example tracking airplanes using radar data

Kalman Filter Example



- Tracking a vehicle's position
- X: current position, V: current velocity
- $P(X'|X,V) = X + V\Delta + \mathcal{N}(0;\sigma_X^2)$
- $P(V'|V) = V + \mathcal{N}(0;\sigma_V^2)$
- Often, there is also a noisy observation; for e.g., the GPS signal as a noisy measurement of X.
- How would you model the above system using a dynamic Bayesian network?

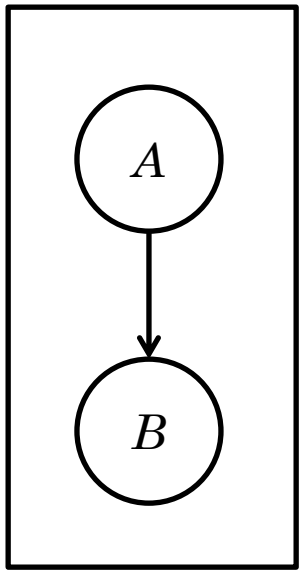
PLATE MODELS

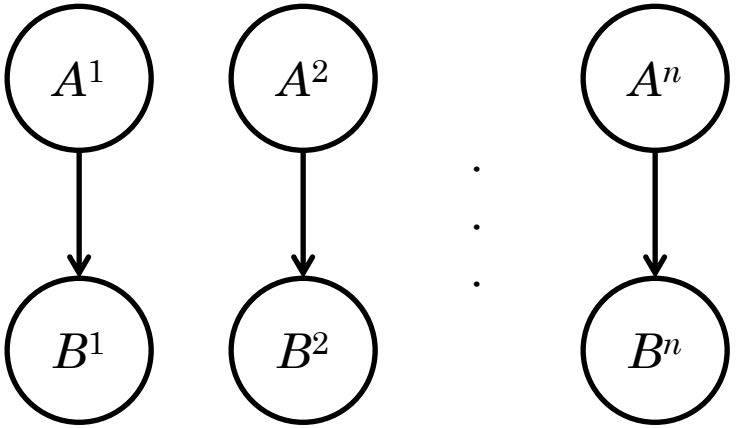
- **Motivation**: Assume we have the following simple network
  - $D \rightarrow G \leftarrow I$
- There are multiple students in a given class, say CS101
- Given the grade of John, we would like to reason about the grade distribution of other students
  - For e.g., if John got an A, we would like to be able to reason that CS101 might be an easy class and thus other students in CS101 have a higher chance of getting As as well

Plate models - informally

- A plate is a rectangle with a Bayesian network inside
- The rectangle tells us to replicate the same structure *and* the parameters multiple times
- Let's see some examples

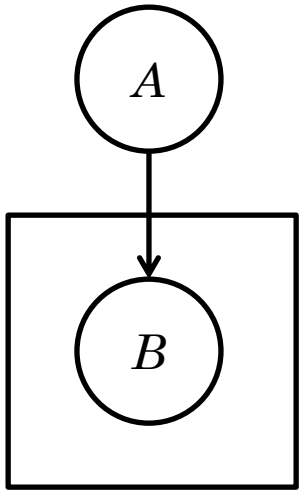
EXAMPLE – ALL INSIDE

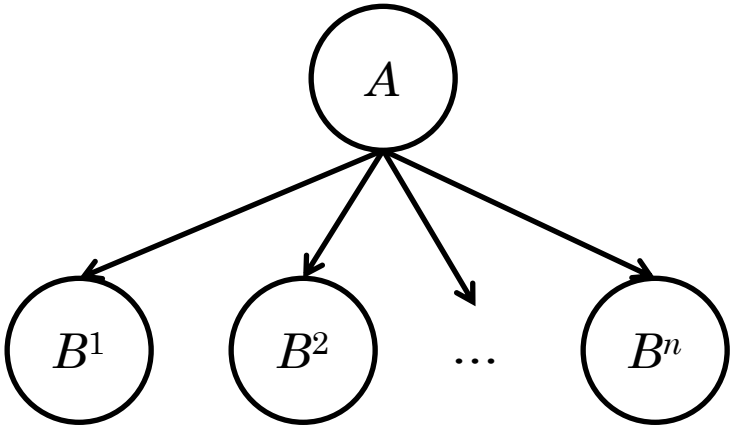




The marginal P(*A*) and the conditional P(*B*|*A*) are copied for each replica.

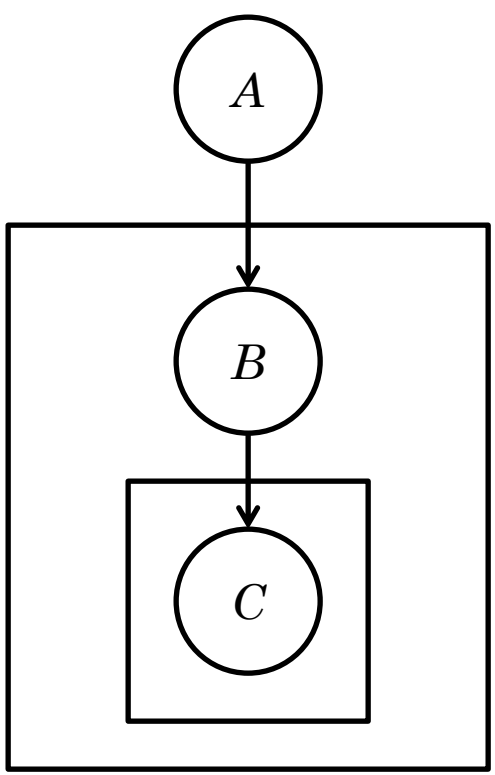
EXAMPLE – SOME OUTSIDE

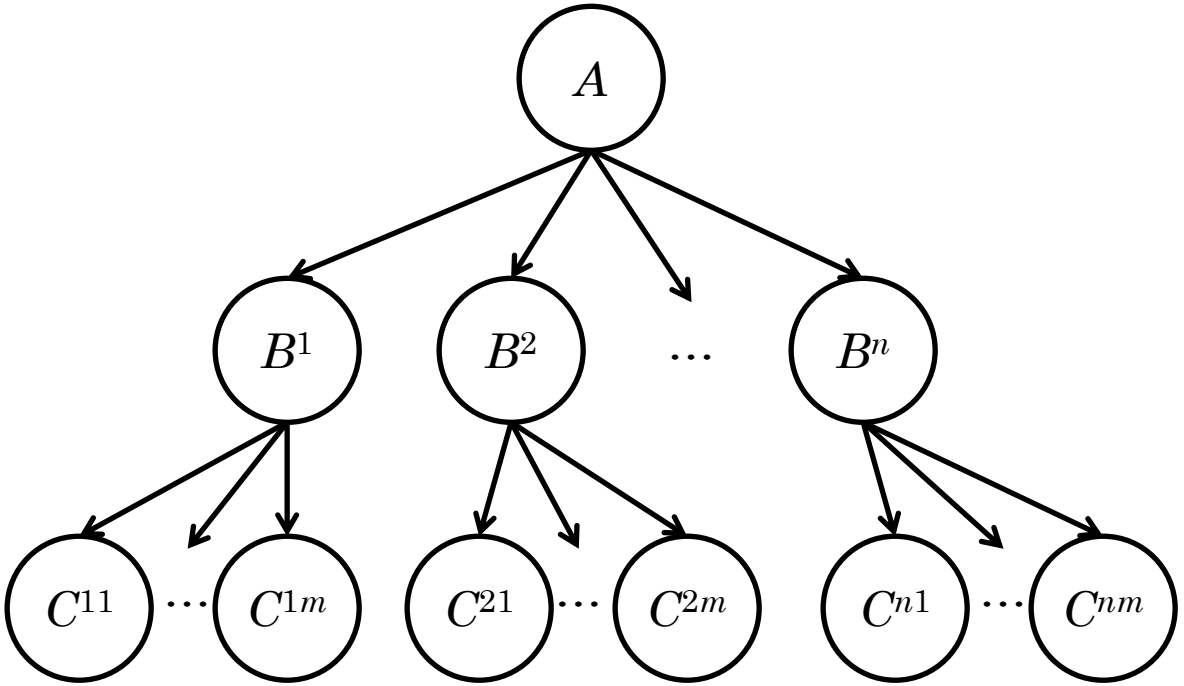




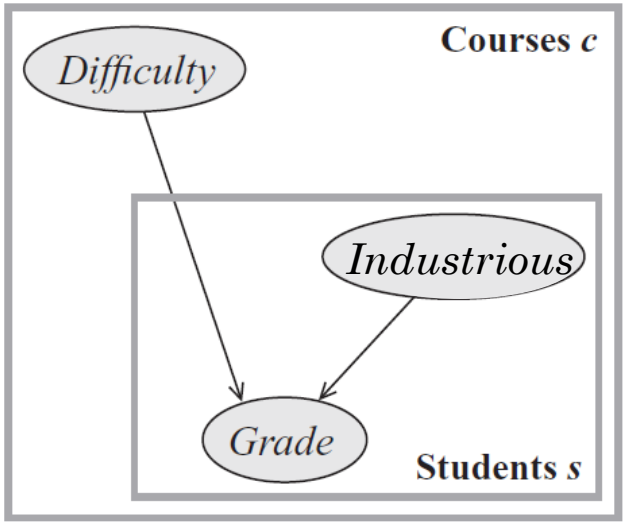
The conditional P(B|A) is copied for each replica.

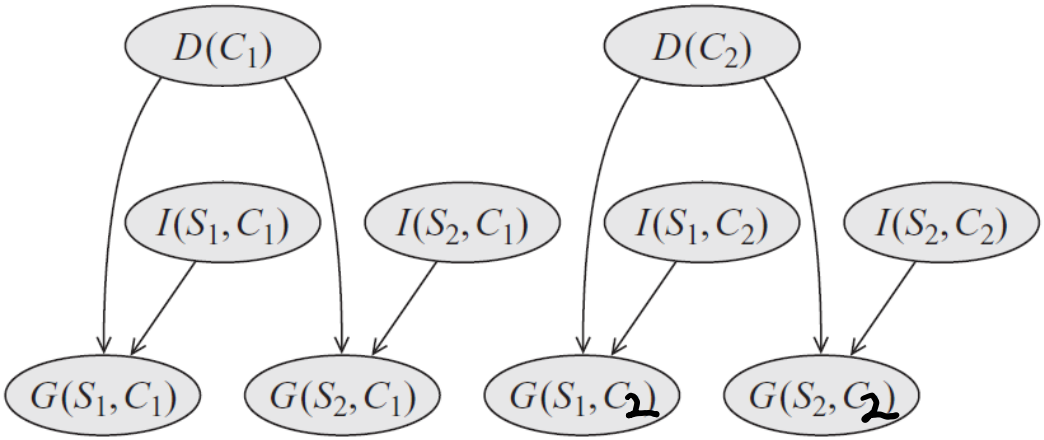
EXAMPLE – NESTED





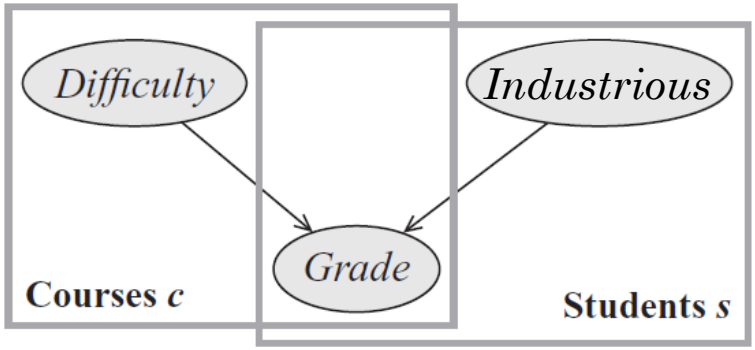
Multiple courses with each its own set of students

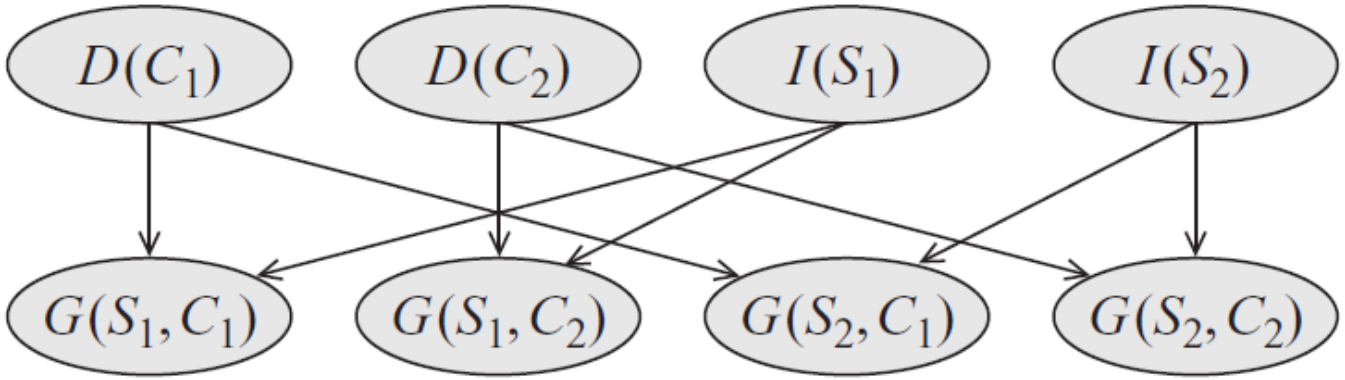




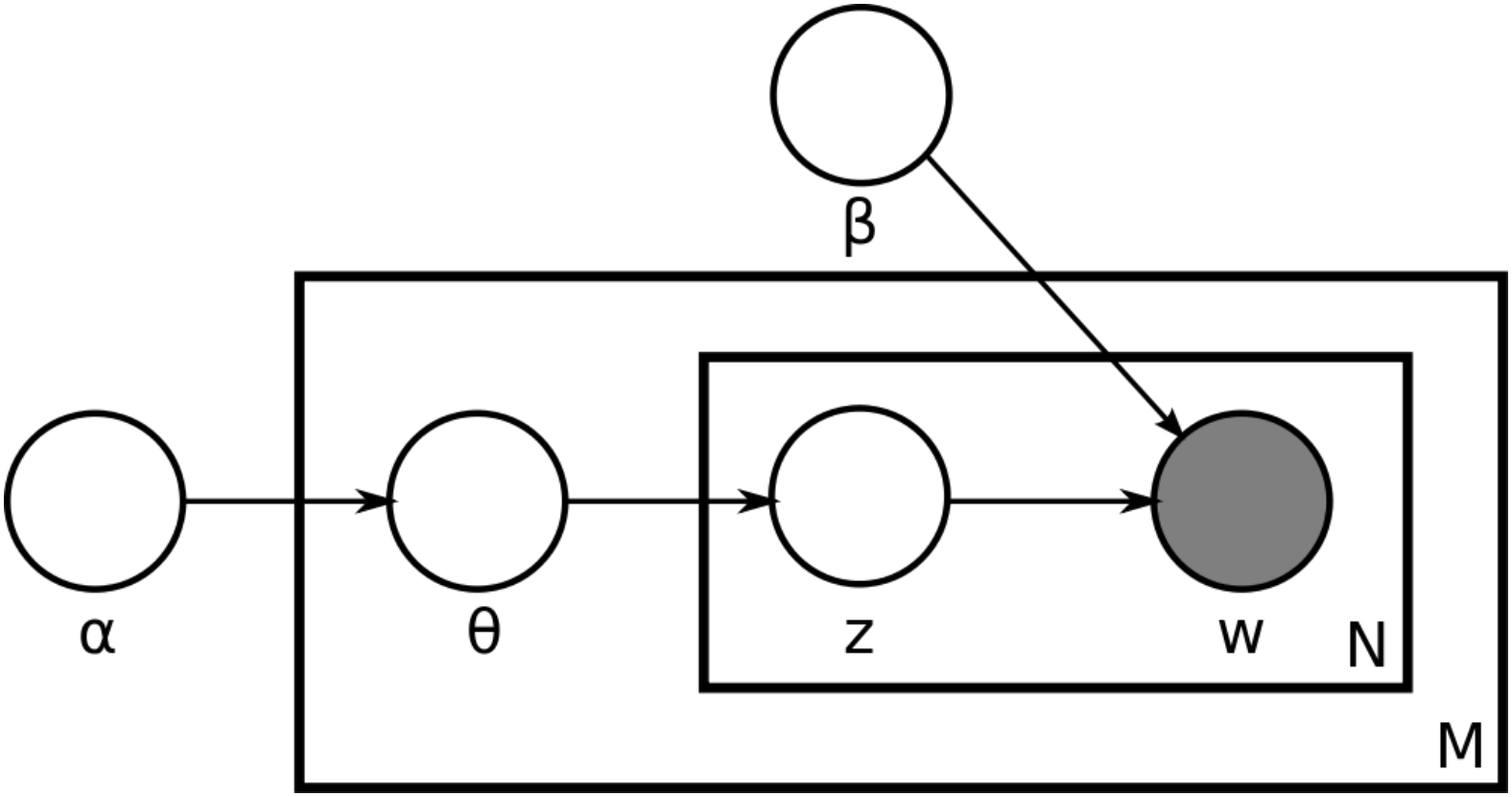
We can reason across students in the same course, but we cannot reason across courses through common students, because the student variables are replicated across courses.

Multiple courses and multiple students





Latent Dirichlet Allocation



https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

Limitations of plate models

- The grade is replicated for all possible combinations of courses and students. What happens when a student takes some classes but not others?
- How can you represent a variable of one object depends on a variable of another object of the same type?
- How can you represent temporal models?

Probabilistic relational models

- Specify when a dependency holds
  - For e.g., the dependency between a course variable and a student variable is contingent upon whether the student registered that course
- Largely based on database concepts
  - Entities and relationships

Statistical relational learning (SRL)

- Last decade has seen tremendous amount of work on probabilistic relational modeling
- Examples
  - Probabilistic Relational Models (PRM)
  - Relational Markov Networks (RMN)
  - Inductive Logic Programming (ILP)
  - Conditional Random Fields (CRF)
  - Relational Dependency Networks (RDN)
  - Bayesian Logic Programming (BLP)
  - Markov Logic Networks (MLN)
- http://www.cs.umd.edu/srl-book/