Bayesian Networks

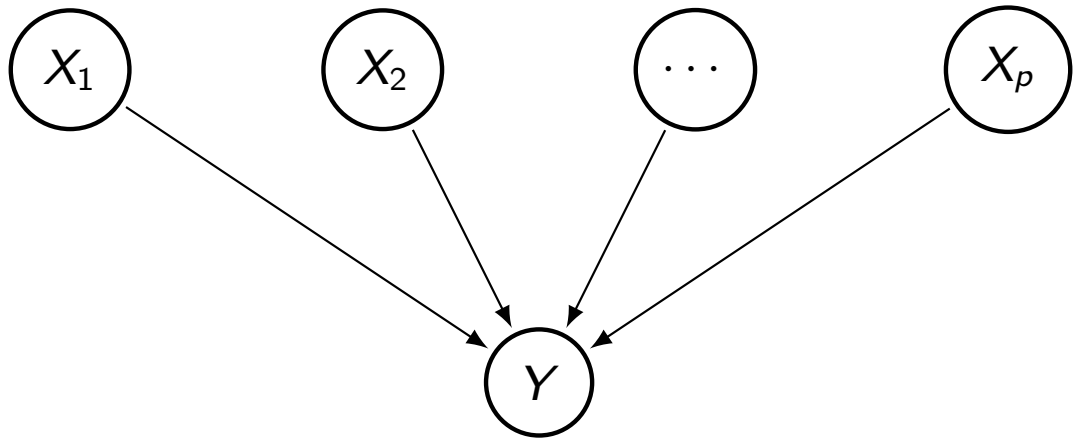
Probabilistic Graphical Models

- 확률 이론은 현대 패턴 인식에서 가장 중심이 되는 이론이다.
- 그래프 도식 방법을 이용하여 확률 분포를 해석하는 방법을 probabilistic graphical model 이라고 부른다.

$$\text{Graph} + \text{Probability} \rightarrow \text{Graphical Model}.$$

- 장점
  - 확률 모델의 구조를 아주 쉽게 시각화 하여 새로운 모델을 디자인하는데 도움을 준다.
  - 그래프화된 구조를 분석함으로써 모델 속성에 대한 직관을 얻을 수 있다.
  - 조건부 분포의 독립 속성 들을 그래프를 통해 손쉽게 파악할 수 있음.
  - 복잡한 학습과 추론 과정을 가지는 모델의 계산 과정을 그래픽적인 요소로 표현이 가능하다.
  - 수학적인 요소들을 명시적인 그래프로 표현 가능하다.

Example: Regression

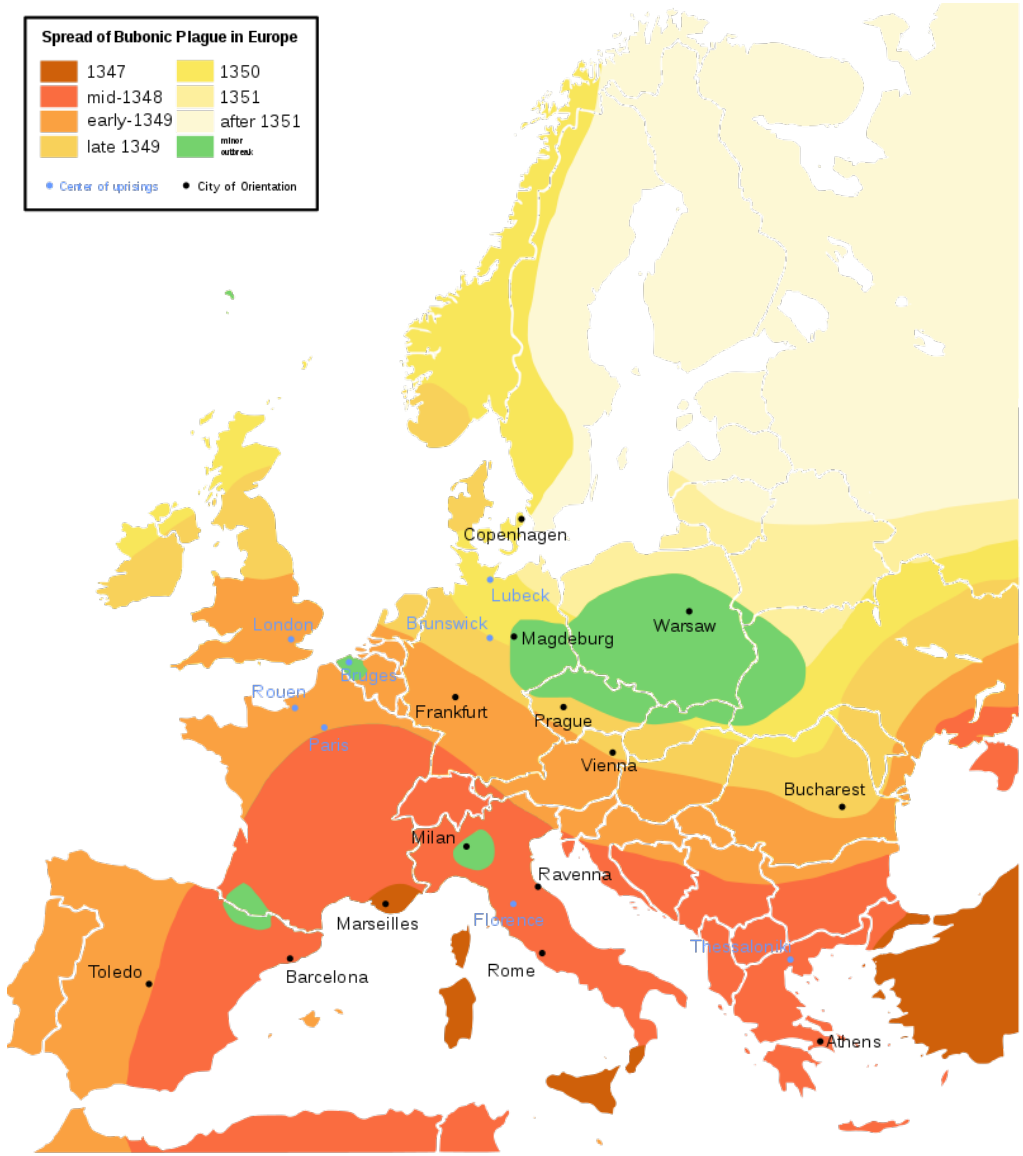


- Linear Model:

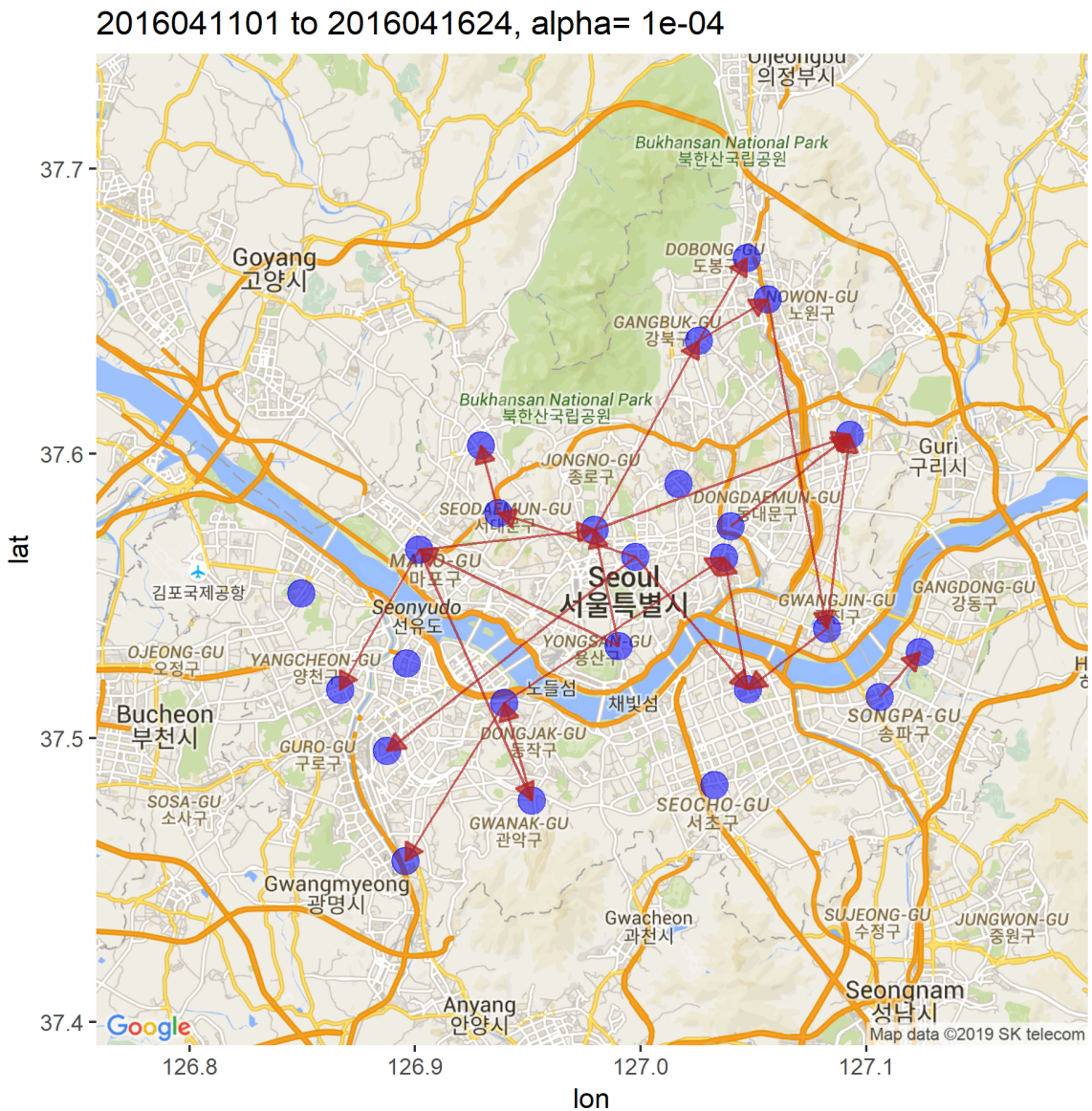
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon.$$

- Variable Selection

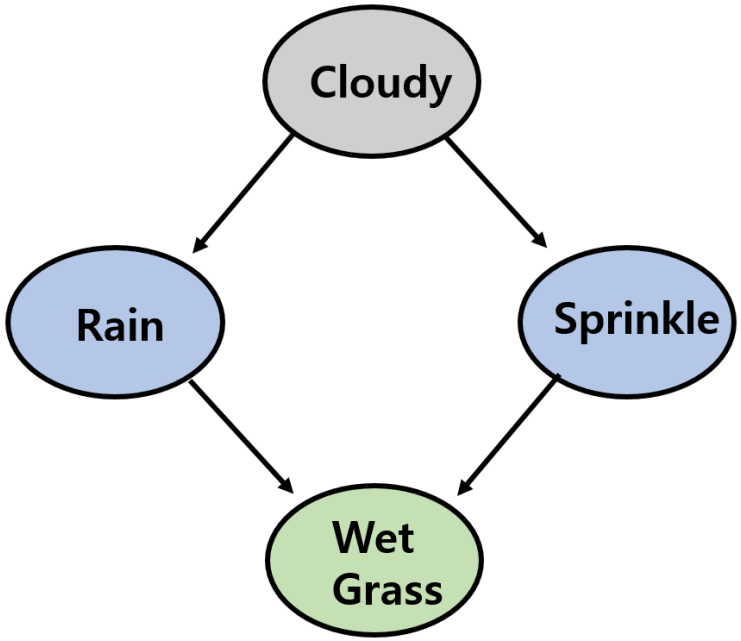
Example: Spatial Analysis I



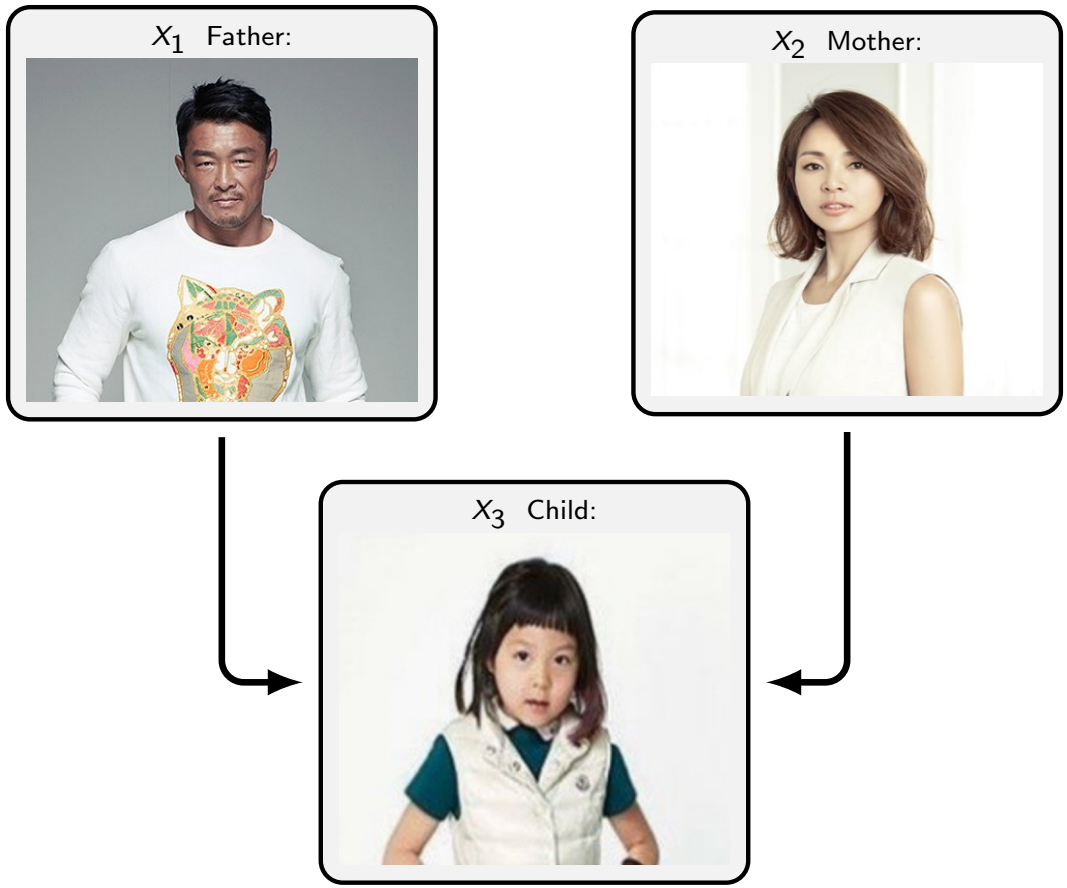
Example: Spatial Analysis II



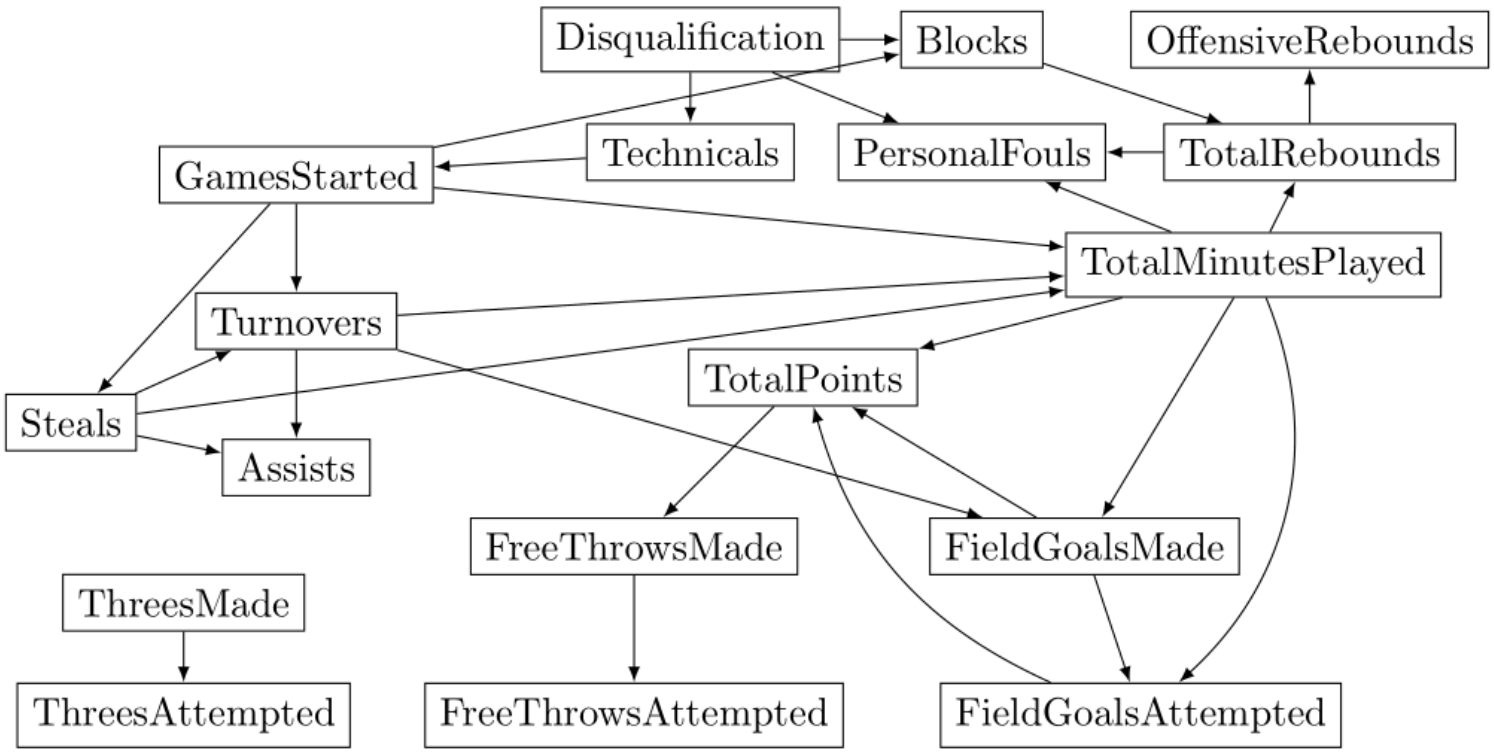
Example: Causal Inference I



Example: Causal Inference II



Example: Directional Dependency Modeling



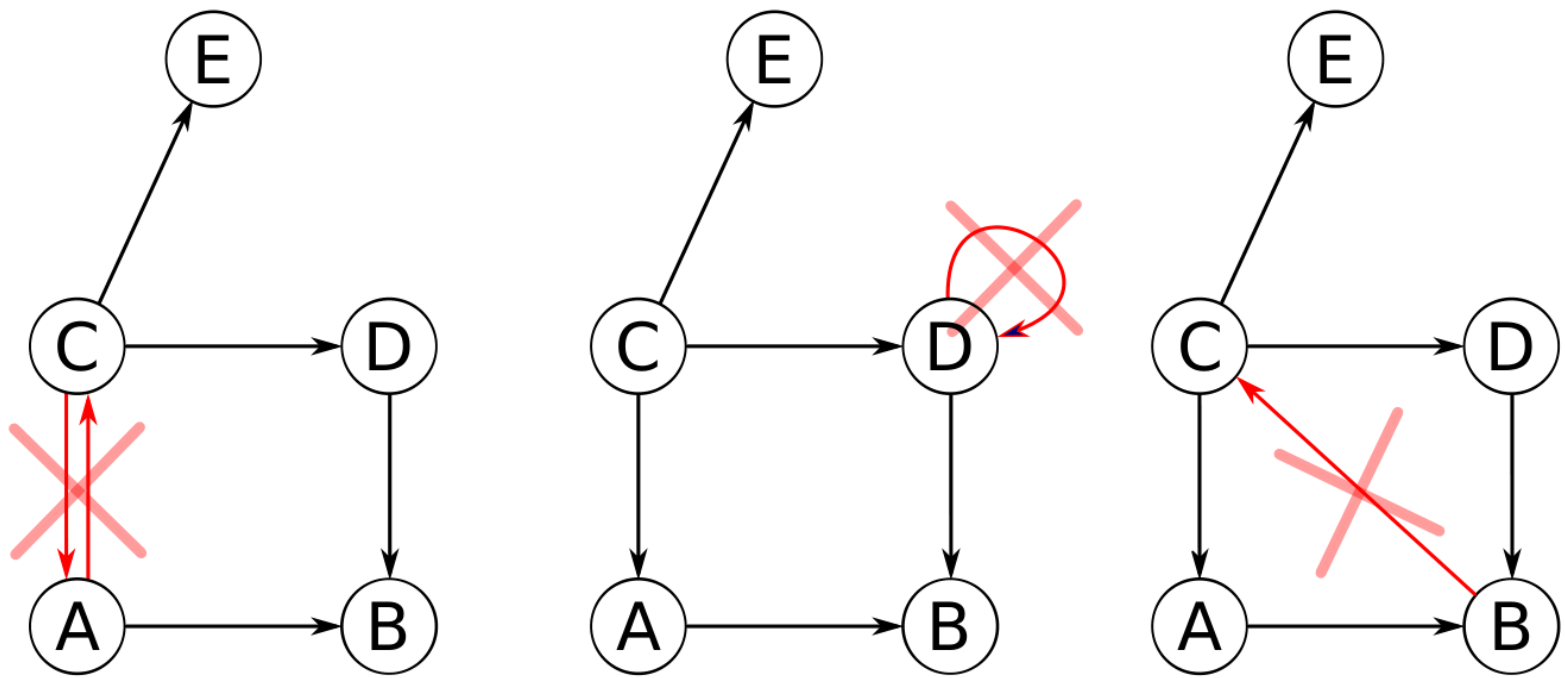
Directed Acyclic Graph (DAG)

- Directed acyclic graph G:

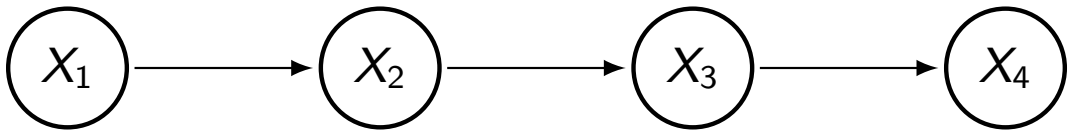
  - $G = (V, E)$.
  - V : a set of nodes, e.g. $V = \{1, 2, 3\}$.
  - E : a set of directed edges, $E = \{(1, 2), (2, 3)\}$.

Directed Acyclic Graphs

- contains only directed edges;
- does not contain any loop (e.g. an edge $v_i \rightarrow v_i$ from a node to itself);
- does not contain any cycle (e.g. a sequence of edges $v_i \rightarrow v_j \rightarrow ... \rightarrow v_k \rightarrow v_i$ that starts and ends in the same node).



Notations for Directed Graphical Models



- $X_{Pa(j)}$: A parents set of X_j, e.g. $X_{Pa(2)} = \{X_1\}$.
- $X_{Ch(j)}$: A children set of X_j, e.g. $X_{Pa(1)} = \{X_2\}$.
- $X_{De(j)}$: A set of all decendants of X_j, e.g. $X_{De(2)} = \{X_3, X_4\}$
- $X_{Nd(j)}$: A set of non-decendatns of X_j,

  e.g. $X_{Nd(2)} = \{X_1\}$, $X_{Nd(3)} = \{X_1, X_2\}$.
- Ordering: (X_1, X_2, X_3, X_4).

DAG models: Factorization

Factorization (Lauritzen, 1996)

$$f_G(X_1, X_2, ..., X_p) = \prod_{j=1}^{p} f_j \left(X_j \mid X_{\text{pa}(j)} \right),$$

where $f_j(X_j \mid X_{\text{pa}(j)})$ refers to the conditional distribution of node X_j given its parents.

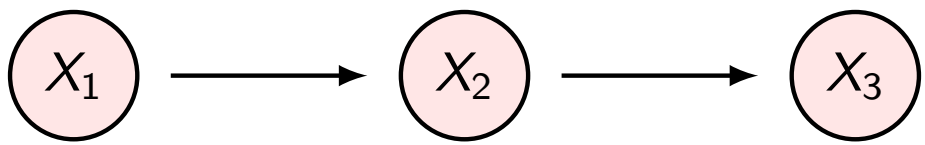


Figure: 3-node Chain Graph with the ordering (X_1, X_2, X_3)

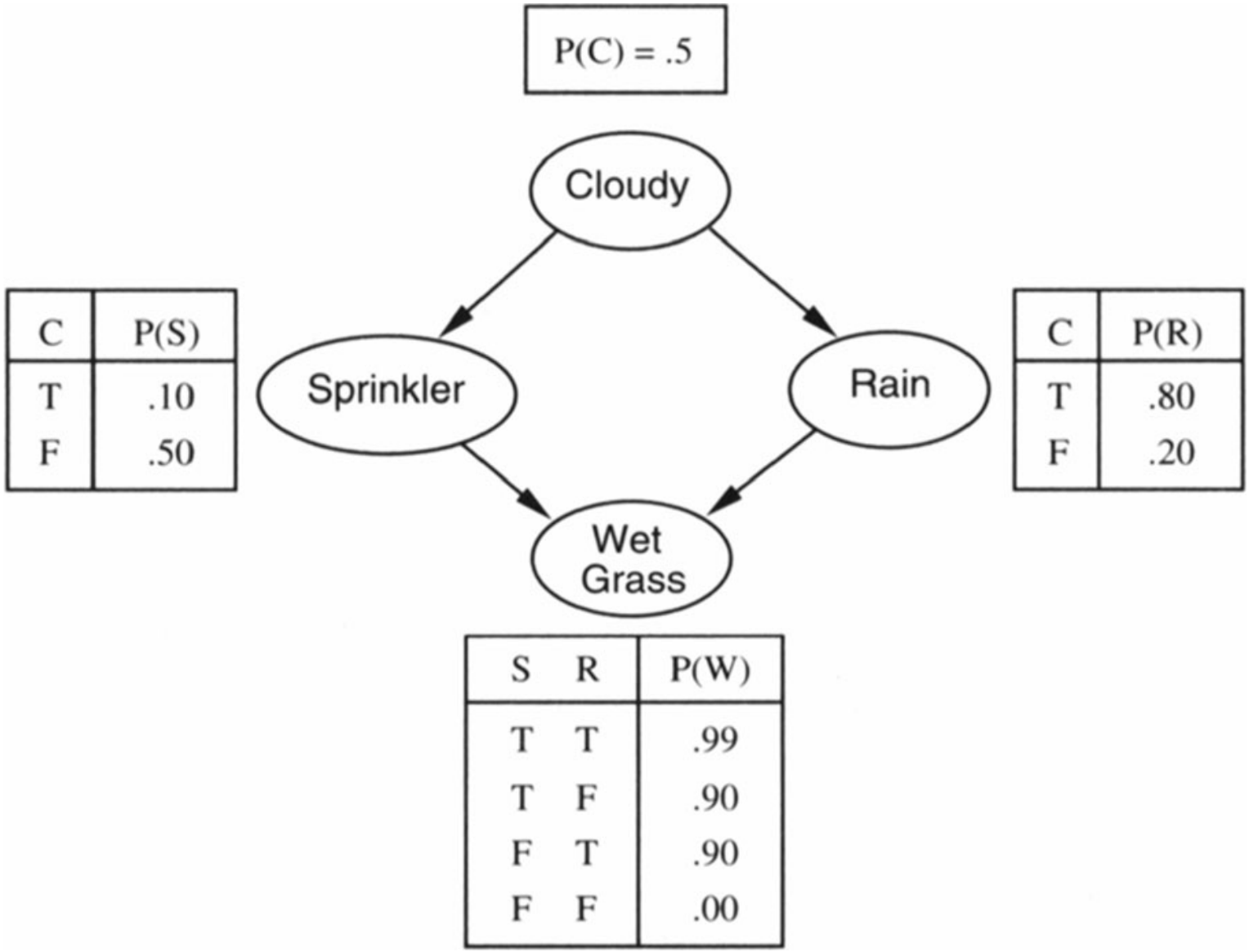
$$f(X_1, X_2, X_3) = f_1(X_1)\ f_2(X_2 \mid X_1)\ f_3(X_3 \mid X_1).$$

Types of DAG Models

The four most common choices in the literature (by far), are:

- Discrete BNs (DBNs), in which X and the $X_j \mid X_{\text{pa}(j)}$ are multinomial;
- Gaussian BNs (GBNs), in which X is multivariate normal and the $X_j \mid X_{\text{pa}(j)}$ are univariate normal;
- Conditional linear Gaussian BNs (CLGBNs), in which X is a mixture of multivariate normals, and $X_j \mid X_{\text{pa}(j)}$ are either multinomial, univariate normal or mixtures of normals.
- Count BNs (CBNs), in which X and the $X_j \mid X_{\text{pa}(j)}$ are Poisson;

Example of Discrete DAG Models



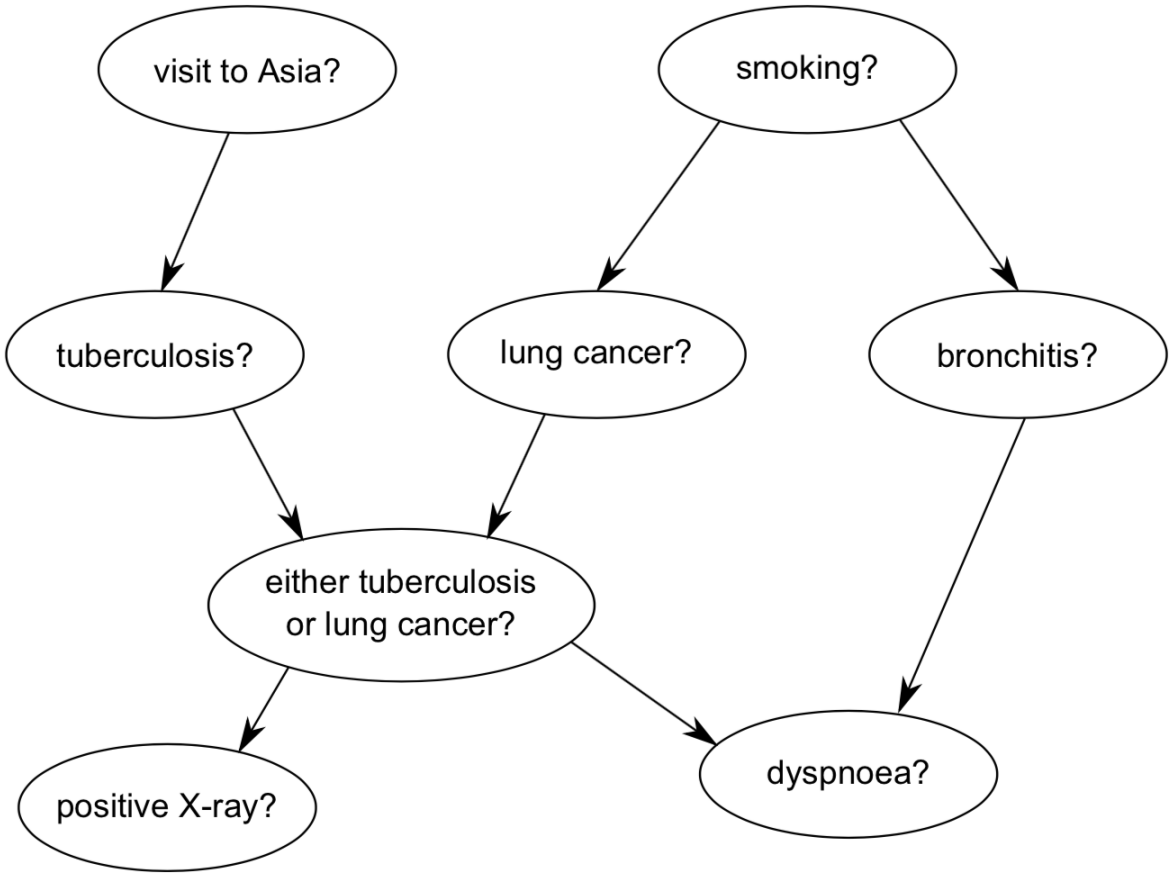
Example of Discrete DAG Models

- The local distributions $X_j \mid X_{\text{pa}(j)}$ take the form of conditional probability tables for each node given all the configurations of the values of its parents.
- Probabilistic inference example: Consider the water sprinkler network, and suppose we observe the fact that the grass is wet. There are two possible causes for this: either it is raining, or the sprinkler is on. Which is more likely?
- We can use Bayes' rule to compute the posterior probability of each explanation. So we see that it is more likely that the grass is wet because it is raining: the ratio is $0.7079/0.4298 = 1.647$.

$$P(S = 1 \mid W = 1) = \frac{P(S = 1, W = 1)}{P(W = 1)} = \frac{\sum P(C, S = 1, R, W = 1)}{P(W = 1)} = \frac{0.2781}{0.6471} = 0.430.$$

$$P(R = 1 \mid W = 1) = \frac{P(R = 1, W = 1)}{P(W = 1)} = \frac{\sum P(C, S, R = 1, W = 1)}{P(W = 1)} = \frac{0.4581}{0.6471} = 0.708$$

Example of Discrete DAG Models: Asia

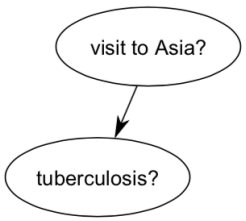


A classic example of DBN is the ASIA network from Lauritzen & Spiegelhalter (1988), which includes a collection of binary variables. It describes a simple diagnostic problem for tuberculosis and lung cancer.

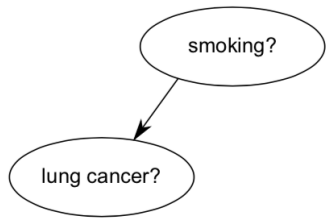
Conditional Probability Tables (CPTs)

yes	no
0.01	0.99

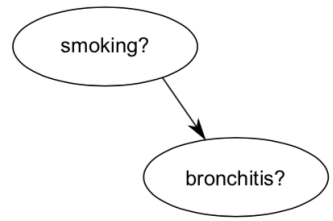
yes	no
0.50	0.50



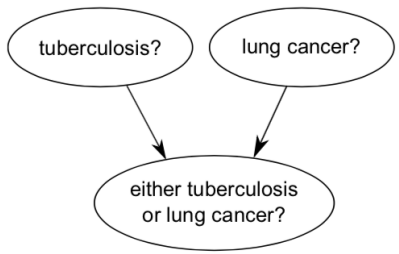
	yes	no
yes	0.05	0.95
no	0.01	0.99



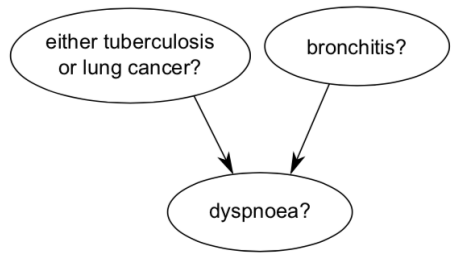
	yes	no
yes	0.10	0.90
no	0.01	0.99



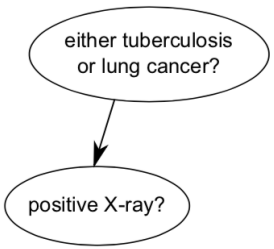
	yes	no
yes	0.60	0.40
no	0.30	0.70



	yes	no
yes:yes	1	0
yes:no	1	0
no:yes	1	0
no:no	0	1



	yes	no
yes:yes	0.90	0.10
yes:no	0.70	0.30
no:yes	0.80	0.20
no:no	0.10	0.90

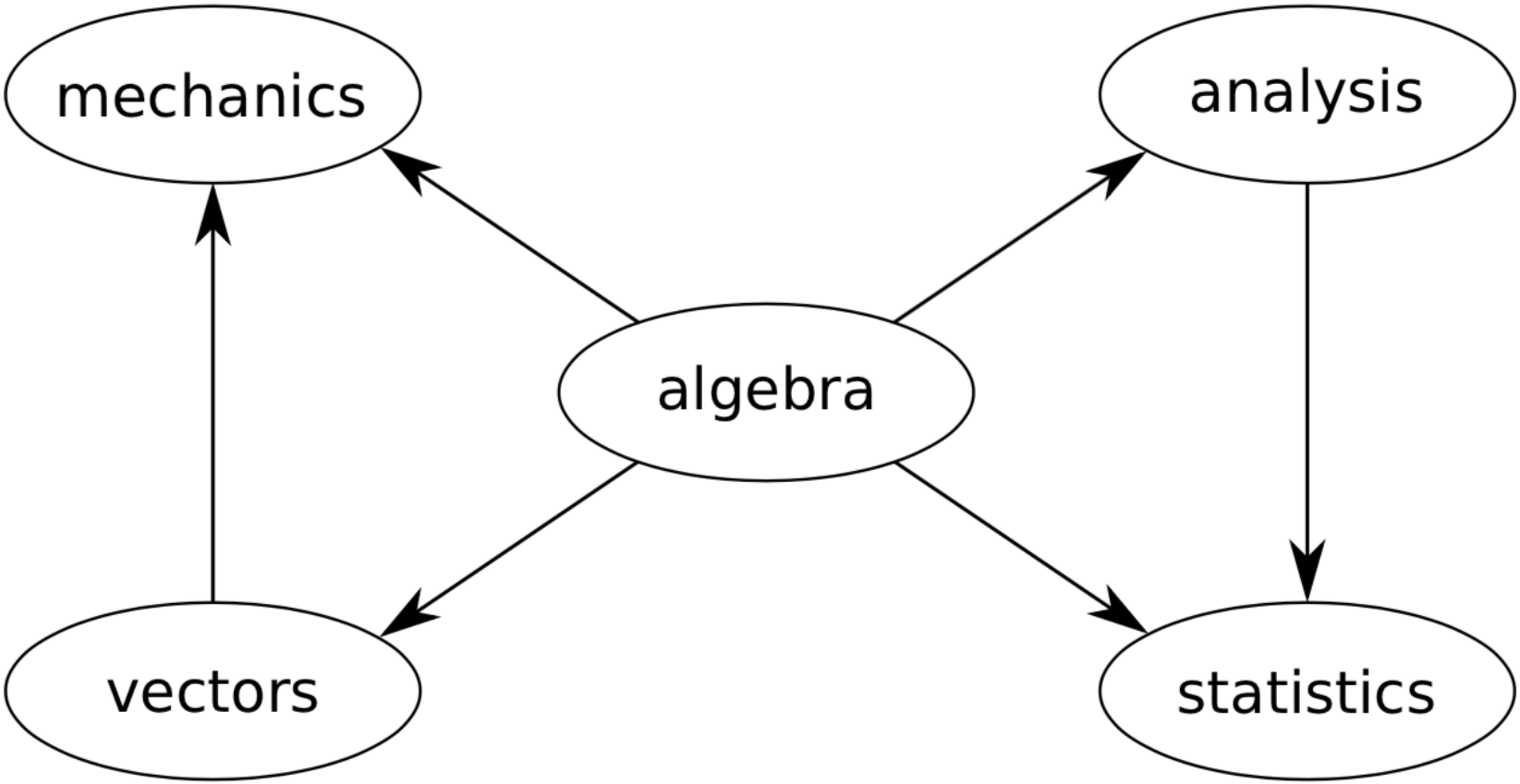


	yes	no
yes	0.98	0.02
no	0.05	0.95

Diagnose

- Asia network could be used to diagnose patients arriving at a clinic.
- Each node in the network corresponds to some condition of the patient, for example, "Visit to Asia" indicates whether the patient recently visited Asia.
- Smoking increases the chances of getting lung cancer and of getting bronchitis.
- Both lung cancer and bronchitis increase the chances of getting dyspnea (shortness of breath).
- Both lung cancer and tuberculosis, but not usually bronchitis, can cause an abnormal lung x-ray.

Gaussian DAG Models



A classic example of GBN is the MARKS networks from Mardia, Kent & Bibby (1979), which describes the relationships between the marks on 5 math-related topics.

Structural Equation Models: Linear Regressions

The local distributions $X_j \mid X_{\text{pa}(j)}$ take the form of linear regression models with the $X_{\text{pa}(j)}$ acting as regressors and with independent error terms.

$$ALG = +50.60 + \varepsilon_{ALG} \sim N(0, 112.8)$$

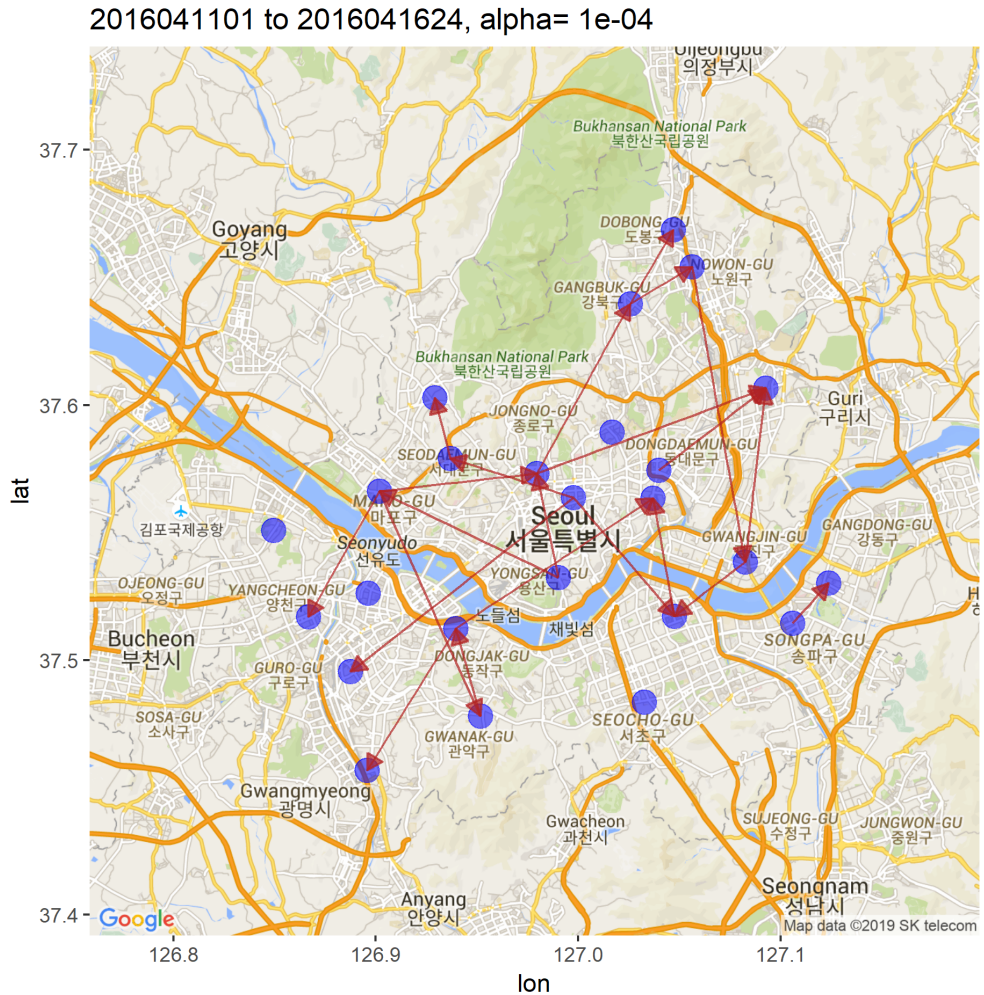
$$ANL = -3.57 + 0.99ALG + \varepsilon_{ANL} \sim N(0, 110.25)$$

$$MECH = -12.36 + 0.54ALG + 0.46VECT + \varepsilon_{MECH} \sim N(0, 195.2)$$

$$STAT = -11.19 + 0.76ALG + 0.31ANL + \varepsilon_{STAT} \sim N(0, 158.8)$$

$$VECT = +12.41 + 0.75ALG + \varepsilon_{VECT} \sim N(0, 109.8)$$

Gaussian DAG Models



A another example of GBN is the Micro Dust Pathway in Seoul, which describes the pathway of PM_{10}.

Structural Equation Models: Linear Regressions

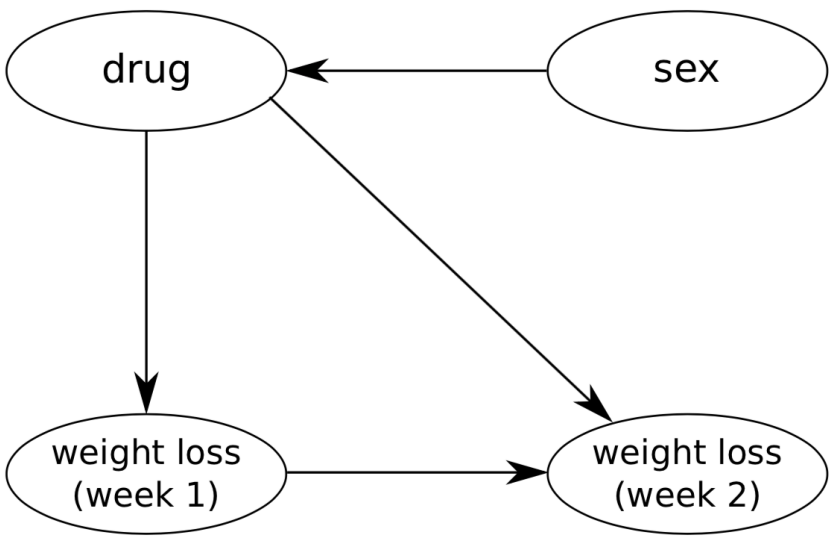
The local distributions $X_j \mid X_{\text{pa}(j)}$ take the form of linear regression models with the $X_{\text{pa}(j)}$ acting as regressors and with independent error terms.

$$X_j = \beta_0 + \sum_{k \in \text{Adjacenct Districts}} \beta_k X_k + \epsilon_j, \qquad \text{where } \epsilon_j \sim N(0, \sigma_j^2)$$

Conditional Linear Gaussian DAG Models

CLGBNs contain both discrete and continuous nodes, and combine DBNs and GBNs as follows to obtain a mixture-of-Gaussians network:

- the local distribution of each discrete node is a CPT;
- the local distribution of each continuous node is a set of linear regression models, one for each configurations of the discrete parents, with the continuous parents acting as regressors.



One of the classic examples is the RATS' WEIGHTS network from Edwards (1995), which describes weight loss in a drug trial performed on rats.

Mixtures of Linear Regressions

The resulting local distribution for the first weight loss for drugs D_1, D_2 and D_3 is:

$$W_{1,D_1} = 7 + \varepsilon_{D_1} \sim N(0, 2.5)$$

$$W_{1,D_2} = 7.50 + \varepsilon_{D_2} \sim N(0, 2)$$

$$W_{1,D_3} = 14.75 + \varepsilon_{D_3} \sim N(0, 11)$$

with just the intercepts since the node has no continuous parents. The local distribution for the second loss is:

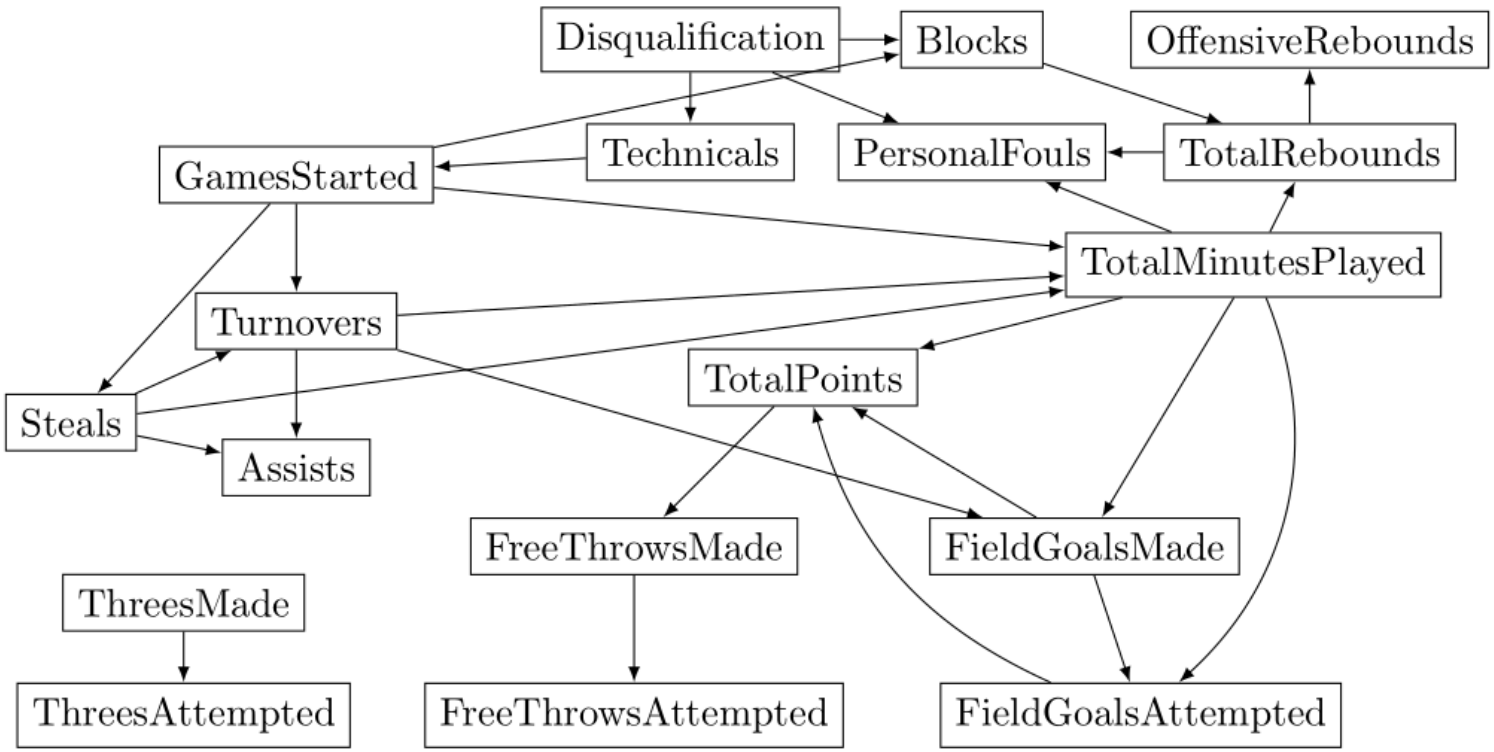
$$W_{2,D_1} = 1.02 + 0.89\beta_{W_1} + \varepsilon_{D_1} \sim N(0, 3.2)$$

$$W_{2,D_2} = -1.68 + 1.35\beta_{W_1} + \varepsilon_{D_2} \sim N(0, 4)$$

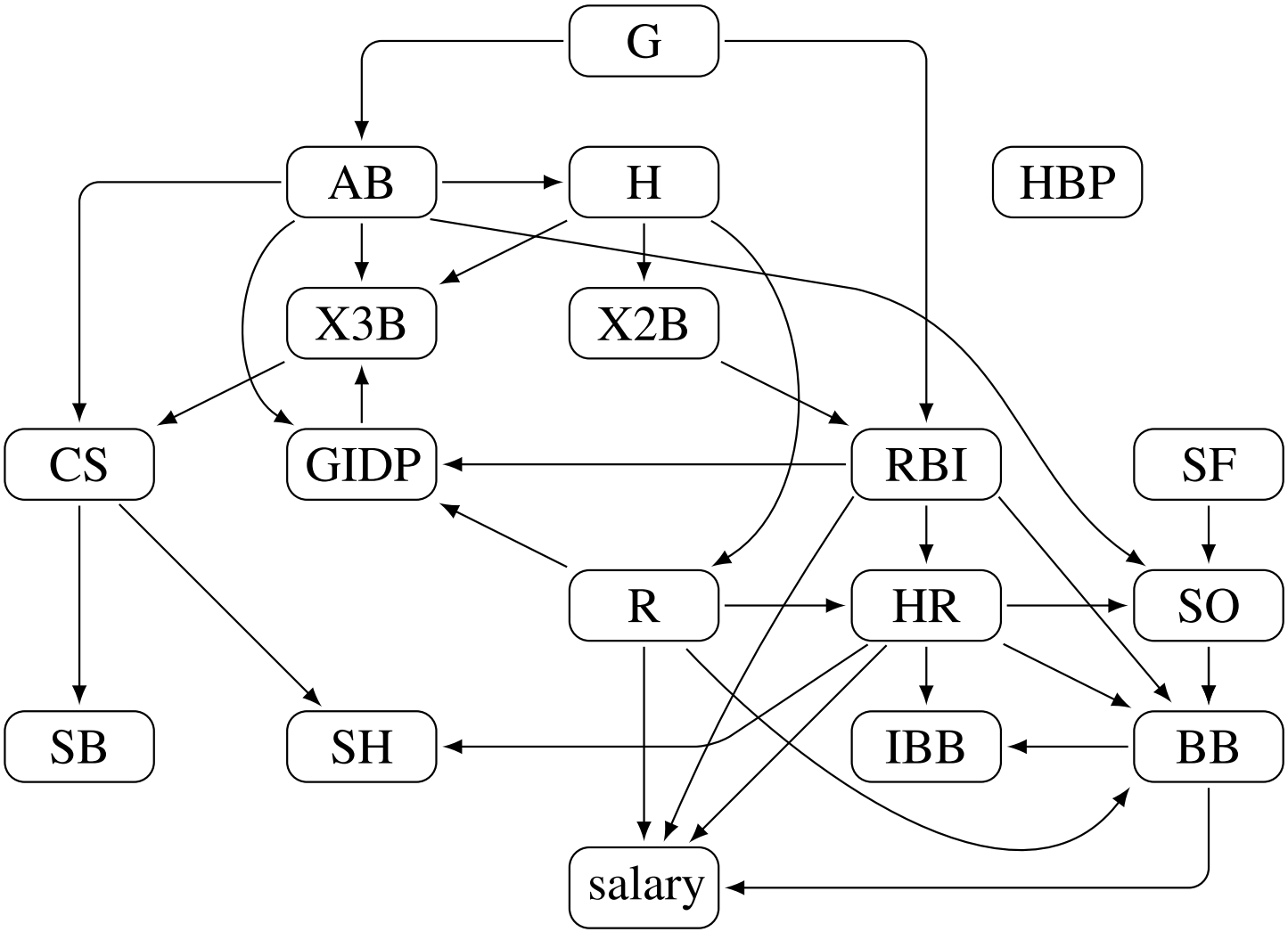
$$W_{2,D_3} = -1.83 + 0.82\beta_{W_1} + \varepsilon_{D_3} \sim N(0, 1.9)$$

Overall, they look like random effect models with random intercepts and random slopes.

Poisson DAG Models: NBA



Poisson DAG Models: MLB

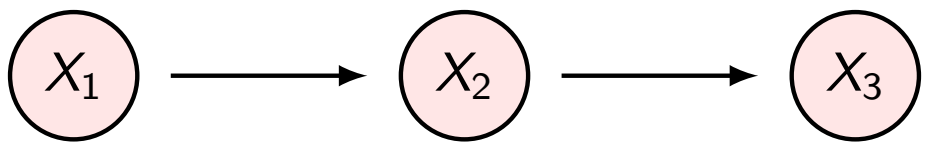


Poisson DAG models

Poisson DAG models is that each conditional distribution is Poisson and the rate parameter depends only on its parents:

$$X_j \mid X_{\text{pa}(j)} \sim \text{Poisson}(g_j(X_{\text{pa}(j)})),$$

where g_j is an arbitrary positive function.



$$P(X_1) \sim Poi(\lambda_1), \quad P(X_2 \mid X_1) \sim Poi(g_2(X_1)), \quad P(X_3 \mid X_2) \sim Poi(g_3(X_2)).$$

Poisson Structural Equation Models

Poisson SEM is that each conditional distribution is Poisson and the rate parameter depends only on its parents:

$$X_j \mid X_{\text{pa}(j)} \sim \text{Poisson}(g_j(X_{\text{pa}(j)})),$$

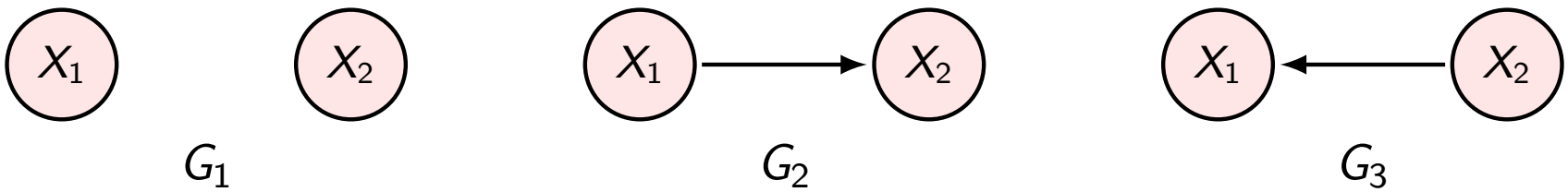
where $g_j(X_{\text{pa}(j)}) = \exp(\theta_j + \sum_{k \in \text{pa}(j)} \theta_{jk} X_k)$.

Using Factorization, the joint distribution of the Poisson SEM can be written as:

$$f(X_{1:p}) = \exp\Big(\sum_{j \in V} \theta_j X_j + \sum_{(k,j) \in E} \theta_{jk} X_j X_k - \sum_{j \in V} \log X_j! - \sum_{j \in V} e^{\theta_j + \sum_{k \in Pa(j)} \theta_{jk} X_k} \Big).$$

Model Identifiability

Is it possible to recover a graph from data? **Partially Yes.**



- We can distinguish G_2 and G_3 from G_1.
- We cannot identify the direction of an edge. Hence, we cannot distinguish G_2 from G_3.

Gaussian Linear SEM

- Form of Gaussian Linear SEMs:
  - Node-wise: $X_j = \sum_{k \in \text{pa}(j)} \beta_{kj} X_k + \epsilon_j \quad \forall j = 1, \ldots, p,$
  - Matrix:

$$(X_1, X_2, ..., X_p)^T = B_0 + B(X_1, ..., X_p)^T + (\epsilon_1, ..., \epsilon_p)^T$$

- Assumptions:
  - Gaussian Noise: $\epsilon_j \sim^{iid} N(0, \sigma_j^2)$ with $\sigma_j^2 > 0$.
  - Causal Minimality: $\beta_{kj} \neq 0$ for all $k \in \text{pa}(j)$, otherwise $\beta_{jk} = 0$.
  - Causal Sufficiency: all variables are observed.

Gaussian Linear SEM

- Form of Gaussian Linear SEMs:
  - Node-wise: $X_j = \sum_{k \in \text{pa}(j)} \beta_{kj} X_k + \epsilon_j \quad \forall j = 1, \ldots, p,$
  - Matrix:

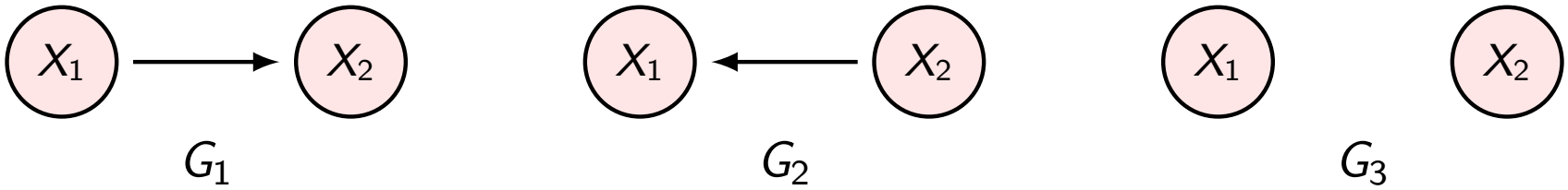
$$(X_1, X_2, ..., X_p)^T = B_0 + B(X_1, ..., X_p)^T + (\epsilon_1, ..., \epsilon_p)^T$$

- Assumptions:
  - Gaussian Noise: $\epsilon_j \sim^{iid} N(0, \sigma_j^2)$ with $\sigma_j^2 > 0$.
  - Causal Minimality: $\beta_{kj} \neq 0$ for all $k \in \text{pa}(j)$, otherwise $\beta_{jk} = 0$.
  - Causal Sufficiency: all variables are observed.
- Joint Distribution:

$$f(G) = \frac{1}{\sqrt{(2\pi)^p \det(\Theta^{-1})}} \exp\Big(-\frac{1}{2}(x_1, ..., x_p)\Theta(x_1, ..., x_p)^T \Big).$$

Motivations of Identifiability: Ordering Recovery

Is it possible to recover a graph from a Gaussian SEM? Yes.



For G_1,

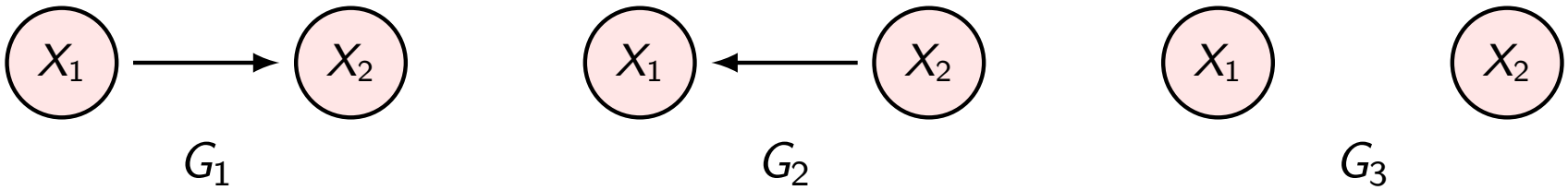
- Forward Selection:

$$\begin{aligned}\mathrm{Var}(X_2) &= \mathbb{E}(\mathrm{Var}(X_2 \mid X_1)) + \mathrm{Var}(\mathbb{E}(X_2 \mid X_1)) = \sigma_2^2 + \beta_1^2\sigma_1^2 \\ &> \sigma_1^2 = \mathrm{Var}(X_1),\end{aligned}$$

- Backward Elimination:

$$\begin{aligned}\mathbb{E}(\mathrm{Var}(X_1 \mid X_2)) &= \mathrm{Var}(X_1) - \mathrm{Var}(\mathbb{E}(X_1 \mid X_2)) \\ &= \sigma_1^2 - \beta_1^2\sigma_1^4/(\beta_1^2\sigma_1^2 + \sigma_2^2) < \sigma_2^2 = \mathbb{E}(\mathrm{Var}(X_2 \mid X_1)).\end{aligned}$$

Motivations of Identifiability: Parent Recovery



Suppose that $\sigma_1^2 < \sigma_2^2$. Then, for G_3,

$$\text{Var}(X_2) = \sigma_2^2 > \sigma_1^2 = \text{Var}(X_1).$$

Under the minimality condition,

$$G_1 : X_1 \not\perp\!\!\!\perp X_2 \quad \text{and} \quad G_3 : X_1 \perp\!\!\!\perp X_2$$

Gaussian DAG Model Identifiability

Let $P(G)$ be generated from a Gaussian linear SEM.

- Π_G is a set of true orderings of graph G,
- $X_{1,2,..,j} = \{X_{\pi_1}, X_{\pi_2}, ..., X_{\pi_j}\}$,
- For any node $m \in V$, let $j = \pi_m$ and $k \in V \setminus \text{nd}(j)$.

Theorem

The DAG G is uniquely identifiable, if there exists $\pi \in \Pi_G$ satisfying either of the two following conditions

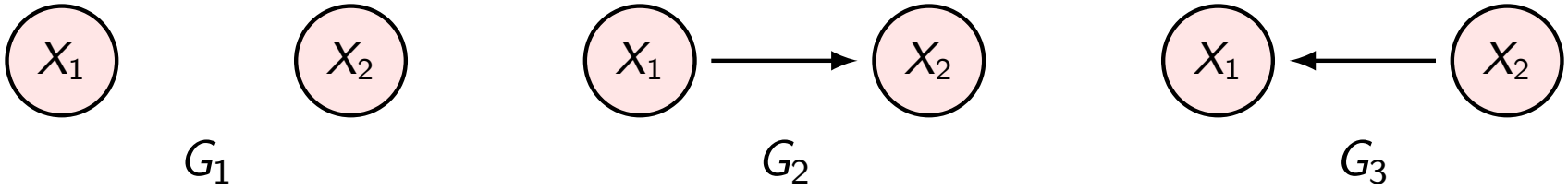
(A) **Forward Selection:** $\sigma_j^2 < \sigma_k^2 + \mathbb{E}(\text{Var}(\mathbb{E}(X_k \mid X_{\text{pa}(k)}) \mid X_{1:(j-1)}))$, or

(B) **Backward Elimination:**

$$\sigma_j^2 > \sigma_\ell^2 + \mathbb{E}(\text{Var}(\mathbb{E}(X_\ell \mid X_{1:(j-1)}) \mid X_{\text{pa}(\ell)})),$$

Poisson DAG Model Identifiability: Overdispersion

Is it possible to recover a graph from count data? **Yes.**



For Poisson random variable X, $\mathrm{Var}(X) = \mathbb{E}(X)$ otherwise the variance is overdispersed relative to the mean.

- For G_1, $\mathrm{Var}(X_1) = \mathbb{E}(X_1)$ and $\mathrm{Var}(X_2) = \mathbb{E}(X_2)$.
- For G_2, $\mathrm{Var}(X_1) = \mathbb{E}(X_1)$, while

$$\begin{aligned}\mathrm{Var}(X_2) &= \mathbb{E}[\mathrm{Var}(X_2 \mid X_1)] + \mathrm{Var}[\mathbb{E}(X_2 \mid X_1)] \\ &> \mathbb{E}[\mathrm{Var}(X_2 \mid X_1)] = \mathbb{E}(X_2),\end{aligned}$$

as long as $\mathrm{Var}[\mathbb{E}(X_2 \mid X_1)] > 0$.

Poisson DAG Model Identifiability: Overdispersion

Theorem: Model Identifiability

For any node $j \in V$, non-empty $\text{pa}_0(j) \subset \text{pa}(j)$ and $S \subset \text{pa}(j) \setminus \text{pa}_0(j)$, if

$$\mathbb{E}\left(\text{Var}\left(\mathbb{E}(X_j \mid X_{\text{pa}(j)}) \mid X_S\right)\right) > 0,$$

the DAG model is identifiable.

- All parents contribute to the variability of the rate parameter.
- The form of link functions (g_j) is not necessarily known.