# Directed Graphical Models I

## Definition and Basic Properties

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

# Recap

- We talked about reasonably weak assumption to facilitate the efficient representation of a probabilistic model.
- Independence assumptions reduce the number of interacting variables, e.g.
  - $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{y})p(\mathbf{z})$
  - $p(x_1, \ldots, x_d) = p(x_1)p(x_2|x_1) \ldots p(x_d|x_{d-1})$
- Chain rule: $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i)$ where $\mathrm{pre}_i = \{x_1, \ldots, x_{i-1}\}$ are the predecessors of $x_i$ in a given ordering of the variables.
- Parametric assumptions, e.g. on $p(x_i|\mathrm{pre}_i)$ in the chain rule, restrict the way the variables may interact.

# Program

1. Visualising factorisations with directed acyclic graphs

2. Directed graphical models

# Program

1. Visualising factorisations with directed acyclic graphs
   - Conditional independences simplify factors in the chain rule
   - Visualisation as a directed acyclic graph
   - Graph concepts

2. Directed graphical models

# Cond independences simplify factors in the chain rule

- We can always express a pdf/pmf $p(\mathbf{x})$ in terms of the chain rule as

$$\begin{aligned} p(\mathbf{x}) &= p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)\dots p(x_d|x_1,\dots x_{d-1}) && (1) \\ &= \prod_{i=1}^{d} p(x_i|\mathrm{pre}_i) && (2) \end{aligned}$$

- Assume that, for each $i$, there is a minimal subset of variables $\mathrm{pa}_i \subseteq \mathrm{pre}_i$ (the "parents" of $x_i$) such that $p(\mathbf{x})$ satisfies

$$x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i \quad \text{for all } i \tag{3}$$

- By conditional independence: $p(x_i|\mathrm{pre}_i) = p(x_i|\mathrm{pa}_i)$
- With the convention $\mathrm{pa}_1 = \varnothing$, we obtain the factorisation

$$p(x_1,\dots,x_d) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i) \tag{4}$$

# What can we do with it?

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i | \mathrm{pa}_i)$$

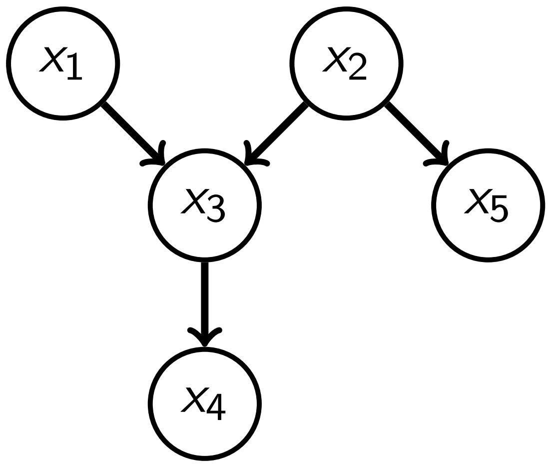
1. $p(x_i|\mathrm{pa}_i)$ involve fewer interacting variables than $p(x_i|\mathrm{pre}_i)$.
   - Makes them easier to model.
   - If specified as a table, fewer numbers are needed for their representation (computational advantage).
2. We can visualise the interactions between the variables with a graph.

# Visualisation as a directed graph

Assume $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i)$ with $\mathrm{pa}_i \subseteq \mathrm{pre}_i$. We visualise the model as a graph with the random variables $x_i$ as nodes, and directed edges that point from the $x_j \in \mathrm{pa}_i$ to the $x_i$. This results in a directed acyclic graph (DAG).

Example:

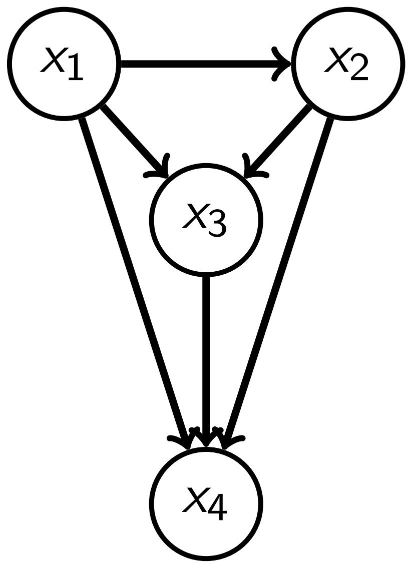
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$$



# Visualisation as a directed graph

Example:

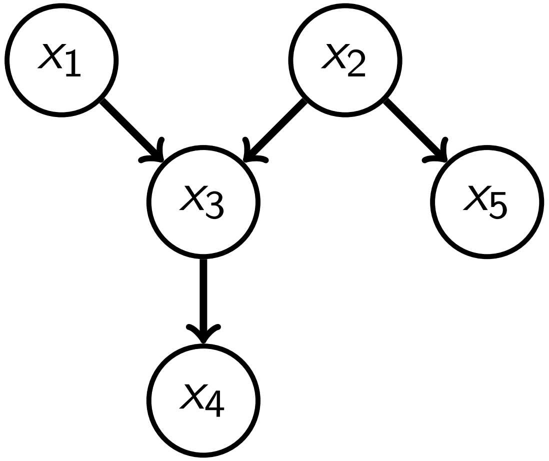
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)$$




Factorisation obtained by chain rule $\equiv$ fully connected directed acyclic graph. Different orderings give different graphs.

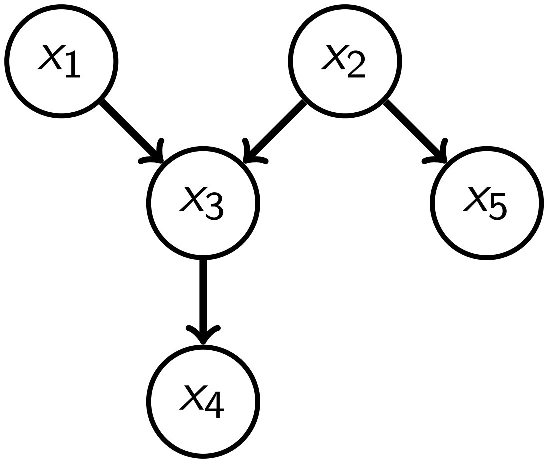
# Graph concepts

- Directed graph: graph where all edges are directed
- Directed acyclic graph (DAG): by following the direction of the arrows you will never visit a node more than once
- $x_i$ is a parent of $x_j$ if there is a (directed) edge from $x_i$ to $x_j$. The set of parents of $x_i$ in the graph is denoted by $\mathrm{pa}(x_i) = \mathrm{pa}_i$, e.g. $\mathrm{pa}(x_3) = \mathrm{pa}_3 = \{x_1, x_2\}$.
- $x_j$ is a child of $x_i$ if $x_i \in \mathrm{pa}(x_j)$, e.g. $x_3$ and $x_5$ are children of $x_2$.



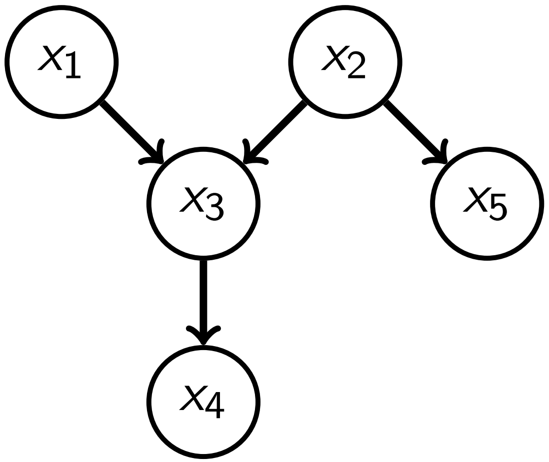
# Graph concepts

- A path or trail from $x_i$ to $x_j$ is a sequence of distinct connected nodes starting at $x_i$ and ending at $x_j$. The direction of the arrows does *not* matter. For example: $x_5, x_2, x_3, x_1$ is a trail.
- A directed path is a sequence of connected nodes where we follow the direction of the arrows. For example: $x_1, x_3, x_4$ is a directed path. But $x_5, x_2, x_3, x_1$ is not a directed path.



# Graph concepts

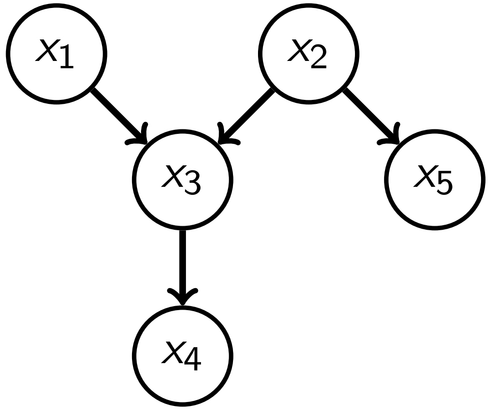
- The ancestors $\mathrm{anc}(x_i)$ of $x_i$ are all the nodes where a directed path leads to $x_i$. For example, $\mathrm{anc}(x_4) = \{x_1, x_3, x_2\}$.
- The descendants $\mathrm{desc}(x_i)$ of $x_i$ are all the nodes that can be reached on a directed path from $x_i$. For example, $\mathrm{desc}(x_1) = \{x_3, x_4\}$.
  (Note: sometimes, $x_i$ is included in the set of ancestors and descendants)
- The non-descendents of $x_i$ are all the nodes in a graph except $x_i$ and the descendants of $x_i$. For example, $\mathrm{nondesc}(x_3) = \{x_1, x_2, x_5\}$



# Graph concepts

- Topological ordering: an ordering $(x_1, \ldots, x_d)$ of some variables $x_i$ is topological relative to a graph if parents come before their children in the ordering.
  (whenever there is a directed edge from $x_i$ to $x_j$, $x_i$ occurs prior to $x_j$ in the ordering.)




- Examples for the graph on the right: (non-exhaustive list)
  - $x_1, x_2, x_3, x_4, x_5$
  - $x_2, x_5, x_1, x_3, x_4$
  - $x_2, x_1, x_3, x_5, x_4$

- There is always at least one ordering that is topological relative to a DAG.

# Program

1. Visualising factorisations with directed acyclic graphs
   - Conditional independences simplify factors in the chain rule
   - Visualisation as a directed acyclic graph
   - Graph concepts

2. Directed graphical models

# Program

1. Visualising factorisations with directed acyclic graphs

2. Directed graphical models
   - Definition
   - Basic properties and ancestral sampling
   - Examples

# Directed graphical model (DGM)

- We started with a factorised pdf/pmf and associated a DAG with it.
- We can also go the other way around and start with a DAG.
- *Definition* A directed graphical model based on a DAG $G$ with $d$ nodes and associated random variables $x_i$ with parents $\mathrm{pa}_i$ is the set of pdfs/pmfs that factorise as

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} k(x_i|\mathrm{pa}_i)$$

  where the $k(x_i|\mathrm{pa}_i)$ are some conditional pdfs/pmfs. (They are sometimes called kernels or factors)
- A pdf/pmf $p(x_1, \ldots, x_d)$ that can be written as above is said to "factorise over the graph $G$". We say that it has property $F(G)$ ("F" for factorisation).

# Why set of pdfs/pmfs?

- The directed graphical model corresponds to a set of probability distributions.
- This is because we did not specify any numerical values for the $k(x_i|\mathrm{pa}_i)$. We only specified which variables the conditionals take as input (namely $x_i$ and $\mathrm{pa}_i$).
- The set includes all distributions that you get by looping, for all variables $x_i$, over all possible $k(x_i|\mathrm{pa}_i)$.
  (e.g. tables or parameter values in parametrised models)
- While a probability distribution corresponds to a probabilistic model, a set of probability distributions (probabilistic models) is often called a statistical model.
- Individual pdfs/pmf in the set are typically also called a directed graphical model.
- Other names for directed graphical models: belief network, Bayesian network, Bayes network.

# Basic properties

$$p(\mathbf{x}) = \prod_{i=1}^{d} k(x_i | \mathrm{pa}_i)$$

Let $x_1, \ldots, x_d$ be any ordering that is topological relative to the graph $G$. Denote by $\mathrm{pre}_i$ the predecessor sets for that ordering. Then:

1. The factors $k(x_i|\mathrm{pa}_i)$ equal the conditionals $p(x_i|\mathrm{pa}_i)$ of $p(\mathbf{x})$.
2. $p(\mathbf{x})$ satisfies the independences:

$$x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i \quad \text{for all } i \tag{5}$$

3. The multivariate marginal of the first $k$ variables in the topological ordering is given by

$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i|\mathrm{pa}_i) \tag{6}$$

# Basic properties: remarks

- When we decomposed a given distribution $p(\mathbf{x})$ with the chain rule and inserted the conditional independences

$$x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i \quad \text{for all } i \tag{7}$$

  we obtained

$$p(\mathbf{x}) = \prod_i p(x_i|\mathrm{pa}_i) \tag{8}$$

  with $p(x_i|\mathrm{pa}_i)$ equal to the conditionals of $x_i$ given $\mathrm{pa}_i$.
- The first two properties say that the reverse holds too: the factorisation in (8) implies the independences in (7) for any topological ordering.
- The factorisation and independences are thus equivalent.
- Since $k(x_i|\mathrm{pa}_i) = p(x_i|\mathrm{pa}_i)$, we will use $p(x_i|\mathrm{pa}_i)$ instead of $k(x_i|\mathrm{pa}_i)$ from now on when we work with DGMs.

# Proof: $k(x_i|\mathrm{pa}_i) = p(x_i|\mathrm{pa}_i)$ (not examinable)

- Assume the variables are labelled such that the ordering $x_1, \ldots, x_d$ is topological relative to the DAG $G$.
- Variable $x_d$ is a leaf variable and not part of any parent set.
- We next compute $p(x_1, \ldots, x_{d-1})$ using the sum rule:

$$
\begin{aligned}
p(x_1, \ldots, x_{d-1}) &= \int p(x_1, \ldots, x_d) \mathrm{d}x_d \\
&= \int \prod_{i=1}^{d} k(x_i|\mathrm{pa}_i) \mathrm{d}x_d \\
&= \int \prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i) k(x_d|\mathrm{pa}_d) \mathrm{d}x_d \quad (x_d \notin \mathrm{pa}_i, i \leq d) \\
&= \prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i) \int k(x_d|\mathrm{pa}_d) \mathrm{d}x_d \\
&= \prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i)
\end{aligned}
$$

# Proof: $k(x_i|\mathrm{pa}_i) = p(x_i|\mathrm{pa}_i)$ (not examinable)

Hence:

$$p(x_d|x_1,\ldots,x_{d-1}) = \frac{p(x_1,\ldots,x_d)}{p(x_1,\ldots,x_{d-1})} = \frac{\prod_{i=1}^{d} k(x_i|\mathrm{pa}_i)}{\prod_{i=1}^{d-1} k(x_i|\mathrm{pa}_i)}$$
$$= k(x_d|\mathrm{pa}_d)$$

Split $(x_1,\ldots,x_{d-1}) = \mathrm{pre}_d$ into non-overlapping sets $\mathrm{pa}_d$ and $\tilde{\mathbf{x}}_d = \mathrm{pre}_d \setminus \mathrm{pa}_d$ so that $p(x_d|x_1,\ldots,x_{d-1}) = p(x_d|\tilde{\mathbf{x}}_d, \mathrm{pa}_d)$.
By the product rule, we have

$$p(x_d, \tilde{\mathbf{x}}_d|\mathrm{pa}_d) = p(x_d|\tilde{\mathbf{x}}_d, \mathrm{pa}_d)p(\tilde{\mathbf{x}}_d|\mathrm{pa}_d)$$
$$= k(x_d|\mathrm{pa}_d)p(\tilde{\mathbf{x}}_d|\mathrm{pa}_d)$$

Next sum out $\tilde{\mathbf{x}}_d$ to obtain

$$p(x_d|\mathrm{pa}_d) = \int p(x_d, \tilde{\mathbf{x}}_d|\mathrm{pa}_d)\mathrm{d}\tilde{\mathbf{x}}_d = k(x_d|\mathrm{pa}_d)\int p(\tilde{\mathbf{x}}_d|\mathrm{pa}_d)\mathrm{d}\tilde{\mathbf{x}}_d$$
$$= k(x_d|\mathrm{pa}_d)$$

where we have used that $x_d$ and $\mathrm{pa}_d$ are not part of $\tilde{\mathbf{x}}_d$.

# Proof: $k(x_i|\mathrm{pa}_i) = p(x_i|\mathrm{pa}_i)$ (not examinable)

Hence:

$$p(x_d|x_1, \ldots, x_{d-1}) = k(x_d|\mathrm{pa}_d) = p(x_d|\mathrm{pa}_d)$$

Next, note that $p(x_1, \ldots, x_{d-1})$ has the same form as $p(x_1, \ldots, x_d)$: apply the same procedure to all $p(x_1, \ldots, x_k)$, for smaller and smaller $k \leq d-1$

Proves that for DGMs, the factors $k(x_i|\mathrm{pa}_i)$ are equal to the conditionals $p(x_i|\mathrm{pa}_i)$ of $p(\mathbf{x})$.

# Proof: basic independences (not examinable)

- In the proof above, we further found that for all $i$,

$$p(x_i|x_1, \ldots, x_{i-1}) = p(x_i|\mathrm{pa}_i), \tag{9}$$

  which means that $x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i$ for all $i$ in the chosen topological ordering.
- Chosen topological ordering was not special: holds for all orderings that are topological relative to the DAG.
- Hence, factorisation $p(\mathbf{x}) = \prod_i p(x_i|\mathrm{pa}_i)$ implies, for any topological ordering

$$x_i \perp\!\!\!\perp (\mathrm{pre}_i \setminus \mathrm{pa}_i) \mid \mathrm{pa}_i \quad \text{for all } i \tag{10}$$

# Proof: marginals (not examinable)

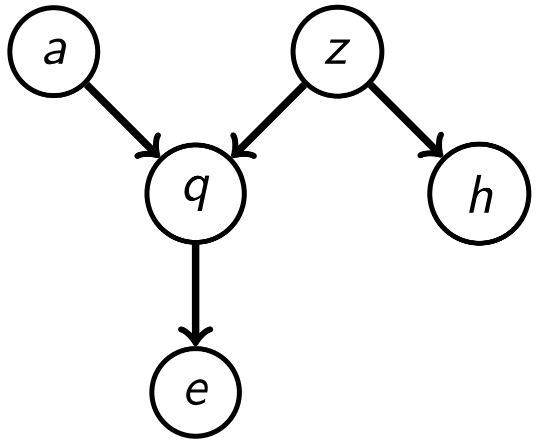
- In the proof, we also found that (for the chosen topological ordering)
$$p(x_1, \ldots, x_k) = \prod_{i=1}^{k} p(x_i|\mathrm{pa}_i) \tag{11}$$
- The marginal joint distribution of the first $k$ variables in the chosen topological ordering is given by the product of the corresponding factors $p(x_i|\mathrm{pa}_i)$.
- Chosen topological ordering was not special: holds for all orderings that are topological relative to the DAG.
- While marginalisation can be very expensive (see later), the above marginals are available for free for DGMs.

# Ancestral sampling

- The DAG not only specifies the joint distribution $p(\mathbf{x}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i)$ but also a sampling/data generating process.
- To generate data from $p(\mathbf{x})$:
  1. Pick an ordering $x_1, \ldots, x_d$ of the random variables that is topological to $G$.
  2. $x_1$ does not have any parents, i.e. set $\mathrm{pa}_1 = \varnothing$ and $p(x_1|\varnothing) = p(x_1)$.
  3. Following the topological ordering, sample from $p(x_i|\mathrm{pa}_i)$, $i = 1, \ldots, d$.
- It's called ancestral sampling because we sample the parents before the children, following the arrows in the DAG.
- The DAG visualises the data generating process, which can be used as modelling tool.

# Example

DAG:




Random variables: $a, z, q, e, h$

Parent sets: $\mathrm{pa}_a = \mathrm{pa}_z = \varnothing, \mathrm{pa}_q = \{a, z\}, \mathrm{pa}_e = \{q\}, \mathrm{pa}_h = \{z\}$.

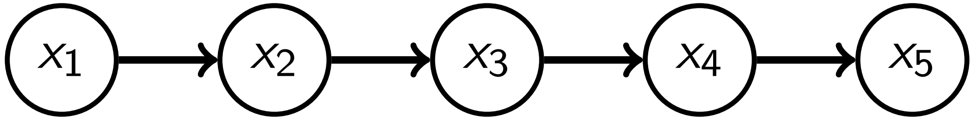
Directed graphical model: set of pdfs/pmfs $p(a, z, q, e, h)$ that factorise as:

$$p(a, z, q, e, h) = p(a)p(z)p(q|a, z)p(e|q)p(h|z)$$

Data generating process: For topological ordering $a, z, q, e, h$:
$a \sim p(a), z \sim p(z), q \sim p(q|a, z), e \sim p(e|q), h \sim p(h|z)$

# Example: Markov chain

DAG:




Random variables: $x_1, x_2, x_3, x_4, x_5$

Parent sets:
$\mathrm{pa}_1 = \varnothing, \mathrm{pa}_2 = \{x_1\}, \mathrm{pa}_3 = \{x_2\}, \mathrm{pa}_4 = \{x_3\}, \mathrm{pa}_5 = \{x_4\}$.

Directed graphical model: set of pdfs/pmfs $p(x_1, \ldots, x_5)$ that factorise as:

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)p(x_5|x_4)$$

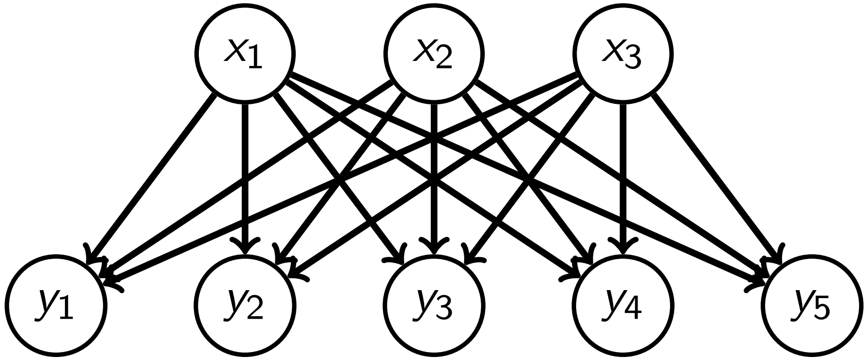
Data generating process: For topological ordering $x_1, \ldots, x_5$:
$x_1 \sim p(x_1), x_2 \sim p(x_2|x_1), x_3 \sim p(x_3|x_2), x_4 \sim p(x_4|x_3), x_5 \sim p(x_5|x_4)$

# Example: Probabilistic PCA, factor analysis, ICA, VAEs

(PCA/ICA: principal/independent component analysis; VAE: var autoencoders)

DAG:




Random variables: $x_1, x_2, x_3, y_1, \ldots, y_5$

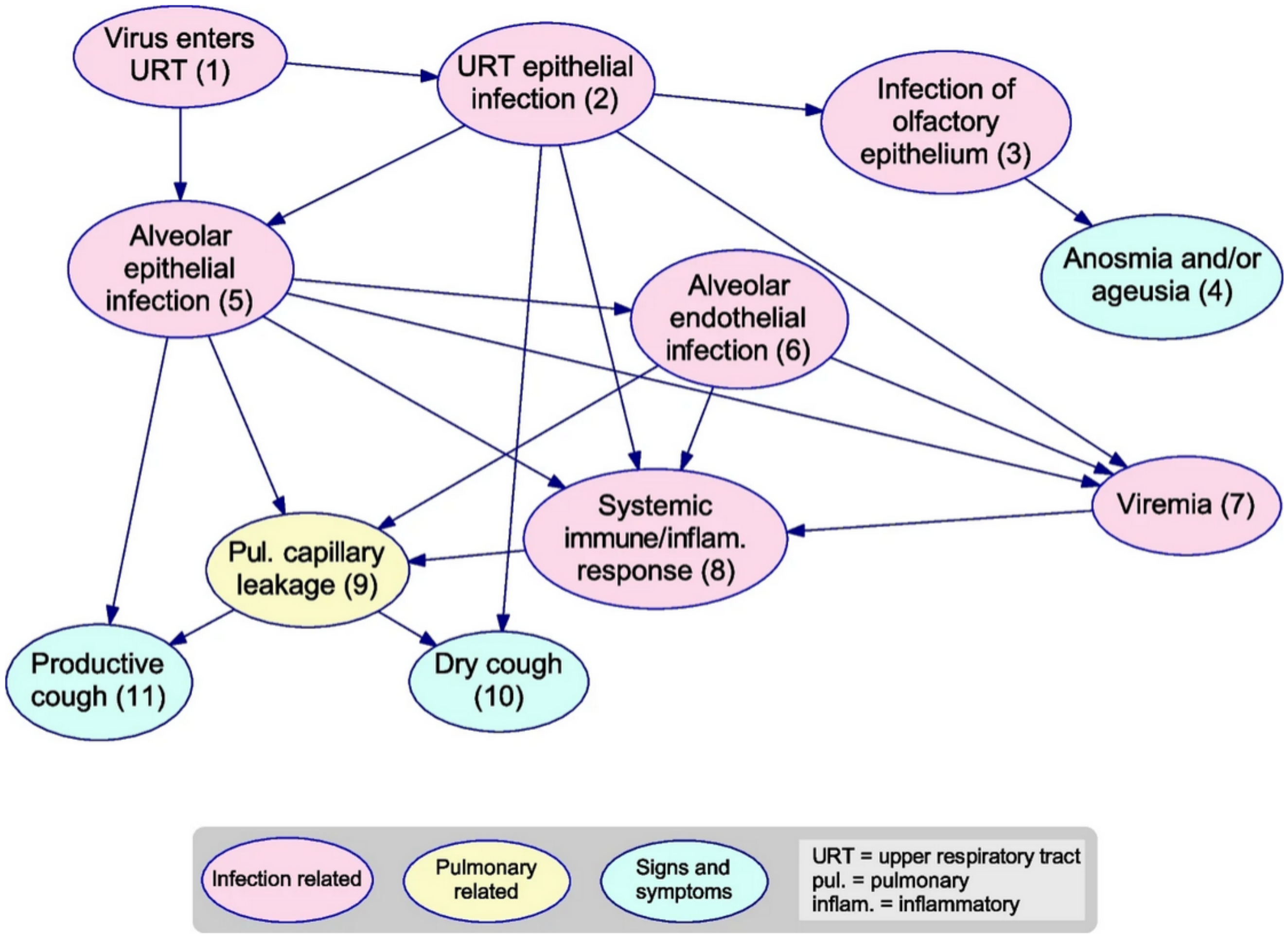
Parent sets: $\mathrm{pa}(x_i) = \varnothing, \mathrm{pa}(y_i) = \{x_1, x_2, x_3\}$ for all $i$.

Directed graphical model: set of pdfs/pmfs $p(x_1, x_2, x_3, y_1, \ldots, y_5)$ that factorise as:

$$
\begin{aligned}
p(x_1, x_2, x_3, y_1, \ldots, y_5) = & p(x_1)p(x_2)p(x_3)p(y_1|x_1, x_2, x_3) \\
& p(y_2|x_1, x_2, x_3) \ldots p(y_5|x_1, x_2, x_3)
\end{aligned}
$$

Data generating process: topological ordering $x_1, x_2, x_3, y_1, \ldots, y_4$
$x_i \sim p(x_i), y_i \sim p(y_i|x_1, x_2, x_3)$

# Example: Modeling COVID-19 disease processes




BMC Med Res Methodol, 2023. `https://doi.org/10.1186/s12874-023-01856-1`

# Program recap

1. Visualising factorisations with directed acyclic graphs
   - Conditional independences simplify factors in the chain rule
   - Visualisation as a directed acyclic graph
   - Graph concepts

2. Directed graphical models
   - Definition
   - Basic properties and ancestral sampling
   - Examples