Combinatorics, Algebra, and Intervention

Liam Solus

KTH Royal Institute of Technology

solus@kth.se

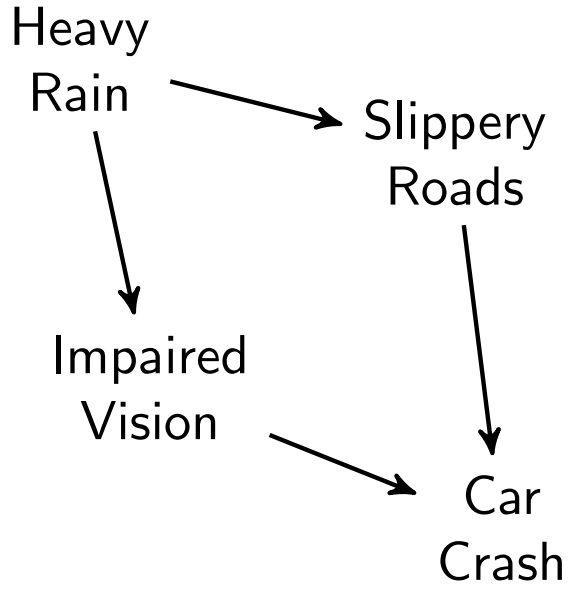
17 September 2020
Applied CATS Seminar
KTH

What can combinatorics tell us about a probability distribution?

Today's Example: Graphical Models and Causality.

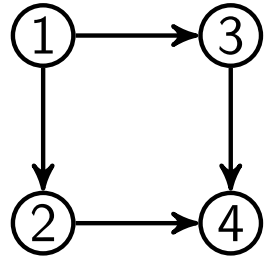
Graphs can help us understand important features of a joint distribution $\mathbb{P} \sim (X_1, \ldots, X_n)$:

1. Conditional independence (CI) relations
2. Causal information.



The Game: Encode some information in a graph. Then

- **(Representation)** What graph represents and how.
- **(Inference)** Use representations to do inference.
- **(Learning)** Use how representations encode information to learn them from data.



How do we associate a distribution to a graph and what does it imply?

$\mathbb{P}$ a distribution over discrete variables $(X_1, \ldots, X_n)$.

X_i has outcomes $[d_i] := \{1, \ldots, d_i\}$.

$\mathbb{P}$ encodes the probability $p(x_1, \ldots, x_n)$ of each $(x_1, \ldots, x_n) \in \mathcal{R} = \prod_{i \in [n]} [d_i]$.

$\mathcal{G} = ([n], E)$ a **directed acyclic graph (DAG)**.

$\mathbb{P}$ is **Markov** to $\mathcal{G}$ if for all $i \in [n]$

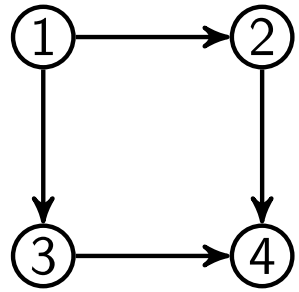
$$X_i \perp\!\!\!\perp X_{\text{nondes}_{\mathcal{G}}(i) \setminus \text{pa}_{\mathcal{G}}(i)} \mid X_{\text{pa}_{\mathcal{G}}(i)}.$$

Theorem. The following are equivalent:

1. $\mathbb{P}$ is Markov to $\mathcal{G}$.
2. $p(x_1, \ldots, x_n) = \prod_{i \in [n]} p(x_i \mid x_{\text{pa}_{\mathcal{G}}(i)})$.
3. $X_A \perp\!\!\!\perp X_B \mid X_C$ in $\mathbb{P}$ for subsets $A, B, C \subset [n]$ if and only if A is **d-separated** from B given C.

X_1, X_2, X_3, X_4 binary.

X_i has outcomes $[2] = \{1, 2\}$



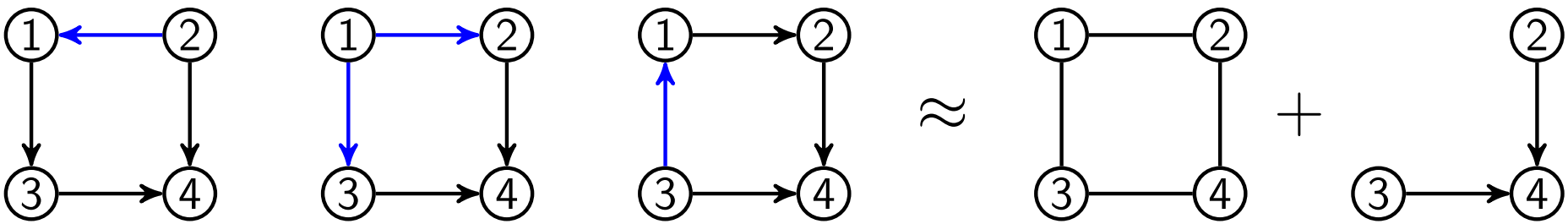
$X_2 \perp\!\!\!\perp X_3 \mid X_1$

$X_4 \perp\!\!\!\perp X_1 \mid X_{\{2,3\}}$

Two different DAGs can encode the same set of CI relations (distributions).
Such DAGs are called **Markov equivalent**.

Theorem. Two DAGs are Markov equivalent if and only if

1. they have the same **skeleton** and **v-structures**. [Verma, Pearl, 1989]
2. they are connected by a sequence of **covered arrow** reversals. [Chickering, 1995]



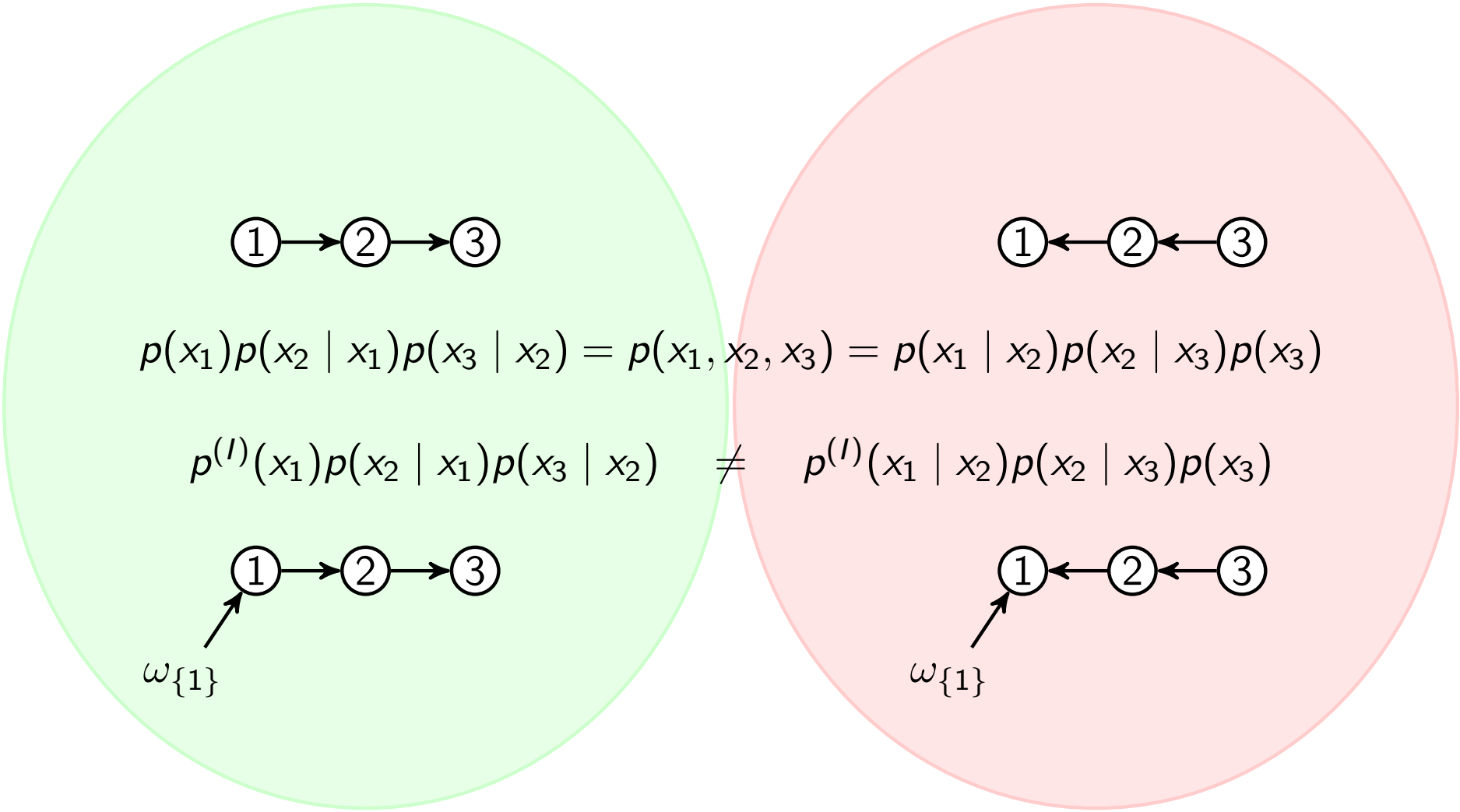
$i \to j$ is **covered** in $\mathcal{G}$ if $\mathrm{pa}_{\mathcal{G}}(j) = \mathrm{pa}_{\mathcal{G}}(i) \cup \{i\}$.

Such **representation** theorems lead to structure **learning** algorithms:

1. SGS and PC algorithms [Spirtes, Glymour and Scheines, 2000]
2. GES, Greedy SP [Chickering, 2002; LS, Wang and Uhler, 2020]

Counting Markov equivalence classes (MECs) and their sizes: [Gillespie, Perlman 2001]
[Ye, Bin, 2016]
[Radhakrishnan, LS and Uhler, 2018]

Interventional DAG models



By **intervening** and comparing the resulting **interventional** and **observational** distributions, we can distinguish between the two **causal networks**.

For $I \subset [n]$, a distribution $\mathbb{P}^{(I)}$ that factorizes as

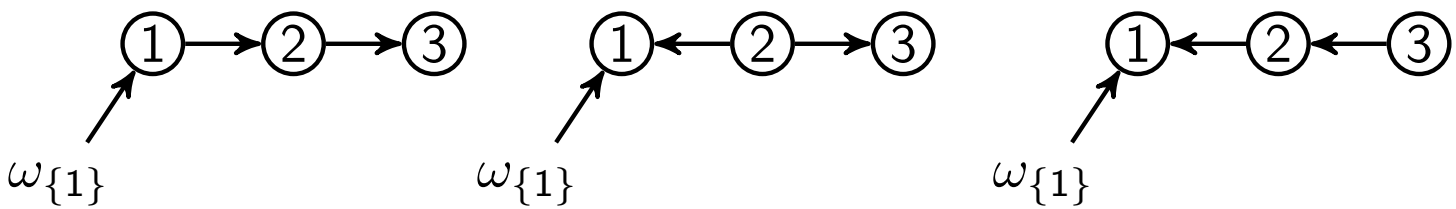
$$p^{(I)}(x_1, \ldots, x_m) = \prod_{i \in I} p^{(I)}(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{i \notin I} p(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)})$$

is an **interventional distribution** for I with respect to the **observational distribution** $\mathbb{P}$ Markov to $\mathcal{G}$.

For a collection of targets $\mathcal{I}$, the set of all $(\mathbb{P}^{(I)})_{I \in \mathcal{I}}$ is the **interventional DAG model** associated to $\mathcal{G}$ and $\mathcal{I}$.

Theorem. Two DAGs are $\mathcal{I}$-Markov equivalent if and only if

1. their $\mathcal{I}$-**DAGs** have the same skeleton and v-structures. [Yang, Katcoff, Uhler, 2018]
2. their $\mathcal{I}$-DAGs are connected by a sequence of covered arrow reversals.



Structure **learning** algorithms: GIES, GISP, etc...

[Hauser, Bühlmann, 2012]
[Wang, LS, Yang, Uhler, 2017]
[Yang, Katcoff, Uhler, 2018]
[Kocaoglu et al., 2019]

What if we also consider the combinatorics (and algebra!) of the parameters?

$\mathbb{P}$ is **Markov** to $\mathcal{G}$ if and only if

$$p(x_1, \ldots, x_n) = \prod_{i \in [n]} p(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}).$$

A naturally associated **algebraic variety**:

- a parameter for each outcome:

$$(x_1, \ldots, x_n) \in \mathcal{R} \longleftrightarrow \mathbb{C}[\mathcal{R}] := \mathbb{C}[p_{x_1 \cdots x_n} : (x_1, \ldots, x_n) \in \mathcal{R}].$$

- a parameter for each outcome of each node and its parents:

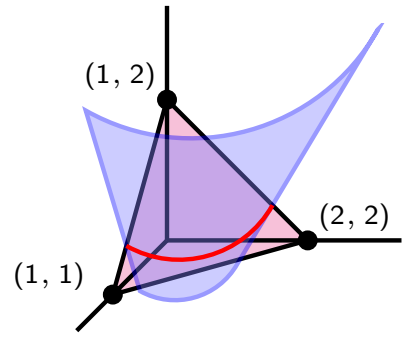
$$p(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}) \longleftrightarrow q_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}} \in \mathbb{C}[U] \qquad \text{for } x_i \in [d_i], \quad x_{\mathrm{pa}_{\mathcal{G}}(i)} \in \prod_{k \in \mathrm{pa}_{\mathcal{G}}(i)} [d_k].$$

- $\sum_{x_i \in [d_i]} p(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}) = 1$ for all $i \in [n]$ and $x_{\mathrm{pa}_{\mathcal{G}}(i)}$.

- So, we quotient the ring $\mathbb{C}[U]$ by the ideal $\mathfrak{q} := \langle \sum_{x_i \in [d_i]} q_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}} - 1 \rangle$.

The zero locus of $\ker(\Phi_{\mathcal{G}})$ is the **DAG model variety**, where

$$\Phi_{\mathcal{G}} : \mathbb{C}[\mathcal{R}] \longrightarrow \mathbb{C}[U]/\mathfrak{q}; \qquad \Phi_{\mathcal{G}} : p_{x_1 \cdots x_n} \longmapsto \prod_{i \in [n]} q_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}}.$$



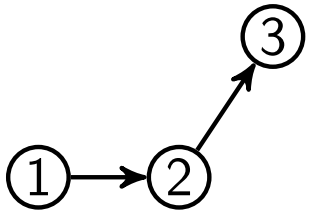
Are nice statistical properties of the DAG model secretly consequences of nice properties of this variety?

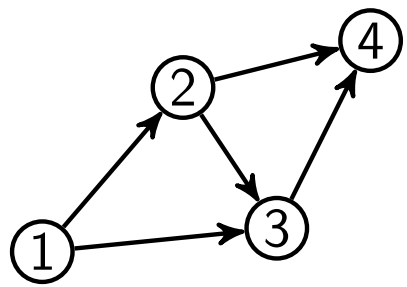
The same map $\Phi_{\mathcal{G}}^{\mathrm{toric}} : \mathbb{C}[\mathcal{R}] \longrightarrow \mathbb{C}[U]$ but into the ring $\mathbb{C}[U]$ yields a **toric ideal**:

$$\ker(\Phi_{\mathcal{G}}^{\mathrm{toric}}) \subset \ker(\Phi_{\mathcal{G}}).$$

DAG models are "nicest" when we can apply the **belief propagation** algorithm for probabilistic inference without sacrificing any information.

This happens when $\mathcal{G}$ is **perfect**: the parents of each node $i \in [n]$ form a clique:





Theorem. The DAG $\mathcal{G}$ is perfect if and only if

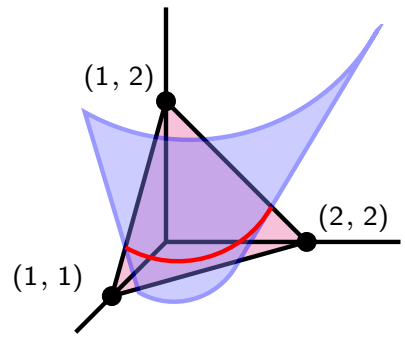
$$\ker(\Phi_{\mathcal{G}}) = \ker(\Phi_{\mathcal{G}}^{\text{toric}}).$$

Moreover, $\ker(\Phi_{\mathcal{G}})$ is toric and has a quadratic and square-free Gröbner basis.

[Hosten, Sullivant, 2002; Geiger et al., 2005; Duarte and LS, 2020]

- The nice property of our statistical model corresponds to a nice property of its defining algebraic structure.
- I.e. "niceness" of our model, locally, within the probability simplex is intrinsic to its variety in the global parameter space $\mathbb{R}^{|\mathcal{R}|}$.

How does this story extend to the interventional framework?



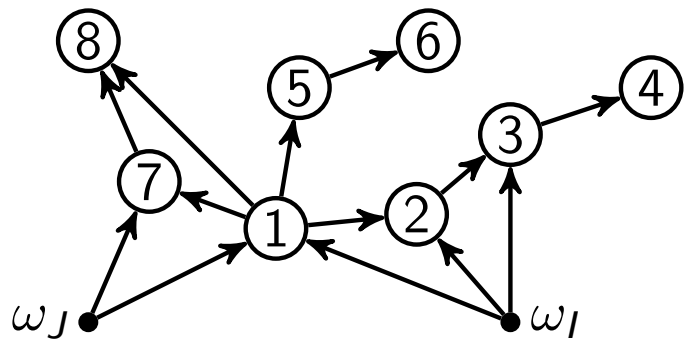
$$p^{(I)}(x_1,\ldots,x_m) = \prod_{i\in I} p^{(I)}(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{i\notin I} p(x_i \mid x_{\mathrm{pa}_{\mathcal{G}}(i)}).$$

For $I \in \mathcal{I}$, and each $p_{x_1\cdots x_m}$ or $q_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}}$ add new parameter $p^{(I)}_{x_1\cdots x_m}$ or $q^{(I)}_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}}$.

The **interventional variety** for $\mathcal{G}$ and $\mathcal{I}$ is the zero locus of $\ker(\Phi_{\mathcal{G},\mathcal{I}})$ where

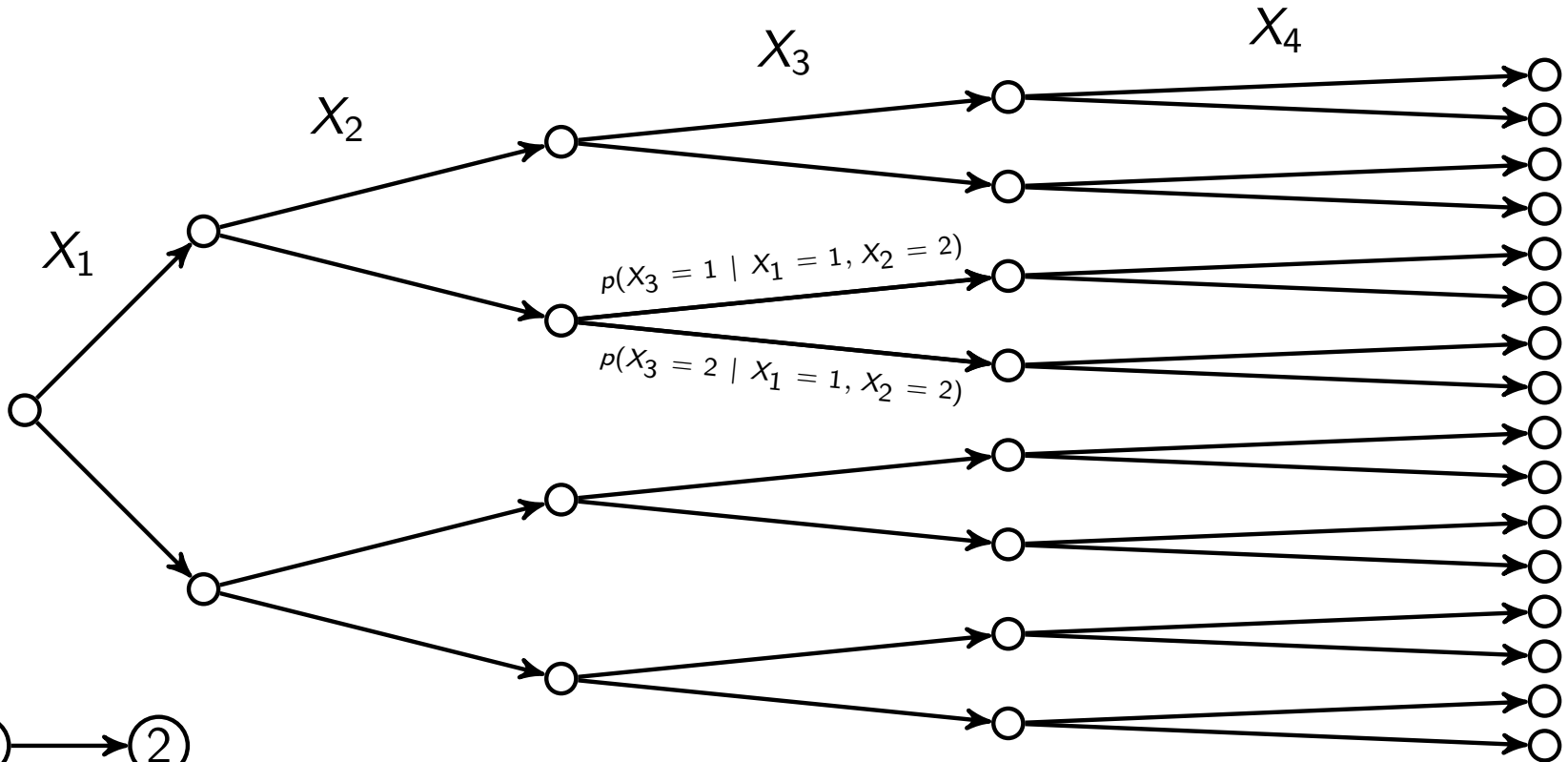
$$\Phi_{\mathcal{G},\mathcal{I}} : \mathbb{C}[\mathcal{R}_{\mathcal{I}}] \longrightarrow \mathbb{C}[U_{\mathcal{I}}]/\mathfrak{q}_{\mathcal{I}};$$
$$\Phi_{\mathcal{G},\mathcal{I}} : p^{(I)}_{x_1\cdots x_m} \longmapsto \prod_{i\in I} q^{(I)}_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}} \prod_{i\notin I} q_{i;x_i;x_{\mathrm{pa}_{\mathcal{G}}(i)}}.$$

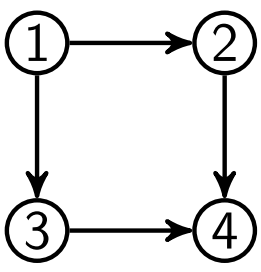
Theorem. $\ker(\Phi_{\mathcal{G},\mathcal{I}}) = \ker(\Phi^{\mathrm{toric}}_{\mathcal{G},\mathcal{I}})$ if and only if $\mathcal{G}$ is perfect, and I is equal to its set of ancestors for all $I \in \mathcal{I}$. Moreover, $\ker(\Phi_{\mathcal{G},\mathcal{I}})$ is toric with a quadratic and square-free Gröbner basis. [Duarte and LS, 2020]



How did we prove it?

An alternative **representation** of the model that embraces the many parameters!



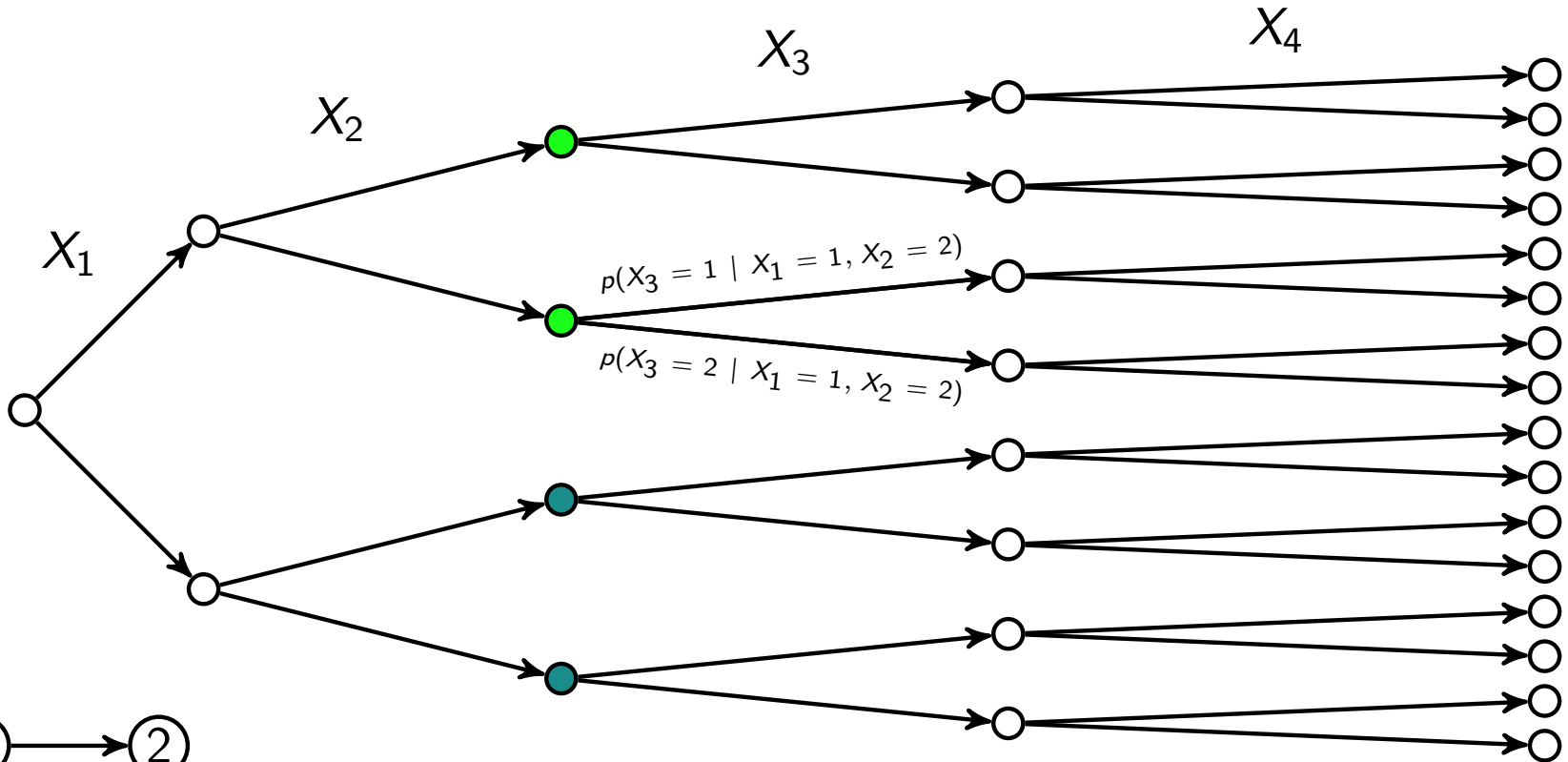


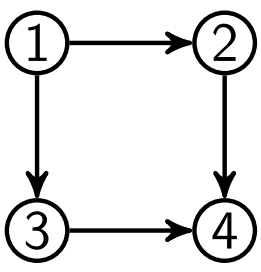
$\mathcal{G}$

$$X_3 \perp\!\!\!\perp X_2 \mid X_1, \quad X_4 \perp\!\!\!\perp X_1 \mid X_{\{2,3\}}$$

How did we prove it?

An alternative **representation** of the model that embraces the many parameters!



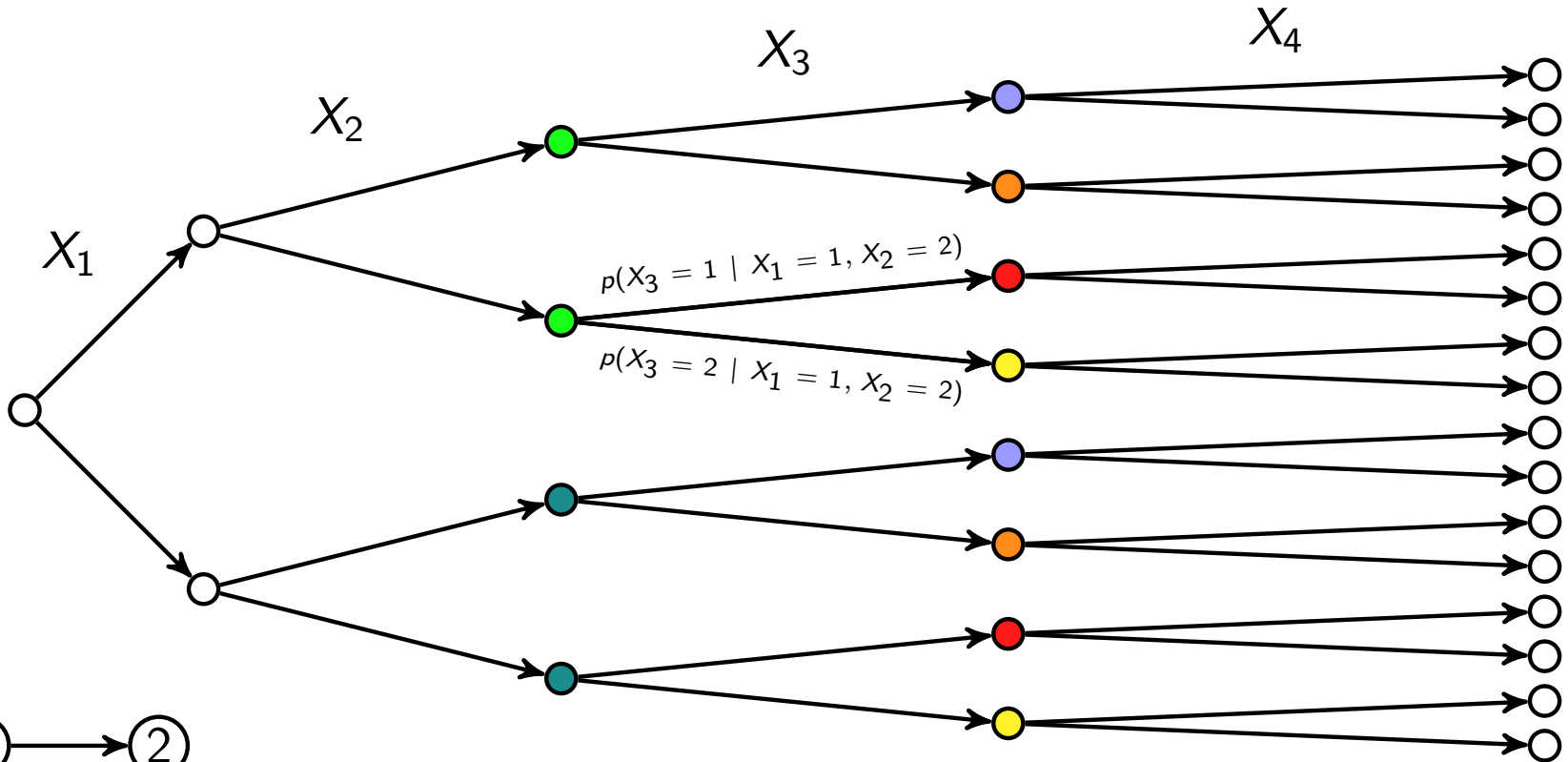


$\mathcal{G}$

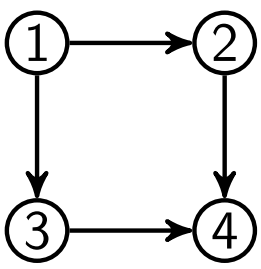
$$X_3 \perp\!\!\!\perp X_2 \mid X_1, \quad X_4 \perp\!\!\!\perp X_1 \mid X_{\{2,3\}}$$

How did we prove it?

An alternative **representation** of the model that embraces the many parameters!



The **staged tree** $\mathcal{T}_{\mathcal{G}}$ for $\mathcal{G}$.

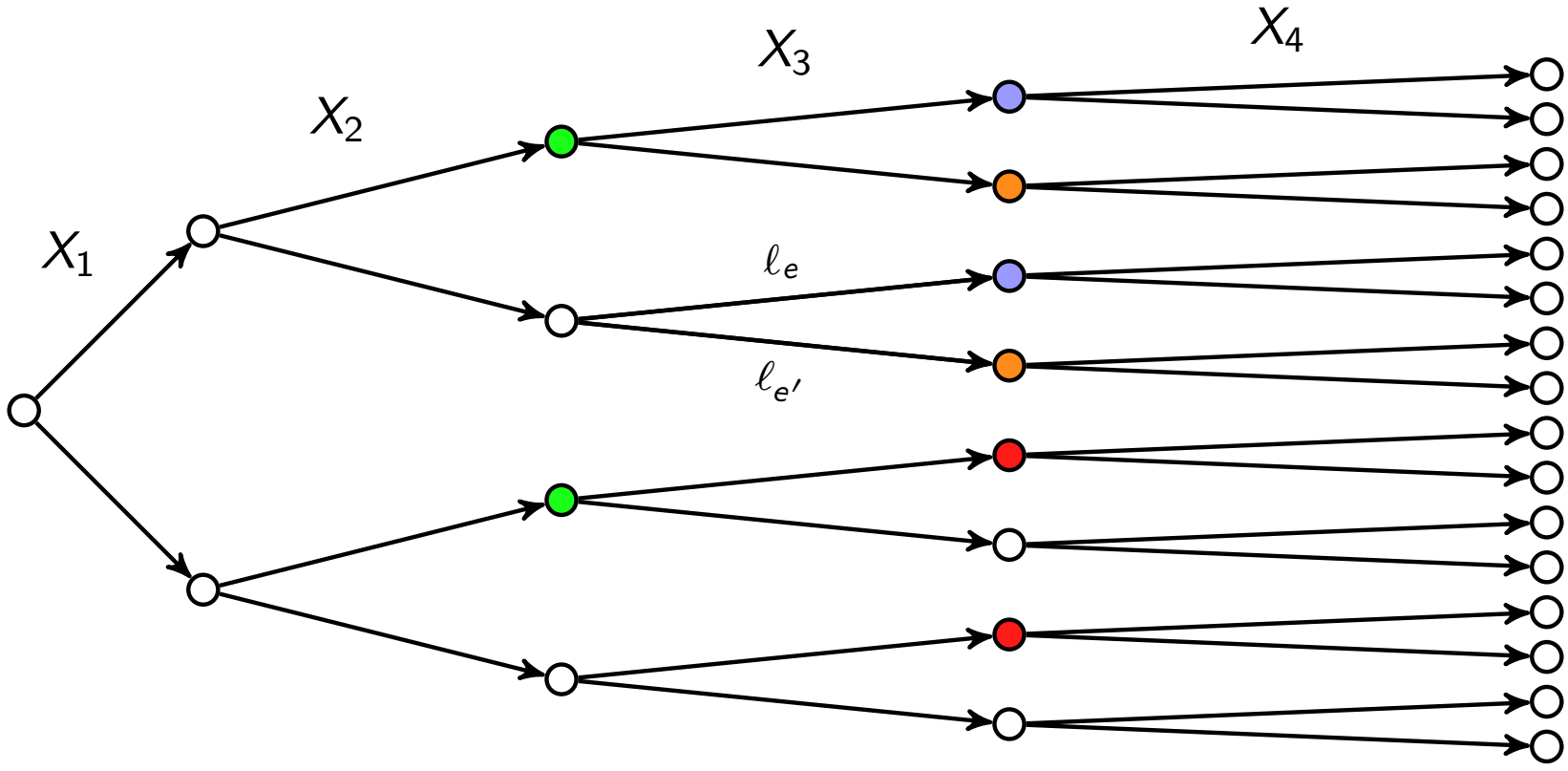


$\mathcal{G}$

$$X_3 \perp\!\!\!\perp X_2 \mid X_1, \quad X_4 \perp\!\!\!\perp X_1 \mid X_{\{2,3\}}$$

How did we prove it?

An alternative **representation** of the model that embraces the many parameters!



A **staged tree** $\mathcal{T}$.

Extend the story!

Given a staged tree $\mathcal{T}$, we can define the **staged tree ideal** $\ker(\Phi_{\mathcal{T}})$ and it's toric ideal $\ker(\Phi_{\mathcal{T}}^{\text{toric}}) \subset \ker(\Phi_{\mathcal{T}})$.

For a DAG $\mathcal{G}$, $\ker(\Phi_{\mathcal{G}}) = \ker(\Phi_{\mathcal{T}_{\mathcal{G}}})$ and $\ker(\Phi_{\mathcal{G}}^{\text{toric}}) = \ker(\Phi_{\mathcal{G}}^{\text{toric}})$.

Theorem. $\mathcal{G}$ is perfect if and only if $\mathcal{T}_{\mathcal{G}}$ is **balanced**. [Duarte and LS, 2020]

Theorem. $\mathcal{T}$ is balanced if and only if $\ker(\Phi_{\mathcal{T}}) = \ker(\Phi_{\mathcal{T}}^{\text{toric}})$ [Duarte and Görgen, 2020]

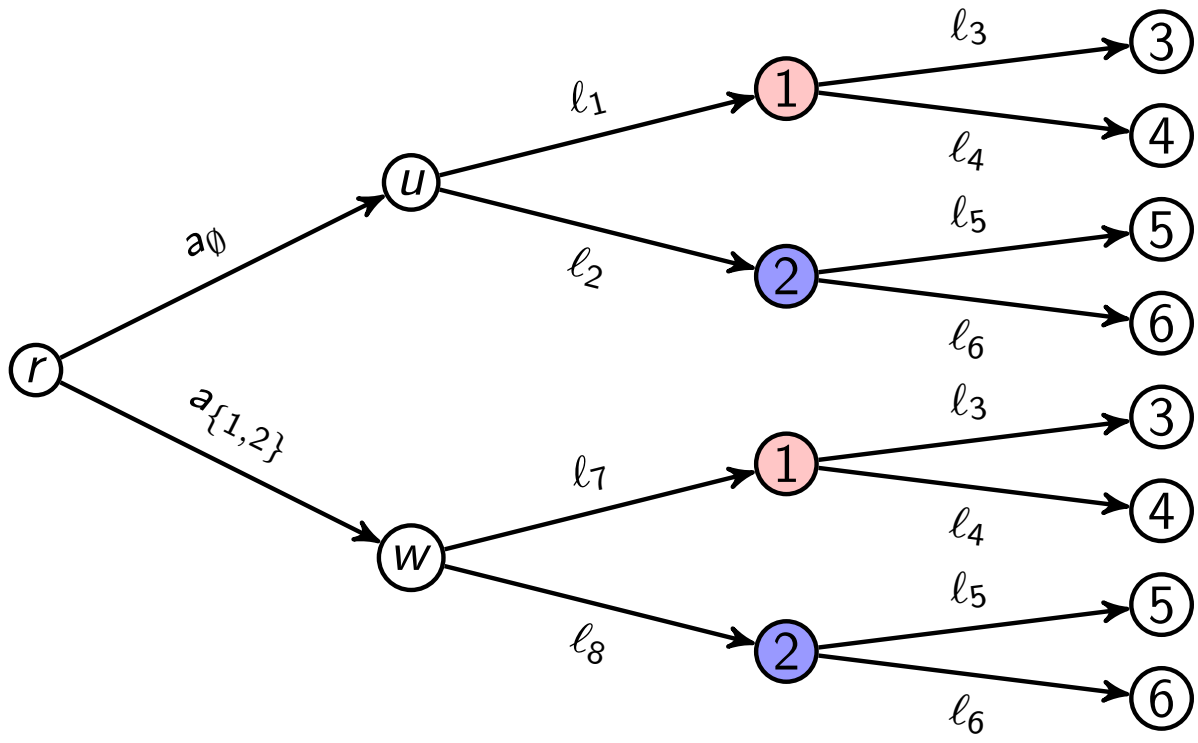
Idea: introduce **interventional staged tree models** $\mathcal{T}_{\mathcal{I}}$ that generalize interventional DAG models $(\mathcal{G}, \mathcal{I})$. Then characterize when such interventional staged trees are balanced!

Theorem. $\ker(\Phi_{\mathcal{G},\mathcal{I}}) = \ker(\Phi_{\mathcal{G},\mathcal{I}}^{\text{toric}})$ if and only if [Duarte and LS, 2020]

1. $\mathcal{G}$ is perfect, and
2. I is equal to its set of ancestors for all $I \in \mathcal{I}$.

Moreover, $\ker(\Phi_{\mathcal{G},\mathcal{I}})$ is toric with a quadratic and square-free Gröbner basis.

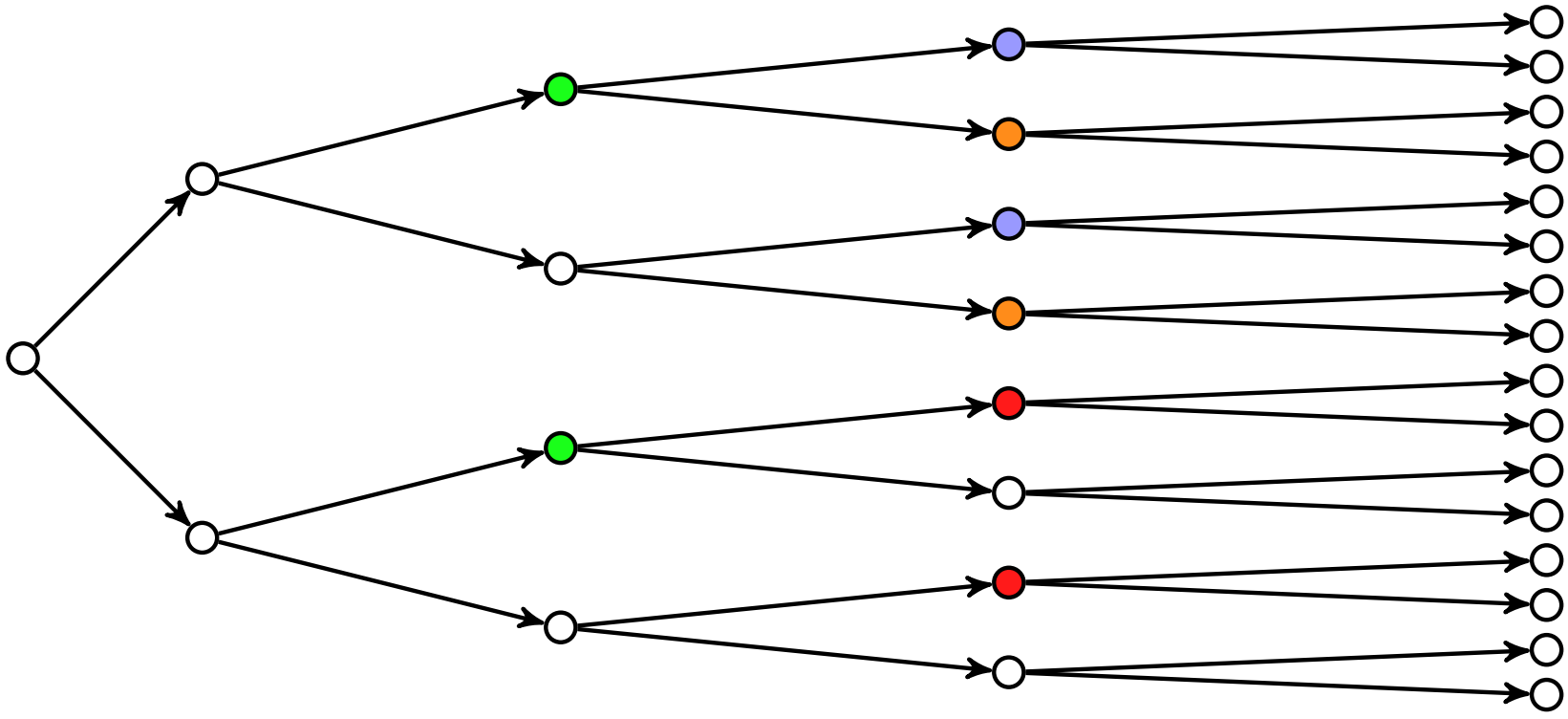
$$p^{(I)}(x_1,\ldots,x_m)=\prod_{i\in I}p^{(I)}(x_i\mid x_{\mathrm{pa}_{\mathcal{G}}(i)})\prod_{i\notin I}p(x_i\mid x_{\mathrm{pa}_{\mathcal{G}}(i)}).$$



An interventional staged tree $\mathcal{T}_{\mathcal{I}}$ where we have targeted outcomes

$$X_1=1 \quad \text{and} \quad X_1=2.$$

A proof in algebra and geometry yielding new statistics...



Such a tree encodes **context-specific information** that cannot be encoded with DAGs.

$$X_4 \perp\!\!\!\perp X_2 \mid X_1, X_3 = 1, \quad X_4 \perp\!\!\!\perp X_2 \mid X_3, X_1 = 1, \quad \text{and} \quad X_3 \perp\!\!\!\perp X_1 \mid X_2 = 1,$$

Interventional staged tree models provide a way to **represent** (and **learn** from data) causal models that encode context-specific information.

Ongoing Work

KTH Matematik

- **On the statistical side:**
  - Characterizations of model equivalence for interventional staged-tree models.
  - Causal structure learning algorithms.

- **On the algebraic side:**
  - Relation to conditional independence ideals.
  - A Macaulay2 package for interventional DAG models.

Vetenskapsrådet

Thank you for listening!