Probabilsitic Dependency Graphs

Oliver E. Richardson Joseph Y. Halpern

Cornell University
Department of Computer Science

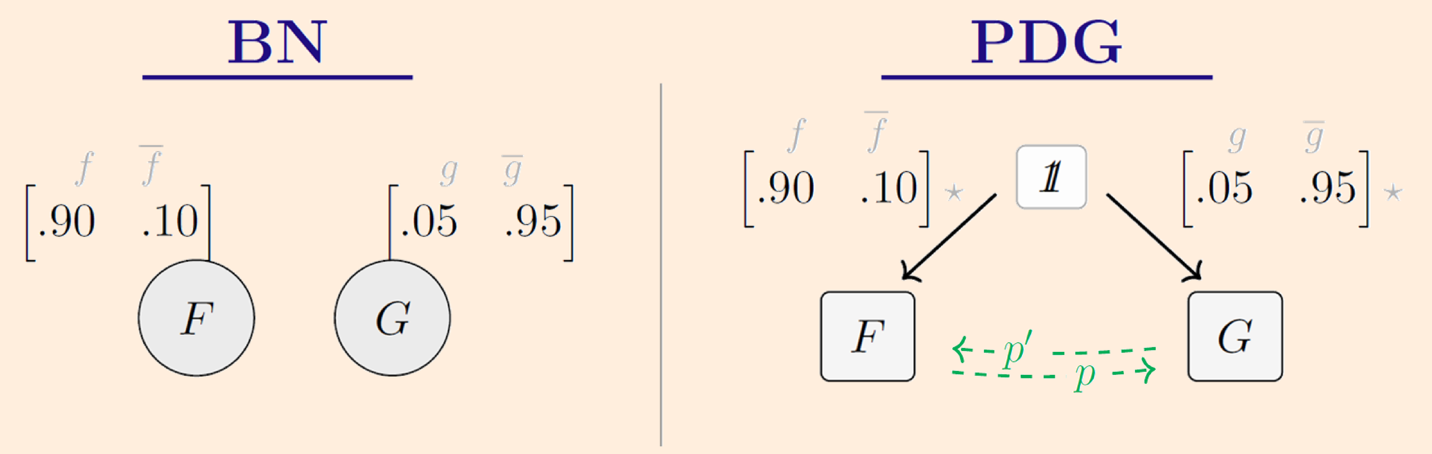
Why Yet Another Graphical Model?

PDGs...

- capture inconsistency, including conflicting information from multiple sources with varying reliability.
- are especially modular; to combine info from two sources, simply take a PDG union. This incorporates new data (edge cpds) and concepts (nodes) without affecting previous information.
- cleanly separate quantitative info (the cpds) from qualitative info (the edges), with variable confidence in both (the weights β and α). This is captured by terms *Inc* and *IDef* in our scoring function.
- have (several) natural semantics; one of them allows us to pick out a unique distribution. Using this distrbution, PDGs can capture BNs and factor graphs.

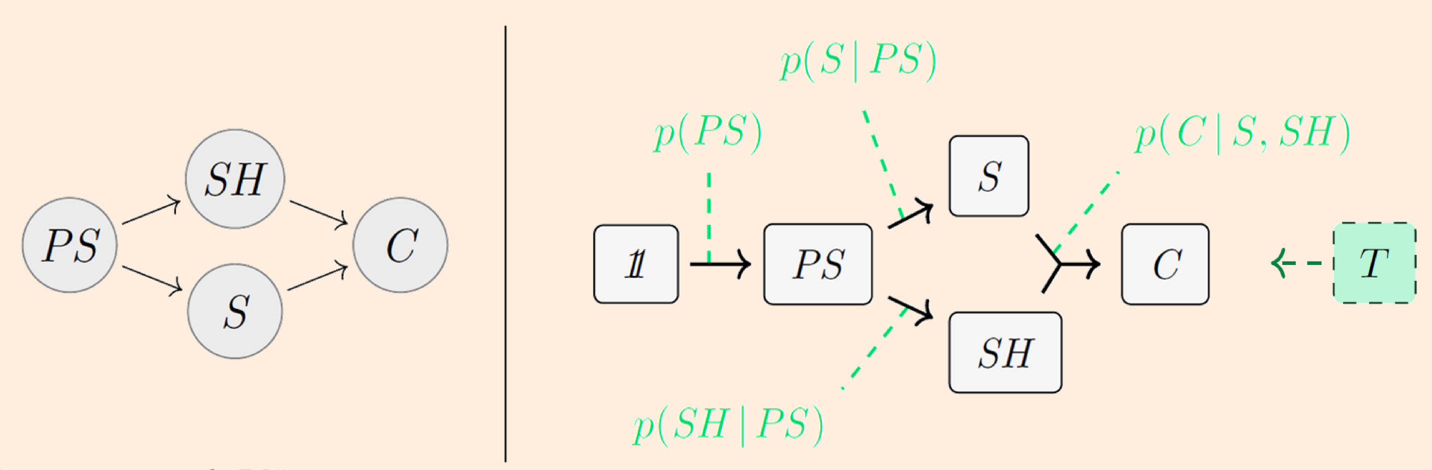
Modeling Examples

A Simple Illustration



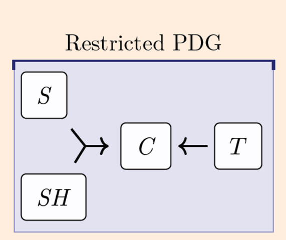
- The cpds of a PDG are attached to edges, not nodes.
- PDGs can incorporate arbitrary new probabilistic information.
- PDGs can be inconsistent
  - ...but BNs must resolve inconsistency first, which may break symmetry and irrecoverably lose information.

Bayesian Networks as PDGs

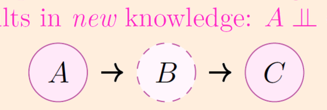


In contrast with BNs:

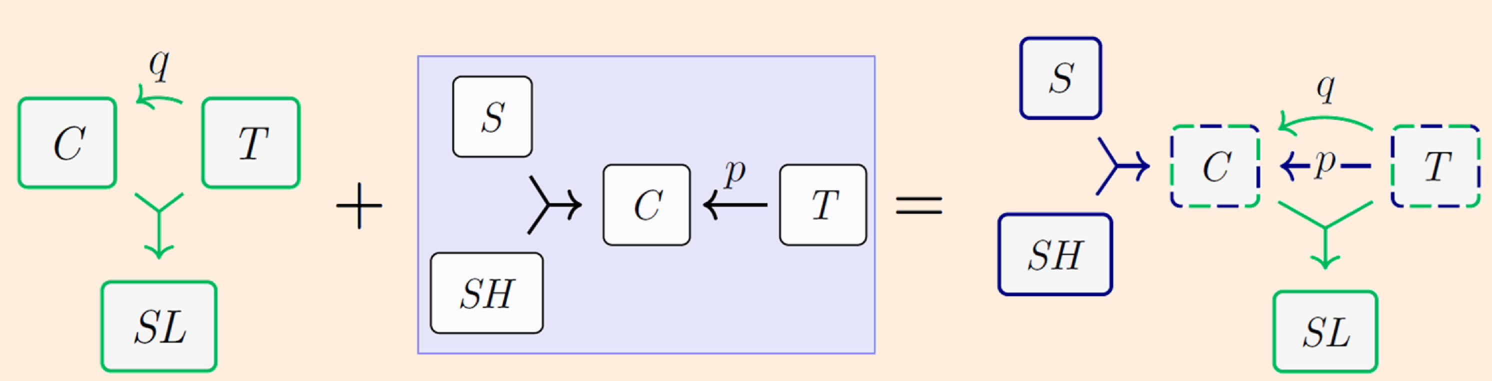
- edge composition has *quantitative* meaning, since edges have cpds;
- a variable can be the target of more than one cpd;
- arbitrary restrictions of PDGs are still PDGs.



In a qualitative BN: *removing* data results in *new* knowledge: $A \perp\!\!\!\perp C$.



Combining PDGs



- Arbitrary PDGs may be combined without loss of information
- They may have parallel edges (e.g., p, q), which directly conflict.

Definition (Probabilistic Dependency Graph)

A PDG is a tuple $\boldsymbol{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$, where

- $\mathcal{N}$ is a finite set of nodes (variables)
- $\mathcal{V}$ gives a set $\mathcal{V}(X)$ of possible values for each X;
- $\mathcal{E}$ is a set of labeled edges $\{X \xrightarrow{L} Y\}$, and associated to each $X \xrightarrow{L} Y$, there is:
  - $\mathbf{p}_L$ a cpd $\mathbf{p}_L(Y \mid X)$;
  - $\alpha_L \in [0, \infty)$ a confidence in the functional dependence $X \to Y$
  - $\beta_L \in (0, \infty)$ a confidence in the reliability of $\mathbf{p}_L$.

PDG Semantics

1. $\{\!\{\boldsymbol{m}\}\!\}$ The set of joint distributions consistent with $\boldsymbol{m}$;
2. $[\![\boldsymbol{m}]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with $\boldsymbol{m}$;

tradeoff parameter $\gamma \geq 0$

$$[\![\boldsymbol{m}]\!]_\gamma(\mu) := \mathit{Inc}_{\boldsymbol{m}}(\mu) + \gamma\, \mathit{IDef}_{\boldsymbol{m}}(\mu)$$

(Quantitative)

Definition (*Inc*)

The *incompatibility* of μ with $\boldsymbol{m}$:

$$\mathit{Inc}_{\boldsymbol{m}}(\mu) := \sum_{X \xrightarrow{L} Y} \beta_L\, \boldsymbol{D}(\mu_{Y|X} \parallel \mathbf{p}_L)$$

The *inconsistency* of $\boldsymbol{m}$ is

$$\mathit{Inc}(\boldsymbol{m}) := \inf_{\mu \in \Delta\mathcal{V}(\boldsymbol{m})} \mathit{Inc}_{\boldsymbol{m}}(\mu).$$

(Qualitative)

Definition (*IDef*)

The $\boldsymbol{m}$-*information deficit* of μ:

$$\mathit{IDef}_{\boldsymbol{m}}(\mu) = \overbrace{\sum_{X \xrightarrow{L} Y} \alpha_L \mathrm{H}_\mu(Y \mid X)}^{\text{\# bits to separately determine each target, knowing the source}} - \underbrace{\mathrm{H}(\mu)}_{\text{\# bits to determine all vars}}$$

3. $[\![\boldsymbol{m}]\!]^*$ The (unique) "best" joint distribution (in the quantitative limit).

$$[\![\boldsymbol{m}]\!]^* := \lim_{\gamma \to 0} \arg\min_\mu [\![\boldsymbol{m}]\!]_\gamma(\mu)$$

Properties of Semantics

Proposition (*the second semantics extends the first*)

$\{\!\{\boldsymbol{m}\}\!\} = \{\mu : [\![\boldsymbol{m}]\!]_0(\mu) = 0\}$.

Proposition (*If there there are distributions consistent with $\boldsymbol{m}$, the best distribution is one of them.*)

$[\![\boldsymbol{m}]\!]^* \in [\![\boldsymbol{m}]\!]_0^*$, *so if $\boldsymbol{m}$ is consistent, then* $[\![\boldsymbol{m}]\!]^* \in \{\!\{\boldsymbol{m}\}\!\}$.

Proposition (uniqueness for small γ)

1. *If* $0 < \gamma \leq \min_L \beta_L^{\boldsymbol{m}}$, *then* $[\![\boldsymbol{m}]\!]_\gamma^*$ *is a singleton.*
2. $\lim_{\gamma \to 0} [\![\boldsymbol{m}]\!]_\gamma^*$ *exists and is unique.*

Capturing BNs as PDGs

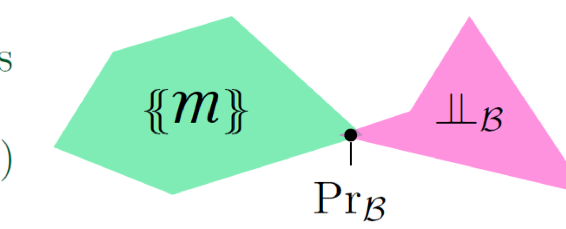
Let $\boldsymbol{m}_{\mathcal{B},\beta}$ be the PDG corresponding to the BN $\mathcal{B}$, with weights β.

Theorem (BNs are PDGs)

If $\mathcal{B}$ is a BN and $\Pr_{\mathcal{B}}$ is the distribution it specifies, then for all $\gamma > 0$ and all vectors β,

$$[\![\boldsymbol{m}_{\mathcal{B},\beta}]\!]_\gamma^* = \Pr_{\mathcal{B}}.$$

space of distributions consistent with $\boldsymbol{m}_{\mathcal{B}}$ (which minimize *Inc*)

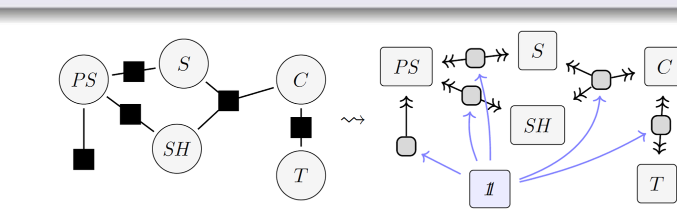


space of distributions with independencies of $\mathcal{B}$ (which can be shown to minimize *IDef*)

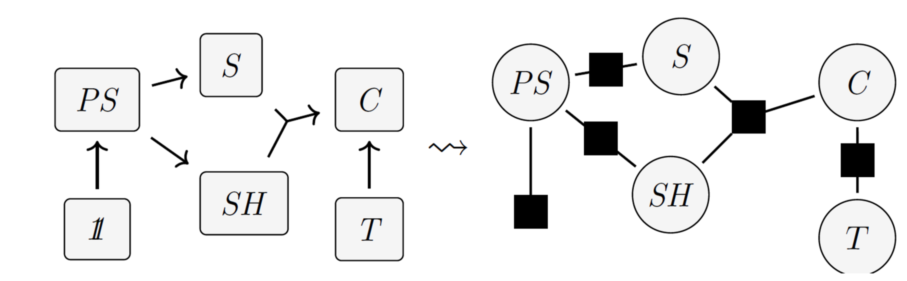
Capturing Factor Graphs as PDGs

Theorem (*PDGs capture factor graphs*)

We can naturally translate factor graphs and their exponential families, into PDGs, in a way which preserves their semantics.



PDGs as Factor Graphs



The cpds of a PDG are essentially factors. Are the semantics the same? Only for $\gamma = 1$.

Theorem

$[\![\boldsymbol{n}]\!]_1^* = \Pr_{\Phi_{\boldsymbol{n}}}$ *for all unweighted PDGs* $\boldsymbol{n}$.

$$[\![\boldsymbol{m}]\!]_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\overbrace{\beta_L \log \frac{1}{\mathbf{p}_L(y^{\mathbf{w}} | x^{\mathbf{w}})}}^{\text{log likelihood / cross entropy}} + \underbrace{(\alpha_L \gamma - \beta_L) \log \frac{1}{\mu(y^{\mathbf{w}} | x^{\mathbf{w}})}}_{\text{local regularization } (\beta_L > \alpha_L \gamma)} \right] - \underbrace{\gamma \log \frac{1}{\mu(\mathbf{w})}}_{\text{global regularization}} \right\}.$$

Inference and Inconsistency: a Glimpse.

Conditioning as inconsistency resolution.

To condition on $Y = y$, in $\boldsymbol{m}$, simply add the edge $1 \xrightarrow{\delta_y} Y$ to get $\boldsymbol{m}_{Y=y}$. Then $[\![\boldsymbol{m}_{Y=y}]\!]^* = [\![\boldsymbol{m}]\!]^* \mid (Y = y)$.

Querying $\Pr(Y \mid X)$ in a PDG $\boldsymbol{m}$.

- We can add $X \xrightarrow{p} Y$ to $\boldsymbol{m}$ with a cpt p, to get $\boldsymbol{m}^{+p}$.
- The choice of cpd p that minimizes the inconsistency of $\boldsymbol{m}^{+p}$ (which is strongly convex and smooth in p) is $[\![\boldsymbol{m}]\!]^*(Y \mid X)$,
- so oracle access to inconsistency yields fast inference by gradient descent.