Probabilsitic Dependency Graphs

Oliver E. Richardson   Joseph Y. Halpern

Cornell University
Department of Computer Science

AAAI, Febuary 2021

Yet Another Probabilistic Graphical Model

We introduce *probabilistic dependency graphs* (PDGs), a new class of graphical models for representing uncertainty.

Yet Another Probabilistic Graphical Model

We introduce *probabilistic dependency graphs* (PDGs), a new class of graphical models for representing uncertainty.

Why do we need another one?

Yet Another Probabilistic Graphical Model

We introduce *probabilistic dependency graphs* (PDGs), a new class of graphical models for representing uncertainty.

Why do we need another one?

- To resolve inconsistency, we must first model it.

Yet Another Probabilistic Graphical Model

We introduce *probabilistic dependency graphs* (PDGs), a new class of graphical models for representing uncertainty.

Why do we need another one?

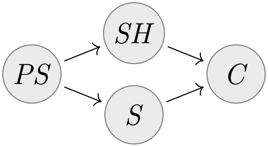
- To resolve inconsistency, we must first model it.
- In doing so, we get much more …

Two aspects of Bayesian Networks (BNs)

Qualitative BN, $\mathcal{G}$

an independence relation on variables

- $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Pa}(X)$, for all non-descendents Y of X



Two aspects of Bayesian Networks (BNs)

Qualitative BN, $\mathcal{G}$

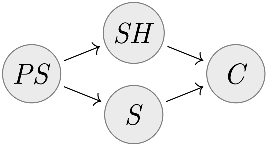
an independence relation on variables

- $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Pa}(X)$, for all non-descendents Y of X

(Quantitative) BN, $\mathcal{B} = (\mathcal{G}, \mathbf{p})$

a qualitative BN ($\mathcal{G}$) and a cpd $p_X(X \mid \mathbf{Pa}(X))$ for each variable X.

- Defines a joint distribution $\Pr_{\mathcal{B}}$ with the independencies $\perp\!\!\!\perp_{\mathcal{G}}$.



Two aspects of Bayesian Networks (BNs)

Qualitative BN, $\mathcal{G}$

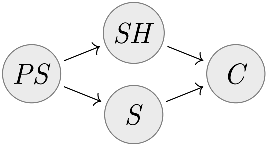
an independence relation on variables

- $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Pa}(X)$, for all non-descendents Y of X

(Quantitative) BN, $\mathcal{B} = (\mathcal{G}, \mathbf{p})$

a qualitative BN ($\mathcal{G}$) and a cpd $p_X(X \mid \mathbf{Pa}(X))$ for each variable X.

- Defines a joint distribution $\Pr_{\mathcal{B}}$ with the independencies $\perp\!\!\!\perp_{\mathcal{G}}$.



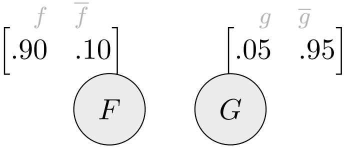
Modeling Example: Floomps and Guns

Grok thinks it likely (.95) that guns are illegal,
but that floomps (local slang) are legal (.90).

Modeling Example: Floomps and Guns

Grok thinks it likely (.95) that guns are illegal,
but that floomps (local slang) are legal (.90).

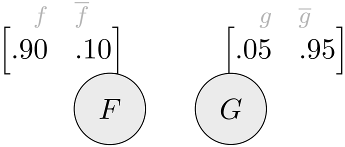
BN



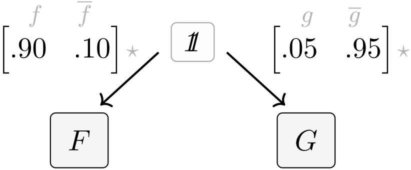
Modeling Example: Floomps and Guns

Grok thinks it likely (.95) that guns are illegal,
but that floomps (local slang) are legal (.90).

BN



PDG

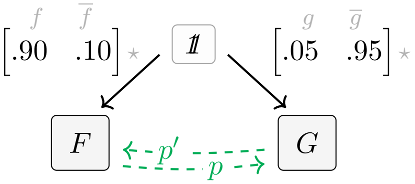


- The cpds of a PDG are attached to edges, not nodes.

Modeling Example: Floomps and Guns

BN

PDG

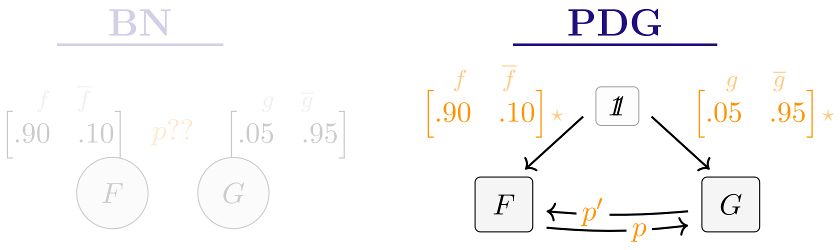


- The cpds of a PDG are attached to edges, not nodes.
- PDGs can incorporate arbitrary new probabilistic information.

Grok learns that Floomps and Guns have the same legal status (92%)

$$p(G \mid F) = \begin{array}{cc} g & \overline{g} \end{array} \begin{bmatrix} .92 & .08 \\ .08 & .92 \end{bmatrix} \begin{array}{c} f \\ \overline{f} \end{array} = \left(p'(F \mid G) \right)^{\mathsf{T}}$$

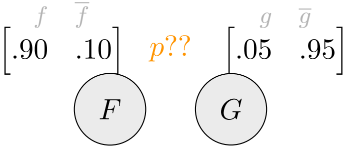
Modeling Example: Floomps and Guns



- The cpds of a PDG are attached to edges, not nodes.
- PDGs can incorporate arbitrary new probabilistic information.
- PDGs can be inconsistent

Modeling Example: Floomps and Guns

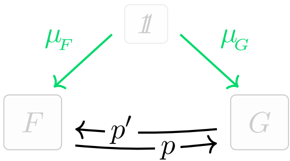
BN



Incorporated:

$\mu_F \quad \mu_G \quad p \quad p'$

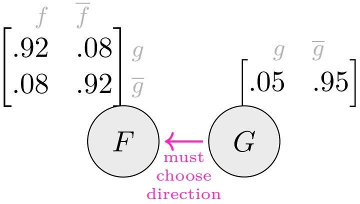
PDG



- 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.

Modeling Example: Floomps and Guns

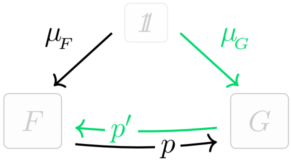
BN



Incorporated:

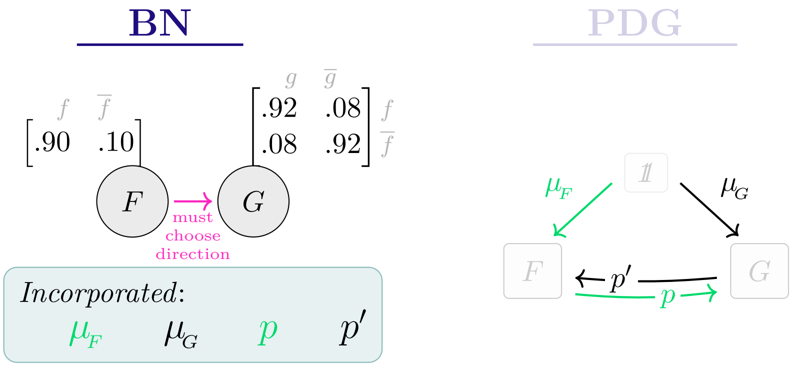
μ_F μ_G p p'

PDG



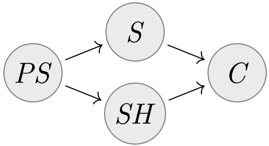
- The cpds of a PDG are attached to edges, not nodes.
- PDGs can incorporate arbitrary new probabilistic information.
- PDGs can be inconsistent
  - ...but BNs must resolve inconsistency first, which may break symmetry and irrecoverably lose information.

Modeling Example: Floomps and Guns

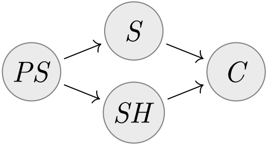


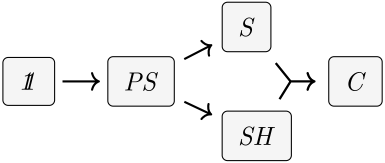
- The cpds of a PDG are attached to edges, not nodes.
- PDGs can incorporate arbitrary new probabilistic information.
- PDGs can be inconsistent
  - ... but BNs must resolve inconsistency first, which may break symmetry and irrecoverably lose information.

Bayesian Networks as PDGs

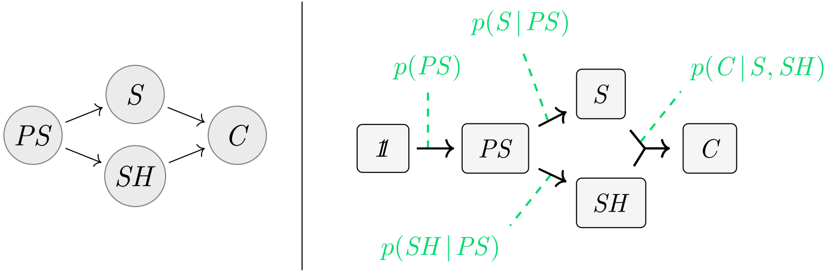


Bayesian Networks as PDGs





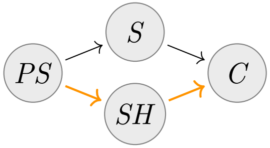
Bayesian Networks as PDGs

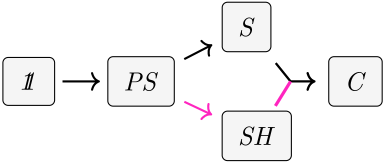


In contrast with BNs:

- edge composition has *quantitative* meaning, since edges have cpds;

Bayesian Networks as PDGs

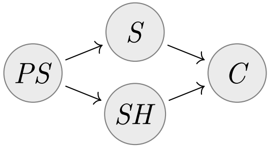


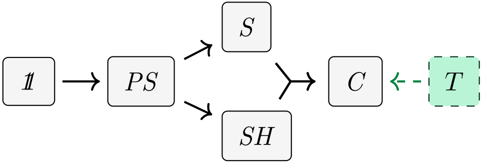


In contrast with BNs:

- edge composition has *quantitative* meaning, since edges have cpds;

Bayesian Networks as PDGs

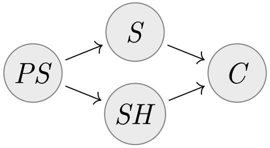


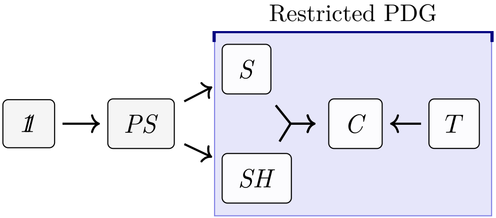


In contrast with BNs:

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

Bayesian Networks as PDGs

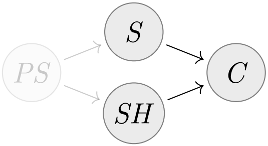




In contrast with BNs:

- 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.

Bayesian Networks as PDGs

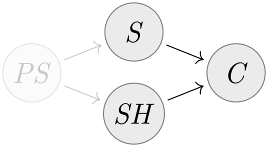


Must now give distributions on SH and S, or distinguish them as "observed" (a *conditional* BN).

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.
  - The analogue is false for BNs!

Bayesian Networks as PDGs



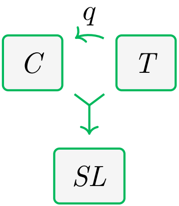
Must now give distributions on SH and S, or distinguish them as "observed" (a *conditional* BN).

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

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.
  - The analogue is false for BNs!

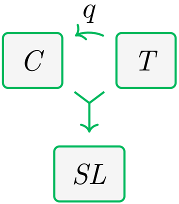
Combining PDGs



Grok wants to be supreme leader (SL).

- She notices that those who use tanning beds have more power,

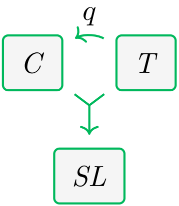
Combining PDGs



Grok wants to be supreme leader (SL).

- She notices that those who use tanning beds have more power,
- ...but mom says $q(C \mid T) = \begin{array}{c c} & \begin{array}{cc} c & \bar{c} \end{array} \\ & \begin{bmatrix} .15 & .85 \\ .02 & .98 \end{bmatrix} \begin{array}{c} t \\ \bar{t} \end{array} \end{array}.$

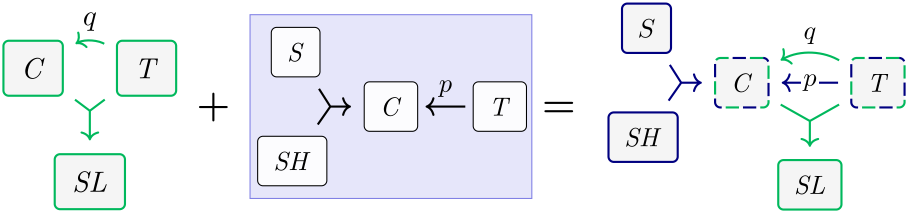
Combining PDGs



Grok wants to be supreme leader (SL).

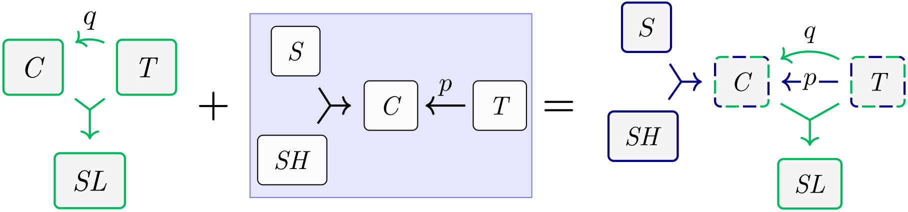
- She notices that those who use tanning beds have more power,
- ...but mom says $q(C \mid T) = \begin{array}{c} \begin{array}{cc} c & \bar{c} \end{array} \\ \begin{bmatrix} .15 & .85 \\ .02 & .98 \end{bmatrix} \end{array} \begin{array}{c} t \\ \bar{t} \end{array}.$
- Grok worries getting cancer from a tanning bed will make SL impossible.

Combining PDGs



- Arbitrary PDGs may be combined without loss of information

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 $\mathcal{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$,

Definition (Probabilistic Dependency Graph)

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

$\mathcal{N}$ is a finite set of nodes (variables)

Definition (Probabilistic Dependency Graph)

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

$\mathcal{N}$ is a finite set of nodes (variables)

$\mathcal{V}$ gives a set $\mathcal{V}(X)$ of possible values for each X;

$$\mathcal{V}(m) := \prod_{X \in \mathcal{N}} \mathcal{V}(X)$$ is the set of possible joint variable settings.

Definition (Probabilistic Dependency Graph)

A PDG is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$, 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 $\{X \xrightarrow{L} Y\}$, (hyper-edges)

Definition (Probabilistic Dependency Graph)

A PDG is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$, 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 $\{X \xrightarrow{L} Y\}$, (hyper-edges)
and associated to each $X \xrightarrow{L} Y$, there is:

Definition (Probabilistic Dependency Graph)

A PDG is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$, 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 $\{X \xrightarrow{L} Y\}$, (hyper-edges)
and associated to each $X \xrightarrow{L} Y$, there is:

$\mathbf{p}_L$ a cpd $\mathbf{p}_L(Y \mid X)$;

Definition (Probabilistic Dependency Graph)

A PDG is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$, 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 $\{X \xrightarrow{L} Y\}$, (hyper-edges)

and associated to each $X \xrightarrow{L} Y$, there is:

$\mathbf{p}_L$ a cpd $\mathbf{p}_L(Y \mid X)$;

$\beta_L \in (0, \infty)$ a confidence in the reliability of $\mathbf{p}_L$.

Definition (Probabilistic Dependency Graph)

A PDG is a tuple $\mathcal{m} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \alpha, \beta)$, 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 $\{X \xrightarrow{L} Y\}$, (hyper-edges)

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

$\{\!\{m\}\!\}$ The set of joint distributions consistent with m;

$[\![m]\!]_\gamma$ A function, scoring distributions by compatibility with m;

$[\![m]\!]^*$ The "best" joint distribution.

PDG Semantics

$\{\!\{m\}\!\}$ The set of joint distributions consistent with m;

$$\Big\{\mu \in \Delta[\mathcal{V}(m)] : \text{ for all } X \xrightarrow{L} Y \in \mathcal{E}.\ \ \mu(Y \,|\, X) = \mathbf{p}_L(Y \,|\, X)\Big\}$$

$[\![m]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with m;

$[\![m]\!]^*$ The "best" joint distribution.

PDG Semantics

$\{\!\{\mathcal{m}\}\!\}$ The set of joint distributions consistent with $\mathcal{m}$;

$$\left\{ \mu \in \Delta[\mathcal{V}(\mathcal{m})] : \text{ for all } X \xrightarrow{L} Y \in \mathcal{E}.\;\; \mu(Y \,|\, X) = \mathbf{p}_L(Y \,|\, X) \right\}$$

$[\![\mathcal{m}]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with $\mathcal{m}$;

tradeoff parameter $\gamma \geq 0$

$$[\![\mathcal{m}]\!]_\gamma(\mu) := \underbrace{\mathit{Inc}_{\mathcal{m}}(\mu)}_{\left(\substack{\text{quantitative} \\ \text{term}}\right)} + \gamma \underbrace{\mathit{IDef}_{\mathcal{m}}(\mu)}_{\left(\substack{\text{qualitative} \\ \text{term}}\right)}$$

$[\![\mathcal{m}]\!]^*$ The "best" joint distribution.

PDG Semantics

$\{\!\{\mathcal{M}\}\!\}$ The set of joint distributions consistent with $\mathcal{M}$;

$$\left\{ \mu \in \Delta[\mathcal{V}(\mathcal{M})] : \text{ for all } X \xrightarrow{L} Y \in \mathcal{E}.\ \ \mu(Y \,|\, X) = \mathbf{p}_{L}(Y \,|\, X) \right\}$$

$[\![\mathcal{M}]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with $\mathcal{M}$;

$$[\![\mathcal{M}]\!]_\gamma(\mu) := \underbrace{\mathit{Inc}_{\mathcal{M}}(\mu)}_{\left(\substack{\text{quantitative}\\ \text{term}}\right)} + \gamma \underbrace{\mathit{IDef}_{\mathcal{M}}(\mu)}_{\left(\substack{\text{qualitative}\\ \text{term}}\right)}$$

$[\![\mathcal{M}]\!]^*_\gamma$ The "best" joint distribution.

$$[\![\mathcal{M}]\!]^*_\gamma := \arg\min_{\mu} [\![\mathcal{M}]\!]_\gamma(\mu)$$

The Scoring Function

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

The Scoring Function

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

Intuition: beyond stating whether or not μ is consistent with $\boldsymbol{m}$, we score μ's compatibility with $\boldsymbol{m}$.

Motivating Examples.

$$\boldsymbol{m} := \begin{tikzcd} \boxed{\mathbb{1}} \arrow[r, bend left, "q"] \arrow[r, bend right, "p"'] & \boxed{X} \end{tikzcd}$$

The Scoring Function

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

Intuition: beyond stating whether or not μ is consistent with $\boldsymbol{m}$, we score μ's compatibility with $\boldsymbol{m}$.

Motivating Examples.

$$\boldsymbol{m} := \begin{tikzcd} \mathbb{1} \arrow[r, bend left, "q"] \arrow[r, bend right, "p"'] & X \end{tikzcd}$$

- If $p = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} .4 & .6 \end{bmatrix} \end{array}^{\star} = q$, then $\boldsymbol{m}$ is consistent, and compatible with the joint distribution $\mu(X) = p$.

The Scoring Function

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

Intuition: beyond stating whether or not μ is consistent with $\mathcal{m}$, we score μ's compatibility with $\mathcal{m}$.

Motivating Examples.

$$\mathcal{m} := \begin{tikzcd} \mathbb{1} \arrow[r, bend left, "q"] \arrow[r, bend right, "p"'] & X \end{tikzcd}$$

- If $p = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} .4 & .6 \end{bmatrix} \end{array}^{\star} = q$, then $\mathcal{m}$ is consistent, and compatible with the joint distribution $\mu(X) = p$.
- If $p = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} .4 & .6 \end{bmatrix} \end{array}^{\star}$ and $q = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} .5 & .5 \end{bmatrix} \end{array}^{\star}$, then $\mathcal{m}$ is not consistent, but $\mu = \begin{bmatrix} .45 & .55 \end{bmatrix}$ matches better than $\mu = \begin{bmatrix} .9 & .1 \end{bmatrix}$.

The Scoring Function

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

Intuition: beyond stating whether or not μ is consistent with $\boldsymbol{m}$, we score μ's compatibility with $\boldsymbol{m}$.

Motivating Examples.

$$\boldsymbol{m} := \begin{tikzcd} \boxed{1} \arrow[r, bend left, "q"] \arrow[r, bend right, "p"'] & \boxed{X} \end{tikzcd}$$

- If $p = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{bmatrix} .4 & .6 \end{bmatrix} \end{array}^{\star} = q$, then $\boldsymbol{m}$ is consistent, and compatible with the joint distribution $\mu(X) = p$.
- If $p = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{bmatrix} .4 & .6 \end{bmatrix} \end{array}^{\star}$ and $q = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{bmatrix} .5 & .5 \end{bmatrix} \end{array}^{\star}$, then $\boldsymbol{m}$ is not consistent, but $\mu = \begin{bmatrix} .45 & .55 \end{bmatrix}$ matches better than $\mu = \begin{bmatrix} .9 & .1 \end{bmatrix}$.
- If $p = \begin{bmatrix} .4 & .6 \end{bmatrix}$ and $q = \begin{bmatrix} 0 & 1 \end{bmatrix}$, then $\boldsymbol{m}$ is much more inconsistent than before, even though $\{\!\{\boldsymbol{m}\}\!\} = \emptyset$ in both cases.

The Scoring Function

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

Definition (*Inc*)

The *incompatibility* of a joint distribution μ with $\boldsymbol{m}$ is given by

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

The Scoring Function

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

Definition (*Inc*)

The *incompatibility* of a joint distribution μ with $\boldsymbol{m}$ is given by

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

The Scoring Function

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

Definition (Inc)

The *incompatibility* of a joint distribution μ with $\mathcal{M}$ is given by

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

$$\boldsymbol{D}(\mu \,\|\, \nu) = \sum_{w \in \mathit{Supp}(\mu)} \mu(w) \log \frac{\mu(w)}{\nu(w)}$$ is the relative entropy from ν to μ.

The Scoring Function

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

Definition (Inc)

The *incompatibility* of a joint distribution μ with $\boldsymbol{m}$ is given by

$$\mathit{Inc}_{\boldsymbol{m}}(\mu) := \sum_{X \xrightarrow{L} Y} \beta_L \mathop{\mathbb{E}}_{x \sim \mu_X} \boldsymbol{D}\Big(\mu(Y \mid X{=}x) \,\Big\|\, \mathbf{p}_L(x)\Big).$$

$$\boldsymbol{D}(\mu \,\|\, \nu) = \sum_{w \in Supp(\mu)} \mu(w) \log \frac{\mu(w)}{\nu(w)}$$ is the relative entropy from ν to μ.

The Scoring Function

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

Definition (Inc)

The *incompatibility* of a joint distribution μ with $\boldsymbol{m}$ is given by

$$\mathit{Inc}_{\boldsymbol{m}}(\mu) := \sum_{X \xrightarrow{L} Y} \beta_L \mathop{\mathbb{E}}_{x \sim \mu_X} \boldsymbol{D}\Big(\mu(Y \mid X{=}x) \;\Big\|\; \mathbf{p}_L(x)\Big).$$

The *inconsistency* of $\boldsymbol{m}$ is the smallest possible incompatibility,

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

The Scoring Function

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

Intuition: each edge $X \xrightarrow{L} Y$ indicates that Y is determined (perhaps noisily) by X alone.

The Scoring Function

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

Intuition: each edge $X \xrightarrow{L} Y$ indicates that Y is determined (perhaps noisily) by X alone.

So a μ with uncertainty in Y after X is known (beyond pure noise) is qualitatively worse.

The Scoring Function

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

Intuition: each edge $X \xrightarrow{L} Y$ indicates that Y is determined (perhaps noisily) by X alone.

So a μ with $\overbrace{\text{uncertainty in } Y \text{ after } X \text{ is known}}^{\textit{measured by } \mathrm{H}(Y \mid X)}$ (beyond $\underbrace{\text{pure noise}}_{\mathrm{H}(\mu)}$) is qualitatively worse.

The Scoring Function

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

Definition (*IDef*)

The *information deficit* of a distribution μ with respect to $\boldsymbol{m}$ is

$$\mathit{IDef}_{\boldsymbol{m}}(\mu) := \sum_{X \xrightarrow{L} Y} \alpha_L \, \mathrm{H}_\mu(Y \mid X) - \mathrm{H}(\mu).$$

The Scoring Function

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

Definition (IDef)

The *information deficit* of a distribution μ with respect to $\boldsymbol{m}$ is

$$\mathit{IDef}_{\boldsymbol{m}}(\mu) := \sum_{X \xrightarrow{L} Y} \alpha_L \, \mathrm{H}_\mu(Y \,|\, X) - \underbrace{\mathrm{H}(\mu)}_{\text{(a) \# bits needed to determine all variables}}.$$

The Scoring Function

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

Definition (IDef)

The *information deficit* of a distribution μ with respect to $\mathcal{m}$ is

(b) # bits required to separately determine each target, knowing the source

$$\mathit{IDef}_{\mathcal{m}}(\mu) := \overbrace{\sum_{X \xrightarrow{L} Y} \alpha_L \, \mathrm{H}_\mu(Y \mid X)} - \underbrace{\mathrm{H}(\mu)}.$$

(a) # bits needed to determine all variables

The Scoring Function

$[\![m]\!]_\gamma(\mu) := Inc_m(\mu) + \gamma\ IDef_m(\mu)$

Definition ($IDef$)

The m-*information deficit* of μ:

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

\# bits to separately determine each target, knowing the source

\# bits to determine all vars

Examples

- $m_0 = \boxed{X} \quad \boxed{Y}$

  $IDef_{m_0}(\mu) = -\mathrm{H}_\mu(X, Y)$
  (optimal μ maximizes entropy of X, Y)

The Scoring Function

$$[\![m]\!]_\gamma(\mu) := Inc_m(\mu) + \gamma\, IDef_m(\mu)$$

Definition (*IDef*)

The m-*information deficit* of μ:

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

Examples

- $m_0 = \boxed{X} \quad \boxed{Y}$

  $IDef_{m_0}(\mu) = -\mathrm{H}_\mu(X, Y)$

  (optimal μ maximizes entropy of X, Y)

- $m_1 = \boxed{X} \longrightarrow \boxed{Y}$

  $IDef_{m_1}(\mu) = -\mathrm{H}_\mu(X)$

  (optimal μ maximizes entropy of X)

The Scoring Function

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

Definition (*IDef*)

The m-*information deficit* of μ:

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

Examples

- $m_0 = \boxed{X} \quad \boxed{Y}$

  $\mathit{IDef}_{m_0}(\mu) = -\mathrm{H}_\mu(X, Y)$
  (optimal μ maximizes entropy of X, Y)

- $m_1 = \boxed{X} \longrightarrow \boxed{Y}$

  $\mathit{IDef}_{m_1}(\mu) = -\mathrm{H}_\mu(X)$
  (optimal μ maximizes entropy of X)

- $m_2 = \boxed{X} \rightrightarrows \boxed{Y}$

  $\mathit{IDef}_{m_2}(\mu) = -\mathrm{H}_\mu(X) + \mathrm{H}_\mu(Y \mid X)$
  (optimal μ maximizes entropy for X, and makes Y a function of X)

The Scoring Function

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

Definition (IDef)

The m-*information deficit* of μ:

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

Examples

- $m_0 = \boxed{X} \qquad \boxed{Y}$

  $\mathit{IDef}_{m_0}(\mu) = -\mathrm{H}_\mu(X, Y)$

  (optimal μ maximizes entropy of X, Y)

- $m_1 = \boxed{X} \longrightarrow \boxed{Y}$

  $\mathit{IDef}_{m_1}(\mu) = -\mathrm{H}_\mu(X)$

  (optimal μ maximizes entropy of X)

- $m_2 = \boxed{X} \rightrightarrows \boxed{Y}$

  $\mathit{IDef}_{m_2}(\mu) = -\mathrm{H}_\mu(X) + \mathrm{H}_\mu(Y \mid X)$

  (optimal μ maximizes entropy for X, and makes Y a function of X)

- $m_3 = \boxed{X} \rightleftarrows \boxed{Y}$

  $\mathit{IDef}_{m_3}(\mu) = -\mathrm{I}_\mu(X; Y)$

  (opt. μ makes X, Y share information)

The Scoring Function

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

Definition (IDef)

The m-*information deficit* of μ:

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

Examples

- $m_0 = \boxed{X} \quad \boxed{Y}$

  $\mathit{IDef}_{m_0}(\mu) = -\,\mathrm{H}_\mu(X, Y)$

  (optimal μ maximizes entropy of X, Y)

- $m_1 = \boxed{X} \longrightarrow \boxed{Y}$

  $\mathit{IDef}_{m_1}(\mu) = -\mathrm{H}_\mu(X)$

  (optimal μ maximizes entropy of X)

- $m_2 = \boxed{X} \rightrightarrows \boxed{Y}$

  $\mathit{IDef}_{m_2}(\mu) = -\,\mathrm{H}_\mu(X) + \mathrm{H}_\mu(Y \mid X)$

  (optimal μ maximizes entropy for X, and makes Y a function of X)

- $m_3 = \boxed{X} \rightleftarrows \boxed{Y}$

  $\mathit{IDef}_{m_3}(\mu) = -\,\mathrm{I}_\mu(X; Y)$

  (opt. μ makes X, Y share information)

Information Diagrams

The Scoring Function

$$[\![\boldsymbol{m}]\!]_\gamma(\mu) := \mathit{Inc}_{\boldsymbol{m}}(\mu) + \underbrace{\gamma}_{\text{tradeoff parameter } \gamma \geq 0} \mathit{IDef}_{\boldsymbol{m}}(\mu)$$

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

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

Definition (IDef)

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

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

The Scoring Function

- A BN strictly enforces the qualitative picture (large γ)

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

tradeoff parameter $\gamma \geq 0$

Definition (*Inc*)

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

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

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

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

Definition (*IDef*)

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

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

The Scoring Function

- A BN strictly enforces the qualitative picture (large γ)
- we are interested in the quantitative limit (small γ)

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

tradeoff parameter $\gamma \geq 0$

Definition (Inc)

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

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

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

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

Definition (IDef)

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

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

PDG Semantics

$\{\!\{\mathcal{M}\}\!\}$ The set of joint distributions consistent with $\mathcal{M}$;

$$\Big\{ \mu \in \Delta[\mathcal{V}(\mathcal{M})] :\ \text{for all } X \xrightarrow{L} Y \in \mathcal{E}.\ \ \mu(Y \,|\, X) = \mathbf{p}_{_L}(Y \,|\, X) \Big\}$$

$[\![\mathcal{M}]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with $\mathcal{M}$;

$$[\![\mathcal{M}]\!]_\gamma(\mu) := \underbrace{\mathit{Inc}_{\mathcal{M}}(\mu)}_{\begin{pmatrix}\text{quantitative}\\ \text{term}\end{pmatrix}} + \gamma \underbrace{\mathit{IDef}_{\mathcal{M}}(\mu)}_{\begin{pmatrix}\text{qualitative}\\ \text{term}\end{pmatrix}}$$

$[\![\mathcal{M}]\!]^*_\gamma$ The "best" joint distribution.

$$[\![\mathcal{M}]\!]^*_\gamma := \arg\min_\mu [\![\mathcal{M}]\!]_\gamma(\mu)$$

PDG SEMANTICS

$\{\mathcal{m}\}$ The set of joint distributions consistent with $\mathcal{m}$;

$$\left\{ \mu \in \Delta[\mathcal{V}(\mathcal{m})] : \text{ for all } X \xrightarrow{L} Y \in \mathcal{E}.\ \ \mu(Y \mid X) = \mathbf{p}_L(Y \mid X) \right\}$$

$[\![\mathcal{m}]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with $\mathcal{m}$;

Proposition (uniqueness for small γ)

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

$[\![\mathcal{m}]\!]^*_\gamma$ The "best" joint distribution.

$$[\![\mathcal{m}]\!]^*_\gamma := \arg\min_\mu [\![\mathcal{m}]\!]_\gamma(\mu)$$

PDG Semantics

$\{\!\{m\}\!\}$ The set of joint distributions consistent with m;

$$\left\{ \mu \in \Delta[\mathcal{V}(m)] : \text{ for all } X \xrightarrow{L} Y \in \mathcal{E}. \;\; \mu(Y \mid X) = \mathbf{p}_L(Y \mid X) \right\}$$

$[\![m]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with m;

Proposition (uniqueness for small γ)

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

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

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

PDG Semantics

1. $\{\!\!\{ \mathcal{M} \}\!\!\}$ The set of joint distributions consistent with $\mathcal{M}$;
2. $[\![\mathcal{M}]\!]_\gamma$ A loss function (parameterized by γ), scoring a joint distribution's compatibility with $\mathcal{M}$;
3. $[\![\mathcal{M}]\!]^*$ The "best" joint distribution.

Proposition (*the second semantics extends the first*)

$\{\!\!\{ \mathcal{M} \}\!\!\} = \{ \mu : [\![\mathcal{M}]\!]_0(\mu) = 0 \}$.

PDG Semantics

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

Proposition (*the second semantics extends the first*)

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

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

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

Capturing Bayesian Networks

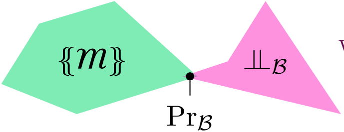
Let $\mathcal{m}_{\mathcal{B},\beta}$ be the PDG corresponding to the BN $\mathcal{B}$, with weights β.

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,

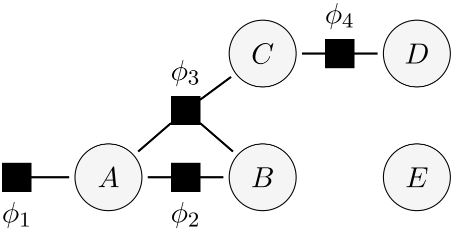
$$[\![\mathcal{m}_{\mathcal{B},\beta}]\!]^*_\gamma = \{\Pr_{\mathcal{B}}\}, \quad \textit{and thus} \quad [\![\mathcal{m}_{\mathcal{B},\beta}]\!]^* = \Pr_{\mathcal{B}}\,.$$

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



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

Factor Graphs



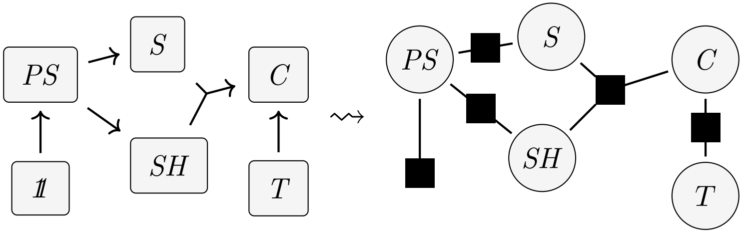
Definition

A *factor graph* Φ is a set of random variables $\mathcal{X} = \{X_i\}$ and *factors* $\{\phi_J : \mathcal{V}(X_J) \to \mathbb{R}_{\geq 0}\}_{J \in \mathcal{J}}$, where $X_J \subseteq \mathcal{X}$; define

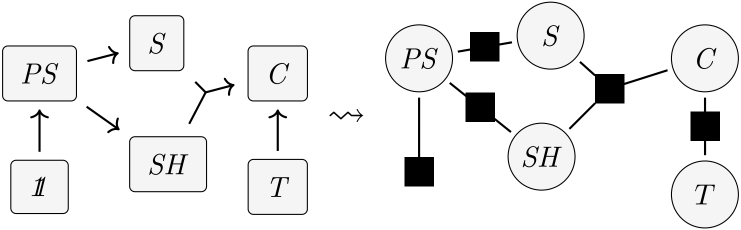
$$\Pr_\Phi(\vec{x}) = \frac{1}{Z_\Phi} \prod_{J \in \mathcal{J}} \phi_J(\vec{x}_J),$$

where Z_Φ is the normalization constant.

PDGs as Factor Graphs

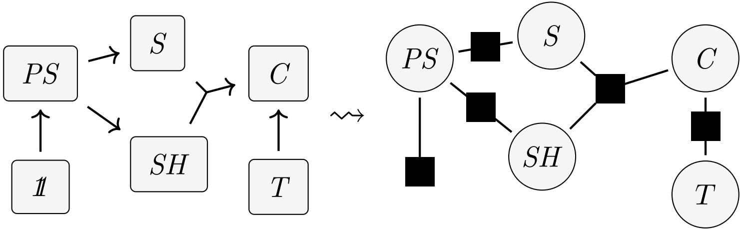


PDGs as Factor Graphs



The cpds of a PDG are essentially factors. Are the semantics different?

PDGs as Factor Graphs

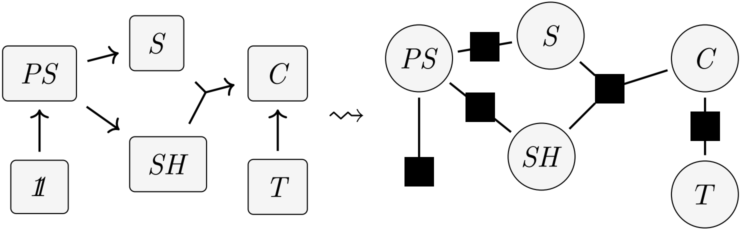


The cpds of a PDG are essentially factors. Are the semantics different? Not for $\gamma = 1$.

Theorem

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

PDGs as Factor Graphs



The cpds of a PDG are essentially factors. Are the semantics different? Not for $\gamma = 1$.

Theorem

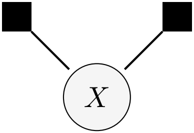
$[\![\boldsymbol{n}]\!]^*_1 = \Pr_{\Phi_{\boldsymbol{n}}}$ *for all unweighted PDGs* $\boldsymbol{n}$.

Theorem

For all unweighted PDGs $\boldsymbol{n}$ *and non-negative vectors* $\mathbf{v}$ *over the edges of* $\boldsymbol{n}$*, and all* $\gamma > 0$*, we have that* $[\![(\boldsymbol{n}, \mathbf{v}, \gamma\mathbf{v})]\!]_\gamma = \gamma\, \mathit{GFE}_{(\Phi_{\boldsymbol{n}}, \mathbf{v})}$*; consequently,* $[\![(\boldsymbol{n}, \mathbf{v}, \gamma\mathbf{v})]\!]^*_\gamma = \{\Pr_{(\Phi_{\boldsymbol{n}}, \mathbf{v})}\}$.

Why not use factor graphs, then?

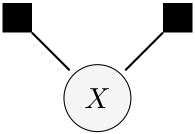
$$m := \begin{tikzcd} \mathbb{1} \arrow[d, bend right, "q"'] \arrow[d, bend left, "p"] \\ X \end{tikzcd}$$



$=: \Phi$

Why not use factor graphs, then?

$$m := \begin{tikzcd} \boxed{\mathbb{1}} \arrow[d, bend right, "q"'] \arrow[d, bend left, "p"] \\ \boxed{X} \end{tikzcd}$$

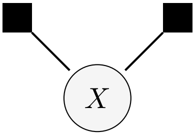


$$=: \Phi$$

- If $p = q$, then $[\![m]\!]^* = p = q \ldots$

Why not use factor graphs, then?

$$\boldsymbol{m} := \begin{tikzcd} \boxed{\mathbb{1}} \arrow[d, bend right, "q"'] \arrow[d, bend left, "p"] \\ \boxed{X} \end{tikzcd}$$

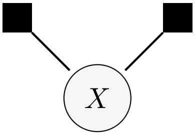


$$=: \Phi$$

- If $p = q$, then $[\![\boldsymbol{m}]\!]^* = p = q \ldots$
- $\ldots$ but $\Pr_\Phi \propto p^2$

Why not use factor graphs, then?

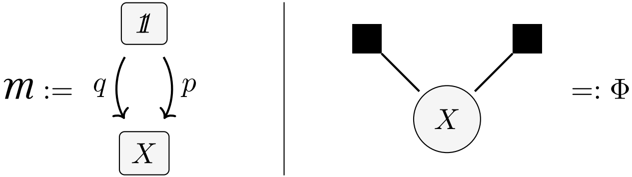
$$\boldsymbol{m} := \begin{tikzcd} \boxed{\mathbb{1}} \arrow[d, bend right, "q"'] \arrow[d, bend left, "p"] \\ \boxed{X} \end{tikzcd}$$



$=: \Phi$

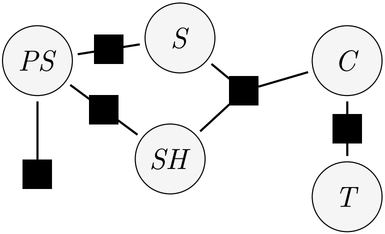
- If $p = q$, then $[\![\boldsymbol{m}]\!]^* = p = q \ldots$
- $\ldots$ but $\Pr_\Phi \propto p^2$
- More generally, (positive) factors individually have *no meaning*,

Why not use factor graphs, then?



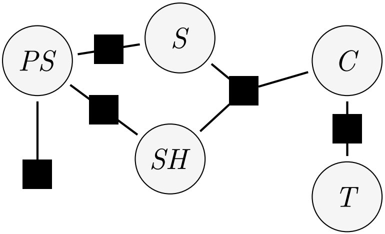
- If $p = q$, then $[\![m]\!]^* = p = q \ldots$
- $\ldots$ but $\Pr_\Phi \propto p^2$
- More generally, (positive) factors individually have *no meaning*,
- a factor graph can fail to normalize, in which case it has no global semantics either.

Factor Graphs as PDGs

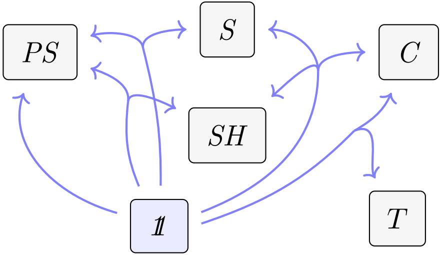


$\rightsquigarrow$

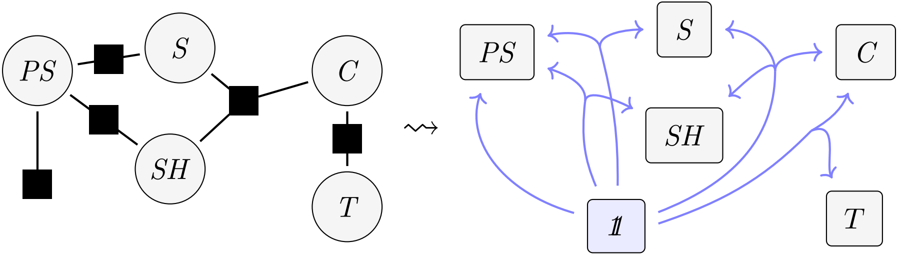
Factor Graphs as PDGs



$\rightsquigarrow$



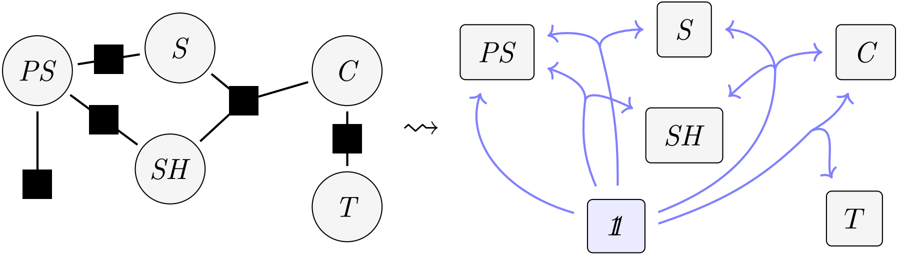
Factor Graphs as PDGs



Theorem

$\Pr_\Phi = [\![\mathcal{N}_\Phi]\!]^*_1$ *for all factor graphs* Φ.

Factor Graphs as PDGs



Theorem

$\Pr_\Phi = [\![m_\Phi]\!]_1^*$ *for all factor graphs* Φ.

Theorem

For all weighted factor graphs $\Psi = (\Phi, \theta)$ *and all* $\gamma > 0$*, we have that* $\mathit{GFE}_\Psi = {}^1\!/\!_\gamma [\![m_{\Psi,\gamma}]\!]_\gamma + C$ *for some constant* C*, so* $\Pr_\Psi$ *is the unique element of* $[\![m_{\Psi,\gamma}]\!]_\gamma^*$.

Letting $x^{\mathbf{w}}$ and $y^{\mathbf{w}}$ denote the values of X and Y, respectively, in $\mathbf{w} \in \mathcal{V}(\mathcal{M})$, we have

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

Inference and Inconsistency: a Glimpse.

Conditioning as inconsistency resolution.

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

Inference and Inconsistency: a Glimpse.

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

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

- We can add $X \xrightarrow{p} Y$ to $\mathcal{m}$ with a cpt p, to get $\mathcal{m}^{+p}$.

Inference and Inconsistency: a Glimpse.

Conditioning as inconsistency resolution.
To condition on $Y = y$, in $\mathcal{M}$, simply add the edge $\mathbb{1} \xrightarrow{\delta_y} Y$ to get $\mathcal{M}_{Y=y}$.
Then $[\![\mathcal{M}_{Y=y}]\!]^* = [\![\mathcal{M}]\!]^* \mid (Y = y)$.

Querying $\Pr(Y \mid X)$ **in a PDG** $\mathcal{M}$**.**

- We can add $X \xrightarrow{p} Y$ to $\mathcal{M}$ with a cpt p, to get $\mathcal{M}^{+p}$.
- The choice of cpd p that minimizes the inconsistency of $\mathcal{M}^{+p}$ (which is strongly convex and smooth in p) is $[\![\mathcal{M}]\!]^*(Y \mid X)$,

Inference and Inconsistency: a Glimpse.

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

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

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

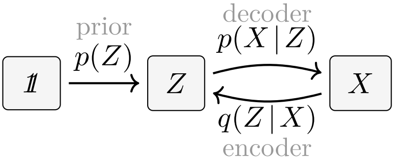
Inference and Inconsistency: a Glimpse.

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

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

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

This is closely related to standard variational techniques!

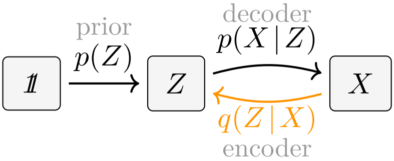


Conditioning as inconsistency resolution.
To condition on $Y = y$, in $\mathcal{M}$, simply add the edge $\mathbb{1} \xrightarrow{\delta_y} Y$ to get $\mathcal{M}_{Y=y}$.
Then $[\![\mathcal{M}_{Y=y}]\!]^* = [\![\mathcal{M}]\!]^* \mid (Y = y)$.

Querying $\Pr(Y \mid X)$ **in a PDG** $\mathcal{M}$**.**

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

This is closely related to standard variational techniques!



Summary

PDGs...

- capture inconsistency, including conflicting information from multiple sources with varying reliability.

Summary

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.

Summary

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.

Summary

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.

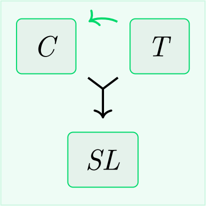
Summary

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.

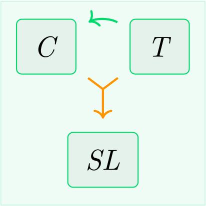
But there is much more to be done!

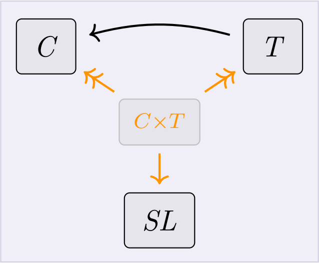
Hyper-graphs? Or merely graphs?



main definition

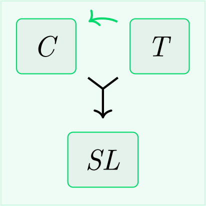
Hyper-graphs? Or merely graphs?

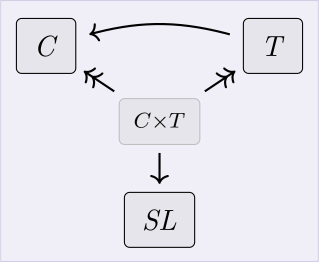




main definition

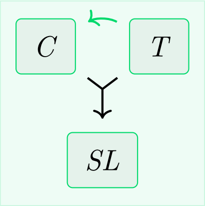
Hyper-graphs? Or merely graphs?

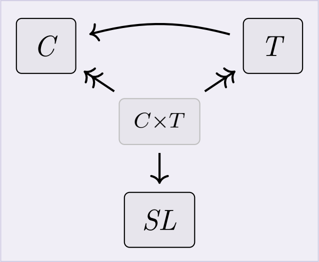




- This widget expands state space, but graphs are simpler.

Hyper-graphs? Or merely graphs?



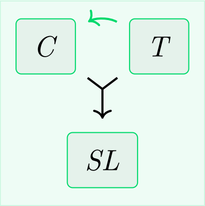


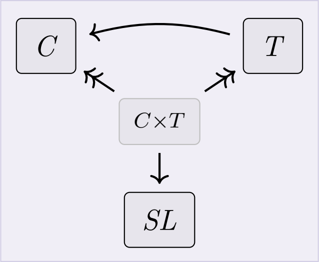
- This widget expands state space, but graphs are simpler.
- There is a natural correspondence

joint distributions $\leftrightarrows$ expanded joint distributions satisfying coherence constraints

main definition

Hyper-graphs? Or merely graphs?





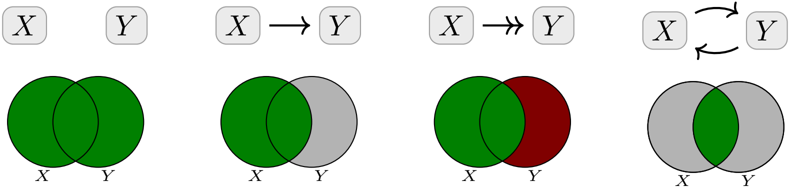
- This widget expands state space, but graphs are simpler.
- There is a natural correspondence

joint distributions $\leftrightarrows$ expanded joint distributions satisfying coherence constraints

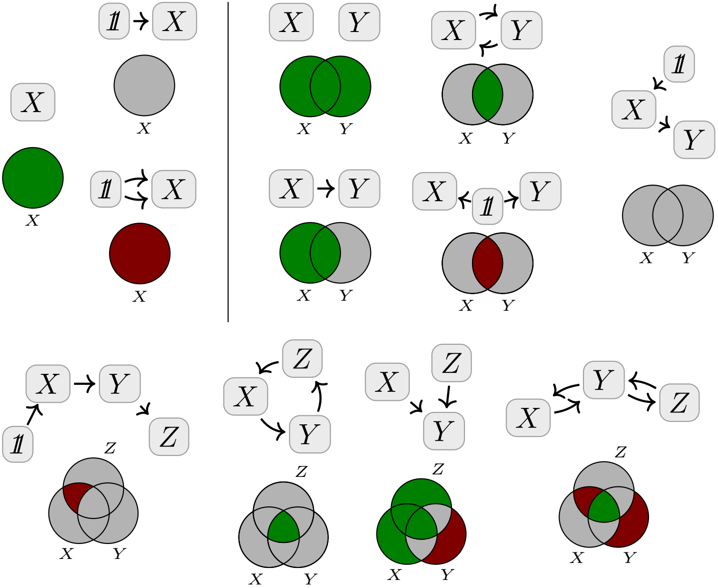
(working directly with hypergraphs is also possible)

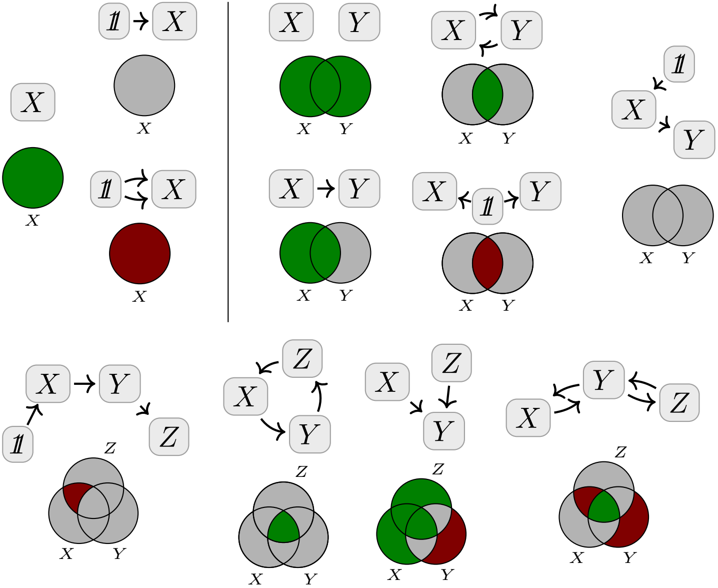
main definition

ILLUSTRATIONS OF *IDef*

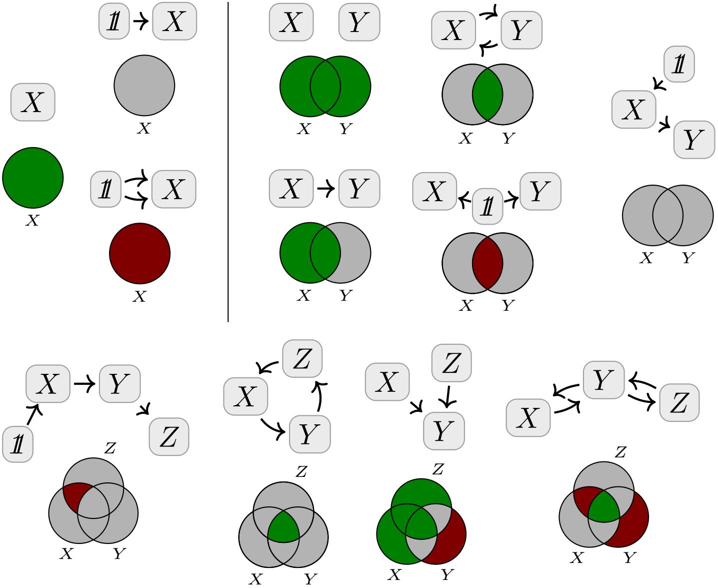


back to semantics

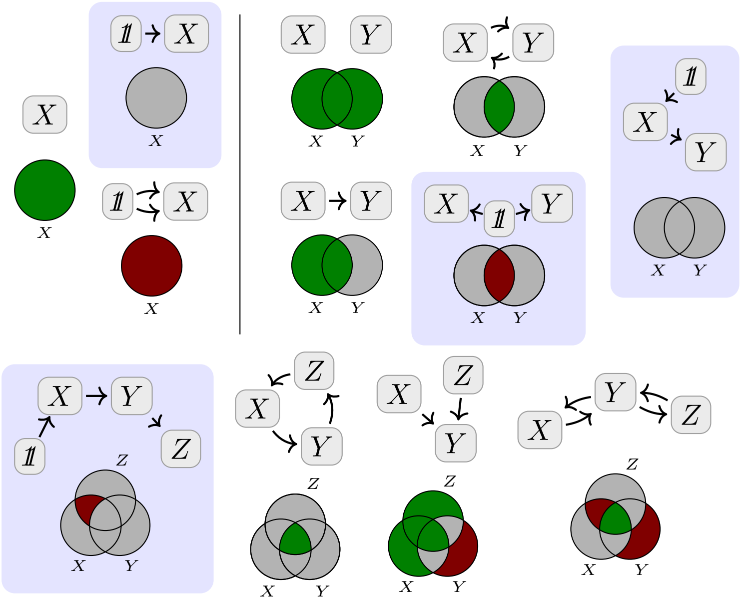




𝟙
X
X
X
𝟙
X
X
X
Y
X
Y
X
Y
X
Y
X
Y
X
Y
X
𝟙
Y
X
Y
𝟙
X
Y
X
Y
X
Y
𝟙
Z
Z
X
Y
Z
X
Y
Z
X
Y
X
Z
Y
Z
X
Y
Y
Z
X
Z
X
Y



X
X
𝟙 → X
X
𝟙 X
X
X Y
X Y
X Y
X Y
X → Y
X Y
X 𝟙 Y
X Y
𝟙
X
Y
X Y
X → Y
𝟙
Z
Z
X Y
Z
X
Y
Z
X Y
X
Z
Y
Z
X Y
Y
X
Z
Z
X Y


X
𝟙
X
Y
Z