**Definition**

- A **Bayesian network** is a graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG).

- Nodes → Random variables
  Edges → Conditional dependencies

- Node (RV) is **conditionally independent** of its non-descendants; given the state of all its parents.  
  or

- Node (RV) is **conditionally independent** of all other nodes j, given its Markov blanket.

- Variables $X$, $Y$ are **conditionally independent (CI)** given $Z$ if
  $\Pr(X \text{ and } Y \mid Z) = \Pr(X \mid Z) \Pr(Y \mid Z)$. 
Examples
Learning Bayesian Network

How to find right DAG?

Scoring criteria!
Scoring Criteria

• A scoring function $Q(G, D)$ evaluates how well a DAG explains the data.
• We will only consider score equivalent and decomposable scoring functions
  - Bayesian Information Criterion (BIC),
  - Bayesian Dirichlet Equivalent (BDE).

• **Score equivalent**: Score of two Markov equivalent graphs are the same.

**WARNING**: Two different DAGs may represent the same probability model! If so, they are called **Markov equivalent**.
Score Decomposable

A scoring function is **score decomposable** if there exists a set of functions (local scores)

\[ q_i|_B : DATA(\{i\} \cup B, d) \rightarrow \mathbb{R} \]

such that

\[ Q(G, D) = \sum_{i \in N} q_{i|pa_G(i)}(D_{i \cup pa_G(i)}) \]

Score DAGs by summing the local score of each node and its parents!
Family Variable Representation

Given DAG $G$ over variables $N$ one has

$$\eta_G(i|B) = 1 \iff B = pa_G(i), \quad \eta_G(T) = 0 \text{ otherwise}$$

Record each node's parent set!

Two graphs are Markov equivalent, but different DAG representations!
Family Variable Polytope (FVP)

\[ \mathcal{P}_\eta := \text{conv}\{\eta_G \mid G \in \text{DAGs on } N\} \]

- Vertex \(\leftrightarrow\) DAG
- Dimension = \(N(2^{(N-1)} - 1)\)
- Facet description for \(N=1,2,3,4\)
- No facet description \(N > 4\)
- Some facets known \(N > 4\)
- Simple IP relaxation
- IP solution gives DAG
Characteristic Imset Representation - Studeny

**Goal:** Unique vector representation of Markov equivalent classes of DAGs.

**Notation:** Suppose $N$ random variables. We index the components of $v \in \mathbb{R}^{|N|}$ using subsets $T \subseteq N$, such as

$$v(T)$$

$N = \{a, b, c\}$

$$v = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

$v(b) = 0$, $v(ac) = 1, \ldots$
Characteristic Imset Representation

Given DAG $G$ over variables $N$ one has $c_G(T) \in \{0, 1\}$ for any $T \subseteq N$, $|T| \geq 2$. Moreover

$$c_G(T) = 1 \text{ if and only if there exists } i \in T \text{ with } T \setminus \{i\} \subseteq pa_G(i).$$

**Theorem** (Studený, Hemmecke, Lindner 2011):

Characteristic imsets $\leftrightarrow$ Markov equivalence classes.
Characteristic Imset Polytope (CIP)

\[ P_c := \text{conv}\{c_G \mid G \in \text{DAGs on } N\} \]

- Vertex \(\leftrightarrow\) Markov Eq. Class
- Dimension = \(2^N - N - 1\)
- Facet description for \(N=1,2,3,4\)
- No facet description \(N > 4\)
- Some facets known \(N > 4\)
- Complex IP relaxation
- IP solution gives eq. class
Geometric Approach to Learning BN

Every reasonable scoring function (BIC, BDE, …) is an affine function of family variable or char imset:

\[ Q(G, D) = \hat{s} \frac{Q}{D} - \langle \hat{t} \frac{Q}{D}, \eta_G \rangle = s \frac{Q}{D} - \langle t \frac{Q}{D}, c_G \rangle \]

Integer and linear programming techniques can be applied!
(Linear relaxations combined with row-generation and Brach-and-Cut)

FVP
Data

CIP
Data

Practical ILP methods & software exist based on FVP and CIP.
FVP: Some Known Faces

- **Order faces** (Cussens et al.)
- **Sink faces** (Cussens et al.)
- **Non-negative inequalities on family variables.** (H., Cussens, Studeny).
- **Modified convexity constraints** (H, Cussens, Studeny).
- **Generalized cluster inequalities** (H, Cussens, Studeny), (Cussens, et al.)
- **Connected matroids on** $C \subseteq N, |C| \geq 2$ (Studeny).
- **Extreme supermodular set functions** (H, Cussens, Studeny).
Super-modular Set Functions

Definition 5.5 (standardized supermodular function). Any vector \( m \in \mathbb{R}^P(N) \) can be viewed as a real set function \( m : \mathcal{P}(N) \to \mathbb{R} \). Such a set function is called *standardized* if \( m(S) = 0 \) for \( S \subseteq N, |S| \leq 1 \), and *supermodular* if

\[
\forall U, V \subseteq N \quad m(U) + m(V) \leq m(U \cup V) + m(U \cap V).
\]

• The set of super-modular vectors is a polyhedral cone.
• Cone is pointed \( \Rightarrow \) it has finitely many extreme rays.
• Extreme rays correspond to faces of FVP.

• Remark: The rank functions of connected matroids are extreme rays of non-decreasing submodular functions (mirrors to supermodular functions).
Cluster Inequalities

\[ \forall C \subseteq N, \ |C| > 1 : \sum_{a \in C} \sum_{B \cap C} \eta(a \mid B) \leq |C| - 1 \]

- Why? Not all nodes in the cluster C can have a parent which is also in the cluster C

Remark 3.2. The cluster inequalities do not allow cycles: consider a digraph G which has a cycle \( i_1 \rightarrow \cdots \rightarrow i_p \rightarrow i_1 \) and note that the cluster inequality with \( C := \{i_1, \ldots, i_p\} \) would be violated.
Generalized Cluster Inequalities

- For every cluster \( C \subseteq \mathbb{N}, |C| > 2 \) and \( k = 1, \ldots, |C| - 1 \) the generalized cluster inequality is

\[
\sum_{a \in C} \sum_{B \subseteq \mathbb{N} \setminus a: \ |B \cap C| \geq k} \eta(a \mid B) \leq |C| - k.
\]

- **Why?** For any DAG G the induced subgraph \( G_C \) is acyclic and the first \( k \) nodes in \( C \) in a total order consistent with \( G_C \) has at most \( k-1 \) parents in \( C \).
Connecting FVP and CIP

Linear map between Family variable and Char Imset: (Studeny, Haws)

\[ c_\eta(S) = \sum_{i \in S} \sum_{B, S \setminus \{i\} \subseteq B \subseteq N \setminus \{i\}} \eta(i \mid B) \quad \text{for any } S \subseteq N, \ |S| \geq 2. \]

\( x \in FVP \) objective is score equivalent (SE) if

\[ \forall G, H \in \text{DAGS} (N) \quad G \sim H \Rightarrow \langle x, \eta_G \rangle = \langle x, \eta_H \rangle. \]

Face \( F \in FVP \) is score equivalent if there exists a SE objective defining \( F \).

A face \( F \in FVP \) is weakly equivalent (WE) if

\[ \forall G, H \in \text{DAGS} (N) \quad G \sim H \quad \eta_G \in F \Rightarrow \eta_H \in F. \]
Score Equivalence, FVP, & CIP

• **Theorem** [H, Cussens, Studeny]
The following conditions are equivalent for a facet $S \subseteq DAGS(N)$
  a) $S$ is closed under Markov equivalence
  b) $S$ contains the whole equivalence class of full graphs
  c) $S$ is SE

• **Corollary** [H, Cussens, Studeny]
There is a 1-1 correspondence between SE faces of FVP and faces of CIP which preserves inclusion.

• **Corollary** [H, Cussens, Studeny]
SE facets of the FVP correspond to those facets of the CIP that contain the 1-imset. None of those facets of CIP include the 0-imset. Moreover, these facets correspond to extreme supermodular functions.
Sufficiency of Score Equivalent Faces

• Learning BN structure = optimizing SE obj over FVP
• Are SE faces of FVP sufficient? Yes!
• **Theorem** [H, Cussens, Studeny]
  Let $o$ be an SE objective. Then the LP problem of

$$\text{maximize } \eta \mapsto \langle o, \eta \rangle \text{ over } \eta \in FVP$$

has the same optimal value as the LP problem to
maximize the same function over the polyhedron

- **SE faces** of FVP corresponding to facets of CIP not containing 0-imset
- Non-negativity and modified convexity constraints.

• **Not true for SE-facets**! 😞 Must use all SE-faces.
Open Conjecture...something to think on

(H, Cussens, Studeny)

**Theorem:** Every weakly equivalent facet of family-variable polytope is a score equivalent facet.

(Haws, Cussens, Studeny)

**Conjecture:** Every weakly equivalent face of family-variable polytope is a score equivalent face.

Believe false, but counter-example must be in $N \geq 4$. Already performed extensive searches in $N=4,5$. 😞
THANK YOU!

David Haws*, Milan Studeny, James Cussens

IBM Watson
dchaws@gmail.com