Markov Networks

- MNs belong to the class of probabilistic graphical models
  Undirected, acyclic graphs of random variables

- Example: Random variables: $X_i$, $i = 1, \ldots, 5$

- Cliques (usually, maximal) in the graph have associated potential functions
  Non-negative real functions

- No conditional probabilities, but initially, local, joint marginal potentials

- Here, three maximal cliques, and three potentials

- Combination of potentials defines joint probability distribution
  Cliques’ potentials become factors of the global joint distribution
Here, three potentials for three cliques:
\[ \psi_1(x_1, x_2, x_3), \psi_2(x_2, x_3, x_4), \psi_3(x_3, x_5) \]

Potentials may have parameters, possibly unknown, so as probability distributions
They could be learned from data

Joint probability distribution (density) for variables in MN:
\[ P(x_1, \ldots, x_5) := \frac{1}{Z} \times \psi_1(x_1, x_2, x_3) \times \psi_2(x_2, x_3, x_4) \times \psi_3(x_3, x_5) \]

\( Z \) is the “partition function”, a normalization factor to obtain a probability distribution
It has to be: \[ \sum_{x_1, \ldots, x_5} P(x_1, \ldots, x_5) = 1 \]

Then: \[ Z := \sum_{x_1, \ldots, x_5} \psi_1(x_1, x_2, x_3) \times \psi_2(x_2, x_3, x_4) \times \psi_3(x_3, x_5) \]

\( Z \) for “Zustandssumme” in German: “sum over states”
(root in Statistical Mechanics, initially largely developed by German speaking scientists)
• RVs \( X_i \) take values on their domains \( \text{Dom}(X_i) \). They reflect outcomes from a real-valued random experiment. They are defined on the sample space \( \Omega \) in common.

• **Example:** (cont.) Assume Bernoulli RVs:

\[
\text{Dom}(X_i) = \{0, 1\}
\]

Random propositional features

• **Potentials:**

1. \( \psi_1(x_1, x_2, x_3) := \text{total number of 1s taken by the variables} \)
   
   E.g. \( \psi_1(1, 0, 1) = 2 \)

2. \( \psi_2(x_2, x_3, x_4) := x_2 + x_3 + x_4 \)
   
   E.g. \( \psi_2(1, 0, 0) = 1 \)

3. \( \psi_3(x_3, x_5) := x_3 \times x_5 \)
   
   E.g. \( \psi_3(0, 1) = 0 \)
Exercise: Compute $Z$ above, the density value $P(1, 0, 0, 1, 1)$, and the marginal value $P_{X_1}(1)$

For $Z$, compute the terms of the summation: $(2^5 \text{ products})$

1. $\psi_1(0, 0, 0) \times \psi_2(0, 0, 0) \times \psi_3(0, 0) = 0 \times 0 \times 0 = 0$
2. $\psi_1(0, 1, 1) \times \psi_2(1, 1, 1) \times \psi_3(1, 1) = 2 \times 3 \times 1 = 6$
3. $\psi_1(0, 1, 0) \times \psi_2(1, 0, 0) \times \psi_3(0, 0) = 1 \times 1 \times 0 = 0$, etc.

$$P(1, 0, 1, 0, 1) := \frac{1}{Z} \times \psi_1(1, 0, 1) \times \psi_2(0, 1, 0) \times \psi_3(1, 1) = \frac{2 \times 1 \times 1}{Z}$$

There could be unknown parameters, to be learned, e.g.

$$\psi'_2(x_2, x_3, x_4) := \alpha \times x_2 + (1 - \alpha) \times x_3 + \theta \times x_4$$
• We have heard about the “Markov Condition”, Markov Processes, etc.

• Main idea and intuition behind MNs:
  The probability distribution of a particular variable (possibly with others in the net) depends only on a “small neighborhood” of the variable

  There are implicit independence assumptions in place that “isolate” it from a large portion of the net

• The way MNs are constructed, via factorized representations, allows to identify certain stochastic (in)dependencies

• There are criteria to identify and exploit them (notion of “d-separation”)

Criteria also applicable to BNs
A common class of MNs comes from Statistical Mechanics (SM):  

\[ P(\bar{x}) := \frac{1}{Z} \times \exp(-\sum_C E(\bar{x}_C)) = \frac{1}{Z} \times \prod_{\bar{x}_C} \frac{1}{e^{E(\bar{x}_C)}} \]

Here, \( \bar{x} \) represents the variables in the graph, and the \( \bar{x}_C \) those in clique \( C \).

A joint probability distribution from potentials:  

\[ \psi_C := \frac{1}{e^{E(\bar{x}_C)}} \]

Think of \( E(\bar{x}_C) \) as an energy function of the variables of sub-state \( \bar{x}_C \).

This distribution makes low energy configurations (states) more likely.

It penalizes high energy states.

It favors higher entropy states (we will come back to this).
• Energy function $E$ may come in different forms

Energy-based models are common in SM, Biochemistry, ML

Whole families of distributions depending on the classes to which potentials belong

• MNs may be easier or more natural to use in some applications than BNs

• Choosing a direction between two variables may not be reasonable

  E.g. in image analysis, with variables representing pixels of a same image

  Also with relational data (think of attributes in a table)

• MNs have symmetries that BNs do not have, and can be exploited

• Inference with MNs tends to be more complex than with BNs
Some More Inference

- Let us see in more general terms what we did on page 12
- Idea: exploit distributive law \( a \times b + a \times c = a \times (b + c) \)

Three operations versus two

- **Example**: A chain model: \( X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_{N-1} \rightarrow X_N \)

With potentials: \( \psi(x_i, x_{i+1}) \)

Joint distribution: \( P(\vec{x}) = \frac{1}{Z} \prod_{i=1}^{N-1} \psi(x_i, x_{i+1}) \)

Marginal of \( X_1 \): \( P_{X_1}(x_1) = \frac{1}{Z} \sum_{x_2, \ldots, x_N} \prod_{i=1}^{N-1} \psi(x_i, x_{i+1}) \)

- Computed naively like this, the computation cost is proportional to \( \prod_{i=1}^{N} |\text{Dom}(X_i)| \)

- By distributivity:

\[
P_{X_1}(x_1) = \frac{1}{Z} \sum_{x_2} [\psi(x_1, x_2) \sum_{x_3} \psi(x_2, x_3) \cdots \sum_{x_{N-1}} \psi(x_{N-2}, x_{N-1}) \sum_{x_N} \psi(x_{N-1}, x_N)]
\]

Now cost proportional to \( \sum_{i=1}^{N-1} |\text{Dom}(X_i)| \times |\text{Dom}(X_{i+1})| \)
• **Exercise:** Consider the MN

Verify that:

\[
P_A(a) := \frac{1}{Z} \sum_{b,c,d,e,f} \psi(a, b) \psi(a, c) \psi(b, d) \psi(c, e) \psi(b, e, f)
\]

\[
= \frac{1}{Z} \sum_b \psi(a, b) \sum_c \psi(a, c) \sum_d \psi(b, d) \sum_e \psi(c, e) \sum_f \psi(b, e, f)
\]

• This *variable elimination algorithm* uses distributivity
  Good for marginal of one variable
Example:

\[ P(x_2) = \frac{1}{Z} \sum_{x_1} \sum_{x_3} \sum_{x_4} \sum_{x_5} \psi(x_1, x_3, x_5) \psi(x_1, x_2) \psi(x_2, x_4) \psi(x_3, x_4) \]

\(O(2^5)\) operations in the naive way

with binary variables

However:

\[ P(x_2) = \frac{1}{Z} \sum_{x_1} \psi(x_1, x_2) \sum_{x_4} \psi(x_2, x_4) \sum_{x_3} \psi(x_3, x_4) \sum_{x_5} \psi(x_1, x_3, x_5) \]

\[ = \frac{1}{Z} \sum_{x_1} \psi(x_1, x_2) \sum_{x_4} \psi(x_2, x_4) \sum_{x_3} \psi(x_3, x_4) m_5(x_1, x_3) \]

\[ = \frac{1}{Z} \sum_{x_1} \psi(x_1, x_2) \sum_{x_4} \psi(x_2, x_4) m_3(x_1, x_4) \]

\(m_i\) are marginals per clique or joins thereof

\[ = \frac{1}{Z} \sum_{x_1} \psi(x_1, x_2) m_4(x_1, x_2) = \frac{1}{Z} m_1(x_2) \]

\(O(2^3)\) now

Summing over \(x_2\) gives \(Z\) (LHS is 1) ("messages" \(m_i\) could be reused, c.f. below)

Not more that 3 variables appear together in any term of a summation
• In general, the maximum number of variables that appear together in a summation term depends on the elimination order.

• The lowest complexity is obtained by the order that minimizes this maximum number.

• Unfortunately, finding the optimal elimination order is NP-hard.

Reduction from SAT

• What about more than one marginal?

• If we want more marginal distributions, we will be repeating operations.

• The algorithm above can be adapted via reuse of precomputations.

• There is a lot more about inference in PGMs ...
Tree-Width of a Graph

- The tree-width (TW) of a graph becomes relevant in many problems of data management and AI
- The TW of a graph measures how close a graph is to a tree
- It is commonly the case that graph problems become easier when the input graph has small TW
- Undirected graph \( G = \langle V, E \rangle \)

A tree-decomposition of \( G \) is a tree \( T = \langle \{ S_1, \ldots, S_n \}, E' \rangle \), such that:

- \( S_1, \ldots, S_n \subseteq V \), i.e each node in \( T \) is a subset of \( V \)
- \( S_1 \cup \cdots \cup S_n = V \)
- \((u, v) \in E \Rightarrow \{u, v\} \subseteq S_i\), for some \( i \)
- If for \( v \in V, \ v \in S_j \cap S_k, \ i \neq k \), then \( v \in S_i \), for every \( S_i \) in the unique (simple) path between \( S_j \) and \( S_k \)
• Width of tree decomposition $\mathcal{T}$:  $\text{width}(\mathcal{T}) := (\max_i |S_i|) - 1$

• The tree-width of graph $\mathcal{G}$:  $\text{tw}(\mathcal{G}) := \min_{\mathcal{T}} \text{width}(\mathcal{T})$
  With $\mathcal{T}$ ranging over all tree decompositions of $\mathcal{G}$

• When $\mathcal{G}$ is already a tree, the edges in $E$ become the $S_i$

The $S_i$ are connected by $E'$ when they share a node in $V$
Chapter 6: Logical + Probabilistic KR

Leopoldo Bertossi
Probabilistic Approaches to KR

- Many logic-based approaches to KR&R have a probabilistic counterpart

- For example, a default rule (as in ASP) may be treated as a probabilistic/statistical statement

  As a conditional probability: \( P(\text{flies}|\text{bird}) = 0.95 \)

  “the probability of flying being a bird is 0.95”

- Consequences may be probabilistic too

- Diagnosis can be stated using conditionals: \( P(\text{flu}|\text{fever}) = \frac{P(\text{flu}) \times P(\text{fever}|\text{flu})}{P(\text{fever})} \) (a priori vs. a posteriori)

- More generally: \( P(\text{cause}|\text{symptom}) = \frac{P(\text{cause})P(\text{symptom}|\text{cause})}{P(\text{symptom})} \)

  \( P(\text{symptom}|\text{cause}) \) easier to estimate by experts than \( P(\text{cause}|\text{symptom}) \)
Probabilistic Reasoning Problems

• We can have PGMs or other probabilistic models
  With features that are random variables subject to some sort of uncertainty

• There are probabilistic approaches that favor representation of:
  • Joint distributions $\sim$ “generative models”
    - MNs
  • Conditional distributions $\sim$ “discriminative models”
    - BNs
    - Regression models: $Y = \alpha \times X + \beta + \epsilon$
      Basically modeling $P(Y|X)$

• In principle, one can pass from one to the other, but there is complexity involved (remember inference)
  We did this with BNs, using the “chain rule” or Bayes formula
• Conditional probabilities allows us to attack several problems in *uncertain knowledge representation and reasoning*

• Probabilistic versions of diagnosis?

Consider an underlying probabilistic model $\mathcal{K}$ (background knowledge) with an associated probability distribution $P_\mathcal{K}$

An observation $O$ (or evidence), and a set of possible hypothesis (basic admissible explanations) $\mathcal{E} = \{E_1, \ldots, E_n\}$

$O$ is the value of a random variable (or several of them) in $\mathcal{K}$, and each $E_i$ is (the value of) a random variable in $\mathcal{K}$

• We can attempt to find the best explanation $E^b \in \mathcal{E}$

$$E^b := \arg \max_{E \in \mathcal{E}} P_\mathcal{K}(E \mid O) \quad (1)$$

The most probable explanation given the evidence

• Usually called MAP-inference: *maximum a posteriori*

After (conditioned on) the observation ...
A different form of probabilistic reasoning: prefer an explanation $E^*$

$$E^* := \arg \max_{E \in E} P_K(O \mid E) \quad (2)$$

The explanation that maximizes the (conditional) probability of the observation

Which is what we observed after all ...

This is similar to maximum-likelihood reasoning in Statistics

Exercise: Verify that under the assumption that the explanations are equally likely (a priori), (1) reduces to (2)

Hint: use Bayes formula

There are model-dependent techniques for these reasoning tasks
Logic + Probability in AI

- Traditionally, the “logical-” and “probabilistic schools” have been separate and competitors
- In the last few years they have become complementary approaches
- Today, KR problems are attacked with mathematical models/techniques that involve simultaneously logic and probability
- Different forms of KR combine logic and probability for KR&R
  Different formalisms, models, underlying assumptions, etc.
- These combined representations (models) can also be learned
  We will see some of them ...
• **Conditional KBs:** Knowledge base $KB$ with

  • Hard knowledge, e.g. $emu \rightarrow bird$
  
  • Soft, conditional, probabilistic rules, of the form $r: (\alpha|\beta)[p]$
  
  E.g. $r_v: (flies|bird)[0.9]$ (a “probabilistic conditional”)

• Semantics? Logical consequences of/from $KB$?

• Possible-worlds semantics: Collection $\mathcal{W}$ of worlds $W$

  • $W$ is a set of propositional (or ground) atoms assumed to be true
    (Herbrand structures, as usual)
  
  • $W$ must satisfy the hard knowledge in $KB$ (as usual)

• $W$ does not have to satisfy $\beta \rightarrow \alpha$, i.e. the conditional as a classical implication

• For this we need the probabilistic component ...
• We start considering a probability distribution $P$ on $\mathcal{W}$, the outcome space: $W \in \mathcal{W} \mapsto P(W)$

• Which probability distribution $P$? (possibly several candidates)

• Since all the worlds in $\mathcal{W}$ satisfy the hard knowledge, consider one that satisfies the conditionals:

For $r: (\alpha|\beta)[p]$, it must hold: $P(\alpha|\beta) = p$ (and $P(\beta) > 0$)

\[
P(\alpha|\beta) := \frac{P(\alpha \land \beta)}{P(\beta)} := \frac{P(\{W \in \mathcal{W} \mid W \models \alpha \land \beta\})}{P(\{W \in \mathcal{W} \mid W \models \beta\})} \quad (**)
\]

• Pick such a distribution $P^*$ (which one?)

• Boolean query $Q$ (expressed in the logical language): It may be true or false in an outcome world $W$

It becomes a Bernoulli RV: $P^*(Q (= 1)) := \sum_{W \in \mathcal{W}} P^*(W)$
• **Example:** Propositional variables: \( \text{yellow, fly, bird, emu, canary, } \ldots \)

\[ KB = \{ \text{bird, emu }\rightarrow \text{bird, (} \text{flies|bird)}[0.9], \text{canary }\rightarrow \text{yellow, } \ldots \} \]

• \( \mathcal{W} \) contains worlds satisfying the hard knowledge:

\( \text{(logical constraints)} \)

\[ W_1 = \{ \text{yellow, bird, canary} \}, \]

\[ W_2 = \{ \text{yellow, bird, fly, canary} \}, \]

\[ W_3 = \{ \text{yellow, emu, bird, fly, canary} \}, \text{ etc.} \]

• Assume there is a distribution \( P \) on \( \mathcal{W} \)

• Query \( Q: \text{yellow }\land \text{bird }\rightarrow \text{fly?} \)

It is true in \( W_2, W_3, \ldots \)

• Event associated to the query: \( E(Q) := \{ W_2, W_3, \ldots \} \)

\[ P(Q) := P(\{ W_2, W_3, \ldots \}) = P(W_2) + P(W_3) + \cdots \]
• More generally: We obtain formulas as consequences with associated probabilities

• We could also define the logico-probabilistic consequences of $KB$ as those with high probability

• For a logical sentence $\varphi$ (or query):

$$KB \models_p \varphi \iff P(\varphi) > 1 - \epsilon$$

As in the previous example, $\varphi$ defines an event

• $\epsilon$ can be pre-specified (and small)

• Which is a good distribution $P$ on $\mathcal{W}$?

A preferred $P^*$?

Some may be “better” or more justified than others
• **ME Distributions:** Prefer a distribution that does not make unjustified, arbitrary assumptions

• One that does not impose unnecessary “structure or complexity” on the model

• Think of Statistical Mechanics: the contents of a gas container tends to reach a state of equilibrium of maximum disorder, with low complexity or structure

• The notion of **Entropy** comes in ...

  Systems tend to reach equilibrium states of maximum entropy (maximum disorder)

  To impose order, structure, complexity, one needs extra energy (an unlikely state)

• Choose a distribution that maximizes the entropy?
• **Entropy**: Probability space \( \langle \Omega, P \rangle \), with \( \Omega = \{ \omega_1, \ldots, \omega_n \} \),
\[ p_i := P(\omega_i) \] (finite case for simplicity)

• **Entropy** of the distribution:
\[
\text{Entropy}(P) := - \sum_{i=1}^{n} p_i \times \log(p_i) \quad (\ast)
\]
\[
= \sum_{i=1}^{n} (p_i \times \log\left(\frac{1}{p_i}\right)) \quad (= H(P))
\]

• Entropy is interpreted as a measure of the level of uncertainty captured by the distribution
  A measure of the degree of disorder it attributes to the system

• This “measure” can be derived from some desirable properties
  As the only function that satisfies them (a theorem)

• Furthermore, one can prove: The uniform distribution maximizes the entropy, i.e. \( p_i = \frac{1}{n} \)
  When there is no extra constraint to satisfy or knowledge to consider
• Back to our problem, it makes sense to choose $P^*$ as the maximum-entropy distribution:

$$P^* := \arg\max_{P \in \mathcal{P}} \text{Entropy}(P)$$

$$= \arg\max_{P \in \mathcal{P}} -\sum_{W \in \mathcal{W}} P(W) \times \ln(P(W))$$

• Conditioned maximization problem over the class $\mathcal{P}$ of probabilities that satisfy the conditions above (c.f. page 8)

• Distribution without arbitrary assumptions/structure, maximum disorder, maximum independence

• Choose a distribution that is as close to the uniform distribution as possible given the conditions

  The one that is the least unjustified ...

• One can define query answering and logico-probabilistic consequences from KB as on pages 8 and 12
• **Example:** Consider a box containing balls and cubes, which can be white or green. We know that all balls are white. Possible distributions?

• We can think this scenario as involving a draw from the box, whose observation gives rise to 2 random variables (features) \( \text{Shape}, \text{Color} \), each taking two values.

• Joint distribution \( P(\text{Shape}, \text{Color}) \) under conditional \( P(\text{Color} = g| \text{Shape} = b) = 0? \)

<table>
<thead>
<tr>
<th>Dist</th>
<th>( w ) Bs</th>
<th>( g ) Bs</th>
<th>( w ) Cs</th>
<th>( g ) Cs</th>
<th>Entropy (in bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \frac{1}{5} )</td>
<td>0</td>
<td>( \frac{2}{5} )</td>
<td>( \frac{2}{5} )</td>
<td>? (compute)</td>
</tr>
</tbody>
</table>

Assuming that 20% of objects are balls
This entails a lot about color, and shape given color: (check!)
\( w = \frac{3}{5}, g = \frac{2}{5}, b|w = \frac{1}{3}, c|w = \frac{2}{3}, b|g = 0, c|g = 1 \)
\( P(g, b) = 0, P(w, b) = \frac{1}{5}, P(g, c) = \frac{2}{5}, P(w, c) = \frac{2}{5} \)
2. 

<table>
<thead>
<tr>
<th>Dist</th>
<th>w Bs</th>
<th>g Bs</th>
<th>w Cs</th>
<th>g Cs</th>
<th>Entropy (in bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(rac{2}{5})</td>
<td>0</td>
<td>(rac{2}{5})</td>
<td>(rac{1}{5})</td>
<td>? (compute)</td>
</tr>
</tbody>
</table>

Assuming 20% of objects to be green, which leads to: (check!) 

\[ w = \frac{4}{5}, \quad g = \frac{1}{5}, \quad b|w = \frac{1}{2}, \quad c|w = \frac{1}{2}, \quad b|g = 0, \quad c|g = 1 \]

3. 

<table>
<thead>
<tr>
<th>Dist</th>
<th>w Bs</th>
<th>g Bs</th>
<th>w Cs</th>
<th>g Cs</th>
<th>Entropy (in bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>(rac{1}{3})</td>
<td>0</td>
<td>(rac{1}{3})</td>
<td>(rac{1}{3})</td>
<td>? (compute)</td>
</tr>
</tbody>
</table>

No assumption determining other properties

<table>
<thead>
<tr>
<th>Dist</th>
<th>w Bs</th>
<th>g Bs</th>
<th>w Cs</th>
<th>g Cs</th>
<th>Entropy (in bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(rac{1}{5})</td>
<td>0</td>
<td>(rac{2}{5})</td>
<td>(rac{2}{5})</td>
<td>1.522</td>
</tr>
<tr>
<td>2</td>
<td>(rac{2}{5})</td>
<td>0</td>
<td>(rac{2}{5})</td>
<td>(rac{1}{5})</td>
<td>1.522</td>
</tr>
<tr>
<td>3</td>
<td>(rac{1}{3})</td>
<td>0</td>
<td>(rac{1}{3})</td>
<td>(rac{1}{3})</td>
<td>1.585</td>
</tr>
</tbody>
</table>

Last row corresponds to maximum entropy distribution ...
Markov Logic Networks

- MLNs combine FO logic and Markov Networks (MNs) in the same logico-probabilistic representation
- They are used for uncertain Knowledge Representation and Reasoning, and also in Machine Learning
  Networks can be learned from data for producing KR models, with new forms of inference
- MLNs belong to Statistical Relational Learning (SRL)
  Handling inherent uncertainty and exploiting compositional structure are fundamental to understanding and designing large-scale systems
  Statistical relational learning builds on ideas from probability theory and statistics to address uncertainty while incorporating tools from logic, databases, and programming languages to represent structure
- We have a knowledge base $KB$ in FO logic, but formulas have “weights” (eventually leading to probabilities)
• Ground atoms of the logical language become the nodes in an undirected graph that is handled as a MN

• The formulas can be used to define cliques, and their weights to define potentials on cliques, and so on ...
• **Example:** (a simplified form of MLN)

Consider the implicitly universally quantified constraint

\[ 3.9: \text{Manager}(M, E) \rightarrow \text{HighlyCompensated}(M) \]  

(*)

• Consider all possible ground atoms built with underlying domain \( \text{Dom} \)

\[ \text{Atoms}_{\text{Dom}} = \{ \text{Man}(m, e) \mid (m, e) \in \text{Dom} \times \text{Dom} \} \cup \{ \text{HC}(m) \mid m \in \text{Dom} \} \]

• Each of these ground atoms becomes a node in a MN

• More precisely, each atom \( A \in \text{Atoms}_{\text{Dom}} \) becomes a Bernoulli random variable \( X_A \) in the MN (it can be true or false)

• These variables are stochastically and mutually dependent with (some of the) other variables \( X_{A'} \)

This will be determined by the edges and potentials in a MN

• The MN has a set of nodes \( V \) of size \( M = 4^2 + 4 \) nodes
• The groundings of the MLN are: (4^2 of them)

1. \neg M(d_1, d_1) \lor HC(d_1)
2. \neg M(d_1, d_2) \lor HC(d_1) (F_2)

... 

16. \neg M(d_4, d_4) \lor HC(d_4)

• Each grounding represents a factor in the underlying MN:
• The instantiations 1.-16. of (*) become the factors

E.g. the factor or clique $F_2$: $M(d_1, d_2) — HC(d_1)$ in the MN
This will be a (mini) clique which will have an associated potential depending on its weight

• Weight $w(F_2) = 3.9$ (inherited from weight for original formula)
• We do not have potentials yet, only the graph

• Weight of a factor determines potential of associated clique

$$
\psi_{F_2(M(d_1, d_2), HC(d_1))(x_1, x_2)} :=
\begin{cases}
1 & \text{if } x_1 = 1 \text{ and } x_2 = 0 \text{ i.e. } F_2 \text{ false} \\
3.9 & \text{otherwise}
\end{cases}
$$
• Similarly for the other 15 factors (original weight inherited by factors)

• Product of potentials defines distribution $P^m$ over possible worlds
  Indirectly over the Bernouilli RVs $X_i$
  (normalized product of their potentials)

• A possible world $W_1 = \{M(d_1, d_2), M(d_3, d_1), HC(d_1), HC(d_4)\}$

• $W_1$ makes true all factors, except for $\neg M(d_3, d_1) \lor HC(d_3)$

• In compatibility with MNs, its weight (or joint potential):
  $weight(W_1) := \Pi_{F: W_1 \models F} w(F) = (3.9)^{15}$
  Product of the weights of the factors that are true in $W_1$

• Probability of $W_1$: $P^m(W_1) := \frac{weight(W_1)}{Z}$ (also from (*)
  (also from (*))

• Normalization denominator: $Z = \sum_{\text{worlds } W} weight(W)$

• Exercise: How large is the number $M$ of nodes in the MN depending on the size $n$ of $\text{Dom}$ and the predicates?

• Let see now a more common way of presenting MLNs
• **Example:** Real-valued weight $w(\varphi)$ assigned to formulas $\varphi \in KB$

<table>
<thead>
<tr>
<th>Formula</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\forall x (\text{Steal}(x) \rightarrow \text{Prison}(x))$</td>
<td>3</td>
</tr>
<tr>
<td>$\forall x \forall y (\text{CrimePartners}(x, y) \land \text{Steal}(x) \rightarrow \text{Prison}(y))$</td>
<td>1.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Fixed, finite domain, e.g. $\text{Dom} = \{bob, anna, ...\}$

• Producing ground atoms, e.g. $\text{CrimePartners}(bob, anna)$, and instantiated formulae, e.g. $\text{Steal}(bob) \rightarrow \text{Prison}(bob)$

• Edge between two nodes (ground atoms) if they appear in a same instantiated formula

A (local, mini) clique for one instantiation of the second formula
• As on many occasions so far, a world is a set of ground atoms
  A Herbrand structure indicating what is true (and indirectly what is not)

  \[ W_1 = \{ \text{CrimePartners}(bob, anna), \text{Steal}(bob) \} \]
  \[ W_2 = \{ \text{CrimePartners}(bob, anna), \text{Steal}(bob), \text{Prison}(anna) \} \]

• A world may satisfy an instantiation of a formula or not
  For example, \( W_2 \) satisfies “the clique” above, but \( W_1 \) not

• The higher the weight, the higher the difference between a world that satisfies the formula and one that does not (with the rest the same)

• The worlds get associated probabilities through the weights

• A world that violates a formula is not invalid (not non-model), but only less likely

Some “models” (worlds) become more likely than others
The weight of a formula captures the way the probability decreases when a ground instance of the formula is violated.

A high weight for a formula becomes a high penalty on worlds that do not satisfy it.

Given a world $W$, each node $N \in V$ takes the value 0 or 1 if false or true in $W$ (worlds become outcomes).

Then, each node $N$ becomes a Bernoulli random variable $X^N$.

Worlds become instantiations of a random vector

$X = \langle X^{N_1}, X^{N_2}, X^{N_3}, \ldots, X^{N_M} \rangle$

$W_1$ becomes $x_1 = \langle 1, 1, 0, \ldots, 0 \rangle$
• Each instantiation of a formula generates a propositional “feature”, with value 1 if true in a world \( \mathcal{W} \), and 0, otherwise.

• We can assign probabilities to worlds. Equivalently, build a joint probability distribution \( P^m \) for \( \mathcal{X} \).

• As with MNs, we can use a log-linear “potential function”.

• For world \( \mathcal{W} \) associated to \( x \in \{0, 1\}^M \):

\[
P^m(\mathcal{W}) := P^m(\mathcal{X} = x) := \frac{1}{Z} \times e^{\sum_{\varphi \in \mathcal{KB}} w(\varphi) \times n(\varphi, x)}
\]

\( (*) \)

• \( n(\varphi, x) \): number of instantiations of \( \varphi \) true in world \( x \) (or its clique \( x_C \)).

• \( Z \) normalizes over all possible worlds:

\[
Z = \sum_{z \in \{0, 1\}^M} \exp(\sum_{\varphi \in \mathcal{KB}} w(\varphi) \times n(\varphi, z))
\]
• From (*): A (ground) clique \( gc \) associated to a formula \( \varphi \) in the MN has the potential:

\[ \psi_{gc}(\vec{x}) := \exp(w(\varphi) \times I_{gc}(\vec{x})) \]

with \( \vec{x} \) formed by 0s and 1s.

• In the example, \( gc \) could be the three ground atoms in the top-left corner: \( gc = \{ N_1, N_2, N_3 \} \).

\( I_{gc}(\vec{x}) \), the *indicator function*, takes value 1 if \( gc \) true for \( \vec{x} \), and 0 otherwise (with that, \( e^0 = 1 \) gives the right factor).

• This can be seen as a **Gibbs** distribution for MNs.

• Since we divide by all possible satisfaction with possible worlds (the \( Z \)), we can see \( w(\varphi) \) as a penalty for not satisfying it. Because in that case, it is multiplied by 0.

• So, hard or strong constraints that we want to see satisfied should have high weights.
• We obtain a probability distribution over possible worlds. Those that satisfy “more” high-weight (instances of) formulas become more likely.

• **Exercise:** Give an example of a MLN with a model that (logically) violates all the formulas $F$ in KB, as universal ICs, but still has a non-zero probability. Hint: Make sure not all ground instantiations of the ICs become false.

• With a MLN we do not have to create the actual, underlying, ground MN. We have a pattern to produce a concrete one if needed.

• Having the exponential on page 24 allows us to deal with sums instead of products.
• It is possible to extend MLNs with functions symbols
  Using Skolem functions could be used for formulas with existential quantifiers

• One can learn MLNs
  Learn the weights and/or the formulas
  The latter define the structure of the underlying and implicit MN

• How to do inference with MLNs?
Inference in Markov Logic Networks

- Inference under MLNs is of a probabilistic nature
- Similarly, the MLN defines a probability distribution $P^m$ over the possible worlds
- Basic inference task is computing the probability of a world, as on page 24

More interesting is a query in the language of the KB: For a sentence $\psi$:

$$P^m(\psi) := P^m(\{W \mid W \models \psi\}) := P^m(\mathcal{X} \text{ makes } \psi \text{ true})$$

$$= P^m(\{x \in \{0,1\}^M \mid \psi \text{ is true in } x\}) = \sum_{x \in \{0,1\}^M \mid x \models \psi} P^m(x)$$

- Computing the probabilities amounts, directly or not, to counting models (possibly with specific properties)

Here, a form of weighted model counting

A hard computational problem ...
In general in SRL, we want to avoid as much as possible doing the grounding of formulas. Followed by the explicit weighted model counting (bound to be computationally complex).

Can we stay at a higher ("lifted") level?

Different areas converge: model counting in logic (around SAT-related problems), graph theory, and data management.

Each grounding of an attribute, or groups thereof, could be Bernoulli. Related to each other in something like a Bayesian Logic Network. E.g. in the presence of constraints (here a referential constraint).
• Too many variables and groundings, many not related to each other

• SRL is precisely about doing things at the higher, relational or FO logical level

  Representation and reasoning at a “lifted”, more general level of granularity

• Can we do model counting without instantiation?

• Can we approximate model counting (and probabilities) without instantiation?

• Doing what is called “Lifted Inference”

  Lifted up to the FO representation

  Exploiting patterns, independence and symmetries