Independence Relations in Probabilistic Logic

Fabio G. Cozman - University of Sao Paulo, Brazil (fgcozman@usp.br)

(joint work with Cassio Polpo de Campos, José Eduardo Ochoa Luna;

also collaboration with Teddy Seidenfeld, CMU)

Some work on...

- Decision support with Bayesian networks, Bayesian network classifiers.
- Sets of probability distributions (credal sets, credal networks).

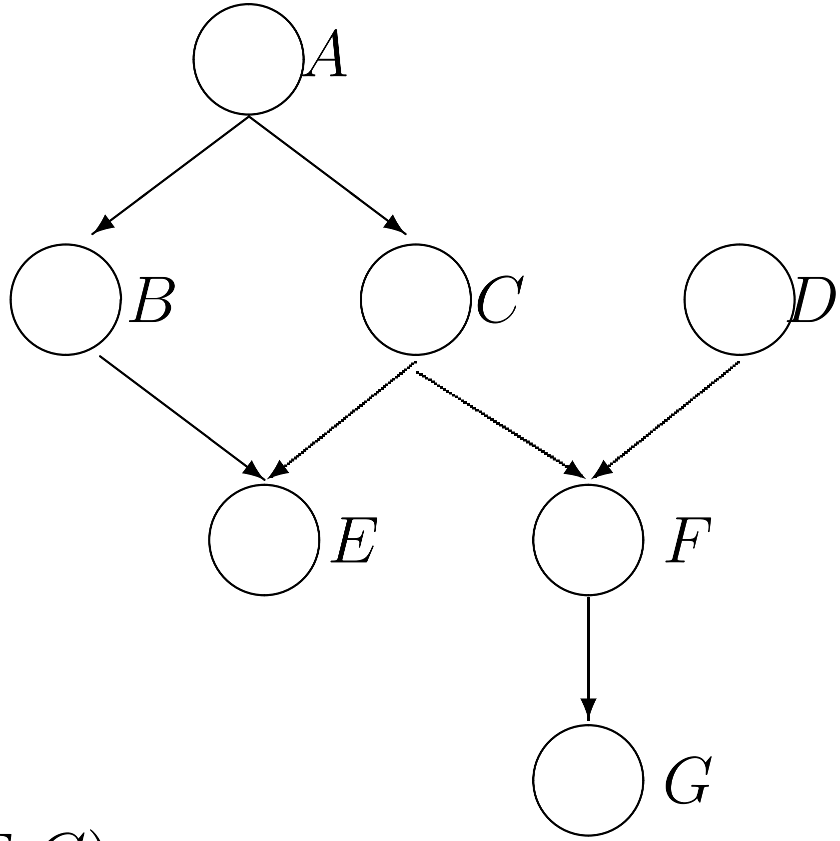
Bayesian Networks

- A Bayesian network encodes $P(X_1, \ldots, X_n)$.
  - This joint distribution is specified through a directed acyclic graph.
  - Each node represents a random variable X_i.
    - *Parents* of X_i: $\mathrm{pa}(X_i)$.
- **Markov condition:** Every variable is independent of its nondescendants nonparents given its parents, implying factorization:

$$p(X_1, \ldots, X_n) \quad = \quad \prod_i p(X_i|\mathrm{pa}(X_i))\,.$$

- Or is it the other way around?
  Should factorization imply Markov condition?

Example



$$P(A, B, C, D, E, F, G) =$$
$$P(A)\, P(D)\, P(B|A)\, P(C|A)\, P(E|B, C)\, P(F|D, E)\, P(G|F)$$

Starting with Markov condition

- More intuitive (matter of taste?).
- Works in the infinite case (where conditional distributions may not exist).
- Easier to generalize.

This talk:

1. Propositional probabilistic logic.
2. A digression: the definition of independence.
3. PPL networks.
4. A digression: credal networks.
5. Extending (a bit) to relational languages.
6. A few applications and challenges.

The propositional case

- Back to Boole... and many others.
- Formula ϕ with propositions, operators ($\neg$, $\wedge$, $\vee$, $\rightarrow$).
- Take Ω as the set of 2^n truth assignments for n propositions.
- Interpret $P(\phi) \geq \alpha$ as

$$\sum_{\omega \models \phi} P(\omega) \geq \alpha.$$

Probabilistic satisfiability

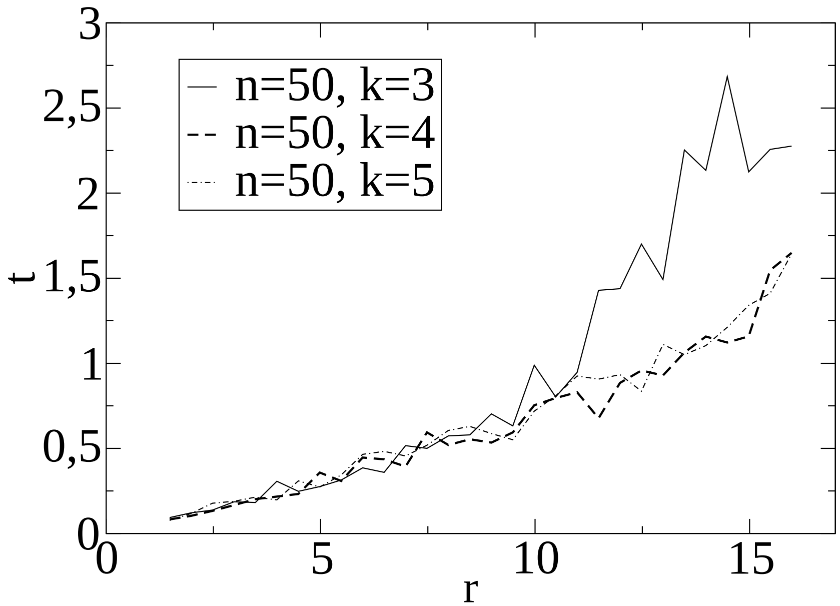
- Given m assessments:
  Is there a probability measure over Ω?
- Must satisfy $P(\omega) \geq 0$ and $\sum_{\omega \in \Omega} P(\omega) = 1$.
- This is a *linear program*.
  - Usually solved with the revised simplex method, where each step is a MAX-SAT problem.
  - Complexity: NP-complete.

Example (inspired by Jaeger 1994)

- Take:
  AntarticBird $\rightarrow$Bird,
  FlyingBird $\rightarrow$Bird,
  Penguim $\rightarrow$Bird,
  FlyingBird $\rightarrow$Flies,
  Penguim $\rightarrow \neg$Flies,
  $P(\text{FlyingBird}|\text{Bird}) = 0.95$,
  $P(\text{AntarticBird}|\text{Bird}) = 0.01$,
  $P(\text{Bird}) \geq 0.2$,
  $P(\text{FlyingBird} \vee \text{Penguim}|\text{AntarticBird}) \geq 0.2$,
  $P(\text{Flies}|\text{Bird}) \geq 0.8$.
- Then
  $P(\text{FlyingBird}|\text{Bird} \wedge \neg\text{AntarticBird}) \in [0.949, 0.960]$,
  $P(\text{Penguim}|\neg\text{AntarticBird}) \in [0.000, 0.050]$.

Difficulties

1. Computational complexity (no phase transition?)



2. Inferential vacuity:
   A, B have no logical relation, $P(A) = 1/2$, $P(B) = 1/2$;
   then $P(A \wedge B) \in [0, 1/2]$.

Independence

- Introduce independence to reduce inferential vacuity.
  - Obvious example:
    A and B independent, $P(A) = 1/2$, $P(B) = 1/2$;
    then $P(A \wedge B) = 1/4$.

- "Unconditional" independence leads to
  - *nonlinear* constraints;
  - even higher complexity (satisfiability is ??-hard).

A digression: defining independence

- We wish to introduce formulas such as

$$\text{independence}(\phi, \theta)$$

  that indicate a believed "independence" between formulas ϕ and θ.
- But what is the translation for it?
- In probability theory, "independence" means that for our *unique* probability distribution we must have

$$P(\phi \wedge \theta) = P(\phi) \times P(\theta)\,.$$

- However, here we may have many distributions.

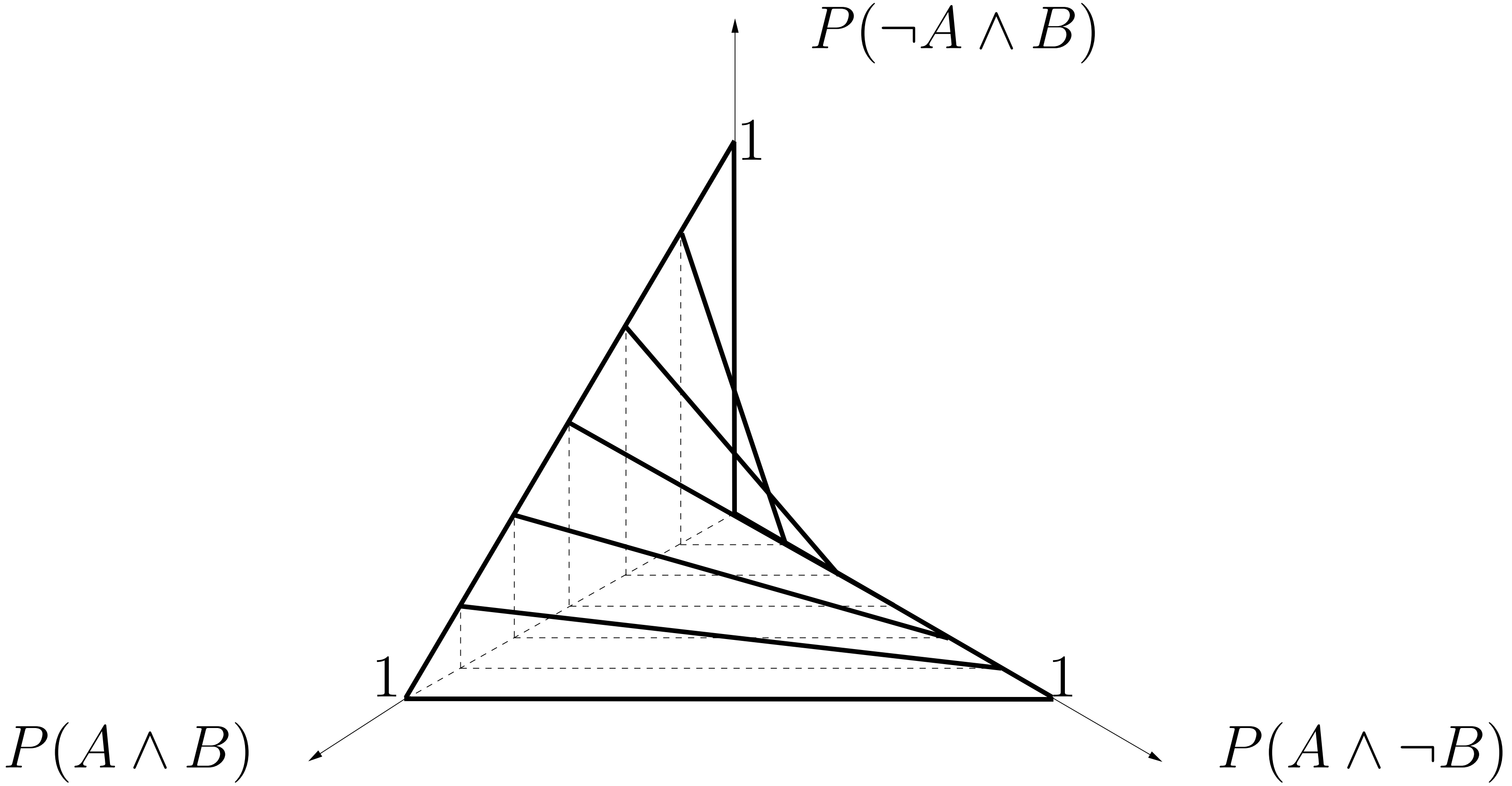
Concepts of independence

Problem: define independence for a *set of probability distributions*.

Or, rather: define independence for *credal sets*.

- Most straightforward:
  every distribution in credal set $K(X, Y)$ factorizes.
- Levi's strong independence:
  Each vertex of our *convex* credal set $K(X, Y)$ factorizes.

Lack of convexity



What is the matter with convexity?

- Note linearity of constraints such as

$$P(A) \geq \beta.$$

- Indeed, theories of credal sets/imprecise probabilities usually attach behavioral meaning *only* to

  convex credal sets.

- Similarly, constraints in “classic” probabilistic logic are all linear.

Other concepts

- Walley's epistemic independence:

$$\underline{E}[f(X)|Y] = \underline{E}[f(X)] \qquad \text{for all } f(X),$$

and

$$\underline{E}[g(Y)|X] = \underline{E}[g(Y)] \qquad \text{for all } g(Y).$$

- Kuznetsov's independence:

$$\mathbf{E}[f(X)g(Y)] = \mathbf{E}[f(X)]\mathbf{E}[g(Y)].$$

Some difficulties

- The conditional version of Walley's epistemic independence fails the *contraction* property:

  Contraction:

  $$(X \perp\!\!\!\perp Y \,|\, Z) \;\&\; (X \perp\!\!\!\perp W \,|\, (Y, Z)) \Rightarrow (X \perp\!\!\!\perp (W, Y) \,|\, Z)$$

- The relationship between Markov condition and factorization fails (also d-separation, etc).

- The conditional version of Kuznetsov's independence also fails contraction.

The solution: Seidenfeld's theory

- Teddy Seidenfeld has presented a theory that allows for non-convex credal sets.
  - Seidenfeld's breakthrough:
    to consider a behavioral axiomatization of choice amongst *sets* of options (rather than just two options).
- Holy grail of the (post-?) Bayesian approach: beliefs can be translated into arbitrary sets of distributions.

Ok, problem solved, but...

- There are problems in dealing with zero probabilities.
- Namely, how do we interpret

$$P(\phi|\varphi)$$

  when we find that $P(\varphi)$ *may* be equal to zero?
- There are several proposals...
  - ...discard the zero, opening the set?
  - ...cut probability at some ϵ?
  - ...abort calculations?

A few points

- Kolmogorov's theory simply ignores zero probabilities: $P(A|B)$ is defined only if $P(B) > 0$.

- Maybe reasonable for single-distribution theory, less appealing in probabilistic logic.
- Indeed, even for single-distribution theory, many voices have called for a better treatment of zero probabilities.
  - There is considerable interest in ways to learn an event of probability zero.
  - Also, Kolmogorov's approach generates a great deal of misery in infinite domains.

The alternative: full conditional measures

- A function $P(A|B)$ where A belongs to a Boolean algebra, and B to the same algebra (minus the empty element $\emptyset$), such that
  - $P(A|A) = 1$;
  - $P(A|B) \geq 0$ for all A;
  - $P(A \vee B|C) = P(A|C) + P(B|C)$ whenever $A \wedge B \neq \emptyset$ (and C is not $\emptyset$);
  - $P(A \wedge B|C) = P(A|B \wedge C)\, P(B|C)$ for all A and B such that $B \wedge C \neq \emptyset$.

- Full probability measures allow $P(A|B)$ to be defined even if $P(B) = 0$!

Very elegant, but...

- For independence, disaster strikes!
- The condition $P(A \wedge B) = P(A)\,P(B)$ seems too weak.
- The condition $P(A|B) = P(A)$ is not symmetric.
- The condition

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B)$$

  fails

  Weak union: $(X \perp\!\!\!\perp (W, Y) \,|\, Z) \Rightarrow (X \perp\!\!\!\perp W \,|\, (Y, Z))$

- A strenghtening (due to Hammond) fails

  Contraction:

$$(X \perp\!\!\!\perp Y \,|\, Z) \;\&\; (X \perp\!\!\!\perp W \,|\, (Y, Z)) \Rightarrow (X \perp\!\!\!\perp (W, Y) \,|\, Z)$$

A summary on independence

1. For this talk, “independence” just means:

   usual “Kolmogorovian” independence for each distribution in our credal set $K(X, Y)$...

2. But this basic concept does deserve more discussion.

3. For much more: look at the *International Symposium on Imprecise Probability: Theories and Applications*...

Back to our problem

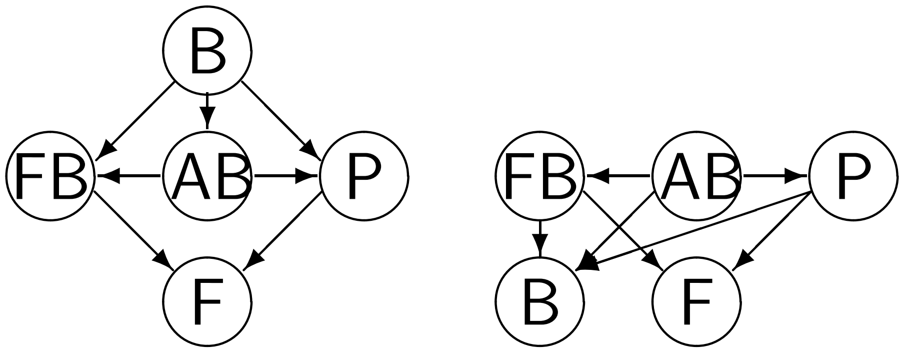
- We want to have propositional formulas, possibly associated with probabilistic assessments.
- We want to have independence (to avoid vacuity), but with enough structure so as to allow efficient reasoning.

Basic idea:

- Organize independence relations using graphs.
- But: do not require that assessments must follow the structure of the graph, nor the structure of the formulas.
- That is, we have: logical formulas, probabilistic assessments, *and* a directed acyclic graph with the usual Markov condition.

That is:

- Binary variables $\{X_1, \ldots, X_n\}$, one per proposition.
- A directed acyclic graph $\mathcal{G}$ is associated with these variables.



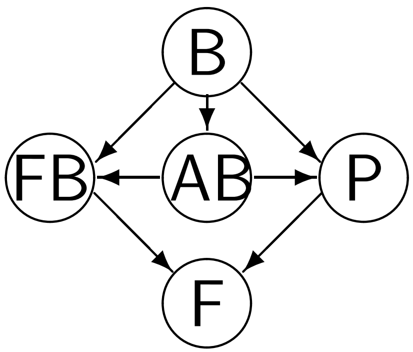
(FlyingBird and Penguim are parents of Flies.)

Markov condition

- Assume:
  X_i is conditionally independent from its nondescendants nonparents given its parents.
- Implies:

$$P(\mathbf{X}) = \prod_i P(X_i|\mathrm{pa}(X_i))$$

  for every possible distribution.



Markov condition comes first...

- The graph imposes independence relations (through Markov condition).
- Probabilistic assessments need not “follow” the edges of the graph.

"PPL" networks

We have:

1. A set of variables $\mathbf{X}$.
2. A graph $\mathcal{G}$.
3. Logical constraints (in CNF) for $\mathbf{X}$.
4. Probabilistic assessments for $\mathbf{X}$.
5. A Markov condition implying the factorization

$$P(\mathbf{X}) = \prod_i P(X_i|\mathrm{pa}(X_i)) .$$

Inference

- Inference: to compute the lower/upper probability for some formula ϕ given some other formula φ:

$$\min / \max P(\phi|\varphi) \,.$$

  - An inference now leads to a nonlinear program subject to factorization.

- Problem deals with $P(X_i|\mathrm{pa}(X_i))$, not with 2^n values...

- Logical constraints and probabilistic assessments can still be arbitrary.

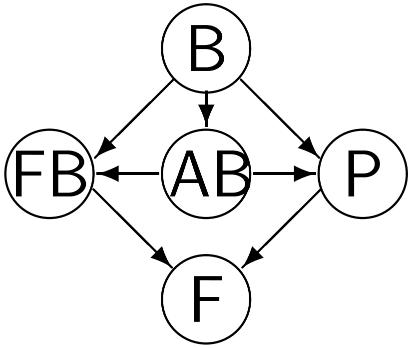
Complexity

- Complexity of inferences:
  - Still NP-complete on graphs with limited treewidth.
  - NP^{PP}-complete in general.

- However: the “graph” in these results is not the “real” graph, but a graph enlarged with connections for assessments.

Example

- Recall:
  AntarticBird$\rightarrow$Bird,
  FlyingBird$\rightarrow$Bird,
  Penguim$\rightarrow$Bird,
  FlyingBird$\rightarrow$Flies,
  Penguim$\rightarrow \neg$Flies,
  $P(\text{FlyingBird}|\text{Bird}) = 0.95$,
  $P(\text{AntarticBird}|\text{Bird}) = 0.01$,
  $P(\text{Bird}) \geq 0.2$,
  $P(\text{FlyingBird} \vee \text{Penguim}|\text{AntarticBird}) \geq 0.2$,
  $P(\text{Flies}|\text{Bird}) \geq 0.8$.



- Nonlinear programming produces
  $P(\text{FlyingBird}|\text{Bird} \wedge \neg\text{AntarticBird}) \in [0.949, 0.960]$
  $P(\text{Penguim}|\neg\text{AntarticBird}) = 0$.

The possibilities...

- A PPL network may be
  1. unsatisfiable;
  2. satisfiable by a single distribution;
  3. satisfiable by a set of distributions.
- Just like “traditional” probabilistic logic: we have a *deductive* language, where we may even discover inconsistency.
- Without any additional effort, probability intervals, credal sets and qualitative probabilities can be directly represented in a PPL network.

Another point

- Inconsistency must obviously be detected.
- But there is also a subtle possibility:
  - Two propositions are independent (by the graph).
  - Yet they are logically dependent.

  This usually leads one of them (or both) to get probability zero.
- Indeed, construction of a PPL network must be an interactive process, where inconsistencies (and other situations) are detected, corrected, etc.
- The development of useful support systems for modeling practical problems is a challenge.

A final point

- Clearly, we are not making heroic efforts to make sure we always have a *unique* distribution.

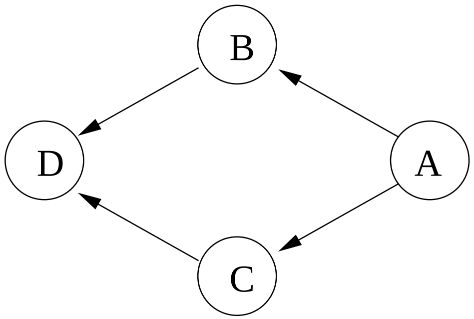
- The question here is: with *these* formulas and assessments, what is the largest set of probabilities that is *coherent* with them?

- But can we compute?

A digression: credal networks

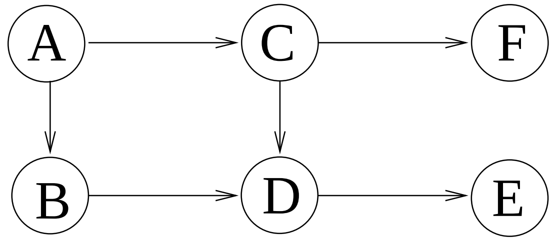
Credal network: directed acyclic graph where each node is associated with

- a random variable X_i,
- sets of conditional probability distributions $K(X_i|\mathrm{pa}(X_i))$



$$
\begin{aligned}
P(a) &\in [0.6, 0.8] \\
P(b|a) &\in [0.3, 0.5] \\
P(b|\neg a) &\in [0.4, 0.7] \\
&\dots
\end{aligned}
$$

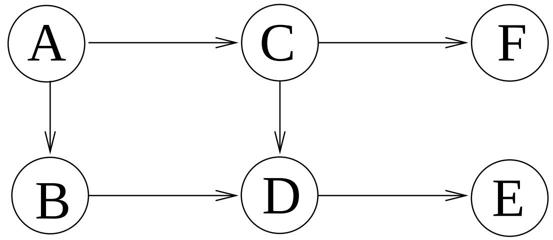
Example



Credal sets define the local probability distributions $P(A)\,, P(C|A)\,, P(B|A)\,, P(D|B,C)\,, P(E|D)\,, P(F|C)$.

Suppose one wants to evaluate $\overline{P}(e, f)$.

Naive inference



Using the joint distribution directly:

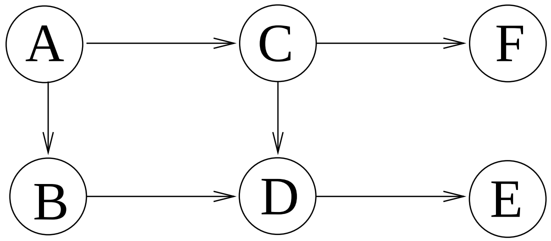
$$\max P(e,f) = \max \sum_{A,B,C,D} P(f|C) \cdot P(e|D) \cdot P(D|B,C) \cdot$$

$$\cdot P(B|A) \cdot P(C|A) \cdot P(A)\,,$$

subject to linear constraints (from local credal sets).

Inference by variable elimination

Write:



$$\max P(e,f) = \max \sum_D P(e|D)\, P(D,f) \quad \text{subject to}$$

$$P(B,C) = \sum_A P(B|A)\, P(C|A)\, P(A)\,, \text{ for all } B, C$$

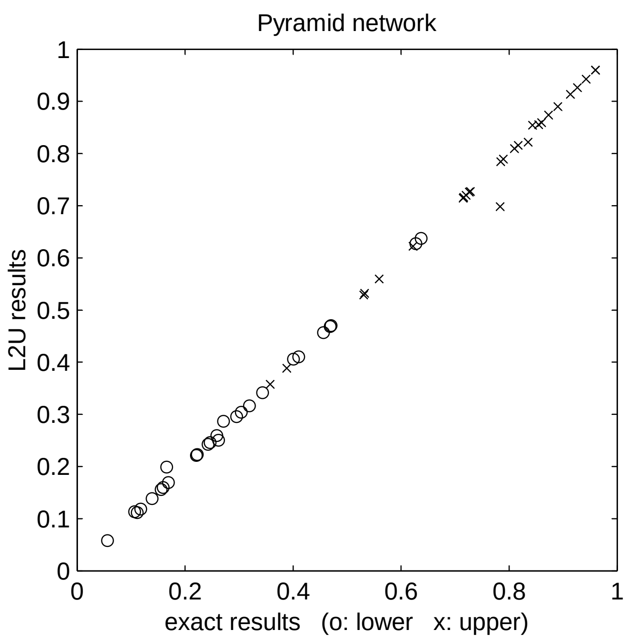
$$P(C,D) = \sum_B P(D|B,C)\, P(B,C)\,, \text{ for all } C, D$$

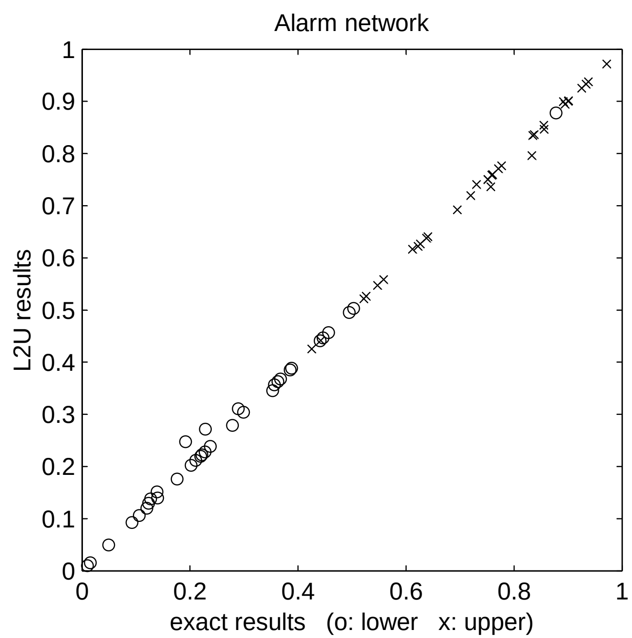
$$P(D,f) = \sum_C P(f|C)\, P(C,D)\,, \text{ for all } D$$

plus the linear constraints. There are many ways to write and to solve (exactly and approximately) such a nonlinear program.

Algorithms: variational and L2U

- Recent work suggests: excellent results with simple gradient-descent, and with branch-and-bound based on linear relaxations.
- Also, some recent work on producing variational approximations (such as a “Loopy Propagation” algorithm for probability intervals).





Main message:

- Algorithms developed for credal networks can be used for PPL networks.

- Much yet to improve, but current status is not so depressing.

Moving to a relational language

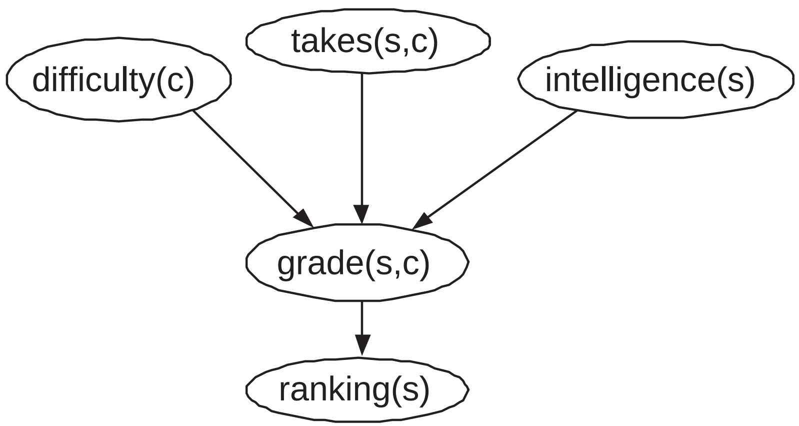
- Simple idea: use the graph-theoretical structure of Jaeger's *relational Bayesian networks*.
  - The graph is now representing just independence relations.
  - Assessments may be general (but complexity depends on the enlarged graph...).
- Again, no heroic efforts to guarantee
  - Consistency.
  - Uniqueness of probabilities.

Inference by propositionalization

- Well, we are assuming a finite domain...
- To do inference, propositionalize to a PPL network.
- Vast open issue is how to do inference (and assess complexity) directly on the relational level...

Example: The university domain

- A student is typically registered in several courses.
- The student's ranking depends on the grades that she receives in all of them.
- Grading depends on intelligence and difficulty.



The university domain

- Consider:

$$P(takes(John, Phil1)) \leq 0.5 \qquad (A_1)$$

$$P(\exists\, x\, \forall\, y\ grade(x, y, A)) \leq 0.001 \qquad (A_2)$$

$$0.1 \leq P(\exists\, x\, \forall\, y\ takes(x, y)) \leq 0.15 \qquad (A_3)$$

$$0.05 \leq P(\exists\, y\, \forall\, x\ takes(x, y)) \leq 0.1 \qquad (A_4)$$

- Also, suppose:

$$\forall\, x, y\ (takes(x, y) \iff \exists\, z\ grade(x, y, z))$$

$$takes(John, Math1)$$

$$takes(Mary, Phil1)$$

$$intelligence(John, Low)$$

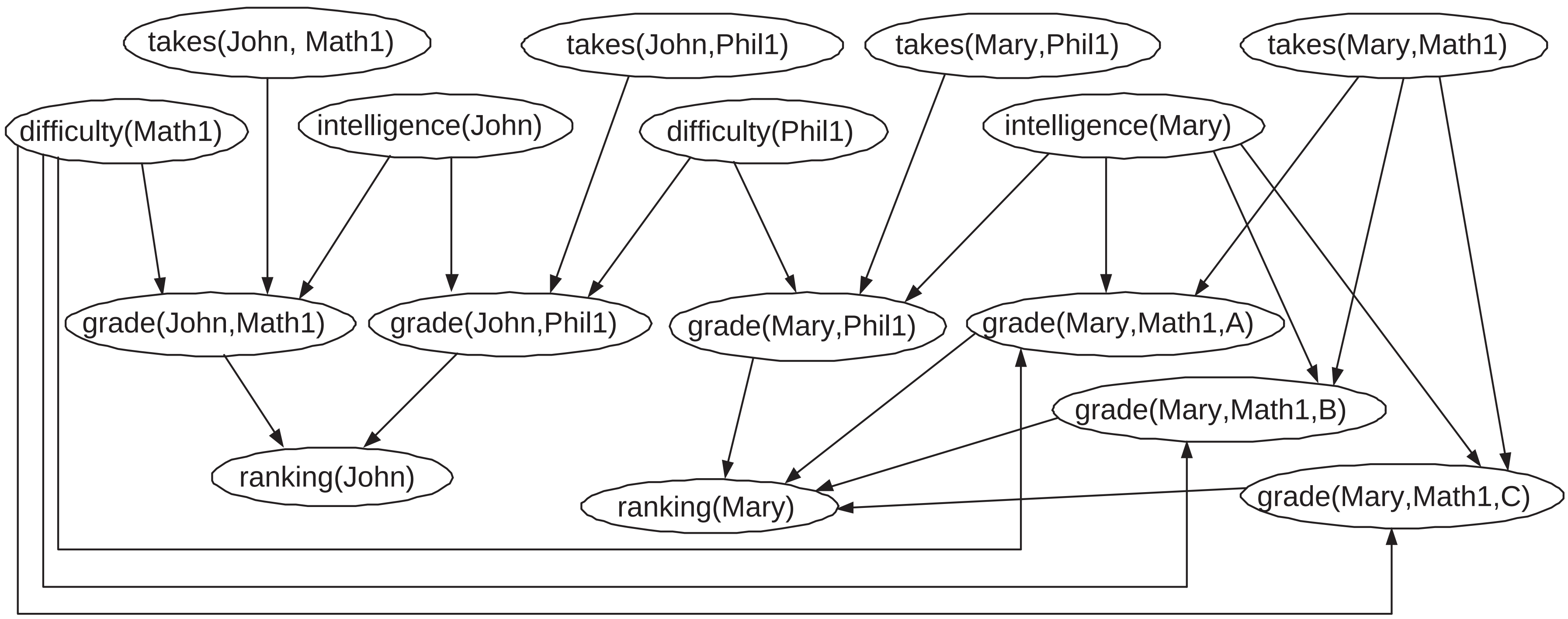
$$intelligence(Mary, High)$$

$$difficulty(Math1, High)$$

$$difficulty(Phil1, Low)$$

The university domain network

(Note: simplified network, without some nodes/edges!)



A few inferences

- As given, problem is inconsistent.
- Removing the assessment A_2, we have a feasible problem.
- Some inferences:
  - The probability of John receiving grade A in Phil1 given he takes all courses:

    $$P(grade(John, Phil1, A)|\forall\ y\ takes(John, y)) \in [0.0, 0.03].$$

  - The probability that Phil1 is taken only by smart students:

    $$P(\forall\ x\ takes(x, Phil1) \rightarrow intelligence(x, High)) \in [0.9, 1.0].$$

Another inference

- The probability of having a student that achieves grade A on all courses:

$$P(\exists\, x\, \forall\, y\ grade(x, y, A)) \in [0.003, 0.04].$$

Last inference shows why A_2 must be removed!

Example: planning

Blocks world in PPDDL.

```
(define (domain blocks-domain)
  (:requirements :probabilistic-effects :equality :typing)
  (:types block)
  (:predicates (holding ?b - block) (emptyhand) (on-table ?b - block)
               (on ?b1 ?b2 - block) (clear ?b - block))
  (:action pick-up
    :parameters (?b1 ?b2 - block)
    :precondition (and (not (= ?b1 ?b2)) (emptyhand)
                  (clear ?b1) (on ?b1 ?b2))
    :effect
      (probabilistic
        3/4 (and (holding ?b1) (clear ?b2) (not (emptyhand))
                 (not (clear ?b1)) (not (on ?b1 ?b2)))
        1/4 (and (clear ?b2) (on-table ?b1) (not (on ?b1 ?b2))))
  )
)
```

Inserting disjunctions into PPDDL

- But suppose the effect is a disjunction (not allowed by PPDDL!): failure causes “release block b_1” OR “nothing happens”.

- Result is entirely within the scope of previous discussion.

Some methodology is needed

- Again, necessary to detect inconsistencies and other situations.
- Need to have solid methodology and support systems.

To do...

- Inference algorithms, particularly approximate (and algorithms that operate on first-order).
- Deal with undirected graphs — main problem here is failure of factorization from Markov condition in the presence of zero probabilities.
- All of this focuses on finite domains (in fact, this is propositional logic...). Move to infinite domains.

Conclusion

- Independence seems to be necessary for realistic probabilistic logic.
- The concept of conditional independence should get more discussion.
- Graph-theoretical models are useful and (relatively) efficient.
- PPL networks and their relational counterparts seem to be flexible and intuitive tools.
- Main strategy:
  - to have a graph, formulas, and assessments separately specified;
  - not to worry much about consistency and uniqueness;
  - and to have inference to check consistency.

Thank you

Acknowledgements:

- FAPESP

Email:

- fgcozman@usp.br
- cassiopc@usp.br
- eduardo.ol@gmail.com