Déterminer la possibilité en XML probabiliste

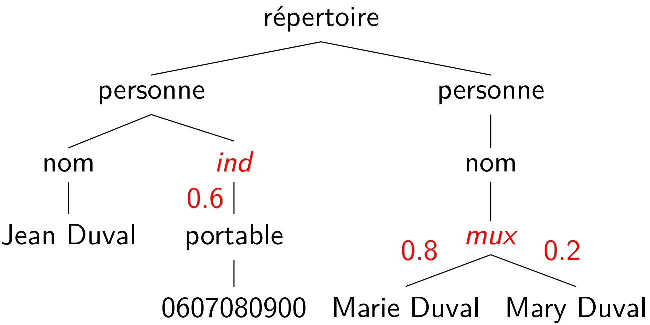
Antoine Amarilli

Télécom ParisTech

17 octobre 2014

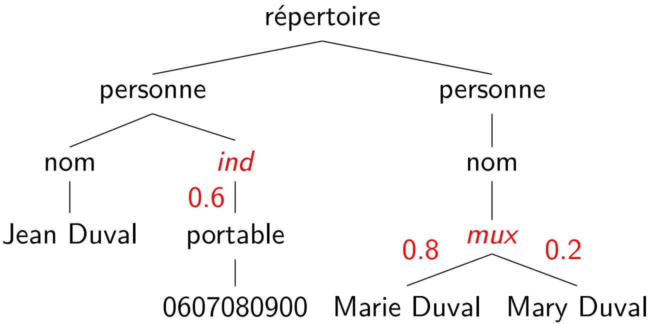
XML probabiliste

Représenter l'incertitude sur le contenu d'un document XML.



XML probabiliste

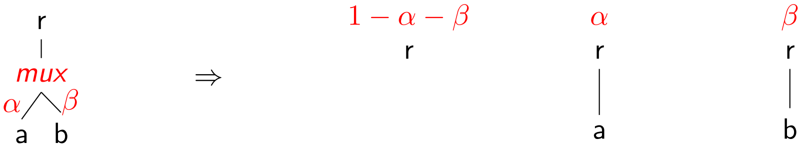
Représenter l'incertitude sur le contenu d'un document XML.



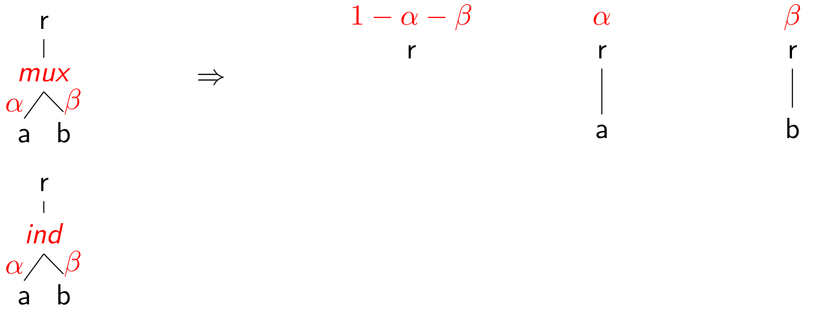
Sémantique : distribution de probabilités sur des documents.

Formalismes locaux : sémantique des mondes possibles

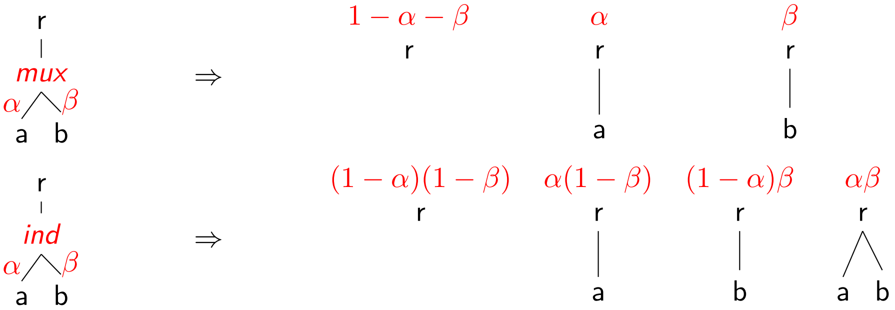
Formalismes locaux : sémantique des mondes possibles



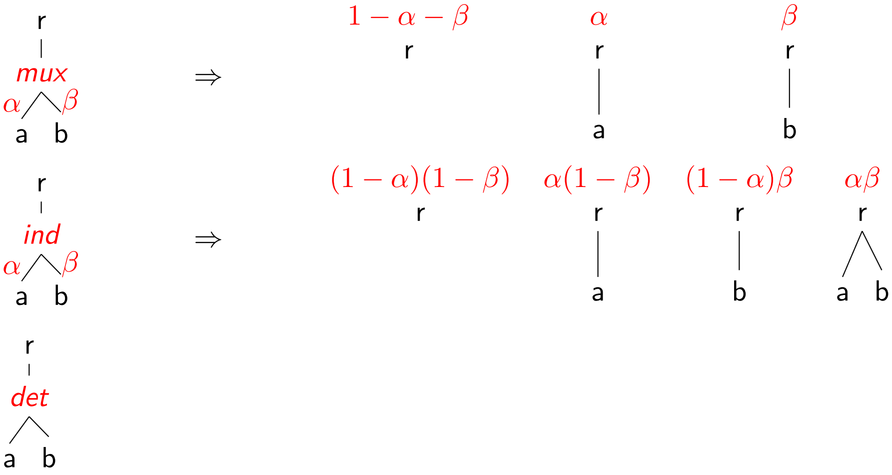
Formalismes locaux : sémantique des mondes possibles



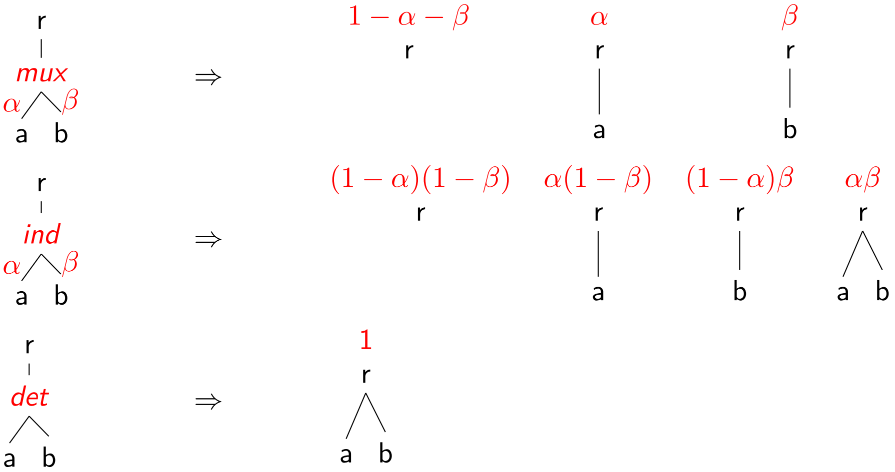
Formalismes locaux : sémantique des mondes possibles



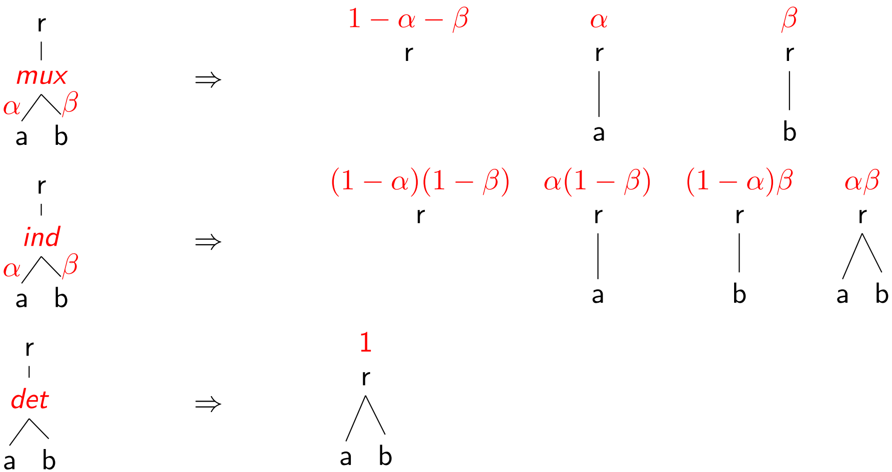
Formalismes locaux : sémantique des mondes possibles



Formalismes locaux : sémantique des mondes possibles



Formalismes locaux : sémantique des mondes possibles



Attention : on impose $\alpha < 1$, $\beta < 1$ pour *ind*.

Problème de la possibilité (Poss)

- Étant donné :
  - un document probabiliste D
  - un document déterministe W
- W est-il un monde possible de D ?
- Si oui, avec quelle probabilité ?

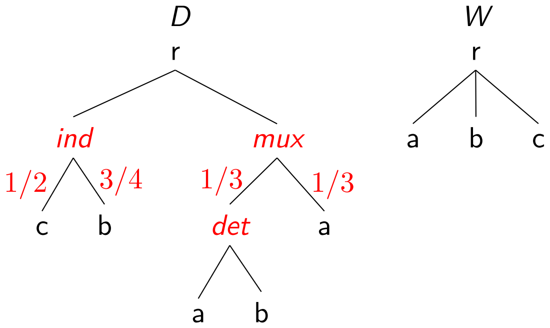
Problème de la possibilité (Poss)

- Étant donné :
  - un document probabiliste D
  - un document déterministe W
- W est-il un monde possible de D ?
- Si oui, avec quelle probabilité ?

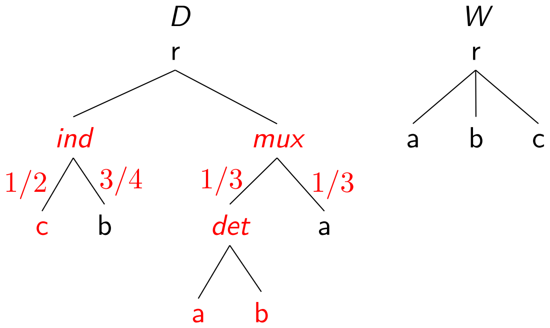
→ Complexité de ce problème ?

  - Types de nœuds autorisés
  - Documents ordonnés ou non
  - Décision ou calcul

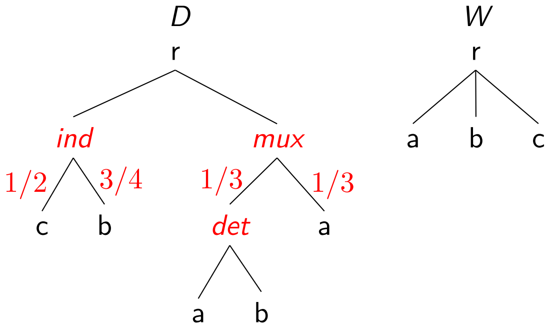
Exemple



Exemple



Exemple



$\rightarrow$ OK si *D* et *W* sont non ordonnés

Table des matières

Appartenance à NP et $\mathrm{FP}^{\#\mathrm{P}}$

- Deviner une occurrence de W dans D
- Deviner une issue pour chaque choix probabiliste
- Vérifier que l'occurrence est réalisée par la valuation

Appartenance à NP et $FP^{\#P}$

- Deviner une occurrence de W dans D
- Deviner une issue pour chaque choix probabiliste
- Vérifier que l'occurrence est réalisée par la valuation

→ La décision de la possibilité est dans NP

→ Le calcul de la probabilité est dans $FP^{\#P}$ (similaire)

Tractable pour les documents avec ordre

- Algo PTIME pour calculer la probabilité
- Intuitivement :
  - tester la correspondance entre les séquences de nœuds frères
  - algorithme dynamique pour tester à chaque niveau

→ Découle des résultats sur les automates d'arbres déterministes sur XML probabiliste : COHEN, KIMELFELD et SAGIV 2009

Tractable pour les documents avec ordre

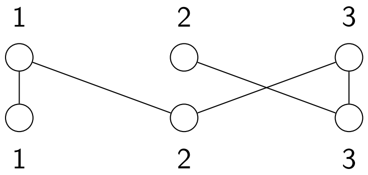
- Algo PTIME pour calculer la probabilité
- Intuitivement :
  - tester la correspondance entre les séquences de nœuds frères
  - algorithme dynamique pour tester à chaque niveau

→ Découle des résultats sur les automates d'arbres déterministes sur XML probabiliste : Cohen, Kimelfeld et Sagiv 2009

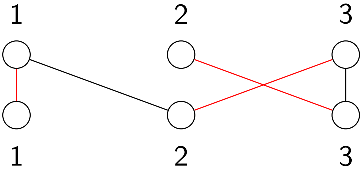
→ Seulement pour les documents ordonnés !

Table des matières

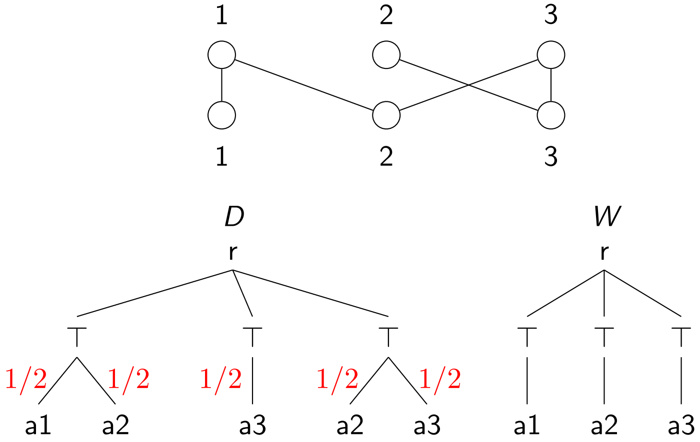
Le calcul est #P-dur pour *ind* ou *mux*



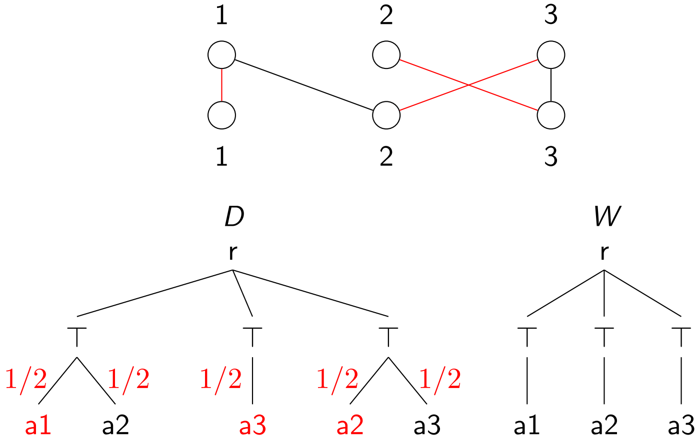
Le calcul est #P-dur pour *ind* ou *mux*



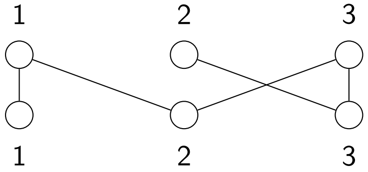
Le calcul est #P-dur pour *ind* ou *mux*

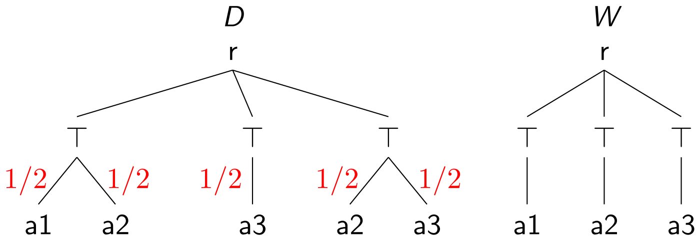


Le calcul est #P-dur pour *ind* ou *mux*



Le calcul est #P-dur pour *ind* ou *mux*





→ Probabilité d'une correspondance : nombre de couplages parfaits divisé par 2^n

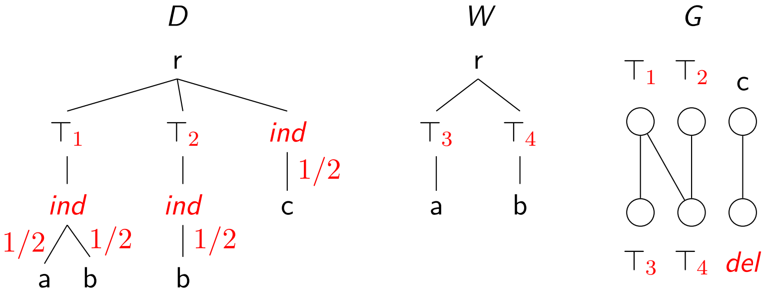
→ Calcul #P-dur pour *ind* ou *mux* sans ordre

Décision en PTIME pour *ind* ou *mux*

- Vérification dynamique entre les paires de nœuds de D et W
  - → Construire un graphe biparti selon la compatibilité des enfants
  - → Ajout de nœuds fictifs pour représenter les suppressions
  - → Vérifier si le graphe a un couplage parfait (PTIME)

Décision en PTIME pour *ind* ou *mux*

- Vérification dynamique entre les paires de nœuds de D et W
  - → Construire un graphe biparti selon la compatibilité des enfants
  - → Ajout de nœuds fictifs pour représenter les suppressions
  - → Vérifier si le graphe a un couplage parfait (PTIME)



Décision NP-dure pour deux parmi *ind*, *mux*, *det*

- Avec *det*, réduction depuis la couverture
  - $S = \{S_i\}$, $S_i = \{s_j^i\}$
  - Y a-t-il $T \subseteq S$ tel que $\bigcup T = \bigcup S$?

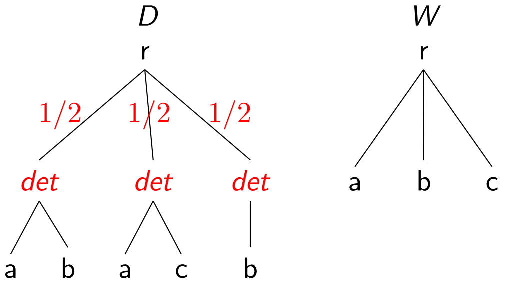
Décision NP-dure pour deux parmi *ind*, *mux*, *det*

- Avec *det*, réduction depuis la couverture exacte
  - $S = \{S_i\}$, $S_i = \{s_j^i\}$
  - Y a-t-il $T \subseteq S$ tel que $\bigcup T = \bigcup S$ sans doublons ?

Décision NP-dure pour deux parmi *ind*, *mux*, *det*

- Avec *det*, réduction depuis la couverture exacte
  - $S = \{S_i\}$, $S_i = \{s_j^i\}$
  - Y a-t-il $T \subseteq S$ tel que $\bigcup T = \bigcup S$ sans doublons ?

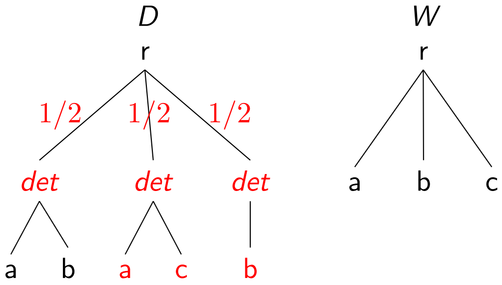
$$S = \{\{a, b\},\ \{a, c\},\ \{b\}\}$$



Décision NP-dure pour deux parmi *ind*, *mux*, *det*

- Avec *det*, réduction depuis la couverture exacte
  - $S = \{S_i\}$, $S_i = \{s_j^i\}$
  - Y a-t-il $T \subseteq S$ tel que $\bigcup T = \bigcup S$ sans doublons ?

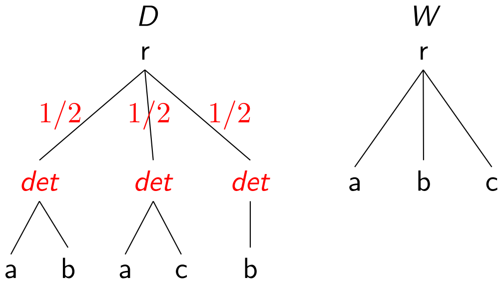
$$S = \{\{a, b\}, \{a, c\}, \{b\}\}$$



Décision NP-dure pour deux parmi *ind*, *mux*, *det*

- Avec *det*, réduction depuis la couverture exacte
  - $S = \{S_i\}$, $S_i = \{s_j^i\}$
  - Y a-t-il $T \subseteq S$ tel que $\bigcup T = \bigcup S$ sans doublons ?

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



Décision NP-dure pour deux parmi *ind*, *mux*, *det* (suite)

- Avec *ind* et *mux*, réduction depuis SAT

Décision NP-dure pour deux parmi *ind*, *mux*, *det* (suite)

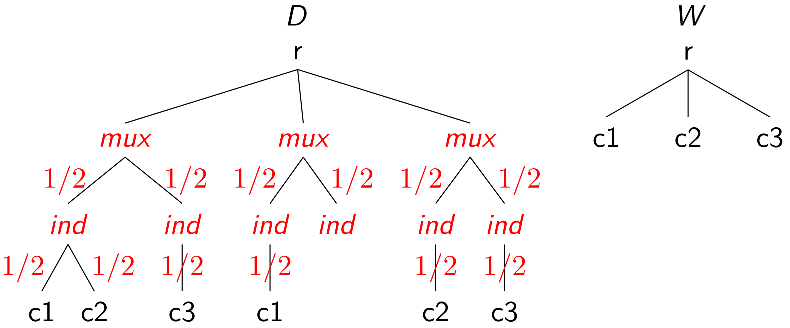
- Avec *ind* et *mux*, réduction depuis SAT
- $F = (a \lor b \lor \neg c) \land (a \lor c) \land (\neg a)$

Décision NP-dure pour deux parmi *ind*, *mux*, *det* (suite)

- Avec *ind* et *mux*, réduction depuis SAT
- $F = (a \vee b \vee \neg c) \wedge (a \vee c) \wedge (\neg a)$
  - a : clauses 1 et 2, ou clause 3
  - b : clause 1, ou rien
  - c : clause 2, ou clause 3

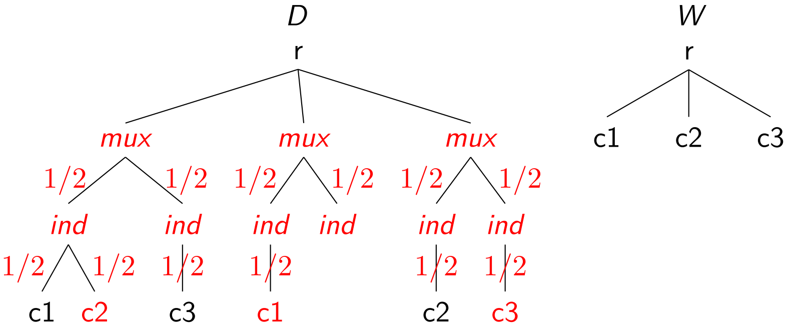
Décision NP-dure pour deux parmi *ind*, *mux*, *det* (suite)

- Avec *ind* et *mux*, réduction depuis SAT
- $F = (a \vee b \vee \neg c) \wedge (a \vee c) \wedge (\neg a)$
  - a : clauses 1 et 2, ou clause 3
  - b : clause 1, ou rien
  - c : clause 2, ou clause 3



Décision NP-dure pour deux parmi *ind*, *mux*, *det* (suite)

- Avec *ind* et *mux*, réduction depuis SAT
- $F = (a \vee b \vee \neg c) \wedge (a \vee c) \wedge (\neg a)$
  - a : clauses 1 et 2, ou clause 3
  - b : clause 1, ou rien
  - c : clause 2, ou clause 3



Décision NP-dure pour deux parmi *ind*, *mux*, *det* (suite)

- Avec *ind* et *mux*, réduction depuis SAT
- $F = (a \vee b \vee \neg c) \wedge (a \vee c) \wedge (\neg a)$
  - a : clauses 1 et 2, ou clause 3
  - b : clause 1, ou rien
  - c : clause 2, ou clause 3

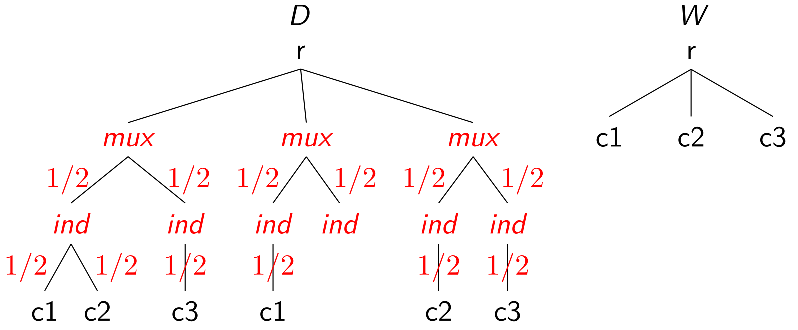


Table des matières

Non-ambiguïté

- D est non-ambigu si les nœuds portent des étiquettes uniques

$\rightarrow$ Il y a une seule façon au plus d'obtenir W !

Non-ambiguïté

- D est non-ambigu si les nœuds portent des étiquettes uniques

$\rightarrow$ Il y a une seule façon au plus d'obtenir W!

$\rightarrow$ Tous les modèles locaux sont tractables (imposer un ordre)

Non-ambiguïté

- D est non-ambigu si les nœuds portent des étiquettes uniques

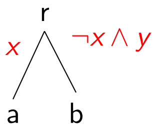
$\rightarrow$ Il y a une seule façon au plus d'obtenir W !

$\rightarrow$ Tous les modèles locaux sont tractables (imposer un ordre)

$\rightarrow$ Peut-on autoriser des corrélations ?

Modèle avec événements : *cie*

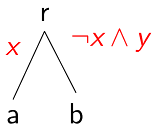
x	0.7
y	0.4



- Distribution de probabilités sur les événements
- Tirer les événements indépendamment
- Arêtes avec des conjonctions d'événements
- Supprimer les arêtes avec des formules fausses

Modèle avec événements : *cie*

x	0.7
y	0.4



- Distribution de probabilités sur les événements
- Tirer les événements indépendamment
- Arêtes avec des conjonctions d'événements
- Supprimer les arêtes avec des formules fausses

$\rightarrow$ Capture *ind*, *mux*, *det*

Poss NP-dur pour *cie* même sans ambiguïté

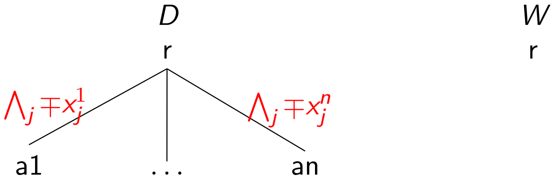
- $F = \bigwedge_i \bigvee_j \pm x_j^i$ in CNF

POSS NP-dur pour *cie* même sans ambiguïté

- $F = \bigwedge_i \bigvee_j \pm x_j^i$ in CNF
- Équivalent à : $\bigwedge_i \neg \bigwedge_j \mp x_j^i$

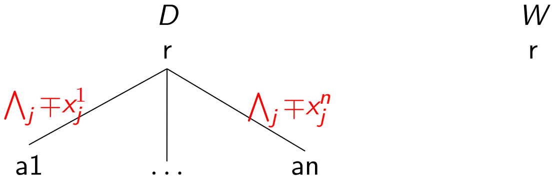
Poss NP-dur pour *cie* même sans ambiguïté

- $F = \bigwedge_i \bigvee_j \pm x_j^i$ in CNF
- Équivalent à : $\bigwedge_i \neg \bigwedge_j \mp x_j^i$



Poss NP-dur pour *cie* même sans ambiguïté

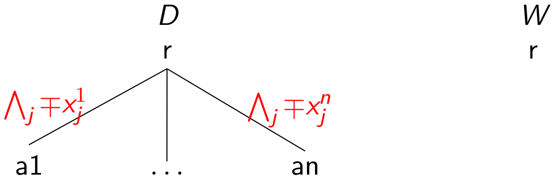
- $F = \bigwedge_i \bigvee_j \pm x_j^i$ in CNF
- Équivalent à : $\bigwedge_i \neg \bigwedge_j \mp x_j^i$



$\rightarrow$ W est un monde possible de D ssi F est satisfiable

POSS NP-dur pour *cie* même sans ambiguïté

- $F = \bigwedge_i \bigvee_j \pm x^i_j$ in CNF
- Équivalent à : $\bigwedge_i \neg \bigwedge_j \mp x^i_j$



$\rightarrow$ *W* est un monde possible de *D* ssi *F* est satisfiable

$\rightarrow$ Décider POSS est NP-dur, même sans ambiguïté

La classe *mie*

Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5

- Événements indépendants multivalués
- Pas de conjonctions

La classe *mie*

Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5

- Événements indépendants multivalués
- Pas de conjonctions
- Capture *mux*
- Ne capture pas *det* ou les hiérarchies *ind*

La classe *mie*

Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5

- Événements indépendants multivalués
- Pas de conjonctions
- Capture *mux*
- Ne capture pas *det* ou les hiérarchies *ind*
- Intractable avec l'ambiguïté

La classe *mie*

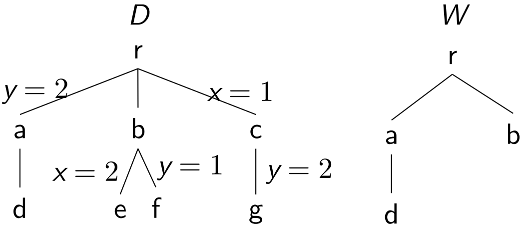
Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5

- Événements indépendants multivalués
- Pas de conjonctions
- Capture *mux*
- Ne capture pas *det* ou les hiérarchies *ind*
- Intractable avec l'ambiguïté

$\rightarrow$ Tractabilité si non-ambiguïté ?

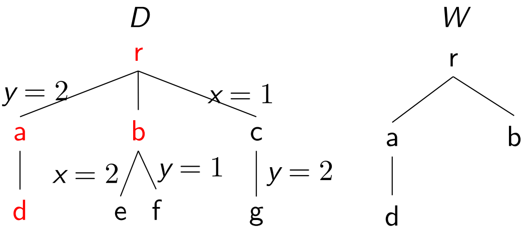
mie est tractable sur les documents non-ambigus

Var	**Val**	**Prob**
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5



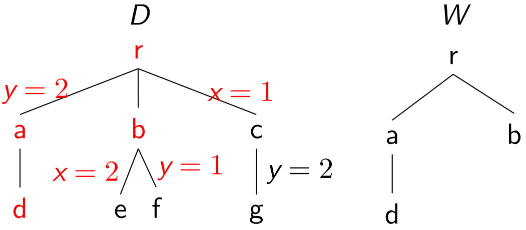
mie est tractable sur les documents non-ambigus

Var	**Val**	**Prob**
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5



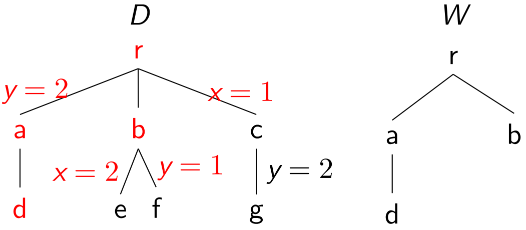
mie est tractable sur les documents non-ambigus

Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5



mie est tractable sur les documents non-ambigus

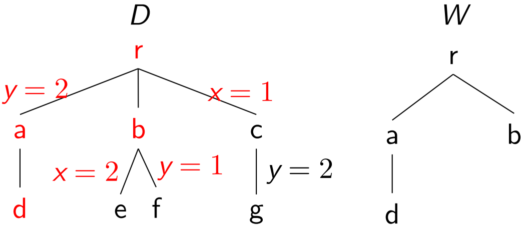
Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5



- $x \neq 2$, $x \neq 1$, $y = 2$, $y \neq 1$
- $x \in \{3, 4\}$, $y \in \{2\}$.

mie est tractable sur les documents non-ambigus

Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5



- $x \neq 2$, $x \neq 1$, $y = 2$, $y \neq 1$
- $x \in \{3, 4\}$, $y \in \{2\}$.

$\rightarrow$ Probabilité 0.1.

Table des matières

Conclusion

- Les modèles locaux avec ordre sont tractables
- Les modèles locaux sans ordre sont tractables
  - → Seulement pour la décision et
  - → Seulement avec *mux* ou *ind*
- Modèles locaux et *mie* tractables si non-ambiguïté
- Les autres cas sont durs

Conclusion

- Les modèles locaux avec ordre sont tractables
- Les modèles locaux sans ordre sont tractables
  - → Seulement pour la décision et
  - → Seulement avec *mux* ou *ind*
- Modèles locaux et *mie* tractables si non-ambiguïté
- Les autres cas sont durs

→ La hauteur n'a aucun impact

→ La valeur des probabilités n'a aucun impact

Conclusion

- Les modèles locaux avec ordre sont tractables
- Les modèles locaux sans ordre sont tractables
  - → Seulement pour la décision et
  - → Seulement avec *mux* ou *ind*
- Modèles locaux et *mie* tractables si non-ambiguïté
- Les autres cas sont durs

→ La hauteur n'a aucun impact

→ La valeur des probabilités n'a aucun impact

→ Peut-on raffiner *mie*, la non-ambiguïté, l'interaction *mux*–*ind* ?

→ Et si D était partiellement ordonné ?

Conclusion

- Les modèles locaux avec ordre sont tractables
- Les modèles locaux sans ordre sont tractables
  - → Seulement pour la décision et
  - → Seulement avec *mux* ou *ind*
- Modèles locaux et *mie* tractables si non-ambiguïté
- Les autres cas sont durs

→ La hauteur n'a aucun impact

→ La valeur des probabilités n'a aucun impact

→ Peut-on raffiner *mie*, la non-ambiguïté, l'interaction *mux*–*ind* ?

→ Et si D était partiellement ordonné ?

Merci pour votre attention !

Bibliographie

COHEN, Sara, Benny KIMELFELD et Yehoshua SAGIV (2009). "Running tree automata on probabilistic XML". In : *Proc. PODS*. ACM, p. 227–236.

Lien avec l'évaluation de requêtes

- Pourquoi l'évaluation de requêtes donne de mauvaises bornes ?
  - $\rightarrow$ Inégalités : "ne pas fusionner deux nœuds"
  - $\rightarrow$ Négation : "pas de nœuds en plus"
  - $\rightarrow$ La requête dépend de l'entrée W