Relational Probabilistic Models

- flat or modular or hierarchical
- explicit states or features or individuals and relations
- static or finite stage or indefinite stage or infinite stage
- fully observable or partially observable
- deterministic or stochastic dynamics
- goals or complex preferences
- single agent or multiple agents
- knowledge is given or knowledge is learned
- perfect rationality or bounded rationality

Relational Probabilistic Models

Often we want a random variable for each individual in a population

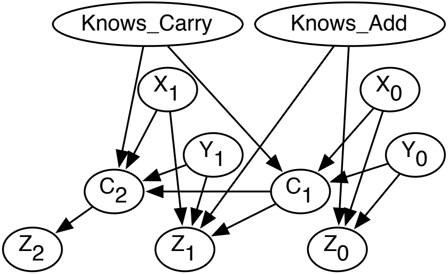
- build a probabilistic model before knowing the individuals
- learn the model for one set of individuals
- apply the model to new individuals
- allow complex relationships between individuals

Predicting students errors

$$\begin{array}{rccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$

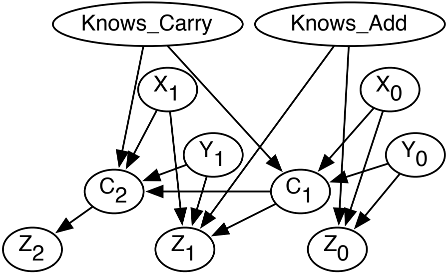
Predicting students errors

$$\begin{array}{rccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



Predicting students errors

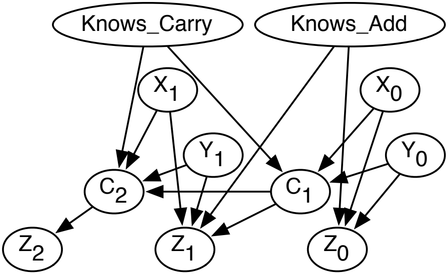
$$\begin{array}{cccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits

Predicting students errors

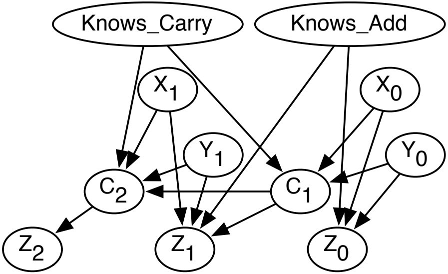
$$\begin{array}{rccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits, problems

Predicting students errors

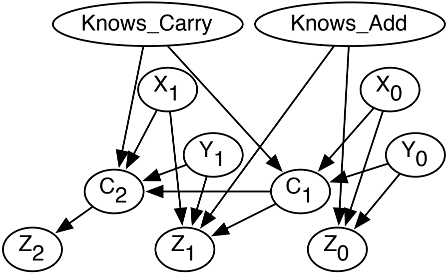
$$\begin{array}{rrrr} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits, problems, students

Predicting students errors

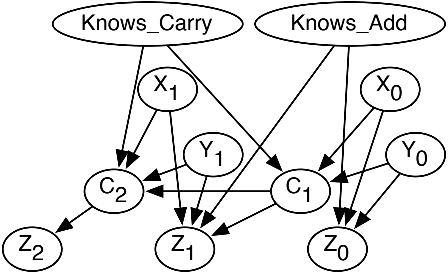
$$\begin{array}{rrrr} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits, problems, students, times?

Predicting students errors

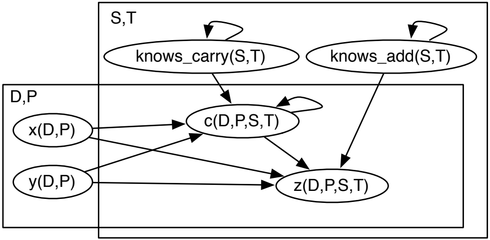
$$\begin{array}{cccc} & & x_2 & x_1 \\ + & & y_2 & y_1 \\ \hline & z_3 & z_2 & z_1 \end{array}$$



What if there were multiple digits, problems, students, times? How can we build a model before we know the individuals?

Multi-digit addition with parametrized BNs / plates

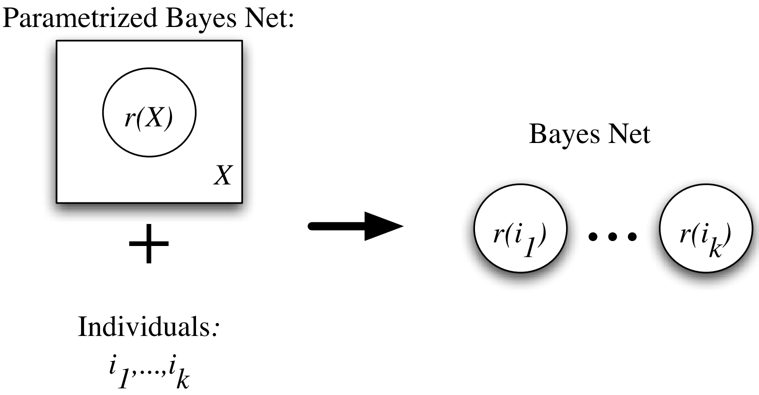
$$\begin{array}{cccc} & x_{j_x} & \cdots & x_2 & x_1 \\ + & y_{j_z} & \cdots & y_2 & y_1 \\ \hline & z_{j_z} & \cdots & z_2 & z_1 \end{array}$$



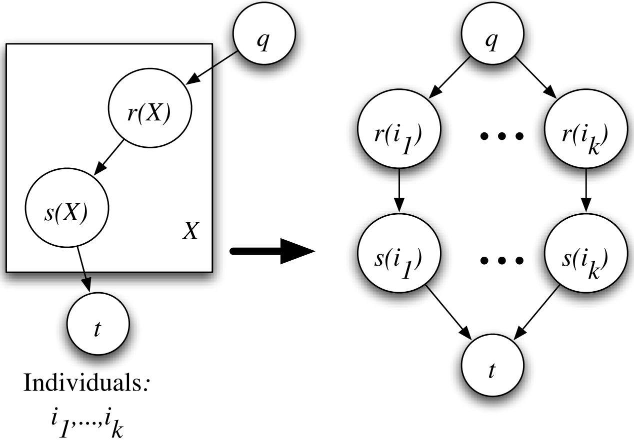
Parametrized Random Variables: $x(D,P)$, $y(D,P)$, $knows\_carry(S,T)$, $knows\_add(S,T)$, $c(D,P,S,T)$, $z(D,P,S,T)$ for digit D, problem P, student S, time T.

There is a random variable for each assignment of a value to D and a value to P in $x(D,P)$. . . .

Parametrized Bayesian networks / Plates



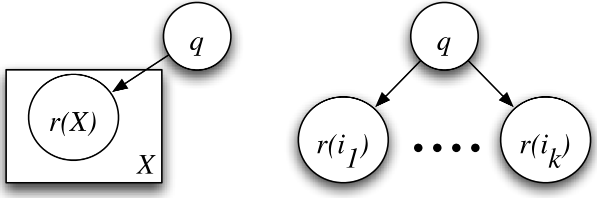
Parametrized Bayesian networks / Plates (2)



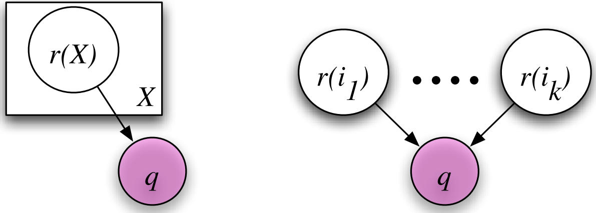
Creating Dependencies

Instances of plates are independent, except by common parents or children.

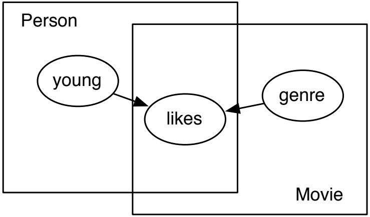
Common Parents



Observed Children

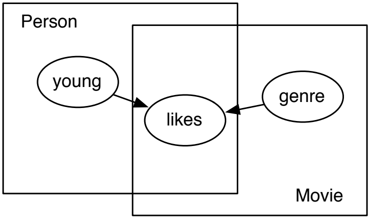


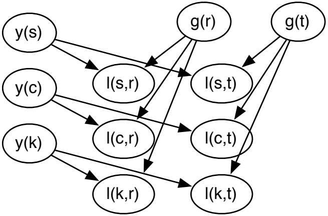
Overlapping plates



Relations: *likes*(P, M), *young*(P), *genre*(M)
likes is Boolean, *young* is Boolean,
genre has range $\{action, romance, family\}$

Overlapping plates



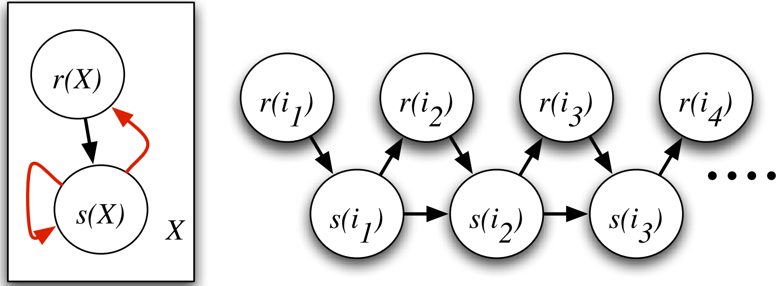


Relations: $likes(P, M)$, $young(P)$, $genre(M)$
likes is Boolean, *young* is Boolean,
genre has range $\{action, romance, family\}$
Three students: sam (s), chris (c), kim (k)
Two movies: rango (r), terminator (t)

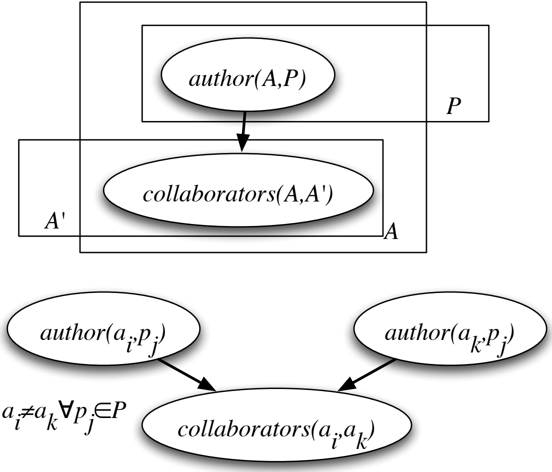
Representing Conditional Probabilities

- $P(likes(P, M)|young(P), genre(M))$ — parameter sharing — individuals share probability parameters.
- $P(happy(X)|friend(X, Y), mean(Y))$ — needs aggregation — $happy(a)$ depends on an unbounded number of parents.
- the carry of one digit depends on carry of the previous digit
- probability that two authors collaborate depends on whether they have a paper authored together

Creating Dependencies: Exploit Domain Structure



Creating Dependencies: Relational Structure



$\forall a_i \in A \; \forall a_k \in A \; a_i \neq a_k \; \forall p_j \in P$

Independent Choice Logic

- A language for first-order probabilistic models.
- Idea : combine logic and probability, where all uncertainty in handled in terms of Bayesian decision theory, and a logic program specifies consequences of choices.
- Parametrized random variables are represented as logical atoms, and plates correspond to logical variables.

Independent Choice Logic

- An alternative is a set of ground atomic formulas.
  $\mathcal{C}$, the choice space is a set of disjoint alternatives.
- $\mathcal{F}$, the facts is a logic program that gives consequences of choices.
- P_0 a probability distribution over alternatives:

$$\forall A \in \mathcal{C} \ \sum_{a \in A} P_0(a) = 1.$$

Meaningless Example

$$\mathcal{C} = \{\{c_1, c_2, c_3\}, \{b_1, b_2\}\}$$

$$\begin{aligned}
\mathcal{F} = \{ \; & f \leftarrow c_1 \wedge b_1, & & f \leftarrow c_3 \wedge b_2, \\
& d \leftarrow c_1, & & d \leftarrow {\sim}c_2 \wedge b_1, \\
& e \leftarrow f, & & e \leftarrow {\sim}d\}
\end{aligned}$$

$$P_0(c_1) = 0.5 \quad P_0(c_2) = 0.3 \quad P_0(c_3) = 0.2$$
$$P_0(b_1) = 0.9 \quad P_0(b_2) = 0.1$$

Semantics of ICL

- There is a possible world for each selection of one element from each alternative.
- The logic program together with the selected atoms specifies what is true in each possible world.
- The elements of different alternatives are independent.

Meaningless Example: Semantics

$$\mathcal{F} = \{\ f \leftarrow c_1 \wedge b_1,\quad f \leftarrow c_3 \wedge b_2,$$
$$d \leftarrow c_1,\quad d \leftarrow {\sim}c_2 \wedge b_1,$$
$$e \leftarrow f,\quad e \leftarrow {\sim}d\}$$

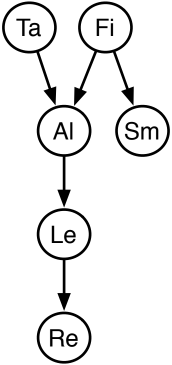
$P_0(c_1) = 0.5$ $P_0(c_2) = 0.3$ $P_0(c_3) = 0.2$
$P_0(b_1) = 0.9$ $P_0(b_2) = 0.1$

		selection		logic program			
w_1	$\models$	c_1	b_1	f	d	e	$P(w_1) = 0.45$
w_2	$\models$	c_2	b_1	${\sim}f$	${\sim}d$	e	$P(w_2) = 0.27$
w_3	$\models$	c_3	b_1	${\sim}f$	d	${\sim}e$	$P(w_3) = 0.18$
w_4	$\models$	c_1	b_2	${\sim}f$	d	${\sim}e$	$P(w_4) = 0.05$
w_5	$\models$	c_2	b_2	${\sim}f$	${\sim}d$	e	$P(w_5) = 0.03$
w_6	$\models$	c_3	b_2	f	${\sim}d$	e	$P(w_6) = 0.02$

$P(e) = 0.45 + 0.27 + 0.03 + 0.02 = 0.77$

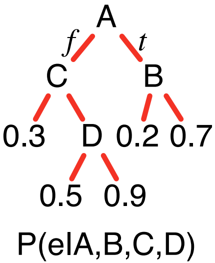
Belief Networks, Decision trees and ICL rules

- There is a local mapping from belief networks into ICL.



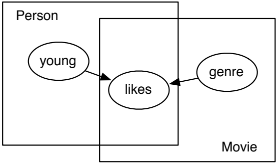
prob ta : 0.02.
prob $fire$: 0.01.
$alarm \leftarrow ta \wedge fire \wedge atf$.
$alarm \leftarrow {\sim}ta \wedge fire \wedge antf$.
$alarm \leftarrow ta \wedge {\sim}fire \wedge atnf$.
$alarm \leftarrow {\sim}ta \wedge {\sim}fire \wedge antnf$.
prob atf : 0.5.
prob $antf$: 0.99.
prob $atnf$: 0.85.
prob $antnf$: 0.0001.
$smoke \leftarrow fire \wedge sf$.
prob sf : 0.9.
$smoke \leftarrow {\sim}fire \wedge snf$.
prob snf : 0.01.

- Rules can represent decision tree with probabilities:



$e \leftarrow a \wedge b \wedge h_1.$	$P_0(h_1) = 0.7$
$e \leftarrow a \wedge \sim b \wedge h_2.$	$P_0(h_2) = 0.2$
$e \leftarrow \sim a \wedge c \wedge d \wedge h_3.$	$P_0(h_3) = 0.9$
$e \leftarrow \sim a \wedge c \wedge \sim d \wedge h_4.$	$P_0(h_4) = 0.5$
$e \leftarrow \sim a \wedge \sim c \wedge h_5.$	$P_0(h_5) = 0.3$

Movie Ratings



prob $young(P) : 0.4.$

prob $genre(M, action) : 0.4, genre(M, romance) : 0.3,$
$\quad genre(M, family) : 0.4.$

$likes(P, M) \leftarrow young(P) \wedge genre(M, G) \wedge ly(P, M, G).$

$likes(P, M) \leftarrow \sim young(P) \wedge genre(M, G) \wedge lny(P, M, G).$

prob $ly(P, M, action) : 0.7.$

prob $ly(P, M, romance) : 0.3.$

prob $ly(P, M, family) : 0.8.$

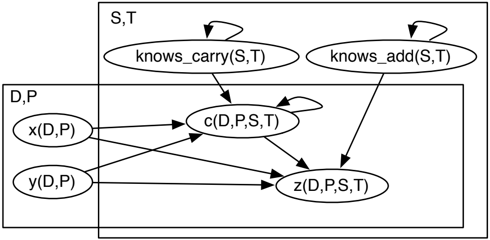
prob $lny(P, M, action) : 0.2.$

prob $lny(P, M, romance) : 0.9.$

prob $lny(P, M, family) : 0.3.$

Example: Multi-digit addition

$$\begin{array}{cccccc} & x_{j_x} & \cdots & x_2 & x_1 \\ + & y_{j_z} & \cdots & y_2 & y_1 \\ \hline & z_{j_z} & \cdots & z_2 & z_1 \end{array}$$



ICL rules for multi-digit addition

$$
\begin{aligned}
z(D,P,S,T) = V \leftarrow\ & \\
& x(D,P) = Vx \land \\
& y(D,P) = Vy \land \\
& c(D,P,S,T) = Vc \land \\
& knows\_add(S,T) \land \\
& \neg mistake(D,P,S,T) \land \\
& V \text{ is } (Vx + Vy + Vc) \text{ div } 10.
\end{aligned}
$$

$$
\begin{aligned}
z(D,P,S,T) = V \leftarrow\ & \\
& knows\_add(S,T) \land \\
& mistake(D,P,S,T) \land \\
& selectDig(D,P,S,T) = V. \\
z(D,P,S,T) = V \leftarrow\ & \\
& \neg knows\_add(S,T) \land \\
& selectDig(D,P,S,T) = V.
\end{aligned}
$$

Alternatives:

$\forall DPST \{noMistake(D,P,S,T), mistake(D,P,S,T)\}$

$\forall DPST \{selectDig(D,P,S,T) = V \mid V \in \{0..9\}\}$

Hidden Variables

$Student$	$Course$	$Grade$
s_1	c_1	A
s_2	c_1	C
s_1	c_2	B
s_2	c_3	B
s_3	c_2	B
s_4	c_3	B
s_3	c_4	?
s_4	c_4	?

Hidden Variables

Student	*Course*	*Grade*
s_1	c_1	*A*
s_2	c_1	*C*
s_1	c_2	*B*
s_2	c_3	*B*
s_3	c_2	*B*
s_4	c_3	*B*
s_3	c_4	?
s_4	c_4	?

$int(S)$	$diff(C)$	$grade(S,C)$		
		A	*B*	*C*
true	*true*	0.5	0.4	0.1
true	*false*	0.9	0.09	0.01
false	*true*	0.01	0.1	0.9
false	*false*	0.1	0.4	0.5

$P(int(S)) = 0.5$

$P(diff(C)) = 0.5$

Hidden Variables

Student	*Course*	*Grade*
s_1	c_1	A
s_2	c_1	C
s_1	c_2	B
s_2	c_3	B
s_3	c_2	B
s_4	c_3	B
s_3	c_4	?
s_4	c_4	?

$int(S)$	$diff(C)$	$grade(S,C)$		
		A	B	C
true	*true*	0.5	0.4	0.1
true	*false*	0.9	0.09	0.01
false	*true*	0.01	0.1	0.9
false	*false*	0.1	0.4	0.5

$P(int(S)) = 0.5$
$P(diff(C)) = 0.5$

$P(grade(s3,c4,a)|Obs) = 0.491,$
$P(grade(s3,c4,b)|Obs) = 0.245,\ P(grade(s3,c4,c)|Obs) = 0.264$
$P(grade(s4,c4,a)|Obs) = 0.264,$
$P(grade(s4,c4,b)|Obs) = 0.245,\ P(grade(s4,c4,c)|Obs) = 0.491$

Learning Relational Models with Hidden Variables

User	Item	Date	Rating
Sam	Terminator	2009-03-22	5
Sam	Rango	2011-03-22	4
Sam	The Holiday	2010-12-25	1
Chris	The Holiday	2010-12-25	4
...	...	...	

Netflix: 500,000 users, 17,000 movies, 100,000,000 ratings.

Learning Relational Models with Hidden Variables

User	Item	Date	Rating
Sam	Terminator	2009-03-22	5
Sam	Rango	2011-03-22	4
Sam	The Holiday	2010-12-25	1
Chris	The Holiday	2010-12-25	4
...	...	...	

Netflix: 500,000 users, 17,000 movies, 100,000,000 ratings.

r_{ui} = rating of user u on item i

$\hat{r_{ui}}$ = predicted rating of user u on item i

D = set of (u, i, r) tuples in the training set

Sum squares error:

$$\sum_{(u,i,r)\in D} (\hat{r_{ui}} - r)^2$$

Learning Relational Models with Hidden Variables

- Predict same for all ratings: $\hat{r_{ui}} = \mu$

Learning Relational Models with Hidden Variables

- Predict same for all ratings: $\hat{r_{ui}} = \mu$
- Adjust for each user and item: $\hat{r_{ui}} = \mu + b_i + c_u$

Learning Relational Models with Hidden Variables

- Predict same for all ratings: $\hat{r_{ui}} = \mu$
- Adjust for each user and item: $\hat{r_{ui}} = \mu + b_i + c_u$
- One hidden feature: f_i for each item and g_u for each user

$$\hat{r_{ui}} = \mu + b_i + c_u + f_i g_u$$

Learning Relational Models with Hidden Variables

- Predict same for all ratings: $\hat{r_{ui}} = \mu$
- Adjust for each user and item: $\hat{r_{ui}} = \mu + b_i + c_u$
- One hidden feature: f_i for each item and g_u for each user

$$\hat{r_{ui}} = \mu + b_i + c_u + f_i g_u$$

- k hidden features:

$$\hat{r_{ui}} = \mu + b_i + c_u + \sum_k f_{ik} g_{ku}$$

Learning Relational Models with Hidden Variables

- Predict same for all ratings: $\hat{r_{ui}} = \mu$
- Adjust for each user and item: $\hat{r_{ui}} = \mu + b_i + c_u$
- One hidden feature: f_i for each item and g_u for each user

$$\hat{r_{ui}} = \mu + b_i + c_u + f_i g_u$$

- k hidden features:

$$\hat{r_{ui}} = \mu + b_i + c_u + \sum_k f_{ik} g_{ku}$$

- Regularize

$$\mathit{minimize} \sum_{(u,i)\in K} (\mu + b_i + c_u + \sum_k f_{ik} g_{ku} - r_{ui})^2$$
$$+ \lambda(b_i^2 + c_u^2 + \sum_k f_{ik}^2 + g_{ku}^2)$$

Parameter Learning using Gradient Descent

$\mu \leftarrow$ average rating
assign $f[i,k]$, $g[k,u]$ randomly
assign $b[i], c[u]$ arbitrarily
repeat:
  for each $(u,i,r) \in D$:
    $e \leftarrow \mu + b[i] + c[u] + \sum_k f[i,k] * g[k,u] - r$
    $b[i] \leftarrow b[i] - \eta * e - \eta * \lambda * b[i]$
    $c[u] \leftarrow c[u] - \eta * e - \eta * \lambda * c[u]$
    for each feature k:
      $f[i,k] \leftarrow f[i,k] - \eta * e * g[k,u] - \eta * \lambda * f[i,k]$
      $g[k,u] \leftarrow g[k,u] - \eta * e * f[i,k] - \eta * \lambda * g[k,u]$