Collective Entity Resolution

Shobeir Fakhraei
University of Southern California

Some slides based on material provided by Lise Getoor, Pedro Domingos, Parag Singla, and others.

Where are we?

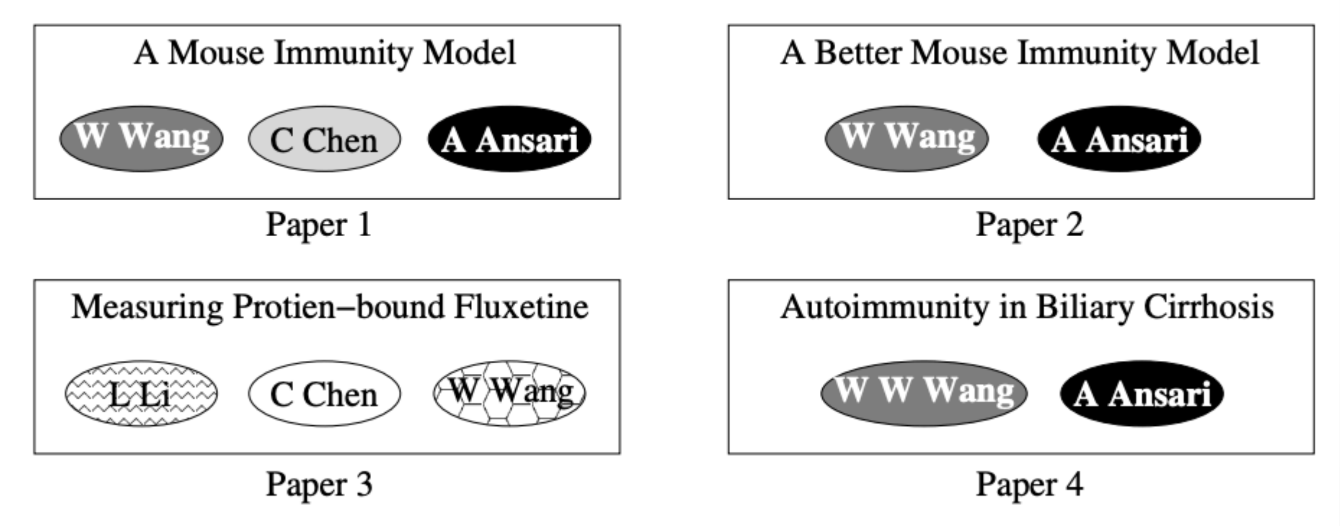
- Information Extraction
  - Semi-structured: Web, HTML
  - Unstructured: Text, Adds, Tweets
- Entity Linkage, Data Cleaning, Normalization
- Logical Data Integration
  - Mediators, Query Rewriting
  - Warehouse, Logical Data Exchange
- Automatic Source Modeling/Learning Schema Mappings
- Semantic Web
  - RDF, SPARQL, OWL, Linked Data
- Advanced Topics
  - Geospatial Data Integration, Knowledge Graphs

String Matching
…
Blocking

Motivating Example

Author Disambiguation:

(1) W. Wang, C. Chen, A. Ansari, “A mouse immunity model”
(2) W. Wang, A. Ansari, “A better mouse immunity model”
(3) L. Li, C. Chen, W. Wang,“Measuring protein-bound fluxetine”
(4) W. W. Wang, A. Ansari, “Autoimmunity in biliar



Probabilistic Graphical Models Background

Background

Bayes Theorem

- Definition of Conditional Probability

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

- Bayes Rule

(Thomas Bayes, 1763)

$$P(B \mid A) = \frac{P(A,B)}{P(A)}$$

$$= \frac{P(A \mid B)P(B)}{P(A)}$$

- Corollary:

The Chain Rule

$$P(A \mid B)P(B) = P(A,B)$$

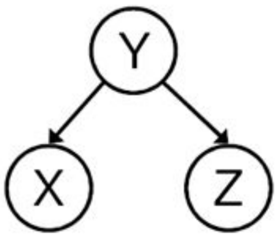
$$P(q_t, q_{t-1}, \ldots q_1) = P(q_t \mid q_{t-1}, \ldots q_1) P(q_{t-1}, \ldots q_1)$$

$$= P(q_t \mid q_{t-1}, \ldots q_1) P(q_{t-1} \mid q_{t-2}, \ldots q_1) P(q_{t-2}, \ldots q_1)$$

$$= P(q_1) \prod_{i=2}^{t} P(q_i \mid q_{i-1}, \ldots q_1)$$

Background
Conditional Independence

- Common cause



Y: Project due

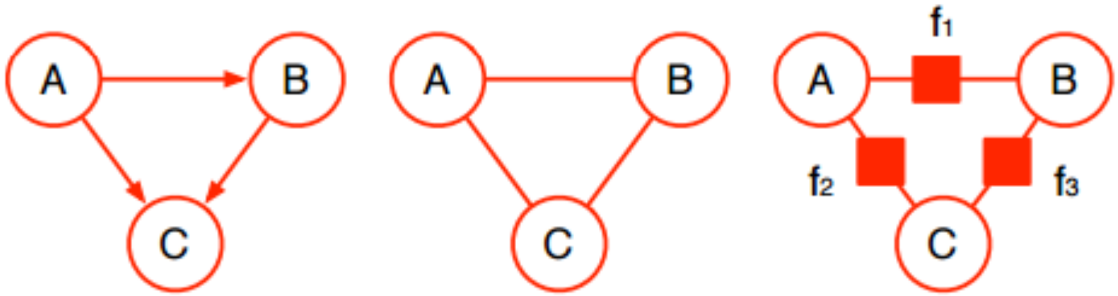
X: Newsgroup busy

Z: Lab full

$$P(X,Z|Y) = P(X|Y).P(Z|Y)$$

- Are X and Z independent?
  - No
- Are they conditionally independent given Y?
  - Yes

Background



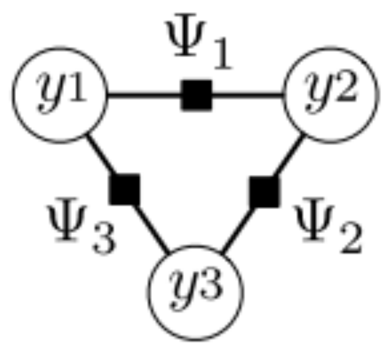
(a) Directed graph (b) Undirected graph (c) Factor graph

- Nodes represent random variables
- Edges reflect dependencies between variables

Background: Factor Graphs

Definition 2.1. A distribution $p(\mathbf{y})$ *factorizes according to a factor graph* G if there exists a set of local functions Ψ_a such that p can be written as

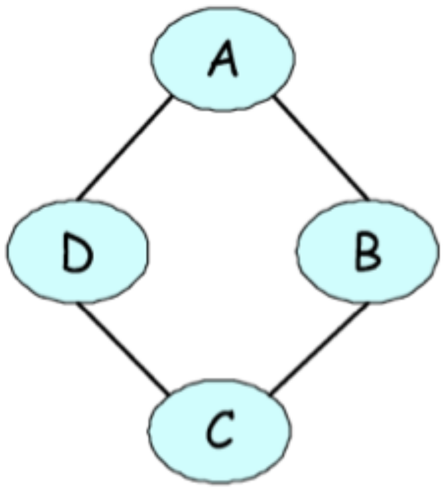
$$p(\mathbf{y}) = Z^{-1} \prod_{a \in F} \Psi_a(\mathbf{y}_{N(a)}) \tag{2.3}$$

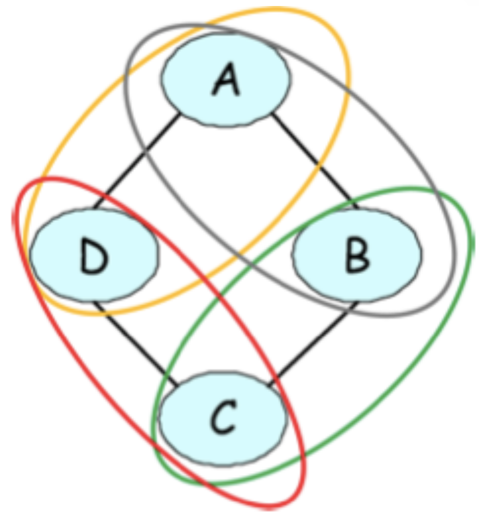


$$p(y_1, y_2, y_3) = \Psi_1(y_1, y_2)\Psi_2(y_2, y_3)\Psi_3(y_1, y_3)$$

For all $\mathbf{y} = (y_1, y_2, y_3)$

Background: Markov Random Fields





$$\tilde{p}(A, B, C, D) = \phi(A, B)\phi(B, C)\phi(C, D)\phi(D, A),$$

Normalize -> $$p(A, B, C, D) = \frac{1}{Z}\tilde{p}(A, B, C, D),$$

$$Z = \sum_{A,B,C,D} \tilde{p}(A, B, C, D)$$

Background: Log-Linear Models

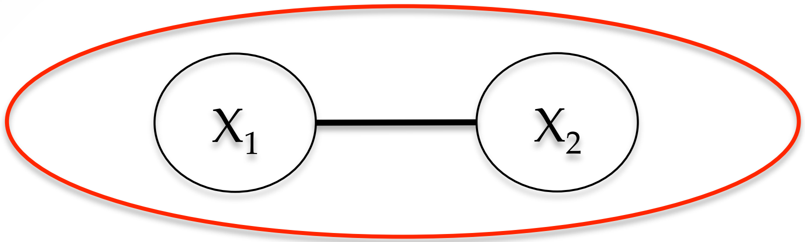
$$\tilde{p}(A,B,C,D) = \phi(A,B)\phi(B,C)\phi(C,D)\phi(D,A),$$

$$\tilde{P} = \prod_i \phi_i(\boldsymbol{D}_i)$$

$$\tilde{P} = \prod_j \exp(\quad w_j f_j(\boldsymbol{D}_j))$$

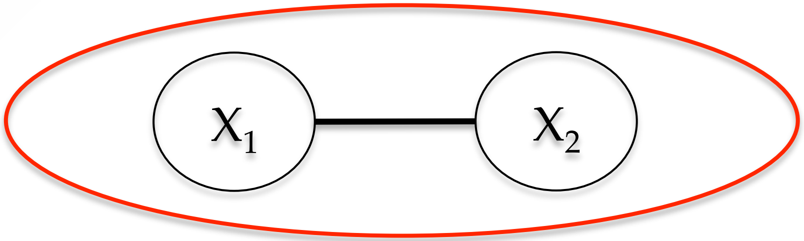
$$\tilde{P} = \exp\left(\quad \sum_j w_j f_j(\boldsymbol{D}_j)\right)$$

Background: Log-Linear Models



$$\phi(X_1, X_2) = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$$

Background: Log-Linear Models



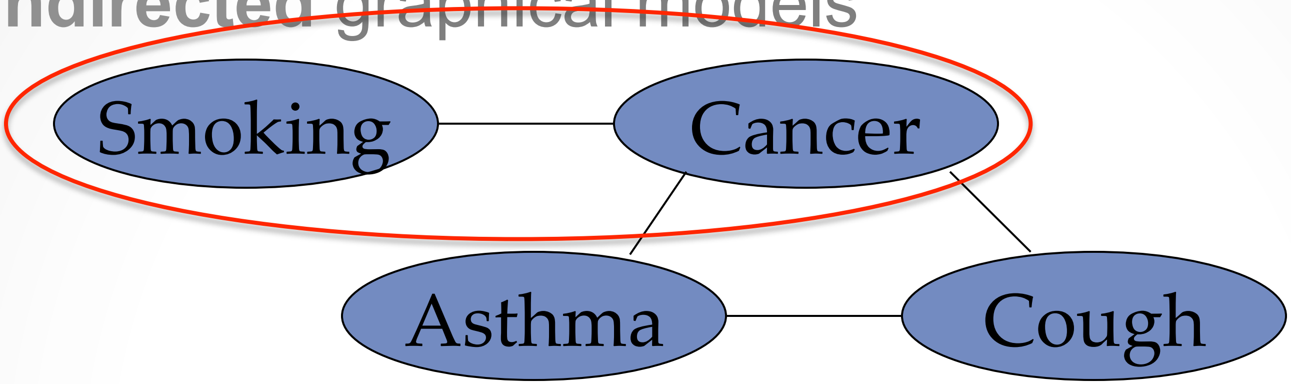
$$\phi(X_1, X_2) = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$$

$$f_{12}^{00} = \mathbb{1}\{X_1 = 0, X_2 = 0\}$$
$$f_{12}^{01} = \mathbb{1}\{X_1 = 0, X_2 = 1\}$$
$$f_{12}^{10} = \mathbb{1}\{X_1 = 1, X_2 = 0\}$$
$$f_{12}^{11} = \mathbb{1}\{X_1 = 1, X_2 = 1\}$$

$$\phi(X_1, X_2) = \exp(\quad \textstyle\sum_{kl} w_{kl} f_{ij}^{kl}(X_1, X_2))$$
$$w_{kl} = \quad \log a_{kl}$$

Markov Networks

- Undirected graphical models



- Potential functions defined over cliques

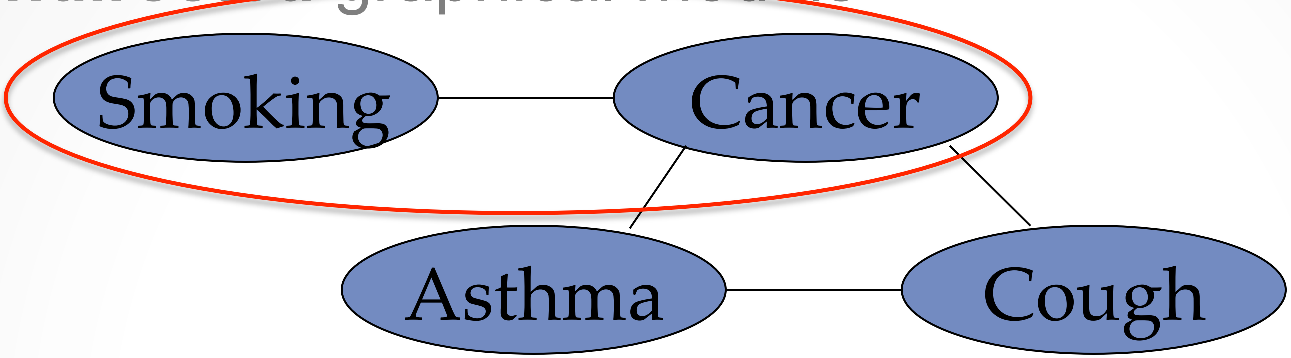
$$P(x) = \frac{1}{Z} \prod_c \Phi_c(x_c)$$

$$Z = \sum_x \prod_c \Phi_c(x_c)$$

Smoking	Cancer	Φ(S,C)
False	False	4.5
False	True	4.5
True	False	2.7
True	True	4.5

Markov Networks

- Undirected graphical models



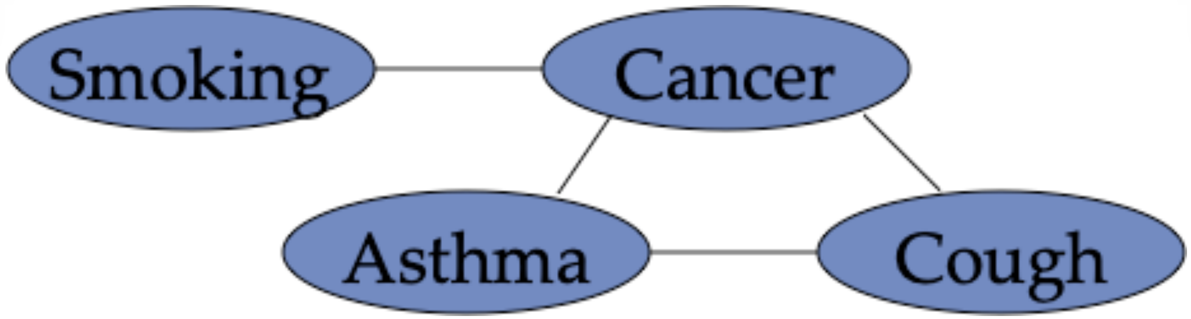
- Log-linear model:

$$P(x) = \frac{1}{Z} \exp\left(\sum_i \underbrace{w_i}_{\text{Weight of Feature } i} \underbrace{f_i(x)}_{\text{Feature } i} \right)$$

$$f_1(\text{Smoking, Cancer}) = \begin{cases} 1 & \text{if } \neg \text{ Smoking } \vee \text{ Cancer} \\ 0 & \text{otherwise} \end{cases}$$

$$w_1 = 1.5$$

How to build the MRF model?



Markov Logic Networks

Artificial Intelligence

Planning

Robotics

Applications

NLP

Multi-Agent Systems

Vision

Interface Layer

Markov Logic

Representation

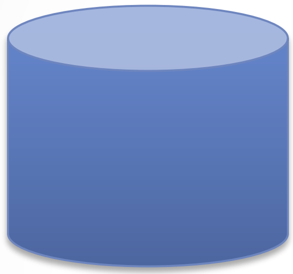
Infrastructure

Inference

Learning

Markov Logic

Data

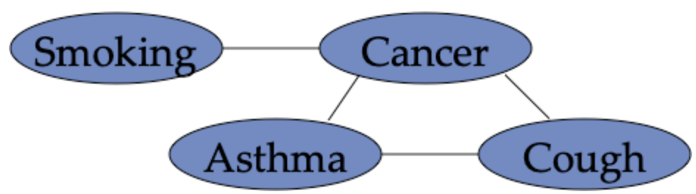


+

Frist Order Logic

1.5	$\forall x \; Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y \; Friends(x,y) \wedge Smokes(x) \Rightarrow Smokes(y)$

MRF



Inference and Learning

Background: First-Order Logic

- Constants, variables, functions, predicates
  - Anna, x, MotherOf(x), Friends(x,y)
- Grounding: Replace all variables by constants
  - Friends (Anna, Bob)
- Formula: Predicates connected by operators
  - Smokes(x) $\Rightarrow$ Cancer(x)
- Knowledge Base (KB): A set of formulas
  - Can be equivalently converted into a clausal form
- World: Assignment of truth values to all ground predicates

Markov Logic

- A logical KB is a set of **hard constraints** on the set of possible worlds
- Let's make them **soft constraints**:
  When a world violates a formula,
  It becomes less probable, not impossible
- Give each formula a **weight**
  (Higher weight $\Rightarrow$ Stronger constraint)

$$P(\text{world}) \propto \exp\left(\sum \text{weights of formulas it satisfies}\right)$$

Definition

- A Markov Logic Network (MLN) is a set of pairs (F, w) where
  - F is a formula in first-order logic
  - w is a real number

- Together with a finite set of constants, it defines a Markov network with
  - One node for each grounding of each predicate in the MLN
  - One feature for each grounding of each formula F in the MLN, with the corresponding weight w

Example: Friends & Smokers

$$\forall x\ Smokes(x) \Rightarrow Cancer(x)$$

$$\forall x, y\ Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$$

Example: Friends & Smokers

1.5	$\forall x \; Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y \; Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

Example: Friends & Smokers

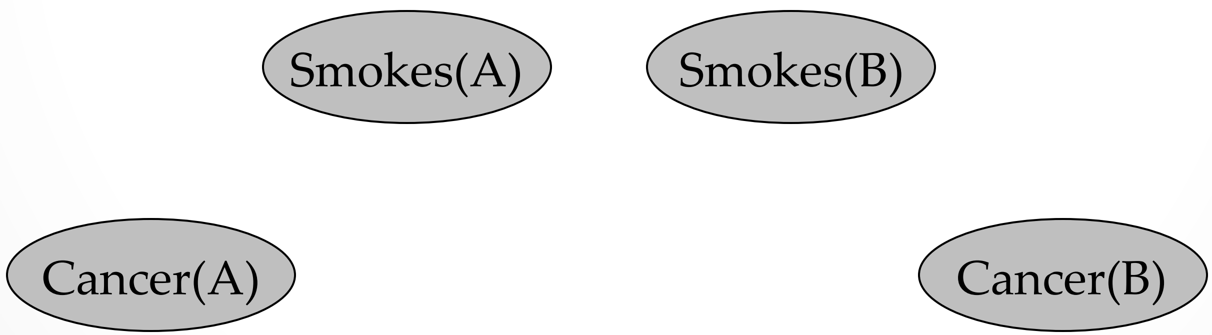
1.5	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y\ Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

Two constants: **Ana** (A) and **Bob** (B)

Example: Friends & Smokers

1.5	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y\ Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

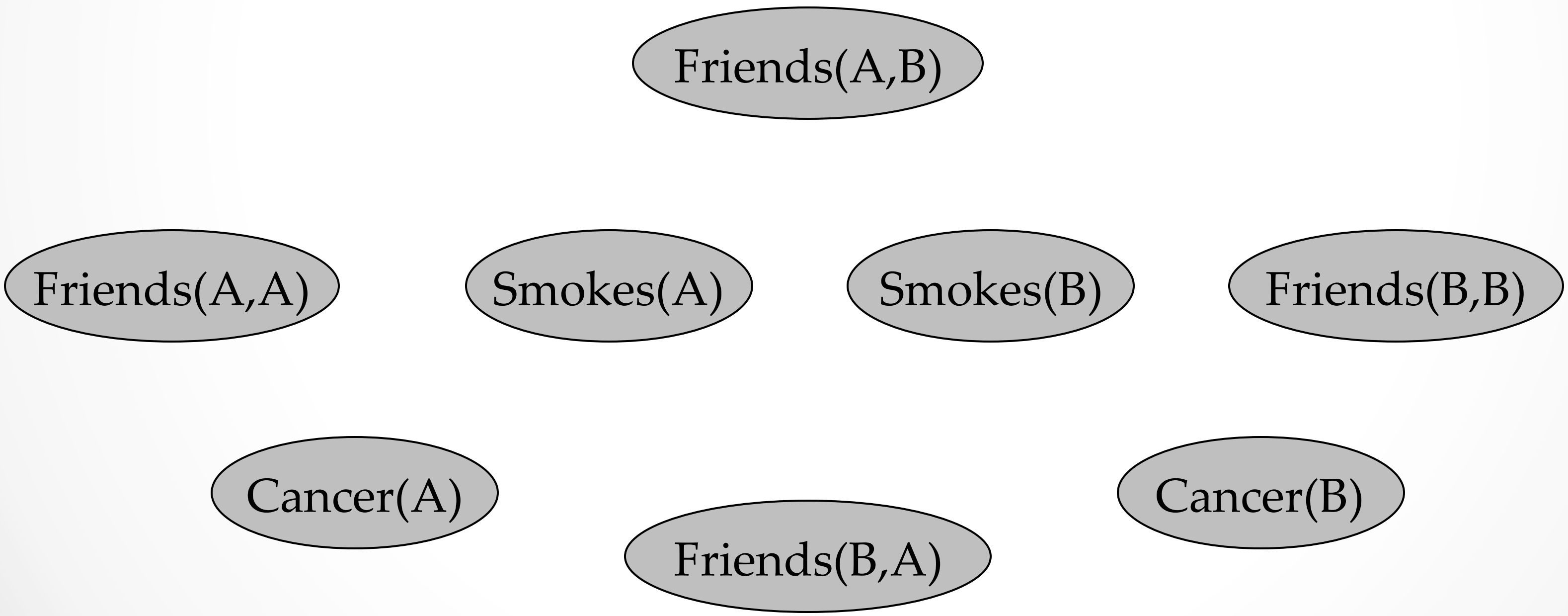
Two constants: **Ana** (A) and **Bob** (B)



Example: Friends & Smokers

1.5	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y\ Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

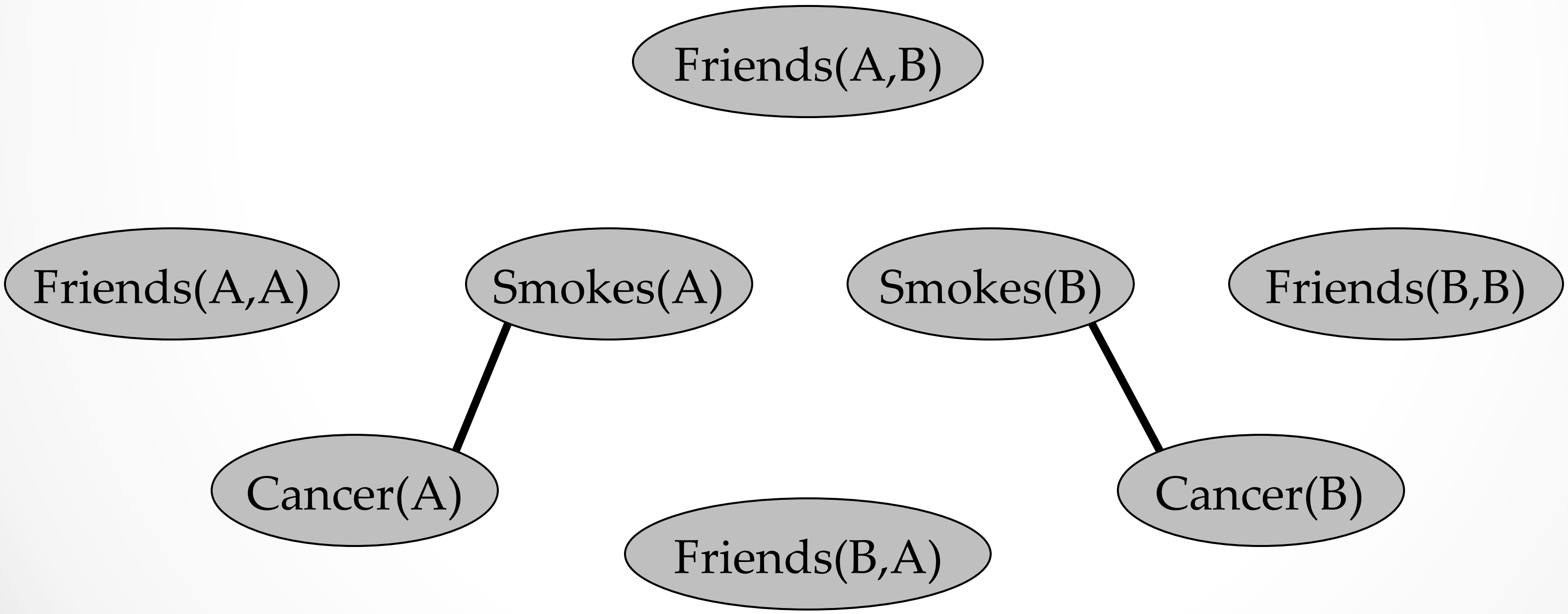
Two constants: **Ana** (A) and **Bob** (B)



Example: Friends & Smokers

1.5	$\forall x \; Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y \; Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

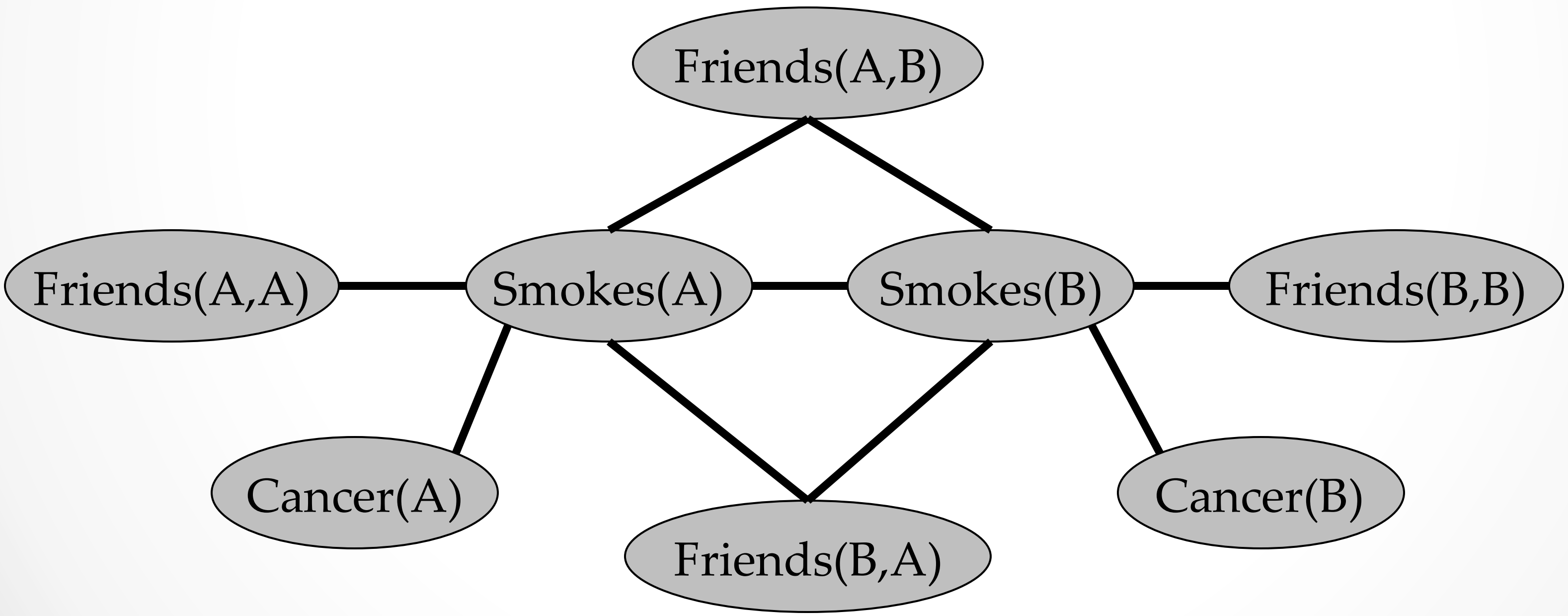
Two constants: **Ana** (A) and **Bob** (B)



Example: Friends & Smokers

1.5	$\forall x \; Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y \; Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

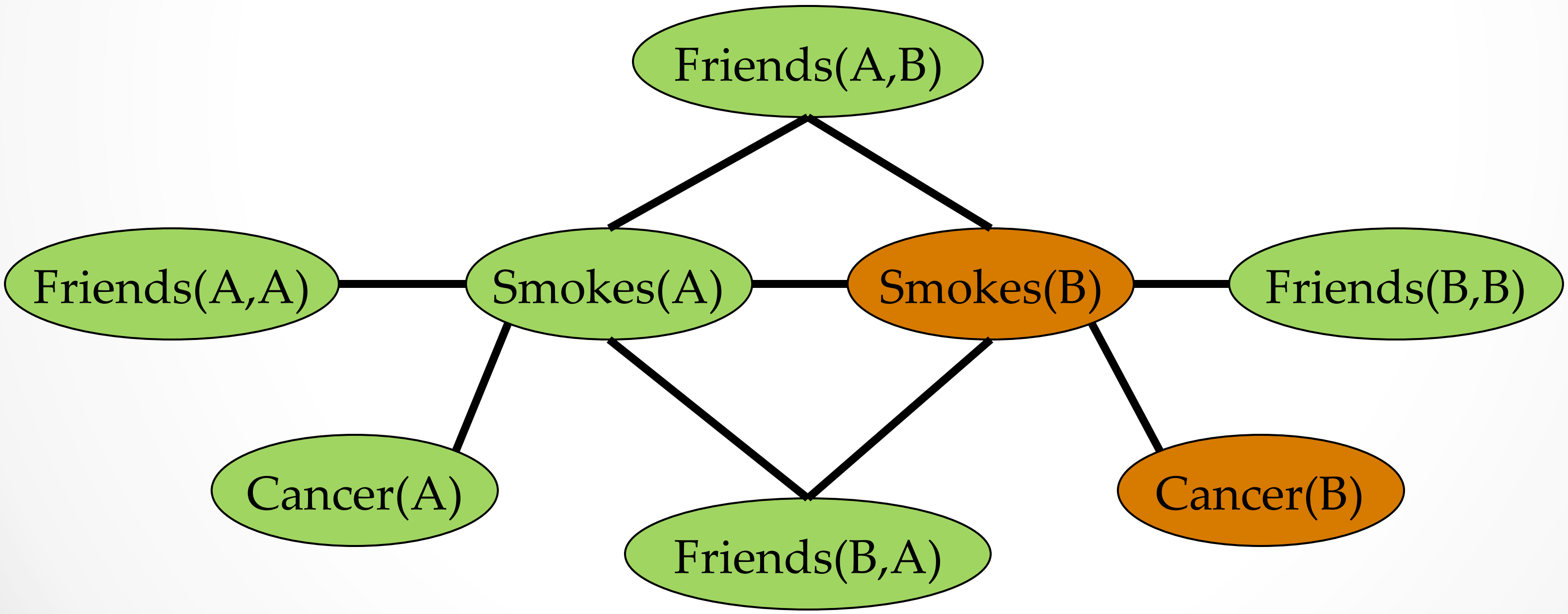
Two constants: **Ana** (A) and **Bob** (B)



Example: Friends & Smokers

1.5	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
1.1	$\forall x, y\ Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

Two constants: **Ana** (A) and **Bob** (B)



State of the World ≡ {0,1} Assignment to the nodes

Markov Logic Networks

- MLN is **template** for ground Markov networks
- Probability of a world x:

$$P(x) = \frac{1}{Z} \exp\left(\sum_{k \in ground\ formulas} w_k f_k(x) \right)$$

- One feature for each ground formula

$$f_k(x) = \begin{cases} 1 \text{ if } kth \text{ formula is satisfied given x} \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

Markov Logic Networks

- MLN is **template** for ground Markov nets
- Probability of a world x:

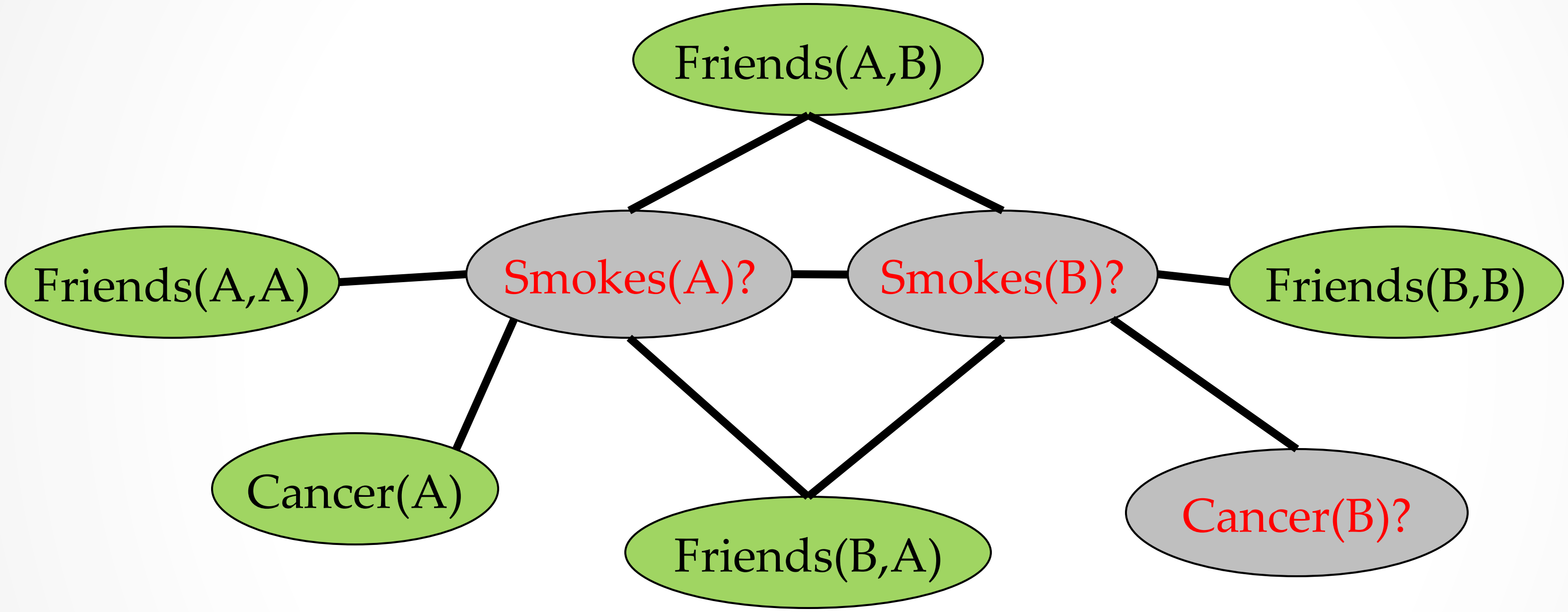
$$P(x) = \frac{1}{Z}\exp\left(\sum_{k \in ground\ formulas} w_k f_k(x)\right)$$

$$P(x) = \frac{1}{Z}\exp\left(\sum_{i \in \text{MLN formulas}} \boxed{w_i}\boxed{n_i(x)}\right)$$

Weight of formula *i*

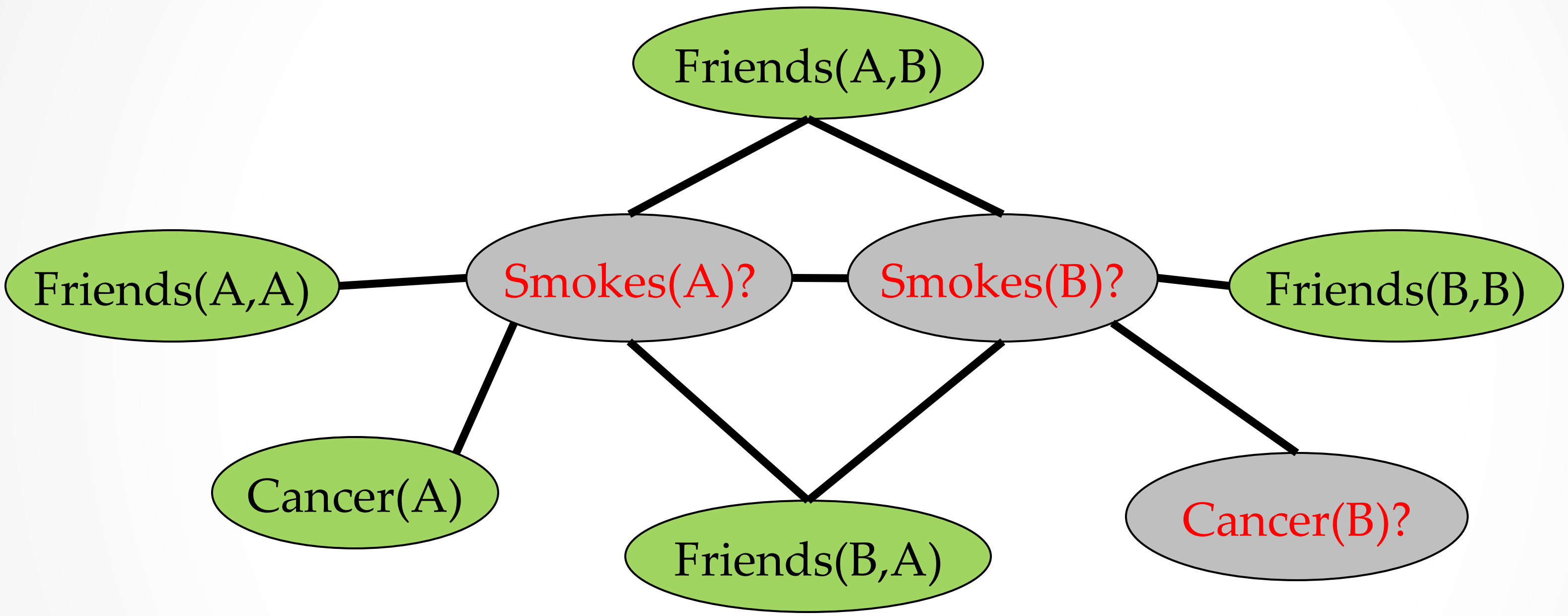
No. of true groundings of formula *i* in *x*

Inference



Friends(A,B)
Friends(A,A)
Smokes(A)?
Smokes(B)?
Friends(B,B)
Cancer(A)
Friends(B,A)
Cancer(B)?

Most Probable Explanation (MPE) Inference



What is the most likely state of Smokes(A), Smokes(B), Cancer(B)?

MPE Inference

- **Problem:** Find most likely state of world given evidence

$$\arg\max_{y} P(y \mid x)$$

Query Evidence

MPE Inference

- **Problem:** Find most likely state of world given evidence

$$\arg\max_{y} P(y \mid x) =$$

$$= \arg\max_{y} \frac{1}{Z_x} \exp\left(\sum_i w_i n_i(x, y) \right)$$

MPE Inference

- **Problem:** Find most likely state of world given evidence

$$\arg\max_{y} P(y \mid x) =$$

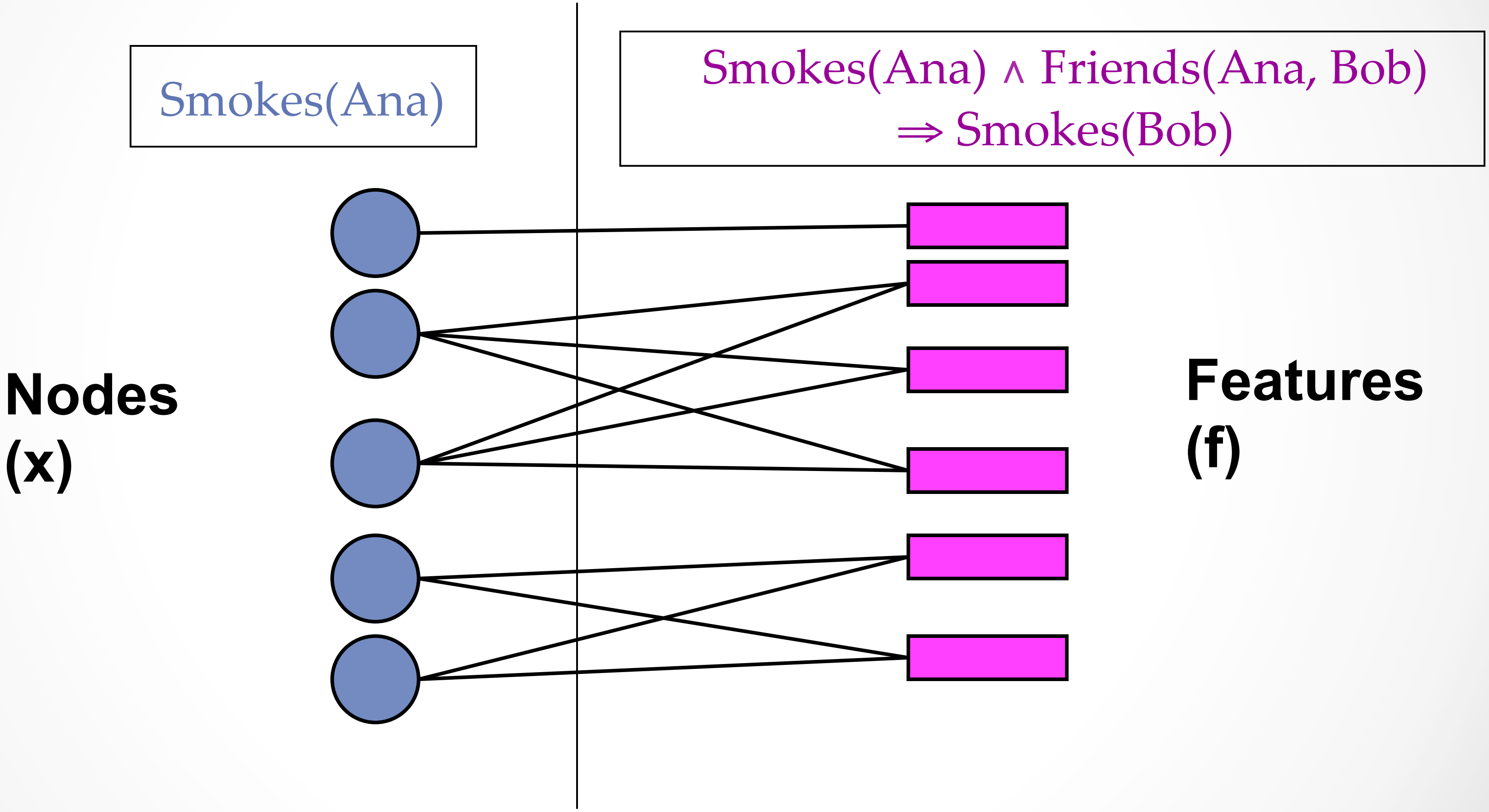
$$= \arg\max_{y} \frac{1}{Z_x} \exp\left(\sum_i w_i n_i(x, y) \right)$$

$$= \arg\max_{y} \sum_i w_i n_i(x, y)$$

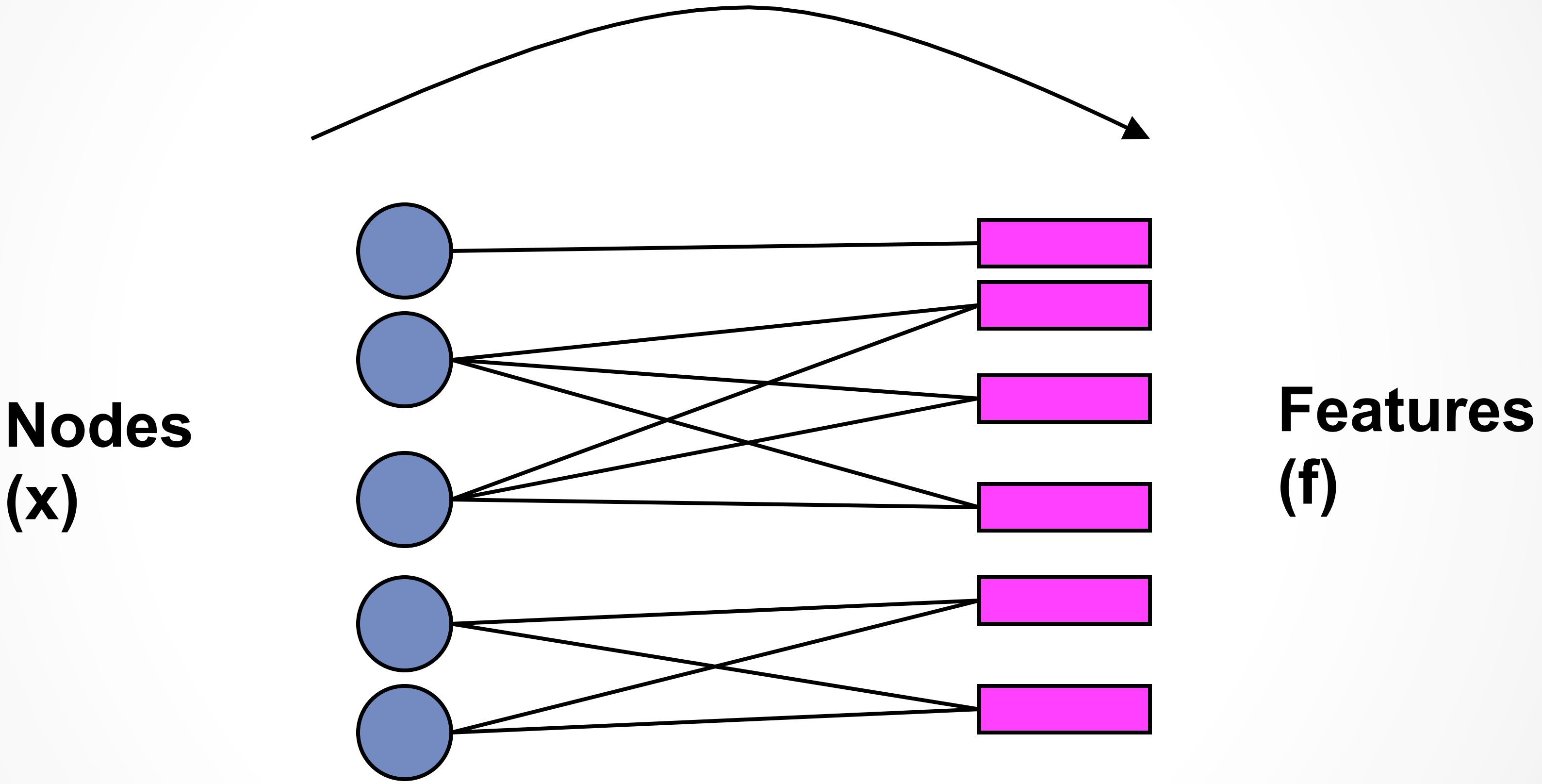
Belief Propagation

- Bipartite network of nodes (variables) and features
- In Markov logic:
  - Nodes = Ground atoms
  - Features = Ground clauses
- Exchange messages until convergence
- Messages
  - Current approximation to node marginals

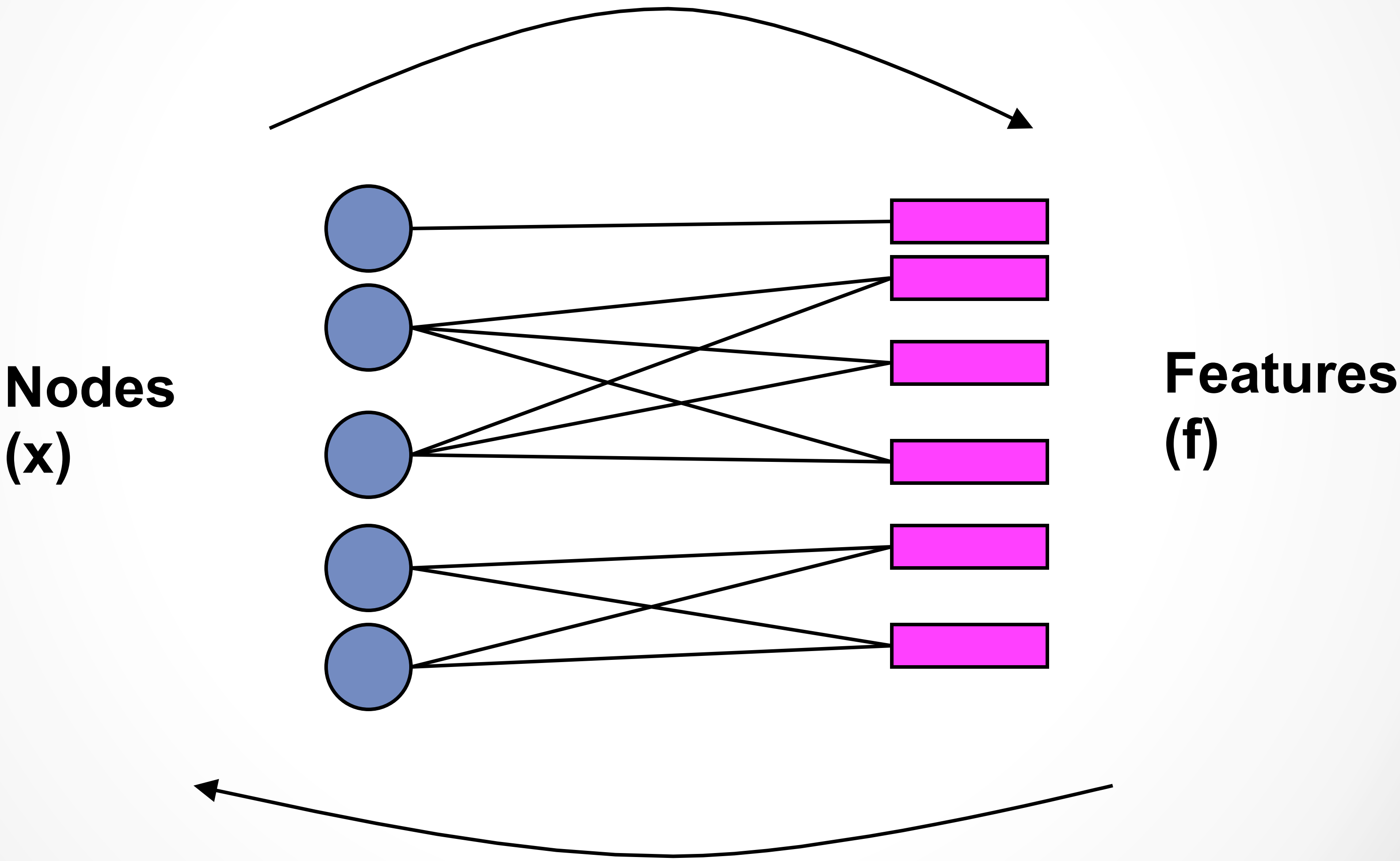
Belief Propagation



Belief Propagation



Belief Propagation



Learning Parameters (Weights)

$w_1?$	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
$w_2?$	$\forall x, y\ Friends(x, y) \wedge Smokes(x) \Rightarrow Smokes(y)$

Three constants: **Ana, Bob, John**

Learning Parameters (Weights)

$w_1?$	$\forall x\ Smokes(x) \Rightarrow Cancer(x)$
$w_2?$	$\forall x, y\ Friends(x,y) \wedge Smokes(x) \Rightarrow Smokes(y)$

Three constants: **Ana, Bob, John**

Smokes
Smokes(Ana)
Smokes(Bob)

Cancer
Cancer(Ana)
Cancer(Bob)

Friends
Friends(Ana, Bob)
Friends(Bob, Ana)
Friends(Ana, John)
Friends(John, Ana)

Closed World Assumption:
Anything not in the database is assumed false.

Learning Parameters (Weights)

- Given training data
- Maximize conditional likelihood of query (y) given evidence (x)
- Use gradient ascent
- Requires inference at each step (slow!)
- Approximate expected counts by counts in MPE state of y given x

$$\frac{\partial}{\partial w_i} \log P_w(y \mid x) = \boxed{n_i(x, y)} - \boxed{E_w\left[n_i(x, y)\right]}$$

No. of true groundings of clause *i* in data

Expected no. true groundings according to model

Entity Resolution with MLN

Entity Resolution

Author

Title

Venue

Parag Singla and Pedro Domingos, "Memory-Efficient Inference in Relational Domains" (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies", in Proc. AAAI-06, Boston, MA, 2006.

Poon H. (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

Entity Resolution

Author
Title
Venue

Parag Singla and Pedro Domingos, “Memory-Efficient Inference in Relational Domains” (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies”, in Proc. AAAI-06, Boston, MA, 2006.

Poon H. (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

Entity Resolution

Author

Title

Venue

Parag Singla and Pedro Domingos, "Memory-Efficient Inference in Relational Domains" (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies", in Proc. AAAI-06, Boston, MA, 2006.

Poon H. (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

ER: Predicates

```
HasWord(field,token)
HasField(record,field)
SameField(field,field)
SameRecord(record,record)
```

ER: Predicates & Formulas

```
HasWord(field,token)
HasField(record,field)
SameField(field,field)
SameRecord(record,record)
```

$\mathtt{HasWord}(f_1,t) \wedge \mathtt{HasWord}(f_2,t) \Rightarrow \mathtt{SameField}(f_1,f_2)$

(Similarity of fields based on shared tokens)

ER: Predicates & Formulas

```
HasWord(field,token)
HasField(record,field)
SameField(field,field)
SameRecord(record,record)
```

$\mathrm{HasWord}(f_1,t) \wedge \mathrm{HasWord}(f_2,t) \Rightarrow \mathrm{SameField}(f_1,f_2)$

(Similarity of fields based on shared tokens)

$\mathrm{HasField}(r_1,f_1) \wedge \mathrm{HasField}(r_2,f_2) \wedge \mathrm{SameField}(f_1,f_2)$
$\Leftrightarrow \mathrm{SameRecord}(r_1,r_2)$

(Similarity of records based on similarity of fields and vice-versa)

ER: Predicates & Formulas

```
HasWord(field,token)
HasField(record,field)
SameField(field,field)
SameRecord(record,record)
```

$\mathtt{HasWord}(f_1,t) \wedge \mathtt{HasWord}(f_2,t) \Rightarrow \mathtt{SameField}(f_1,f_2)$
(Similarity of fields based on shared tokens)

$\mathtt{HasField}(r_1,f_1) \wedge \mathtt{HasField}(r_2,f_2) \wedge \mathtt{SameField}(f_1,f_2)$
$\Leftrightarrow \mathtt{SameRecord}(r_1,r_2)$
(Similarity of records based on similarity of fields and vice-versa)

$\mathtt{SameRecord}(r_1,r_2)$ ^ $\mathtt{SameRecord}(r_2,r_3) \Rightarrow \mathtt{SameRecord}(r_1,r_3)$
(Transitivity)

Example: Author Disambiguation

HasTitle$(b_1,t_1) \wedge$ **HasTitle**$(b_2,t_2) \wedge$ **HasWord**$(t_1,\text{word}) \wedge$ **HasWord**(t_2,word)
$\Rightarrow$ **SameBib**(b_1, b_2)

SameBib$(b_1, b_2) \wedge$ **SameBib**(b_2, b_3)
$\Rightarrow$ **SameBib**(b_1, b_3)

HasAuthor$(b_1,a_1) \wedge$ **HasAuthor**$(b_2,a_2) \wedge$ **SameBib**(b_1,b_2)
$\Rightarrow$ **SameAuthor**(a_1,a_2)

HasAuthor$(b_1,a_1) \wedge$ **HasAuthor**$(b_2,a_2) \wedge$ **SameAuthor**(a_1,a_2)
$\Rightarrow$ **SameBib**(b_1,b_2)

Example: Author Disambiguation

HasTitle$(b_1,t_1) \wedge$ **HasTitle**$(b_2,t_2) \wedge$ **HasWord**$(t_1,\text{word}) \wedge$ **HasWord**(t_2,word)
$\Rightarrow$ **SameBib**(b_1, b_2)

SameBib$(b_1, b_2) \wedge$ **SameBib**(b_2, b_3)
$\Rightarrow$ **SameBib**(b_1, b_3)

HasAuthor$(b_1,a_1) \wedge$ **HasAuthor**$(b_2,a_2) \wedge$ **SameBib**(b_1,b_2)
$\Rightarrow$ **SameAuthor**(a_1,a_2)

HasAuthor$(b_1,a_1) \wedge$ **HasAuthor**$(b_2,a_2) \wedge$ **SameAuthor**(a_1,a_2)
$\Rightarrow$ **SameBib**(b_1,b_2)

HasAuthor$(b,a_1) \wedge$ **HasAuthor**(b,a_2)
$\Rightarrow$ **Coauthor**(a_1,a_2)

Coauthor$(a_1, a_2) \wedge$ **Coauthor**(a_1, a_3)
$\Rightarrow$ **SameAuthor**(a_1,a_3)

Experimental Results

Table 1. Experimental results on the Cora database.

System	Citation CLL	Citation AUC	Author CLL	Author AUC	Venue CLL	Venue AUC
NB	-0.637 ± 0.010	0.913 ± 0.000	-0.133 ± 0.021	0.986 ± 0.000	-0.747 ± 0.017	0.738 ± 0.002
MLN(B)	-0.643 ± 0.010	0.915 ± 0.000	-0.131 ± 0.022	0.987 ± 0.000	-0.760 ± 0.017	0.736 ± 0.002
MLN(B+C)	-0.809 ± 0.012	0.891 ± 0.000	-0.386 ± 0.064	0.968 ± 0.000	-1.163 ± 0.034	0.741 ± 0.001
MLN(B+T)	-0.369 ± 0.003	0.949 ± 0.000	-0.213 ± 0.036	0.994 ± 0.000	-1.036 ± 0.029	0.745 ± 0.002
MLN(B+C+T)	-0.597 ± 0.007	0.964 ± 0.000	-0.171 ± 0.043	0.984 ± 0.000	-0.704 ± 0.023	0.828 ± 0.002
MLN(B+C+T+S)	-0.503 ± 0.006	0.988 ± 0.000	-0.100 ± 0.033	0.992 ± 0.000	-0.874 ± 0.027	0.807 ± 0.002
MLN(B+N+C+T)	-0.879 ± 0.008	0.952 ± 0.000	-0.096 ± 0.032	0.992 ± 0.000	-0.781 ± 0.023	0.817 ± 0.002
MLN(G+C+T)	-0.394 ± 0.004	0.973 ± 0.000	-0.263 ± 0.053	0.980 ± 0.000	-1.196 ± 0.031	0.743 ± 0.002

Table 2. Experimental results on the BibServ database.

System	Citation CLL	Citation AUC	Author CLL	Author AUC	Venue CLL	Venue AUC
MLN(B)	-0.008 ± 0.003	0.997 ± 0.001	-0.586 ± 0.114	0.910 ± 0.013	-0.806 ± 0.121	0.908 ± 0.011
MLN(B+C)	-0.001 ± 0.000	0.999 ± 0.000	-0.544 ± 0.113	0.887 ± 0.007	-1.166 ± 0.151	0.876 ± 0.012
MLN(B+T)	-0.006 ± 0.003	0.993 ± 0.003	-0.600 ± 0.116	0.909 ± 0.013	-0.827 ± 0.123	0.898 ± 0.010
MLN(B+C+T)	-0.006 ± 0.004	0.998 ± 0.000	-0.473 ± 0.105	0.928 ± 0.009	-1.146 ± 0.149	0.876 ± 0.012
MLN(B+C+T+S)	-0.006 ± 0.004	0.970 ± 0.020	-0.486 ± 0.107	0.926 ± 0.010	-1.133 ± 0.148	0.876 ± 0.012
MLN(B+N+C+T)	-0.018 ± 0.005	1.000 ± 0.000	-0.363 ± 0.091	0.940 ± 0.008	-0.936 ± 0.133	0.897 ± 0.012
MLN(G+C+T)	-0.735 ± 0.101	0.491 ± 0.000	-4.679 ± 0.256	0.432 ± 0.001	-0.716 ± 0.112	0.906 ± 0.012

Software Packages

Alchemy
University of Washington
http://alchemy.cs.washington.edu

Tufy
Stanford University
http://i.stanford.edu/hazy/tuffy/

Continuous Variables

Probabilistic Soft Logic
University of Maryland/ UC Santa Cruz
https://psl.linqs.org/
Entity Resolution Example:
https://github.com/linqs/psl-examples/tree/master/entity-resolution