Pseudo-Likelihood for Relational Data

Oliver Schulte

School of Computing Science

Simon Fraser University

Vancouver, Canada

To appear at SIAM SDM conference on data mining.

The Main Topic

- In relational data, units are interdependent
  ⇨ no product likelihood function for model.
- How to do model selection?
- Proposal of this talk: use **pseudo likelihood.**
  - *Unnormalized* product likelihood.
  - Like independent-unit likelihood, but with event frequencies instead of event counts.

Overview

- Define pseudo log-likelihood for *directed graphical models* (Bayes Nets).
- Interpretation as *expected log-likelihood* of random small groups of units.
- Learning Algorithms:
  - MLE solution.
  - Model Selection.
- Simulations.

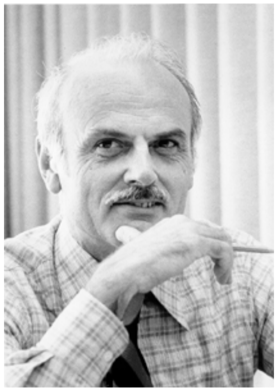
Outline

- Brief intro to relational databases.
- Statistics and Relational Databases.
- Briefer intro to Bayes nets.
- Relational Random Variables.
- Relational (pseudo)-likelihoods.

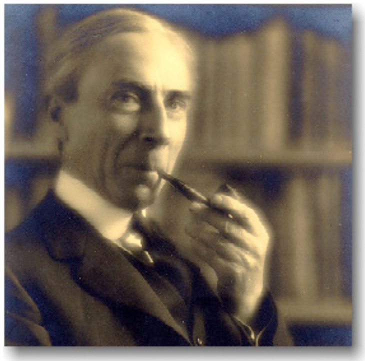
Relational Databases

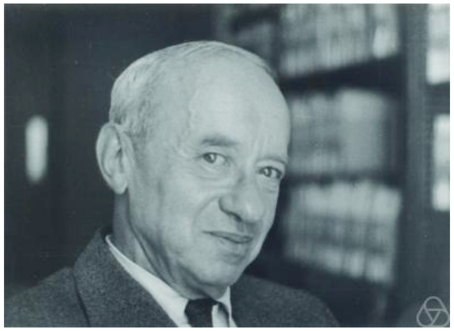
- 1970s: Computers are spreading. Many organizations use them to store their data.
- Ad hoc formats
  - ⇨ hard to build general data management systems.
  - ⇨ lots of duplicated effort.
- The Standardization Dilemma:
  - Too restrictive: doesn't fit users' needs.
  - Too loose: back to ad-hoc solutions.

The Relational Format



- Codd (IBM Research 1970)
- The fundamental question: *What kinds of information do users need to represent?*
- Answered by 1^{st}-order predicate logic!
  (Russell, Tarski).





- The world consists of
  - Individuals/entities.
  - Relationships/links among them.

Tabular Representation

- Tables for Entity Types, Relationships.

Student

s-id	Intelligence	Ranking
Jack	3	1
Kim	2	1
Paul	1	2

Course

c-id	Rating	Difficulty
101	3	1
102	2	2

Professor

p-id	Popularity	Teaching-a
Oliver	3	1
Jim	2	1

RA

s-id	p-id	Salary	Capability
Jack	Oliver	High	3
Kim	Oliver	Low	1
Paul	Jim	Med	2

Registration

s-id	c.id	Grade	Satisfaction
Jack	101	A	1
Jack	102	B	2
Kim	102	A	1
Paul	101	B	1

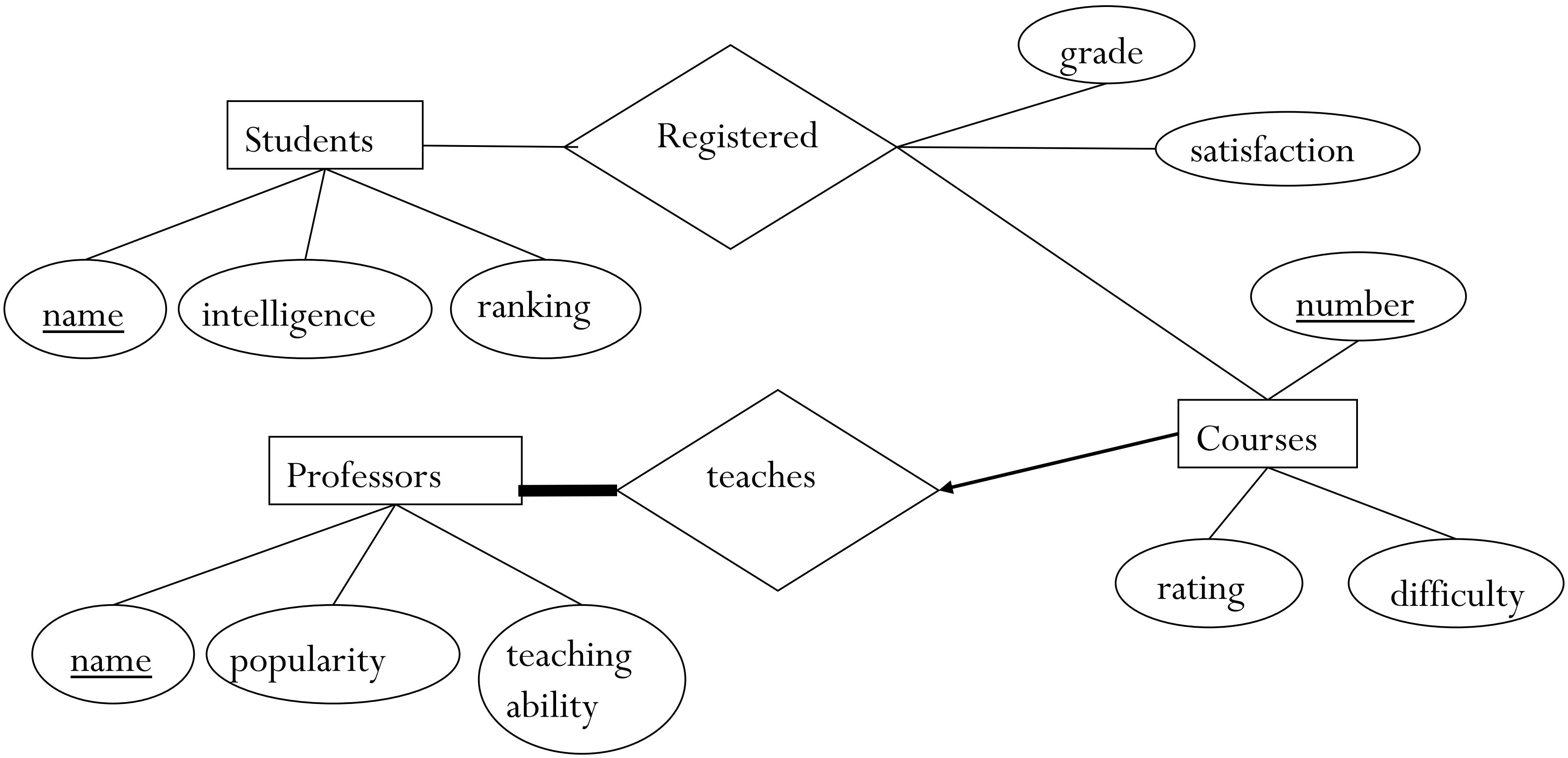
Database Management Systems

- Maintain data in linked tables.
- Structured Query Language (SQL) allows fast *data retrieval*.
  - E.g., find all SFU students who are statistics majors with gpa > 3.0.
- Multi-billion dollar industry, $15+ bill in 2006.
- IBM, Microsoft, Oracle, SAP, Peoplesoft.

Relational Domain Models

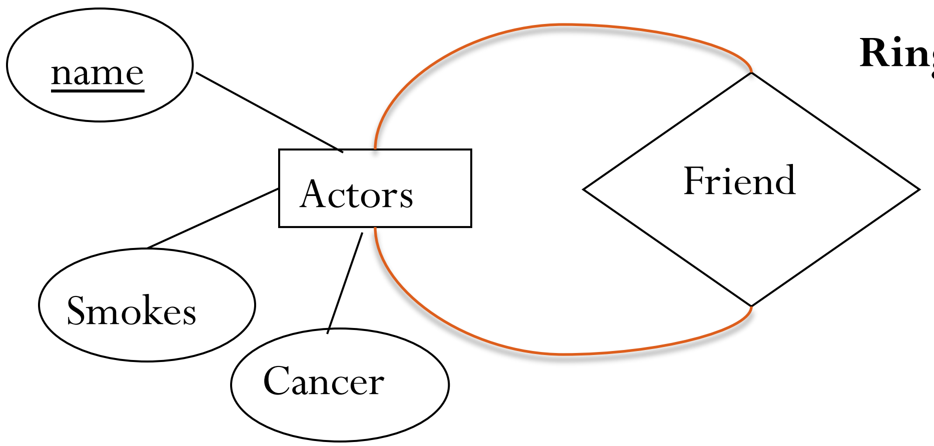
- Visualizing Domain Ontology.
- Active Area of Research.
  - Unified Modelling Language (UML).
  - Semantic Web (XML).
- Classic Tool: The Entity-Relationship (ER) Diagram.

ER Diagram Example

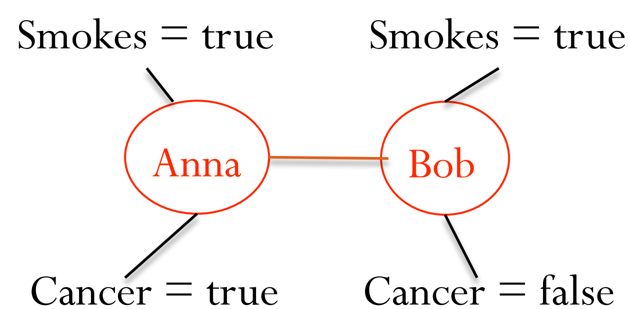


ER Model for Social Network

Ring Diagram



Social Network



Data Tables

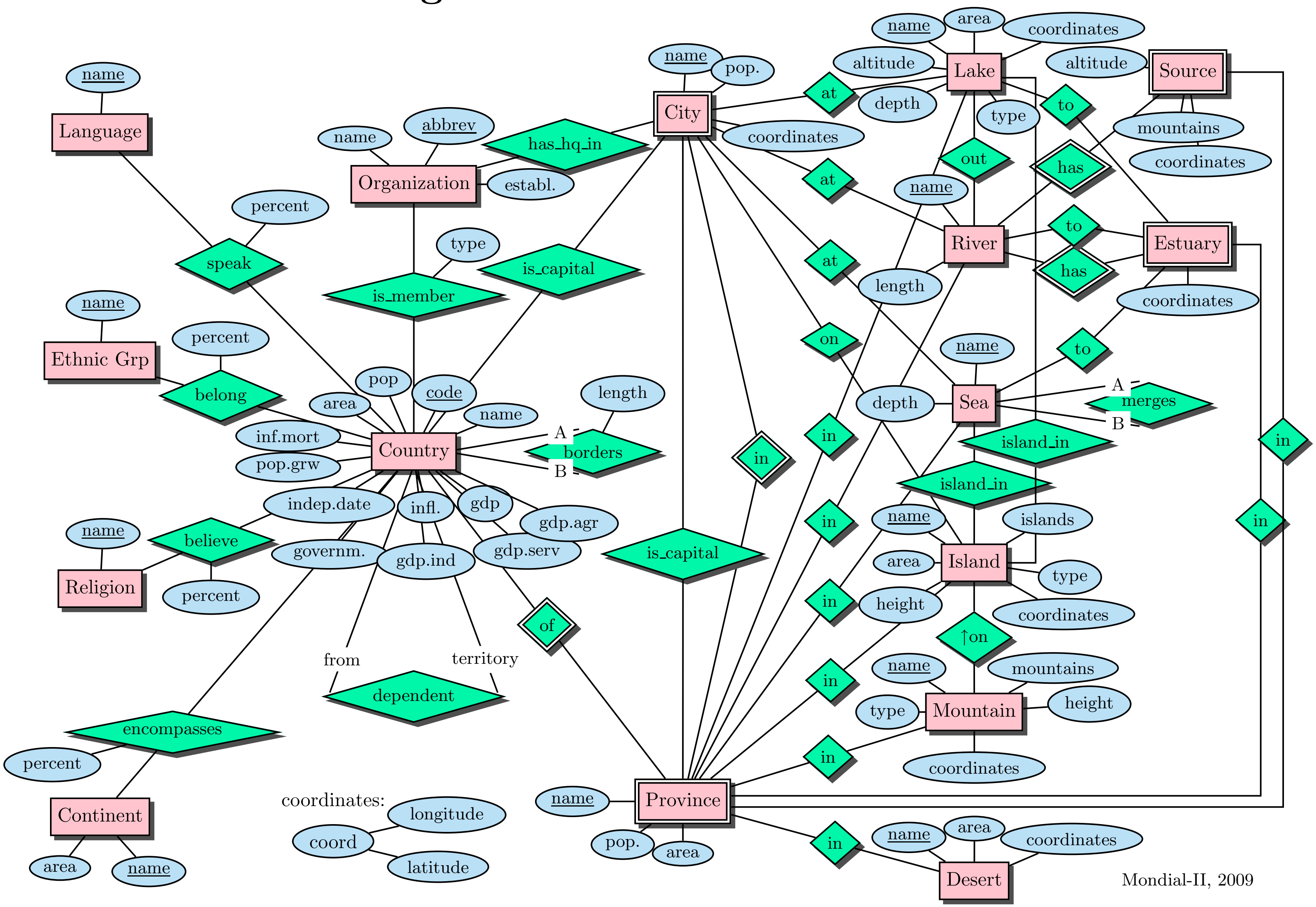
Actors

Name	Smokes	Cancer
Anna	T	T
Bob	T	F

Friend

Name1	Name2
Anna	Bob
Bob	Anna

ER-Diagram of the Mondial Database



Relationship to Social Network Analysis

- A single-relation social network is a simple special case of a relational database.
- Converse also true if you allow:
  - Different types of nodes ("actors").
  - Labels on nodes.
  - Different types of (hyper)edges.
  - Labels on edges.
    - See Newman (2003) SIAM Review.
- **Observation** A relational database is equivalent to a general network as described.

Outline

- ☑ Brief intro to relational databases.
- *Statistics and Relational Databases.*
- Briefer intro to Bayes nets.
- Relational Random Variables.
- Relational (pseudo)-likelihoods.

Beyond storing and retrieving data

- Much new interest in analyzing databases.
  - Data Mining.
  - Data Warehousing.
  - Business Intelligence.
  - Predictive Analytics.

• Fundamental Question: how to combine logic and probability?
• Domingos (U of W, CS): "Logic handles complexity, probability represents uncertainty."

Typical Tasks for Statistical-Relational Learning (SRL)

- **Link-based Classification**: given the links of a target entity and the attributes of related entities, predict the class label of the target entity.
- **Link Prediction**: given the attributes of entities and their other links, predict the existence of a link.

Link-based Classification

- Predict Attributes given Links, other Attributes
- E.g., P(diff(101))?

Student

s-id	Intelligence	Ranking
Jack	3	1
Kim	2	1
Paul	1	

Course

c-id	Rating	Difficulty
101	3	???

Professor

p-id	Popularity	Teaching-a
Oliver	3	1
		1

RA

s-id	p-id	Salary	Capability
Jack	Oliver	High	3
Kim	Oliver	Low	1
Paul	Jim	Med	2

Registration

s-id	c.id	Grade	Satisfaction
Jack	101	A	1
Jack	102	B	2
Kim	102	A	1
Paul	101	B	1

Link prediction

- Predict links given links, attributes.
- E.g.,P(Registered(jack,101))?

Student

s-id	Intelligence	Ranking
Jack	3	1
Kim	2	1
Paul	1	[illegible]

Course

c-id	Rating	Difficulty
101	3	1
[illegible]	[illegible]	[illegible]

Professor

p-id	Popularity	Teaching-a
Oliver	3	1
[illegible]	[illegible]	1

RA

s-id	p-id	Salary	Capability
Jack	Oliver	High	3
Kim	Oliver	Low	1
Paul	Jim	Med	2

Registration

s-id	c.id	Grade	Satisfaction
Jack	101	A	1
Jack	102	B	2
Kim	102	A	1
Paul	101	B	1

Generative Models

- Model the joint distribution over links and attributes.
- Today's Topic.
- We'll use Bayes nets as the model class.

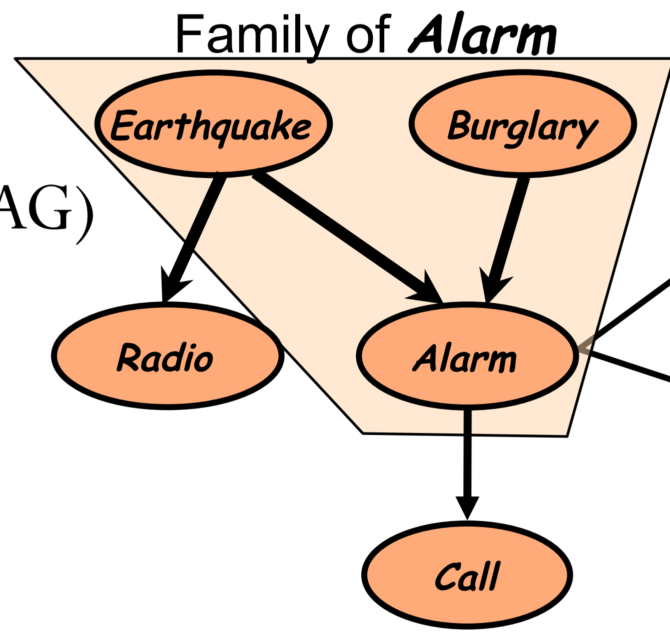
What is a Bayes (belief) net?

Compact representation of joint probability distributions via conditional independence

Qualitative part:

Directed acyclic graph (DAG)

- Nodes - random vars.
- Edges - direct influence



E	B	$P(A \mid E,B)$	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

Together:
Define a unique distribution in a factored form

Quantitative part:
Set of conditional probability distributions

$$P(B,E,A,C,R) = P(B)P(E)P(A \mid B,E)P(R \mid E)P(C \mid A)$$

Figure from N. Friedman

Why are Bayes nets useful?

- Graph structure supports
  - Modular representation of knowledge
  - Local, distributed algorithms for inference and learning
  - Intuitive (possibly causal) interpretation
- A solution to the relevance problem: Easy to compute "Is X relevant to Y given Z".
- Nice UBC Demo .

Outline

- ☑ Brief intro to relational databases.
- ☑ Statistics and Relational Databases.
- ☑ Briefer intro to Bayes nets.
- *Relational Random Variables.*
- Relational (pseudo)-likelihoods.

Relational Data: what are the random variables?

- Intuitively, the attributes and relationships in the database.
  - i.e., the columns plus link existence.
  - i.e., the components of the ER diagrams.
- Proposal from David Poole (CS UBC): apply the concept of **functors** from Logic Programming.
- I'm combining this with Halpern (CS Cornell) and Bacchus' (CS U of T) random selection probabilistic semantics for logic.

Population Variables

Russell: "A good notation thinks for us".

- Consider a model with multiple populations.
- Let X_1, X_2, Y_1, Y_2, .. be **population variables.**
- Each variable represents a random draw from a population.
- Population variables are jointly independent.
- A **functor** f is a function of one or more population variables.
- A **functor random variable** is written as $f_1(X)$ or $f_2(X,Y)$ or $f_3(X,Y,Z)$.

Unary Functors = Descriptive Attributes of Entities

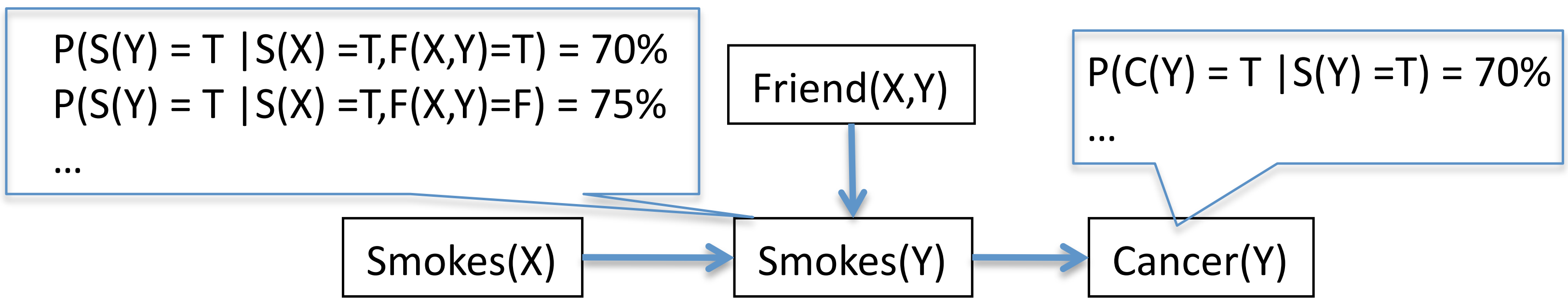
- Population of Students, Professors.
- Population variables S,P.
- Attributes r.v.s $age(S)$, $gpa(S)$, $age(P)$, $rank(P)$.
- Can have several selections $age(S_1), age(S_2)$.
- If S is uniform over students in the database:
  - P(gpa(S)=3.0) = **empirical or database frequency** of 3.0 gpa in student population.
- Can instantiate or *ground* functors with constants.
  - E.g., gpa(jack) returns the gpa of Jack.

Binary Functors = Relationships

- *Registered(S,C)*: indicator function of existence of relationship.
- If *S,C* uniformly distributed over observed population:
  - *P(Registered(S,C)=1)* = #(*s,c*) s.t. Student *s* is registered in course *c*/ #Students *x* #Courses.
    = Database Frequency of Registration.
- Can also form chains:
  P(grade(S,C)=A, Teaches(C,P)=1).

Functor Bayes Nets

- Poole IJCAI 2003: A **functor Bayes Net** is a Bayes net whose nodes are functor random variables.



Likelihood Functions for Functor Bayes Nets: Latent Variables

- Problem: Given a database D and an FBN model B, how to define $P(D|B)$?
- Fundamental Issue: interdependent units, not iid.
- One approach: introduce *latent variables* such that units are independent conditional on hidden "state" (e.g., Kersting et al. IJCAI 2009).
  - Cf. social network analysis Hoff, Rafferty (U of W Stats), Linkletter SFU Stats.
  - Cf. nonnegative matrix factorization----Netflix challenge.

Likelihood Function for Single-Table Data

- For single table T:

$$\ln[P(T \mid B)] = L(T \mid B) = \sum_{\text{nodes } i} \sum_{\text{values } a} \sum_{\text{parent-state } j} n_T(a,j) \times \ln(P_B(a \mid j))$$

Table count of co-occurrences of child node value and parent state

Parameter of Bayes net

Smokes(Y) → Cancer(Y)

Actors

Name	Smokes	Cancer
Anna	T	T
Bob	T	F

Proposed Pseudo Log-Likelihood

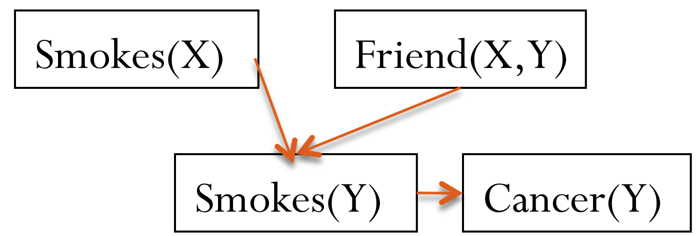
- For database D:

$$\ln[P(T \mid B)] = L(T \mid B) =$$

$$\sum_{\text{nodes } i} \sum_{\text{values } a} \sum_{\text{parent-state } j} p_D(a,j) \times \ln(P_B(a \mid j))$$

Database joint frequency of child node value and parent state

Parameter of Bayes net



Actors

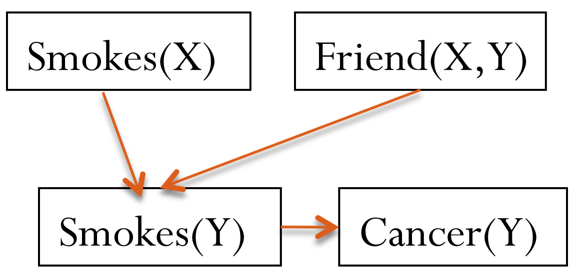
Name	Smokes	Cancer
Anna	T	T
Bob	T	F

Friend

Name1	Name2
Anna	Bob
Bob	Anna

Random Selection Log-Likelihood

1. Randomly select instances $X_1 = x_1, \ldots, X_n = x_n$. for each variable in FBN.
2. Look up their properties, relationships in database.
3. Compute log-likelihood for the FBN assignment obtained from the instances.
4. L^R = expected log-likelihood over uniform random selection of instances.



	Hyperentity		Hyperfeatures						
Γ	X	Y	F(X,Y)	S(X)	C(X)	S(Y)	C(Y)	P_B^R	$ln(P_B^R)$
γ_1	Anna	Bob	T	T	T	T	F	0.105	-2.254
γ_2	Bob	Anna	T	T	F	T	T	0.245	-1.406
γ_3	Anna	Anna	F	T	T	T	T	0.263	-1.338
γ_4	Bob	Bob	F	T	F	T	F	0.113	-2.185

$$L^R = -(2.254+1.406+1.338+2.185)/4 \approx -1.8$$

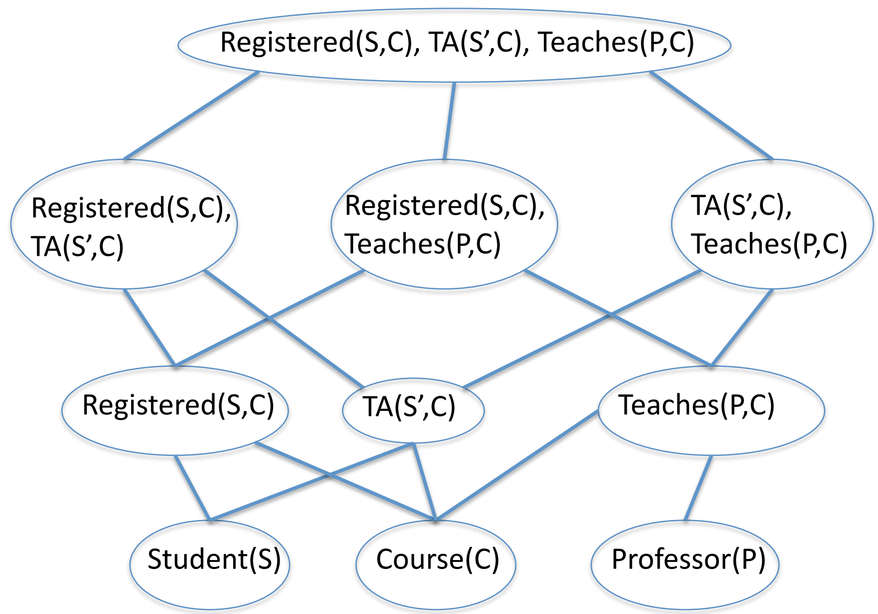
Proposition The random selection log-likelihood equals the pseudo log-likelihood.

Parameter Estimation

Proposition For a given database D, the parameter values that maximize the pseudo likelihood are the empirical conditional frequencies.

Model Selection

- New model selection algorithm (Khosravi, Schulte et al. AAAI 2010).
- Level-wise search through table join lattice.

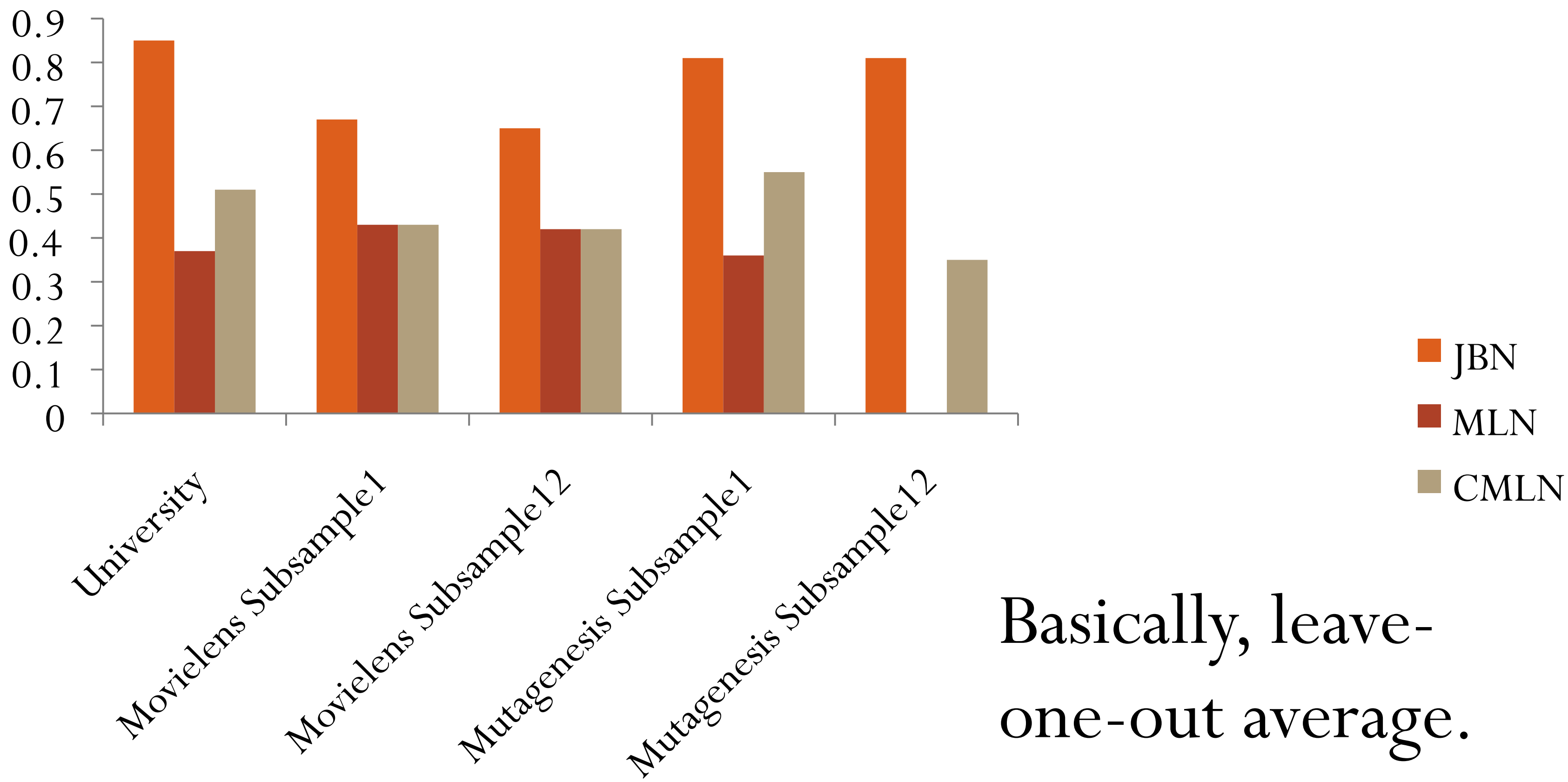


Running time on benchmarks

Dataset	JBN	MLN	CMLN
University	0.03+0.032	5.02	11.44
MovieLens	1.2+120	NT	NT
MovieLens Subsample 1	0.05 + 0.33	44	121.5
MovieLens Subsample 2	0.12 + 5.10	2760	1286
Mutagenesis	0.5 +NT	NT	NT
Mutagenesis subsample 1	0.1 + 5	3360	900
Mutagenesis subsample 2	0.2 +12	NT	3120

- Time in Minutes. NT = did not terminate.
- x + y = structure learning + parametrization (with Markov net methods).
- JBN: Our join-based algorithm.
- MLN, CMLN: standard programs from the U of Washington (Alchemy)

Accuracy



Basically, leave-one-out average.

Future Work: Inference

Prediction is usually based on *knowledge-based model construction* (Ngo and Haddaway, 1997; Koller and Pfeffer, 1997; Haddaway, 1999).

- Basic Idea: instantiate population variables with all population members. Predict using instantiated model.
- With Bayes nets, can lead to cycles.
- My conjecture: cycles can be handled with a normalization constant that has a closed form.
- Help?!

Summary: Likelihood for relational data.

- Combining relational databases and statistics.
  - Very important in practice.
  - Combine logic and probability.
- Interdependent units → hard to define model likelihood.
- Proposal: Consider a randomly selected small group of individuals.
- Pseudo log-likelihood = expected log-likelihood of randomly selected group.

Summary: Statistics with Pseudo-Likelihood

- Theorem: Random pseudo log-likelihood equivalent to standard single-table likelihood, replacing table counts with database frequencies.
- Maximum likelihood estimates = database frequencies.
- Efficient Model Selection Algorithm based on lattice search.
- In simulations, very fast (minutes vs. days), much better predictive accuracy.

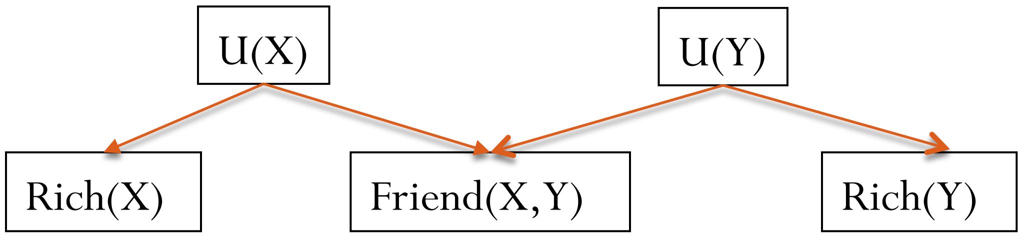
Thank you!

- Any questions?

Choice of Functors

- Can have complex functors, e.g.
  - Nested: *wealth(father(father(X)))*.
  - Aggregate: $AVG_C\{grade(S,C): Registered(S,C)\}$.
- In remainder of this talk, use functors corresponding to
  - Attributes (columns), e.g., *intelligence(S), grade(S,C)*
  - Boolean Relationship indicators, e.g. *Friend(X,Y)*.

Hidden Variables Avoid Cycles



- Assign unobserved values *u(jack), u(jane)*.
- Probability that Jack and Jane are friends depends on their unobserved "type".
- In ground model, *rich(jack)* and *rich(jane)* are correlated given that they are friends, but neither is an ancestor.
- Common in social network analysis (Hoff 2001, Hoff and Rafferty 2003, Fienberg 2009).
- $1M prize in Netflix challenge.
- Also for multiple types of relationships (Kersting et al. 2009).
- Computationally demanding.

Typical Tasks for Statistical-Relational Learning (SRL)

- **Link-based Classification**: given the links of a target entity and the attributes of related entities, predict the class label of the target entity.
- **Link Prediction**: given the attributes of entities and their other links, predict the existence of a link.