Learning Bayesian Networks for Relational Databases

Oliver Schulte
School of Computing Science
Simon Fraser University
Vancouver, Canada

Outline

- Review of relational databases.
- Example Bayesian networks.
- Relational classification with Bayesian networks.
- Fundamental Learning Challenges.
  - Defining model selection scores.
  - Computing sufficient statistics.
- Work in Progress.
  - Anomaly Detection.
  - Homophily vs. social influence.
  - Player contribution to team result.

new causal questions

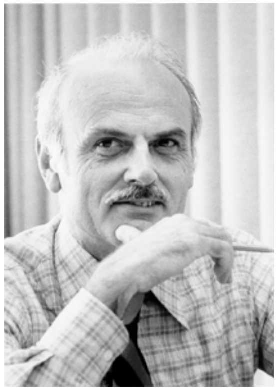
Relational Databases

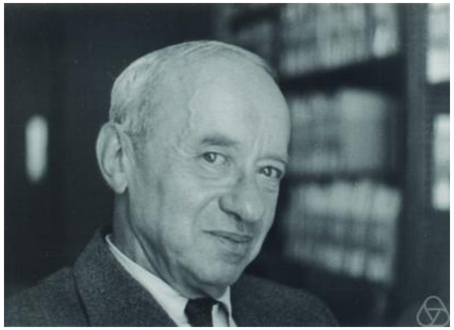
Relational Databases

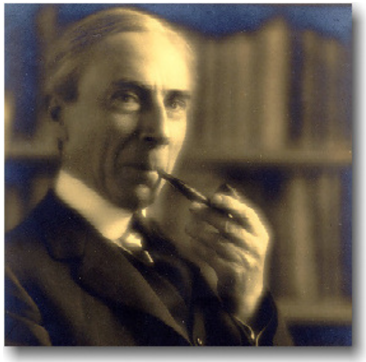
- 1970s: Computers are spreading. Many organizations use them to store their data.
- Ad hoc formats
  - ⇨ hard to build general data management systems.
  - ⇨ lots of duplicated effort.
- The Standardization Dilemma:
  - Too restrictive: doesn't fit users' needs.
  - Too loose: back to ad-hoc solutions.

The Relational Format

- Codd (IBM Research 1970)
- The fundamental question: *What kinds of information do users need to represent?*
- Answered by first-order predicate logic! (Russell, Tarski).
- The world consists of
  - Individuals/entities.
  - Relationships/links among them.







Tabular Representation

A database is a finite model for an *arbitrary* first-order logic vocabulary.

Students S		
<u>Name</u>	intelligence(S)	ranking(S)
Jack	3	1
Kim	2	1
Paul	1	2

Professor P		
<u>Name</u>	popularity(P)	teaching Ability(P)
Oliver	3	1
David	2	1

Registration(S,C)			
<u>Name</u>	<u>Number</u>	grade	satisfaction
Jack	101	A	1
Jack	102	B	2
Kim	102	A	1
Kim	103	A	1
Paul	101	B	1
Paul	102	C	2

Course C			
<u>Number</u>	Prof(C)	rating(C)	difficulty(C)
101	Oliver	3	1
102	David	2	2
103	Oliver	3	2

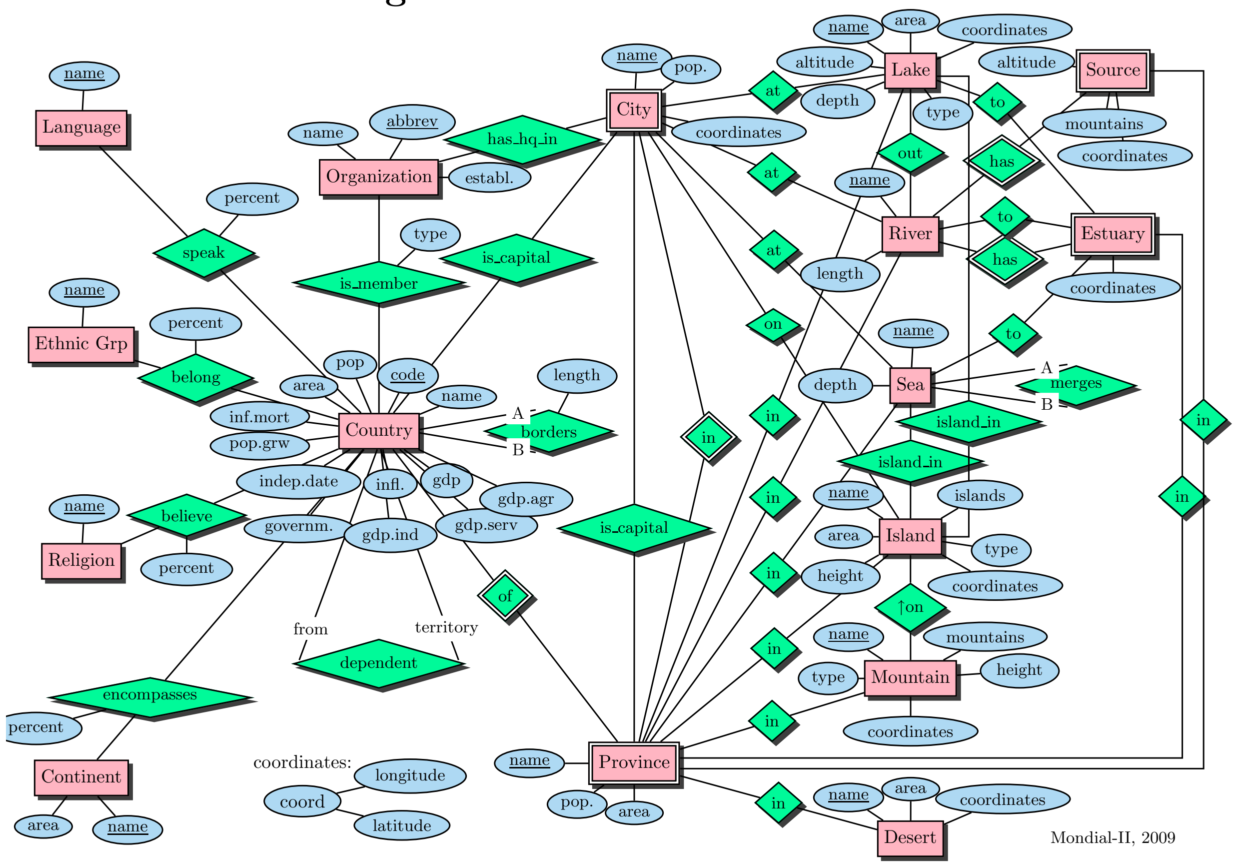
<u>Key fields</u> are underlined.

Nonkey fields are deterministic **functions of key fields**.

Ullman, J. D. (1982), Principles of Database Systems

Data Format Is Complex

ER-Diagram of the Mondial Database



Database Management Systems

- Maintain data in linked tables.
- Structured Query Language (SQL) allows fast *data retrieval*.
  - E.g., find all CMU students who are statistics majors with gpa > 3.0.
- Multi-billion dollar industry, $15+ bill in 2006.
- IBM, Microsoft, Oracle, SAP.
- Much interest in analysis (data mining, business intelligence, predictive analytics, OLAP…)

Relationship to Single Data Table

- Single data table = finite model for *monadic* first-order predicates.
- Single population.

Students S		
<u>Name</u>	**intelligence(S)**	**ranking(S)**
Jack	3	1
Kim	2	1
Paul	1	2

Jack

3 1

Kim

3 2

Paul

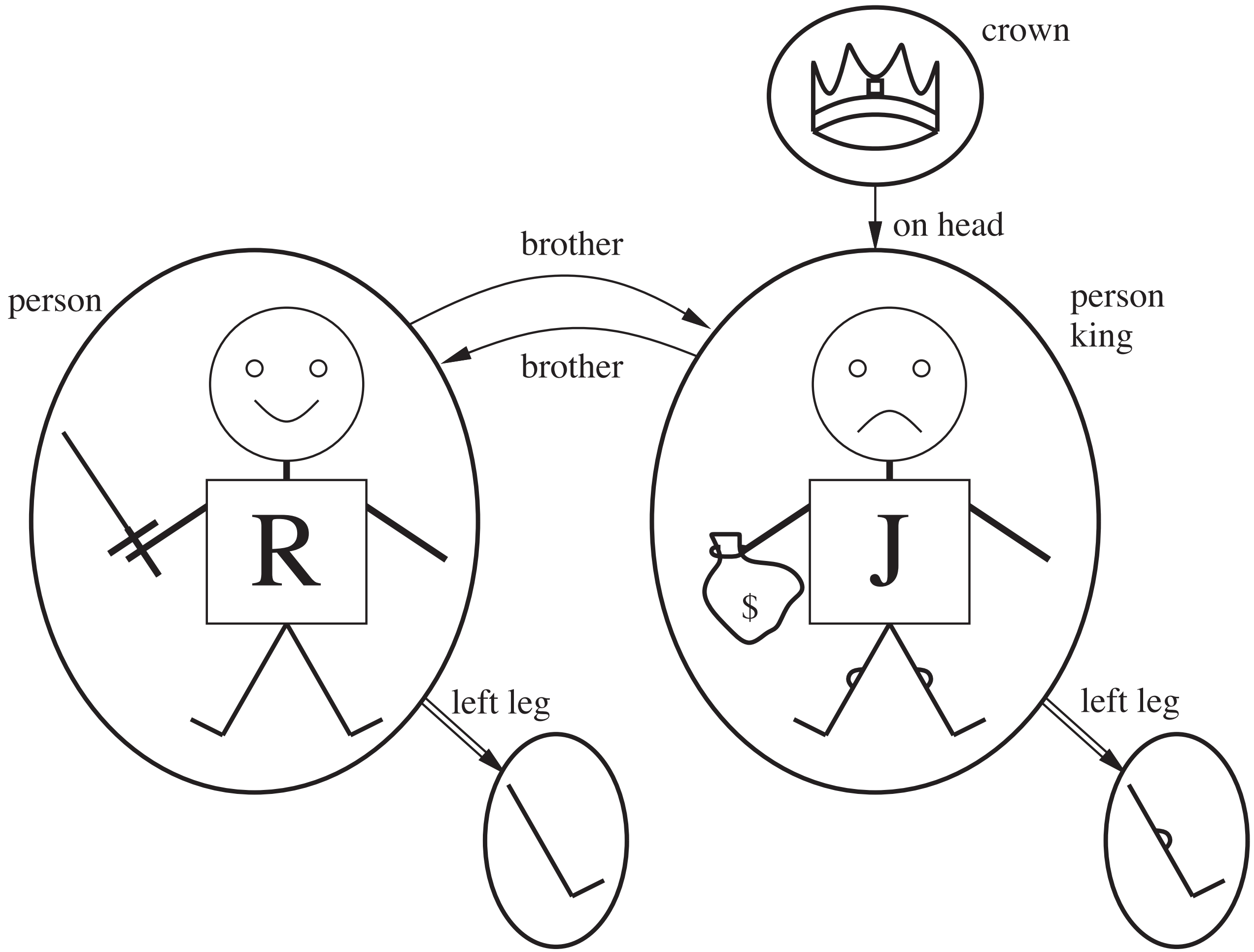
1 2

Relationship to Network Analysis

- A single-relation social network = finite model for single binary predicate (“Friend(X,Y)”).
- General network allows:
  - Different types of nodes (“actors”).
  - Labels on nodes.
  - Different types of (hyper)edges.
  - Labels on edges.
    - See Newman (2003).
- **Observation** A relational database is equivalent to a general network as described.

Newman, M. E. J. 2003. The structure and function of complex networks. SIAM Review 45, 167-256.

Example: First-order model as a network

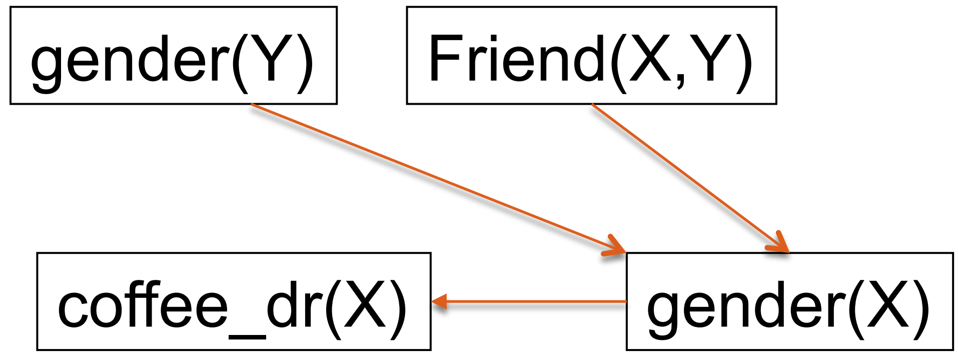


Russell and Norvig, "A Modern Introduction to Artificial Intelligence", Fig. 8.2.

Bayesian Networks for Relational Databases

Russell and Norvig, "Artificial Intelligence", Ch.14.6, 3rd ed.
D.Heckerman, Chris Meek & Koller, D. (2004), 'Probabilistic models for relational data', Technical report, Microsoft Research.
Poole, D. (2003), First-order probabilistic inference, *IJCAI, pp. 985-991.*

Random Selection Semantics for Bayes Nets



$P(gender(X) = male, gender(Y) = male, Friend(X, Y) = true, coffee\_dr(X) = true) = 30\%$

means

"if we randomly select a user X and a user Y, the probability that both are male and that X drinks coffee is 30%.

Bayesian Network Examples

- Mondial Network
- University Network

Random Selection Semantics for Random Variables

Population

Population variables

Parametrized Random Variables

People

Bob

Anna

X
Random Selection from People.
P(X = Anna) = ½.

Y
Random Selection from People.
P(Y = Anna) = ½.

Gender(X)
Gender of selected person.
P(Gender(X) = W) = ½.

Friend(X,Y) =
T if selected people are friends, F otherwise
P(Friend(X,Y) = T) = ½.

Gender(Y)
Gender of selected person.
P(Gender(Y) = W) = ½.

Halpern, "An analysis of first-order logics of probability", AI Journal 1990.
Bacchus, "Representing and reasoning with probabilistic knowledge", MIT Press 1990.

Inference: Relational Classification

Independent Individuals and Direct Inference

P(rank(S)=hi| Intelligence(S) = hi) = 70%

Intelligence(S) → rank(S)

- Query: What is P(rank(bob) = hi | intelligence(bob) = hi)?
- Answer: 70%.

intelligence = hi.
rank = ?

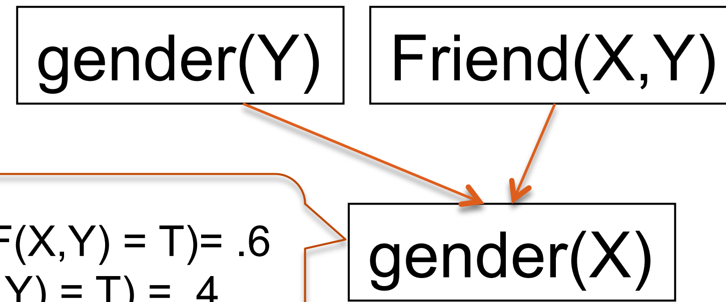
The *direct inference principle*

$P(\phi(X) = p) \rightarrow P(\phi(a)) = p$

where ϕ is a first-order formula with free variable X,
a is a constant.

Halpern, "An analysis of first-order logics of probability", AI Journal 1990.

Direct Inference is insufficient for related individuals

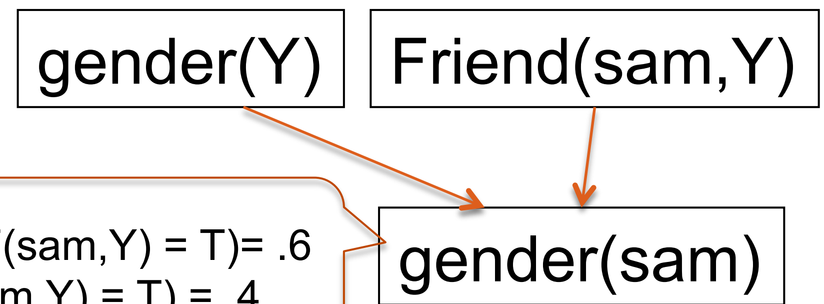


P(gender(X) = Woman |gender(Y) = Woman, F(X,Y) = T)= .6
P(gender(X) = Woman| gender(Y) = Man, F(X,Y) = T) = .4

- Suppose that Sam has friends Alice, John, Kim, Bob,…
- Direct inference specifies
  P(gender(sam) = Man | gender(alice) = Woman) = .6
  but not
  P(gender(sam) = Man | gender(alice), gender(john), gender(kim), gender(bob)….).

Random Selection Classification

- Basic idea: log-conditional probability ➔ **expected** log-conditional probability wrt random instantiation of free first-order variables.
- Good predictive accuracy (Schulte et al. 2012, Schulte et al. 2014).



P(gender(sam) = Woman | gender(Y) = Woman, F(sam,Y) = T)= .6
P(gender(sam) = Woman | gender(Y) = Man, F(sam,Y) = T) = .4

gender(Y)	ln(CP)	proportion	product
female	ln(0.6) = -0.51	40%	-0.51x0.4=-0.204
male	ln(0.4) = -0.92	60%	-0.92x0.6 =-0.552
score	gender(sam) = Woman		-0.204-0.552 = -0.756
score	gender(sam) = Man		=-0.67

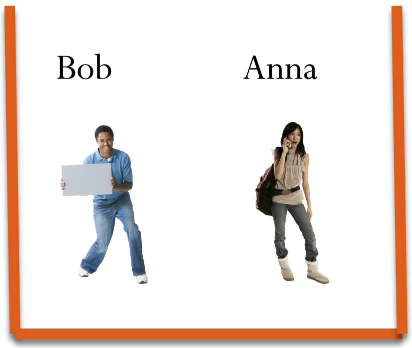
Defining Joint Probabilities

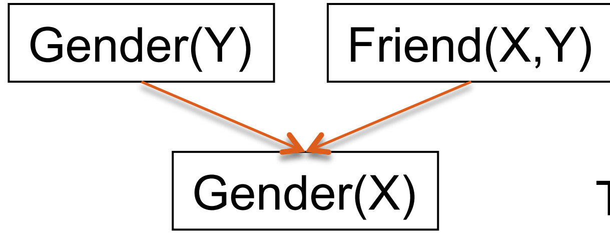
- Knowledge-based Model Construction: Instantiate graph with first-order nodes to obtain graph with instance nodes.
- Fundamental problem: DAGs are not closed under instantiation.
- Alternative: **relational dependency networks**.

Wellman, M.; Breese, J. & Goldman, R. (1992), 'From knowledge bases to decision models', *Knowledge Engineering Review* 7, 35--53.
Neville, J. & Jensen, D. (2007), 'Relational Dependency Networks', *Journal of Machine Learning Research* 8, 653--692.
Heckerman, D.; Chickering, D. M.; Meek, C.; Rounthwaite, R.; Kadie, C. & Kaelbling, P. (2000), 'Dependency Networks for Inference, Collaborative Filtering, and Data Visualization', *Journal of Machine Learning Research* 1, 49—75.

The Cyclicity Problem

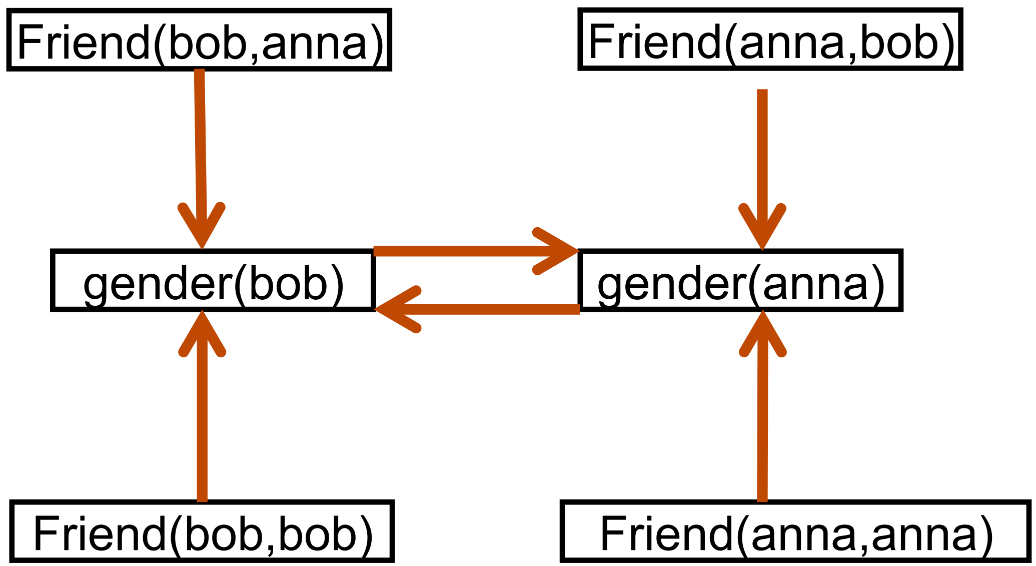
People





Template Bayesian Network

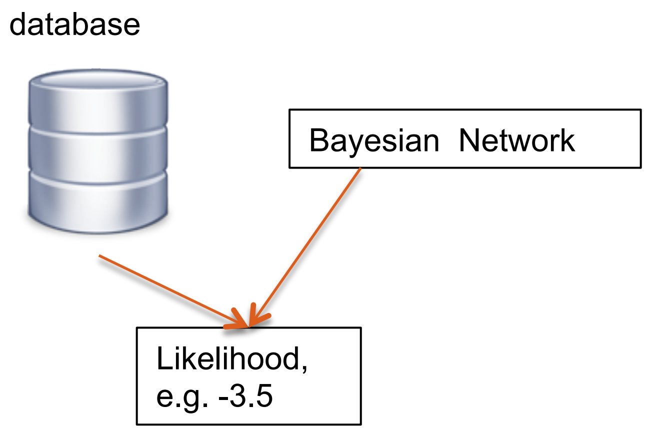
Grounding: Instantiate population variables with constants



Instantiated Inference Graph

Likelihood-Based Learning

Wanted: a likelihood function



Problems

- Multiple Tables.
- Dependent data points
- ➢ Products are not normalized
- ➢ Pseudo-likelihood

Users

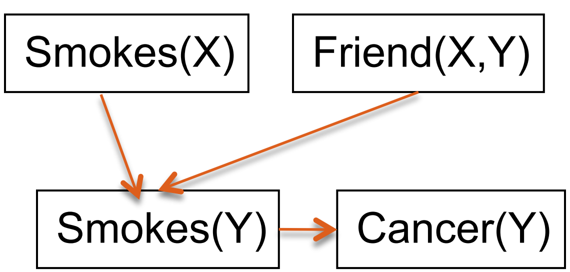
Name	Smokes	Cancer
Anna	T	T
Bob	T	F

Friend

Name1	Name2
Anna	Bob
Bob	Anna

The Random Selection Log-Likelihood

1. Randomly select instances $X_1 = x_1, \ldots, X_n = x_n$ for each first-order variable in BN.
2. Look up their properties, relationships in database.
3. Compute log-likelihood for the BN 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)	S(Y)	C(Y)	P_B^γ	$ln(P_B^\gamma)$
γ_1	Anna	Bob	T	T	T	F	0.105	-2.254
γ_2	Bob	Anna	T	T	T	T	0.245	-1.406
γ_3	Anna	Anna	F	T	T	T	0.263	-1.338
γ_4	Bob	Bob	F	T	T	F	0.113	-2.185

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

Schulte, O. (2011), A tractable pseudo-likelihood function for Bayes Nets applied to relational data, *in 'SIAM SDM', pp. 462-473.*

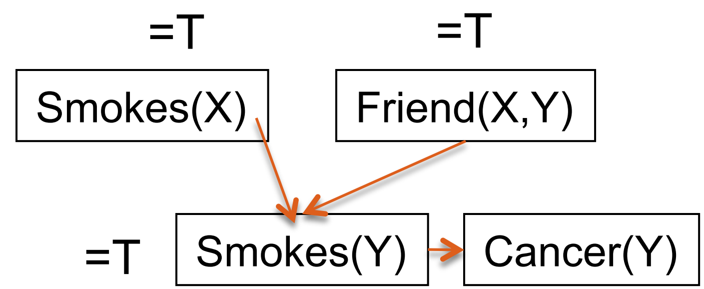
Equivalent Closed-Form

For each node, find the *expected log-conditional probability*, then sum.

$$\ln P^*(D|B) = \sum_{\text{nodes } i} \sum_{\text{values } k} \sum_{\text{parent-states } j} P_D(v_i = k, pa_i = j) \ln P_B(v_i = k | pa_i = j)$$

Database D **frequency** of co-occurrences of child node value and parent state

Parameter of Bayes net



Users

Name	Smokes	Cancer
Anna	T	T
Bob	T	F

Friend

Name1	Name2
Anna	Bob
Bob	Anna

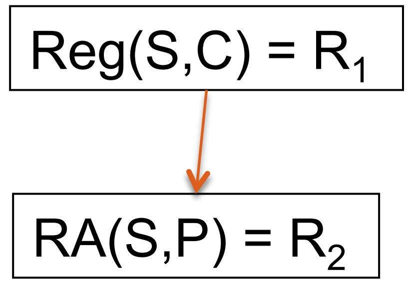
Pseudo-likelihood Maximization

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

The Bad News

- Sufficient Statistics are harder to compute than for i.i.d. data.
  - e.g. find the number of (X,Y) such that **not** *Friend(X,Y) and neither X nor Y has cancer.*
- Scoring models is computationally more expensive than generating candidate models.

The Fast Moebius Transform Finds Negated Relationship Counts



J.P. = joint probability

Initial table with no false relationships

R_1	R_2	J.P.
T	T	0.2
*	T	0.3
T	*	0.4
*	*	1

R_1	R_2	J.P.
T	T	0.2
F	T	0.1
T	*	0.4
F	*	0.6

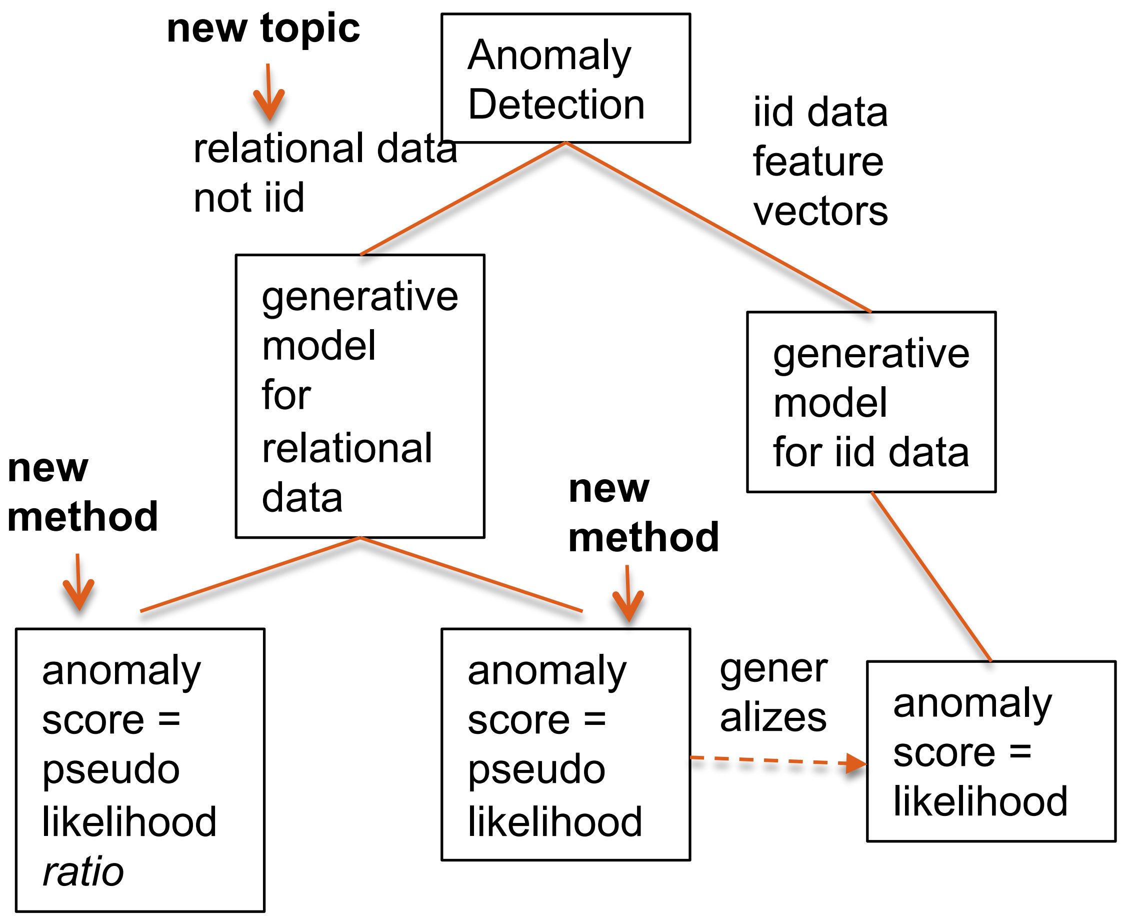
table with joint probabilities

R_1	R_2	J.P.
T	T	0.2
F	T	0.1
T	F	0.2
F	F	0.5

Kennes, R. & Smets, P. (1990), Computational aspects of the Moebius transformation, '*UAI*', *pp. 401-416.*
Schulte, O.; Khosravi, H.; Kirkpatrick, A.; Gao, T. & Zhu, Y. (2014), 'Modelling Relational Statistics With Bayes Nets', *Machine Learning 94, 105-125.*

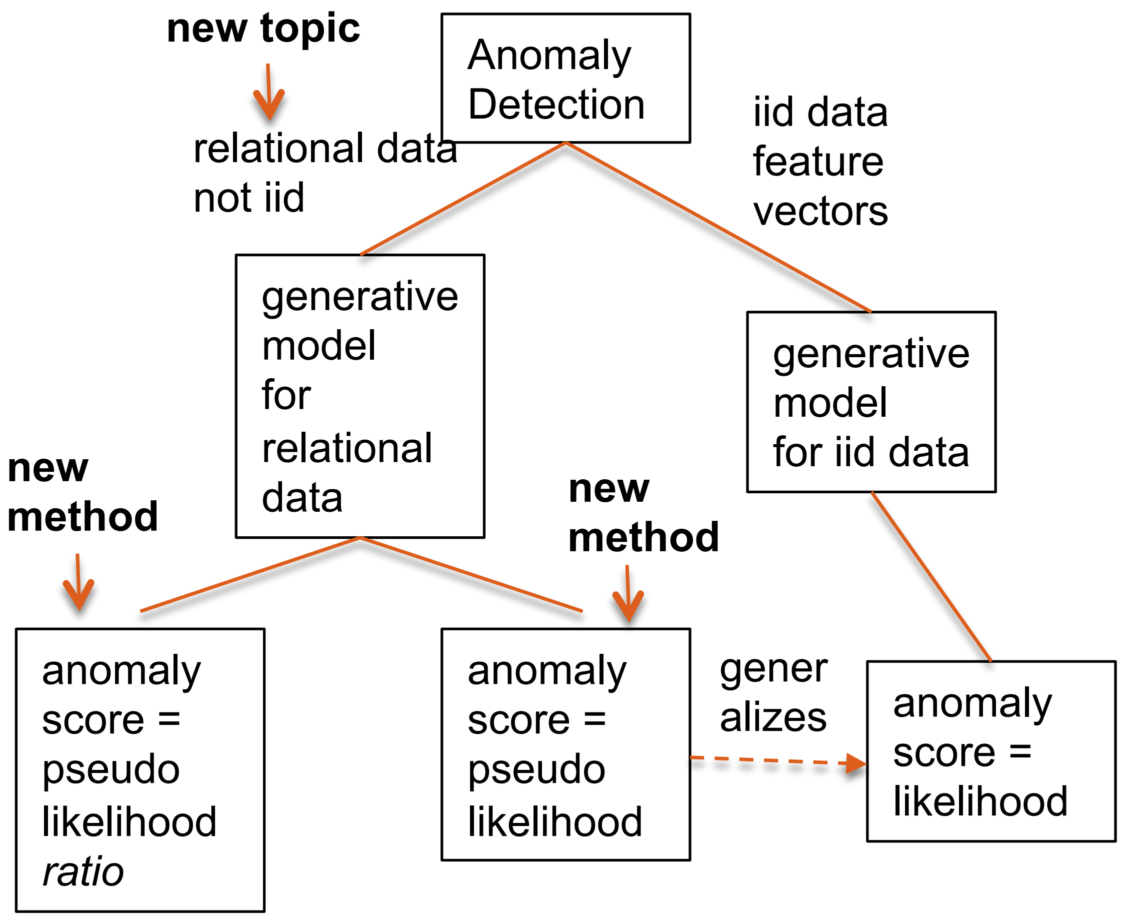
Anomaly Detection

Anomaly Detection with Generative Models



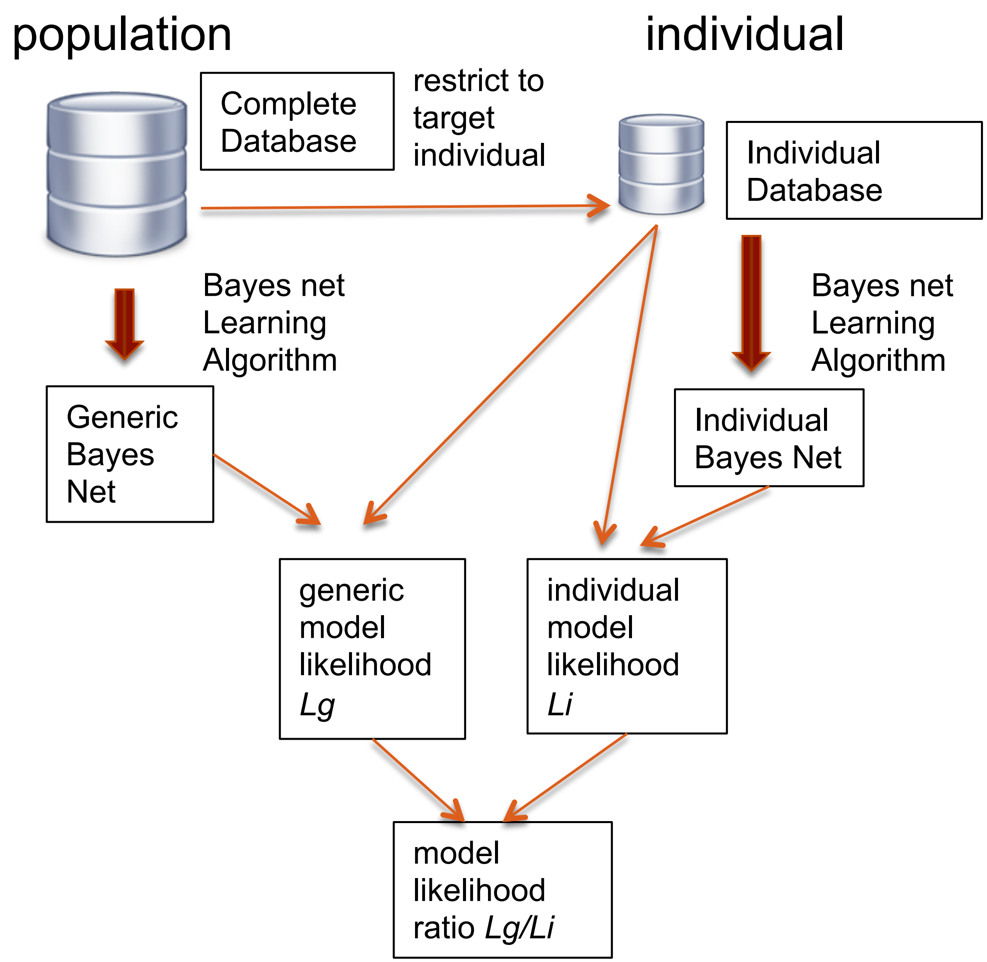
Cansado, A. & Soto, A. (2008), 'Unsupervised anomaly detection in large databases using Bayesian networks', *Applied Artifical Intelligence 22(4), 309—330.*
http://www.bayesserver.com/Techniques/AnomalyDetection.aspx

Anomaly Detection with Generative Models

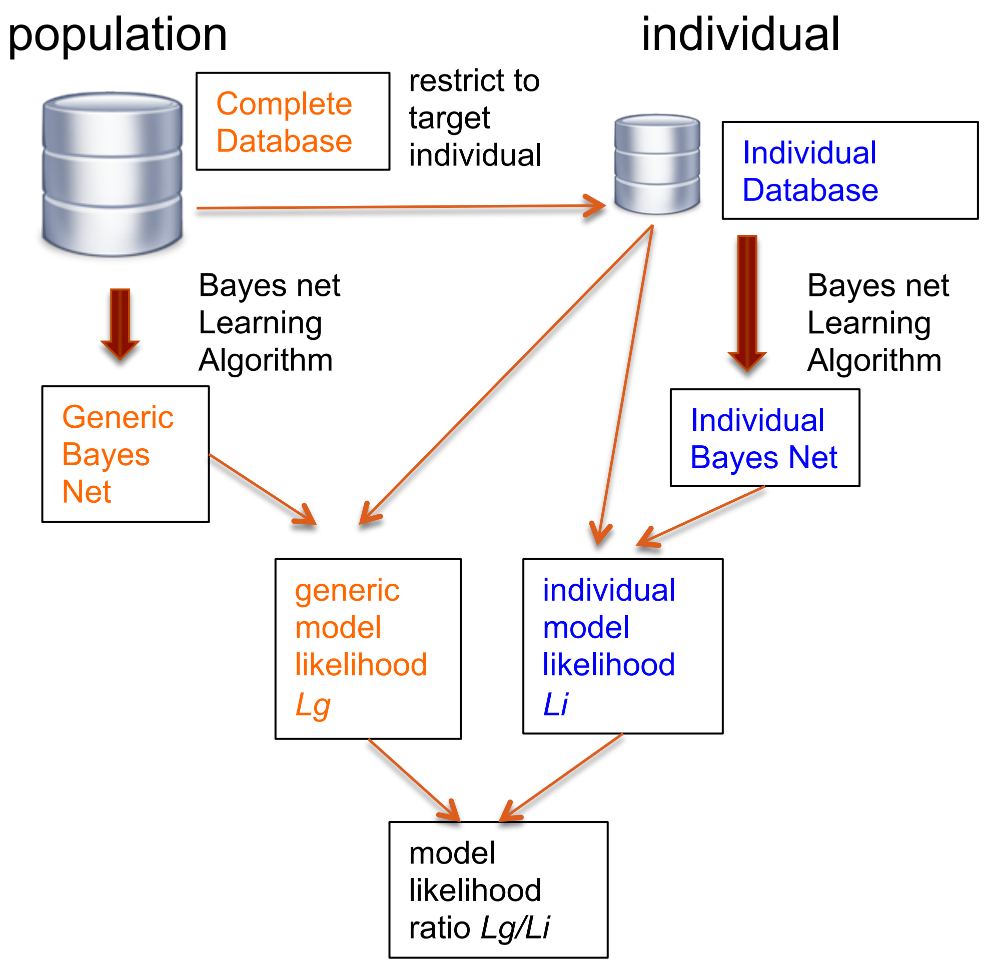


Cansado, A. & Soto, A. (2008), 'Unsupervised anomaly detection in large databases using Bayesian networks', *Applied Artifical Intelligence 22(4), 309—330.*
http://www.bayesserver.com/Techniques/AnomalyDetection.aspx

New Anomaly Measure



New Anomaly Measure



Anomaly Metric Correlates With Success

Unusual Teams have worse standing. N = 20.

	ρ (Likelihood-ratio , Standing)
Top Teams	0.62
Bottom Teams	0.41

Unusual Movies have higher ratings. N = 3060.

Genre	ρ (Likelihood-ratio , avg-rating)
Film-Noir	0.49
Action	0.42
Sci-Fi	0.35
Adventure	0.34
Drama	0.28

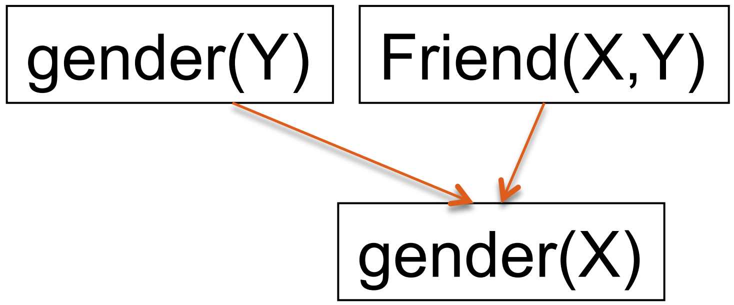
Riahi, F.; Schulte, O. & Liang, Q. (2014), 'A Proposal for Statistical Outlier Detection in Relational Structures', AAAI-StarAI Workshop on Statistical-Relational AI.

Causal Questions

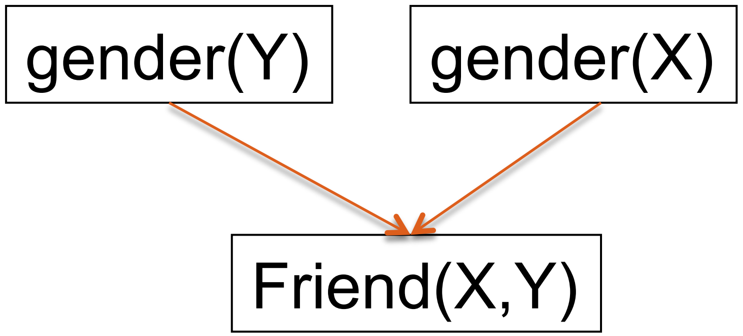
Relationships vs. Attributes

- Do relationships cause attributes? E.g., Homophily.
- Do attributes cause relationships? E.g., social influence.
- Can we tell?

Social Influence



Homophily

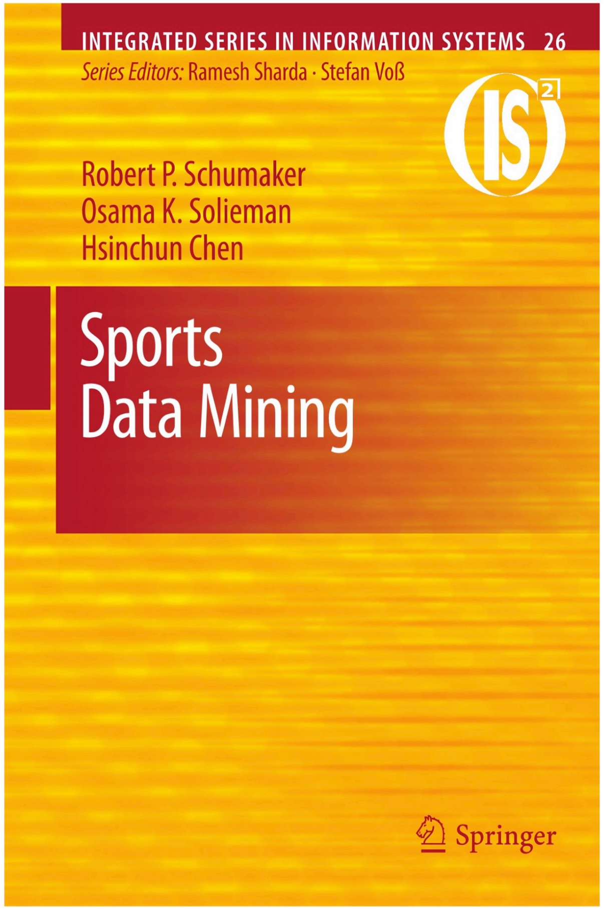


http://www.acthomas.ca/academic/acthomas.htm

Shalizi, C. R. & Thomas, A. C. (2011), 'Homophily and contagion are generically confounded in observational social network studies', *Sociological Methods & Research 40(2), 211--239.*

Individual Causal Contributions to Group Results

- Important Problem in Sports Statistics: How much did a player contribute to a match result?
- Sabermetrics.
- Actual Causation.



Pearl, J. (2000), *Causality: Models, Reasoning, and Inference, Ch. 10.*

Player-Based Approaches: Ice Hockey

- Basic question: what difference does the *presence of a player* make? Examples:
  - Logistic regression of which team scored given a presence indicator variable for each player (Grammacy et al. 2013).
  - Log-linear model of goal-scoring rate given a presence indicator variable for each player (Thomas et al. 2013).
- Major problem: distinguish players from same line.

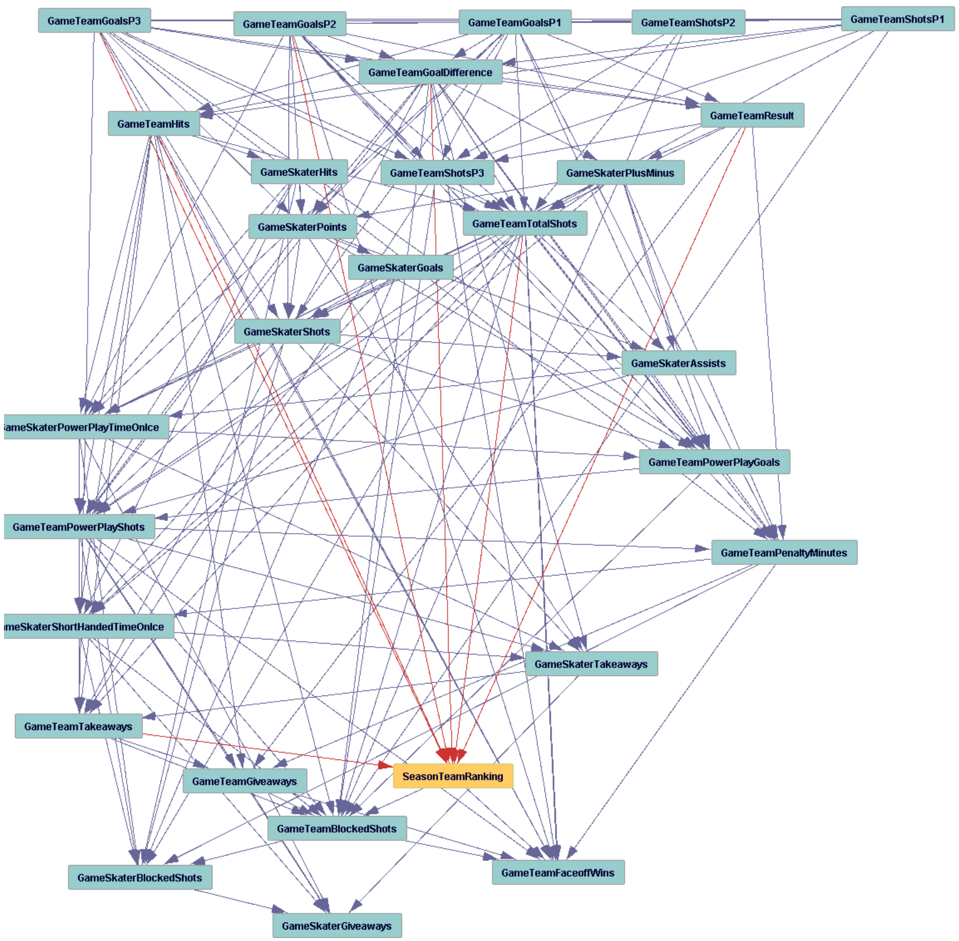
Gramacy, R.; Jensen, S. & Taddy, M. (2013), 'Estimating player contribution in hockey with regularized logistic regression.', *J ournal of Quantitative Analysis in Sports 9, 97-111.*
Thomas, A.; Ventura, S.; Jensen, S. & Ma, S. (2013), 'Competing Process Hazard Function Models for Player Ratings in Ice Hockey', *The Annals of Applied Statistics 7(3), 1497-1524.*

Action-Based Approaches

- Basic question: What difference does an *action* make?
- ➢ Model causal effect of action on goal.
- Player contribution = sum of scores of player's actions.
  - Schuckers and Curro (2013), McHall and Scarf (2005; soccer).
- Can action effect models be improved with causal graphs?
  - Model Selection.
  - Model causal chains.

Schuckers, M. & Curro, J. (2013), Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events, *in '7th Annual MIT Sloan Sports Analytics Conference'*.

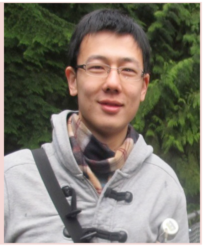
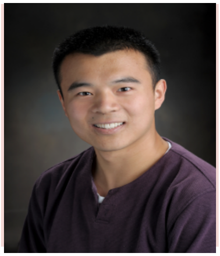
Run Tetrad on NHL data (preliminary)



Summary

- Relational data: common and complex.
- Random selection semantics for logic answers fundamental statistical questions in a principled way.
  - inference.
  - (pseudo)-likelihood function.
- Computing sufficient statistics is hard.
  - Fast Moebius transform helps.
- Anomaly detection as an application in progress.
- New Causal Questions:
  - do attributes cause relationships or vice versa?
  - how much does an individual contribute to a group result (e.g., a goal in sports).

Collaborators

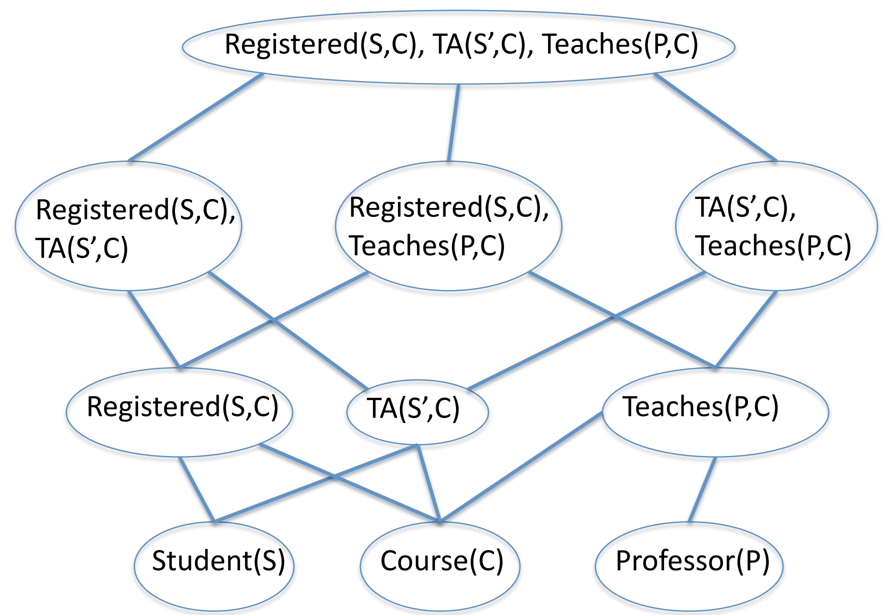
Oliver Schulte	Hassan Khosravi	Arthur Kirkpatrick	Tianxiang Gao	Yuke Zhu	Zhensong Qian	Fatemeh Riahi
						

The End

- Any questions?

Structure Learning

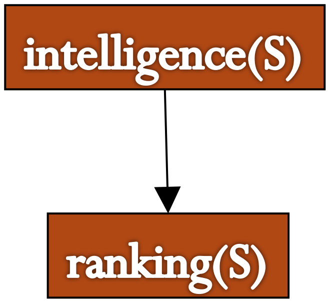
- In principle, just replace single-table likelihood by pseudo likelihood.
- Efficient new algorithm (Khosravi, Schulte et al. AAAI 2010). Key ideas:
  - Use single-table BN learner as black box **module**.
  - **Level-wise search** through table join lattice. Results from shorter paths are propagated to longer paths.



Phase 1: Entity tables

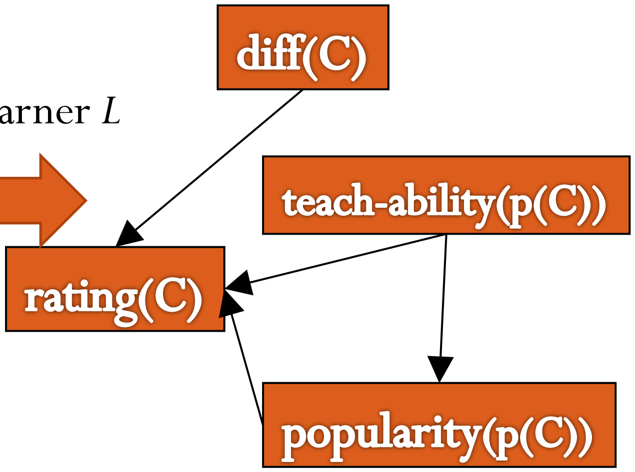
Students		
Name	intelligence	ranking
Jack	3	1
Kim	2	1
Paul	1	2

BN learner L



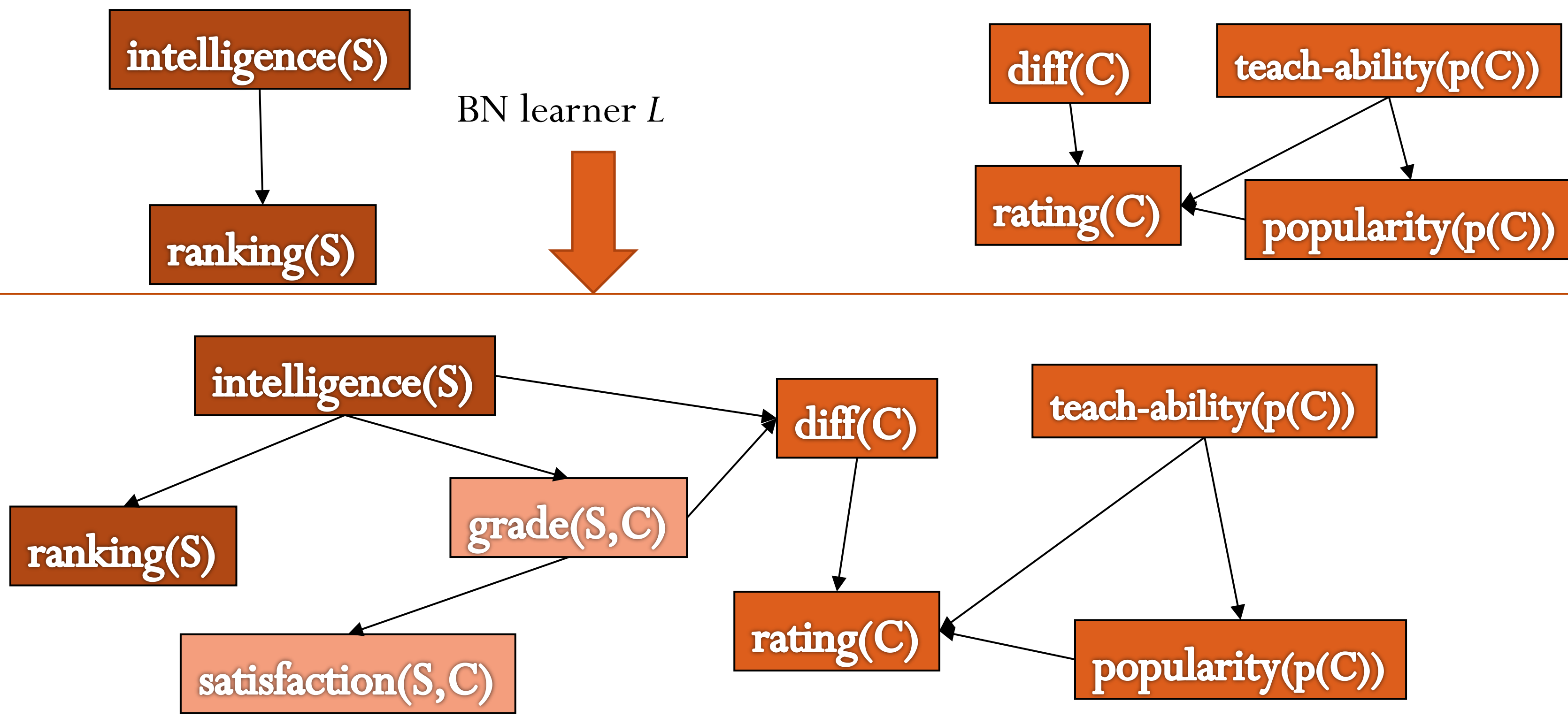
Course				
Number	rating	difficulty	Prof-popularity	Prof-teachablity
101	3	1	1	2
102	2	2	2	2
103	3	2	1	1

BN learner L



Phase 2: relationship tables

Registration				Student		Course			
S.Name	C.number	grade	satisfaction	intelligence	ranking	rating	difficulty	Popularity	Teach-ability
Jack	101	A	1	3	1	3	1	1	2
….	….	…	…	…	…	…	…	…	…



Phase 3: add Boolean relationship indicator variables