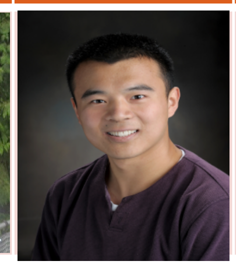
Machine Learning for Information Networks

Oliver Schulte

School of Computing Science

Simon Fraser University

Collaborators

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

Outline

- What are information networks/multi-relational data?
- Why machine learning for information networks?
- Unifying logic and statistics: learning first-order Bayesian networks
- Applications
  - Frequency Modelling/Density Estimation
  - Relational Exception Mining
- How is relational learning different from non-relational learning?

What Are Information Networks?

Representing Relational Data

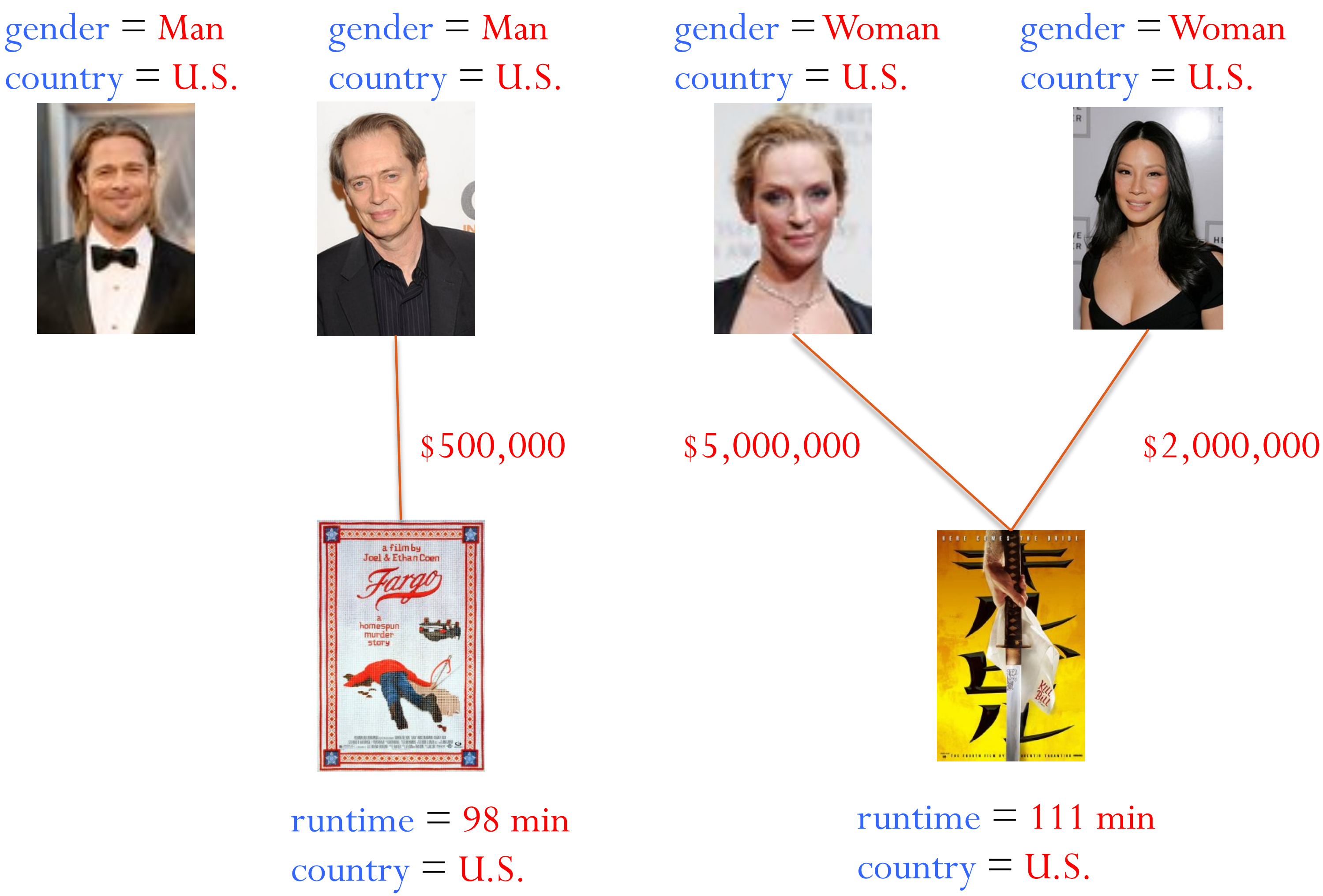
Definition

An information network (Sun and Han 2012) is a graph with

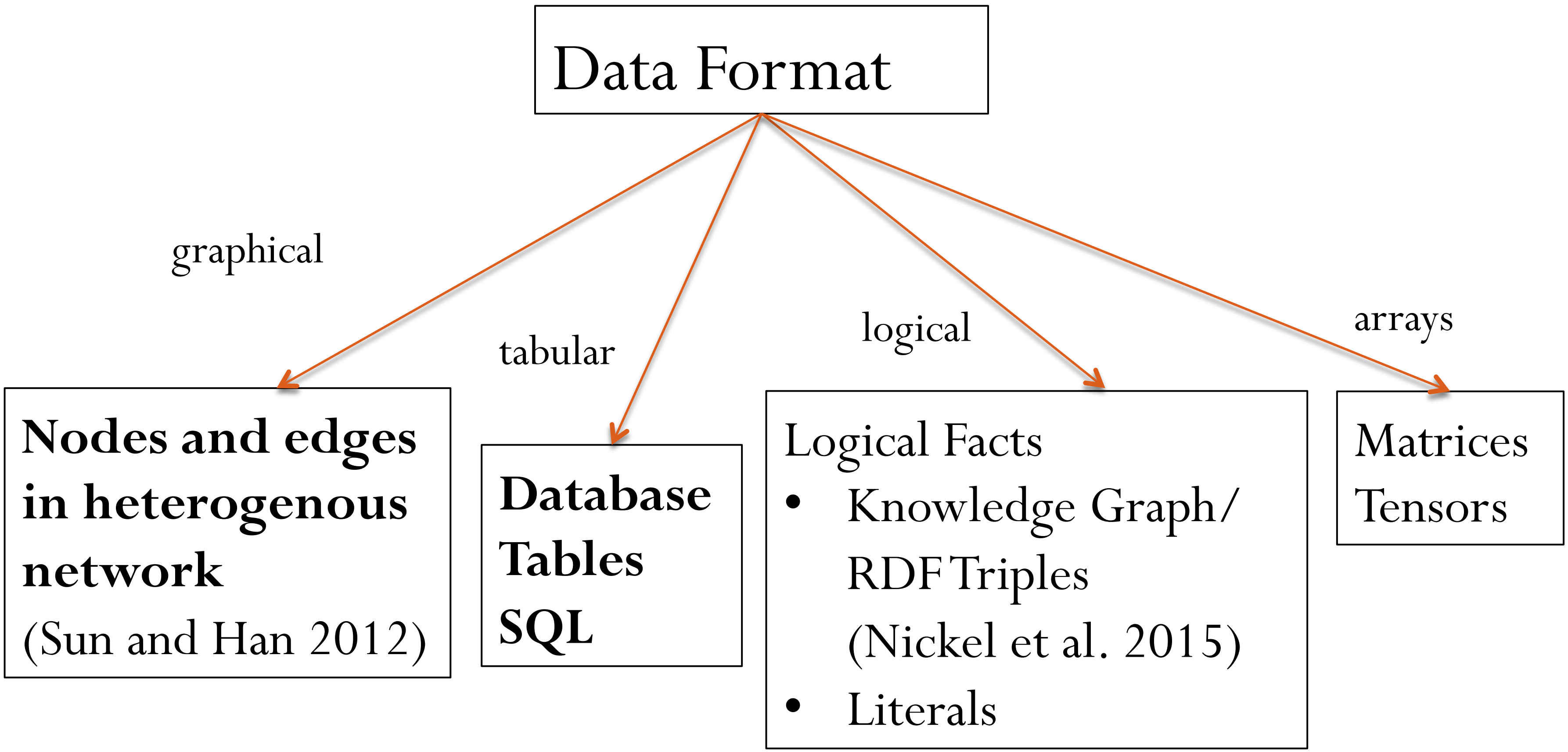
- **nodes** (aka entities)
- **edges** (aka relationships)
  - can be hyperedges
- Nodes and edges
  - can be of different types ➔ heterogeneity
  - can have attributes (aka features)

Sun, Y. & Han, J. (2012), Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers.

Toy Example


gender = Man
country = U.S.
gender = Man
country = U.S.
gender = Woman
country = U.S.
gender = Woman
country = U.S.
$500,000
$5,000,000
$2,000,000
a film by
Joel & Ethan Coen
Fargo
a
homespun
murder
story
runtime = 98 min
country = U.S.
runtime = 111 min
country = U.S.

Different Communities Use Different Formats for Information Network



Nickel, M.; Murphy, K.; Tresp, V. & Gabrilovich, E. (2016), 'A review of relational machine learning for knowledge graphs', Proceedings of the IEEE 104(1), 11--33.

Table Representation

One table for each type of entity/link

Actors

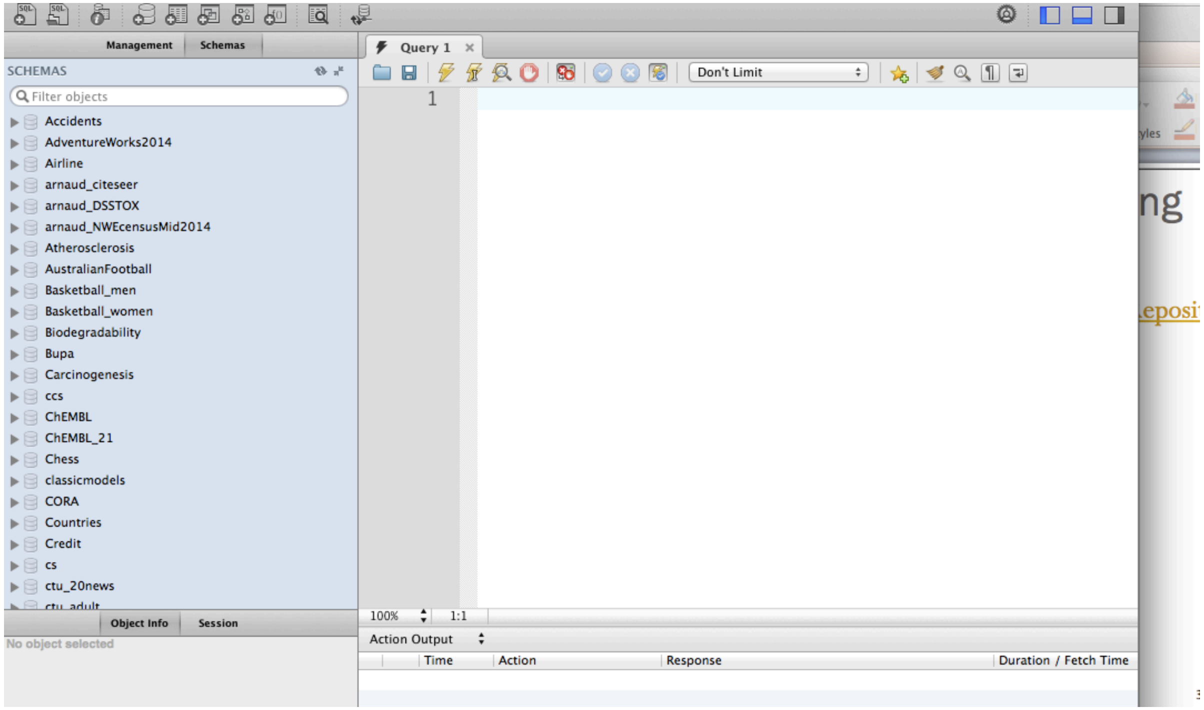
	Attributes	
Name	gender	country
Brad_Pitt	M	U.S.
Lucy_Liu	W	U.S.
Steve_Buscemi	M	U.S.
Uma_Thurman	W	U.S.

ActsIn

Name	Title	salary (M$)
Lucy_Liu	Kill_Bill	2
Steve_Buscemi	Fargo	0.5
Uma_Thurman	Kill_Bill	5

Plug: The Prague Relational Learning Repository

- 80+ relational databases [Repository]
- Can search for different dataset properties.
- Write-up and connection details are [available] http://arxiv.org/abs/1511.03086



Why Machine Learning for Information Networks?

Enterprise Data Are Relational

- Most organizations maintain data in a relational database management system.
- Structured Query Language (SQL) allows fast *data retrieval*.
  - E.g., find all movie ratings > 4 where the user is a woman.
- Multi-billion dollar industry, $Bn 15+ in 2006.
- IBM, Microsoft, Oracle, SAP, Peoplesoft.

Impedance Mismatch

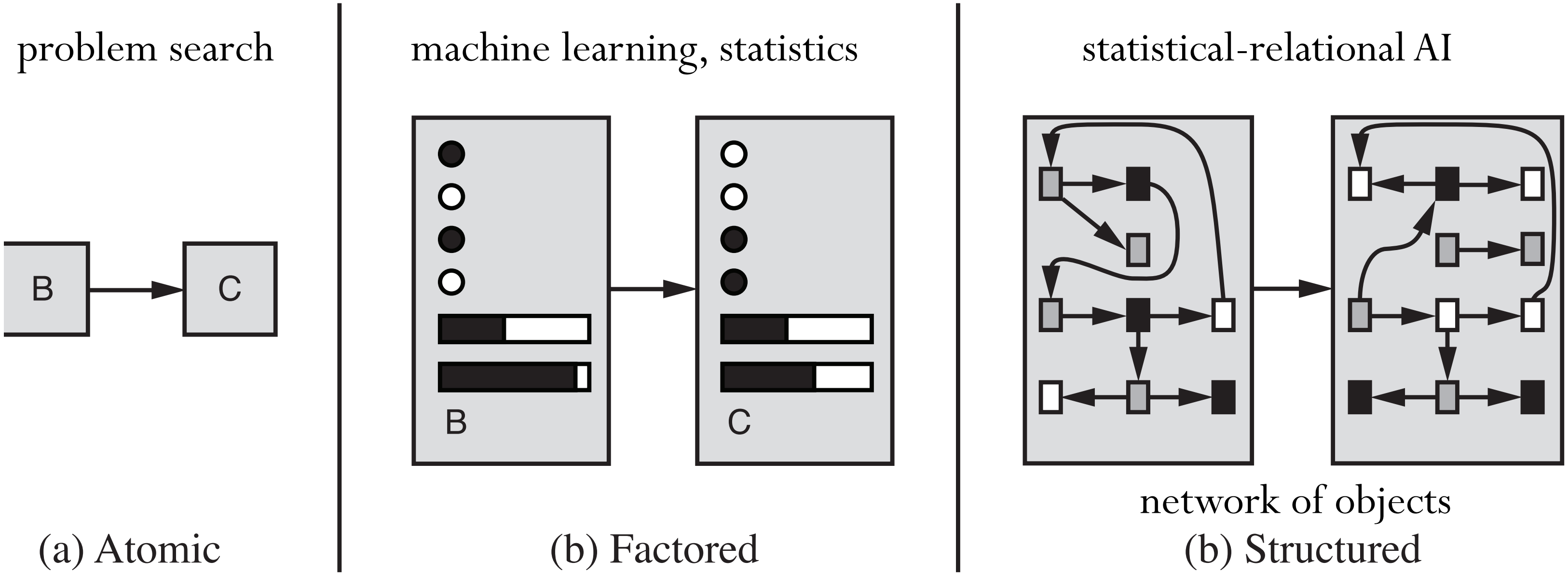
- Standard machine learning packages (R, SciKit, Weka,..) accept a *single* data table as input.
- In a database with *multiple* tables, which table do we input?
- SAP data scientist: "When our customers want to use machine learning, they spend 80% of their time getting the data into the right format".

	Attributes	
Name	gender	country
Brad_Pitt	M	U.S.
Lucy_Liu	W	U.S.
Steve_Buscemi	M	U.S.

Name	Title	salary (M$)
Lucy_Liu	Kill_Bill	2
Steve_Buscemi	Fargo	0.5
Uma_Thurman	Kill_Bill	5

AI Motivation: Expressive Power

- Russell and Norvig: Hierarchy of environment representations
- The more information an agent has about its environment, the better its performance



Russell, S. & Norvig, P. (2010), *Artificial Intelligence: A Modern Approach, Prentice Hall.*

Logic and Probability

- Russell (UC Berkeley): "Their unification holds enormous promise for AI"
- Domingos (U of Washington): "Logic handles complexity, probability represents uncertainty."

Russell, S. (2015), 'Unifying logic and probability', *Communications of the ACM 58(7), 88--97.*

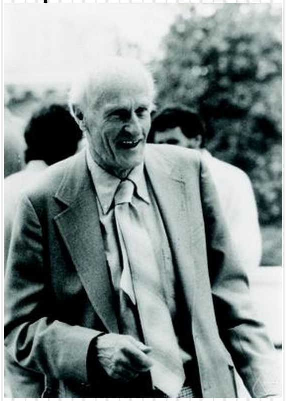
Unifying Logic and Statistics

Lise Getoor

David Poole

Stuart Russsell

Stephen Kleene



Poole, D. (2003), First-order probabilistic inference, '*IJCAI*'.
Getoor, L. & Grant, J. (2006), 'PRL: A probabilistic relational language', *Machine Learning* ***62(1-2), 7-31.***
Russell, S. & Norvig, P. (2010), *Artificial Intelligence: A Modern Approach, Prentice Hall.*
Stephen Kleene, (1952). Introduction to Metamathematics.

Function Representation

- The attributes and relationships in an information network can mathematically be represented using *functions*, e.g.
  - gender
  - ActsIn
  - salary

Example Function Representation

First-Order Logic: Terms

- A constant refers to an individual
  - “Fargo”
- A first-order variable refers to a class of individuals
  - “Movie” refers to Movies

Terms

- A constant or first-order variable is a term.
- The result of applying a function to a term is a term.

contains first-order variables?

first-order term
e.g. salary(Actor, Movie)

ground term
e.g. salary(UmaThurman, Fargo)

Stephen Kleene, (1952). *Introduction to Metamathematics.* North Holland.

Relational Random Variables

- *First-order random variable = First-order term* + probabilistic semantics (Wang et al. 2008)
- Both complex terms and complex random variables are built by function application

Statistics	Logic
Apply function to random variable(s) ➔ new random variable	Apply function to term(s) ➔ new term

Wang, D. Z.; Michelakis, E.; Garofalakis, M. & Hellerstein, J. M. (2008), BayesStore: managing large, uncertain data repositories with probabilistic graphical models, in , Proceedings VLDB Endowment, , pp. 340—351.

Formulas

- A (conjunctive) formula is a **joint assignment** $term_1 = value_1, \ldots, term_n = value_n$
  - e.g., ActsIn(Actor, Movie) = T, gender(Actor) = W
- A *ground* formula contains only constants
  - e.g., ActsIn(UmaThurman, KillBill) = T, gender(UmaThurman) = W

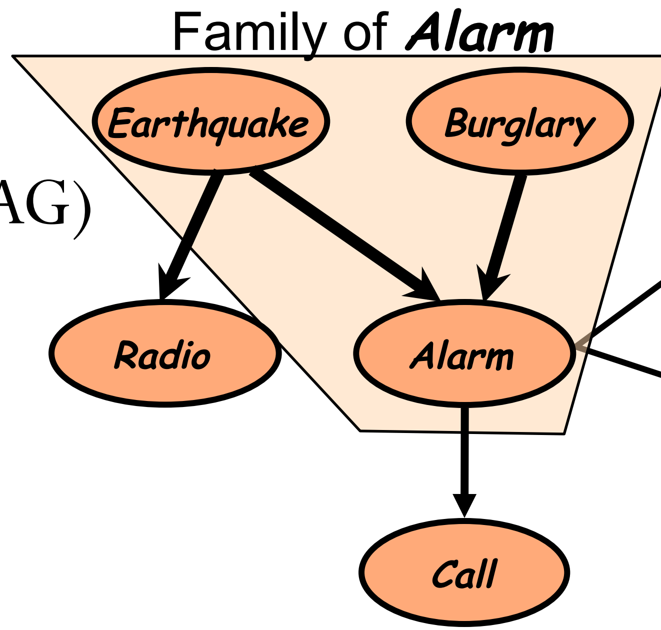
What is a Bayesian network?

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 | 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

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

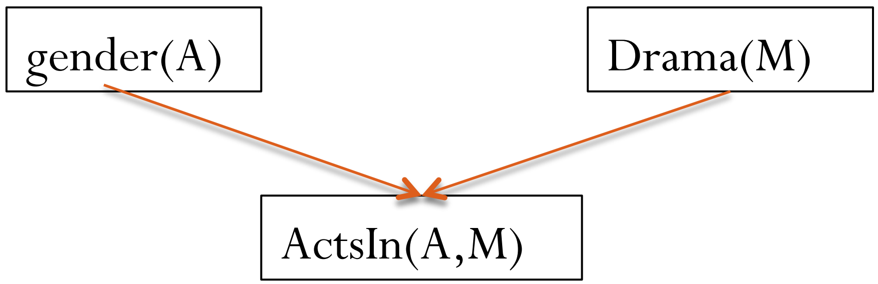
Figure from N. Friedman

Why are Bayes nets useful?

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

Bayesian networks for relational data

- **A first-order Bayesian network** is a Bayesian network whose nodes are first-order terms
  (Wang et al. 2008)
- AKA parametrized Bayesian network
  (Poole 2003, Kimmig et al. 2014)



Wang, D. Z.; Michelakis, E.; Garofalakis, M. & Hellerstein, J. M. (2008), BayesStore: managing large, uncertain data repositories with probabilistic graphical models, in ,VLDB Endowment, , pp. 340--351.
Kimmig, A.; Mihalkova, L. & Getoor, L. (2014), 'Lifted graphical models: a survey', *Machine Learning, 1--45*.

Frequency Semantics for First-Order Bayesian Networks

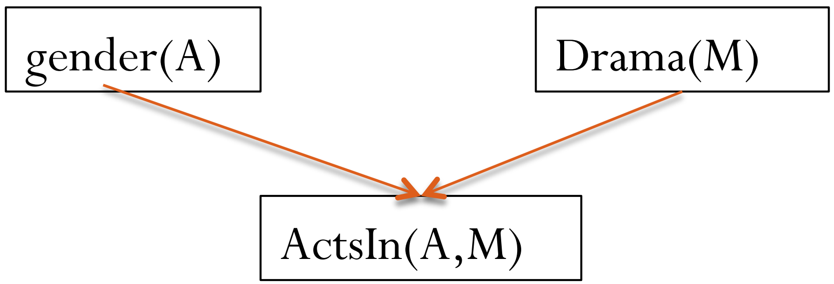
Joe Halpern

Fahim Bacchus

Halpern, J.Y. (1990), 'An analysis of first-order logics of probability', *Artificial Intelligence* ***46(3), 311--350.***
Bacchus, F. (1990), *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach to Probabilities, MIT Press, Cambridge, MA.*

Random Selection Semantics for First-Order Bayesian Networks

- We can compute joint probabilities from a FOBN, e.g.
- P(gender(Actor) = W, ActsIn(Actor,Movie) = T, Drama(Movie) = F) = 2/8
- But what does this represent?



"if we randomly select an actor and a movie, the probability is 2/8 that the actor appears in the movie, the actor is a woman, and the movie is a drama"

Random Selection Semantics

Population

Actors

Movies

Population (first-order) variables

Actor
Random Selection from Actors.
$P(Actor = brad\_pitt) = 1/4$

Movie
Random Selection from Movies.
$P(Movie = Fargo) = 1/2$

First-Order Random Variables (Terms)

gender(Actor)
Gender of selected actor.
$P(gender(Actor) = W) = 1/2$

ActsIn(Actor, Movie) =
T if selected actor appears in selected movie, F otherwise
$P(ActsIn(Actor, Movie) = \mathrm{T}) = 3/8$

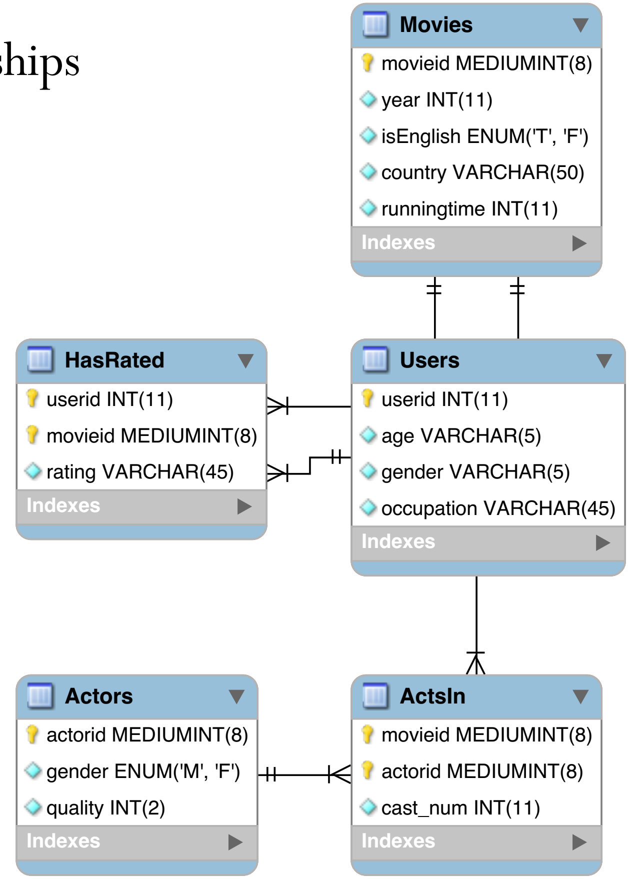
Drama(Movie)
Is the selected movie a drama?
$P(Drama(Movie) = T) = 1/2$

Real-World Examples

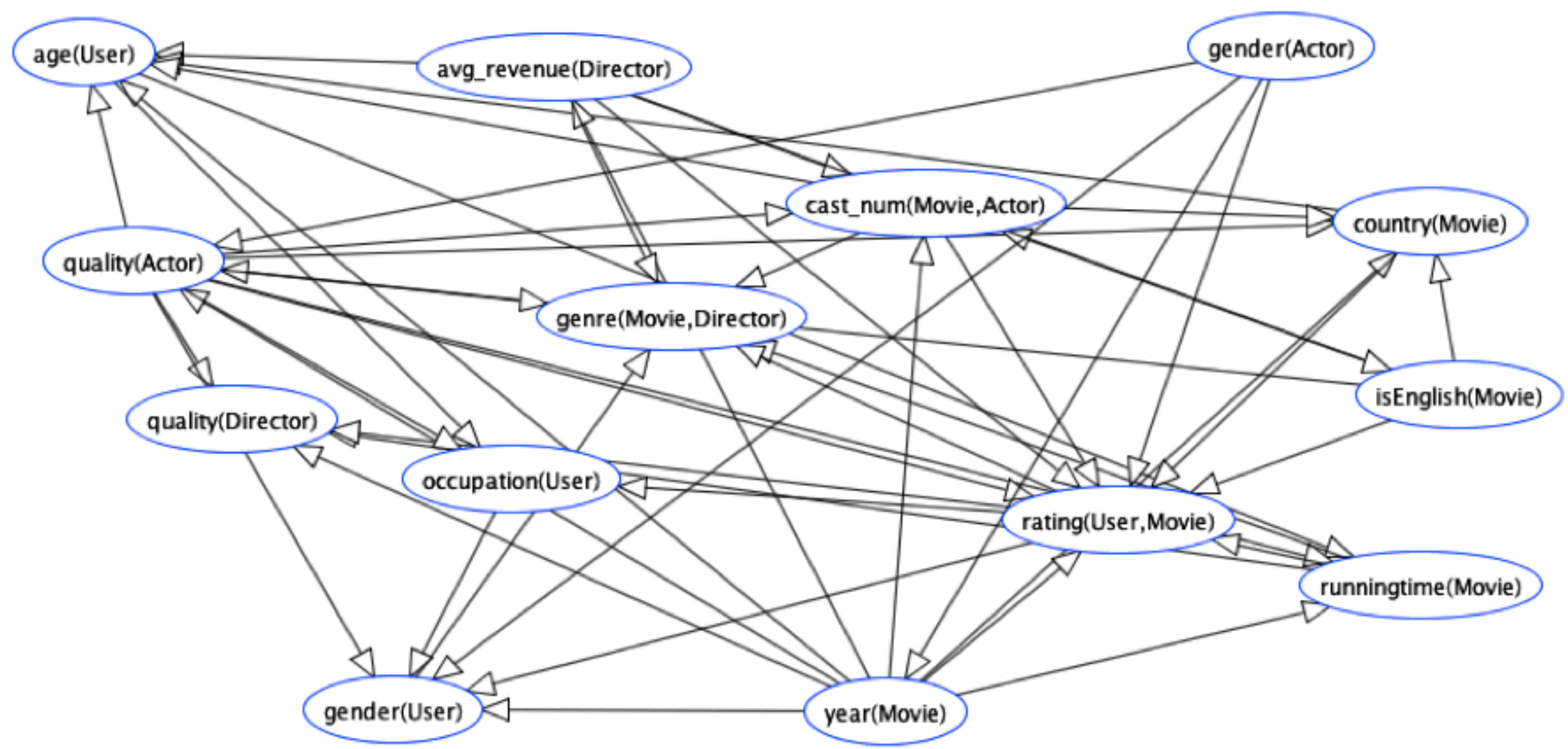
- To illustrate frequency semantics, learn and evaluate on the training set
  - ground truth about frequencies
- We discuss generalization later

IMDb Data Format

data with two relationships

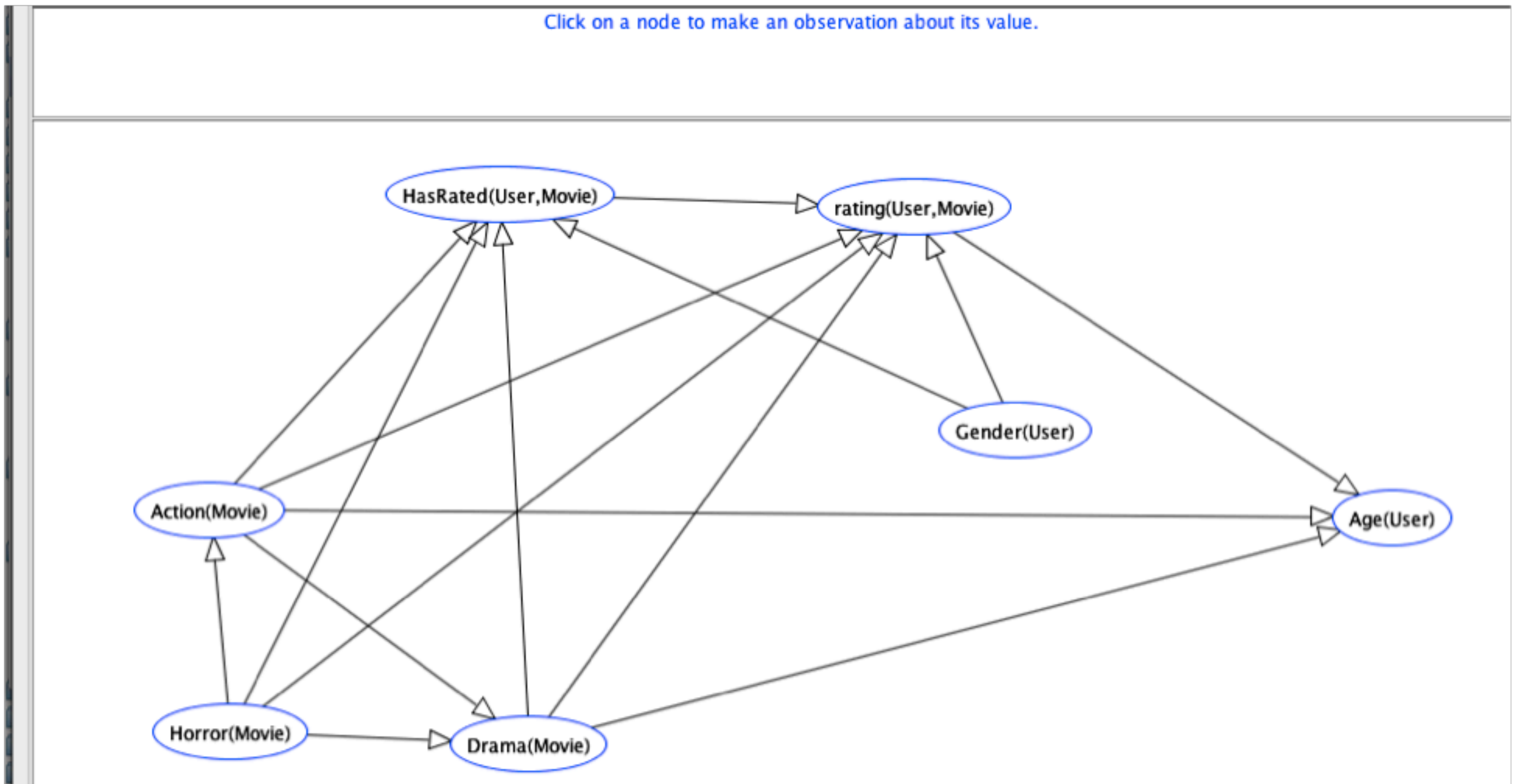


Learned Bayes Net for Full IMDB

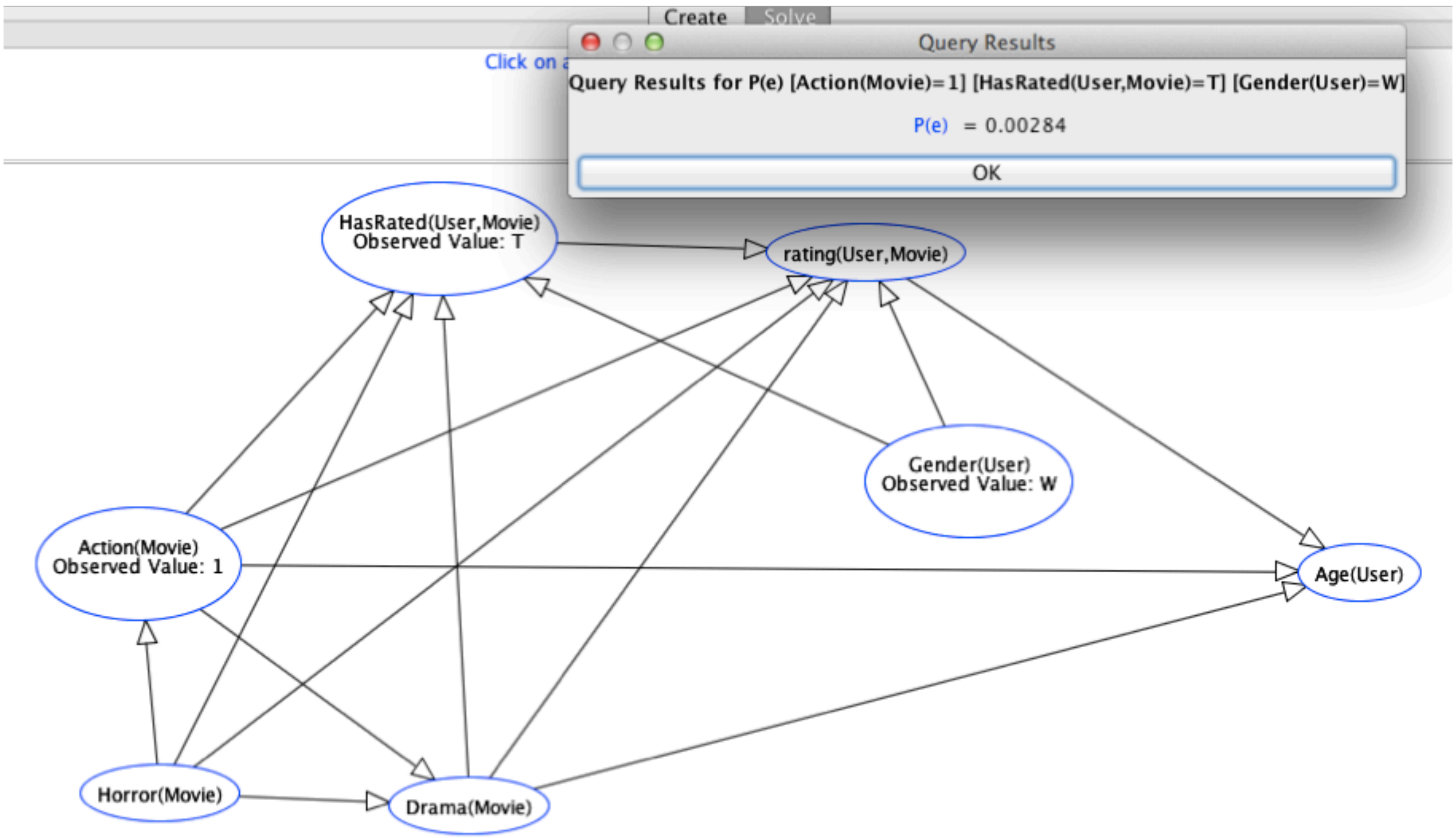


Learned Bayes Net for IMDb

With only 1 relationship HasRated(User,Movie).



Bayes Net Query



Data Query

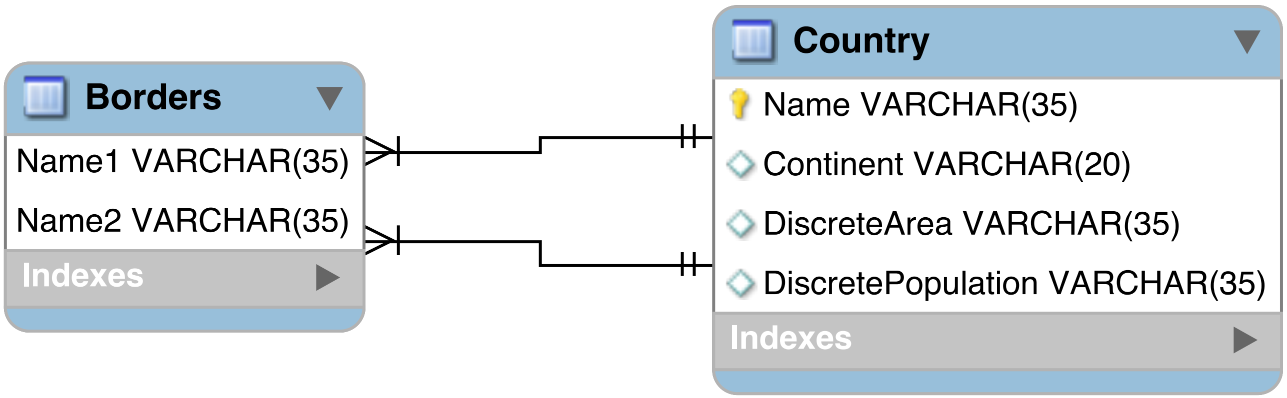
Num Movies	**3883**
Num Users	6039
Num Movie-User Pairs	3883 x 6039 = 23449437

movie-user pairs with action movie, woman user

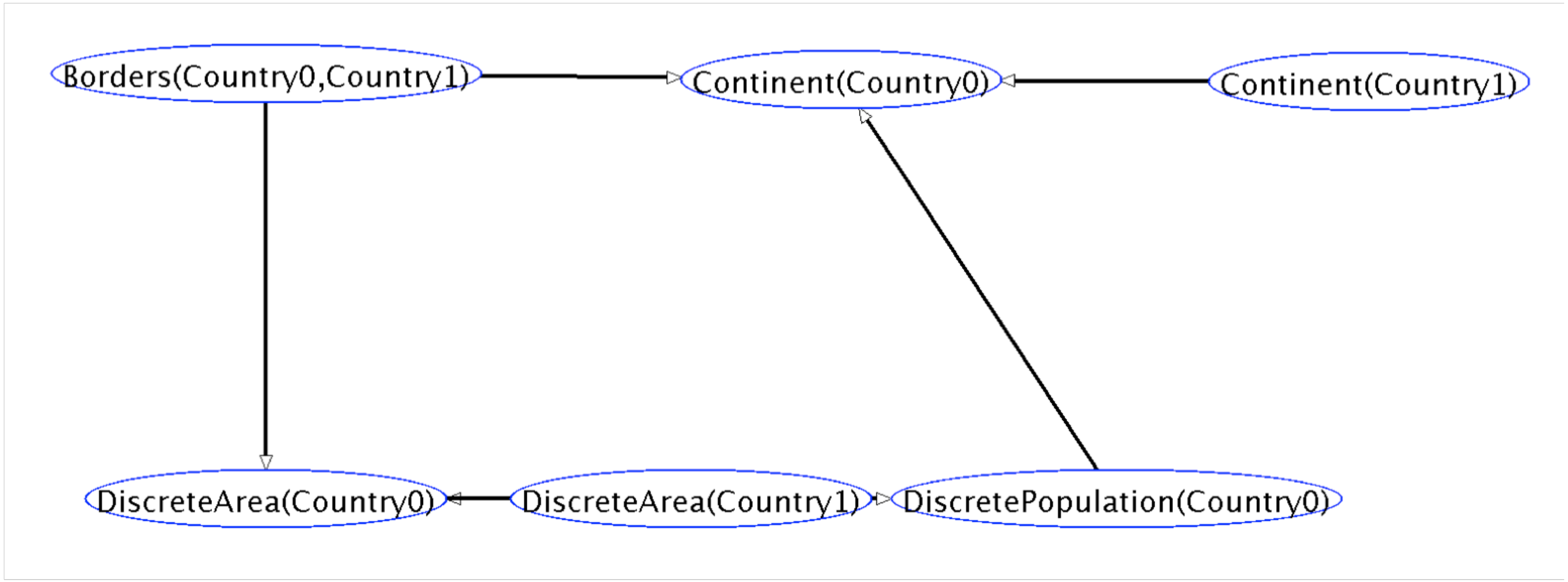
Action(Movie) = T, HasRated(User,Movie) = T, gender(User) = W	**66642**
Frequency	66642/23449437= 0.0028

More Examples in spreadsheet on website

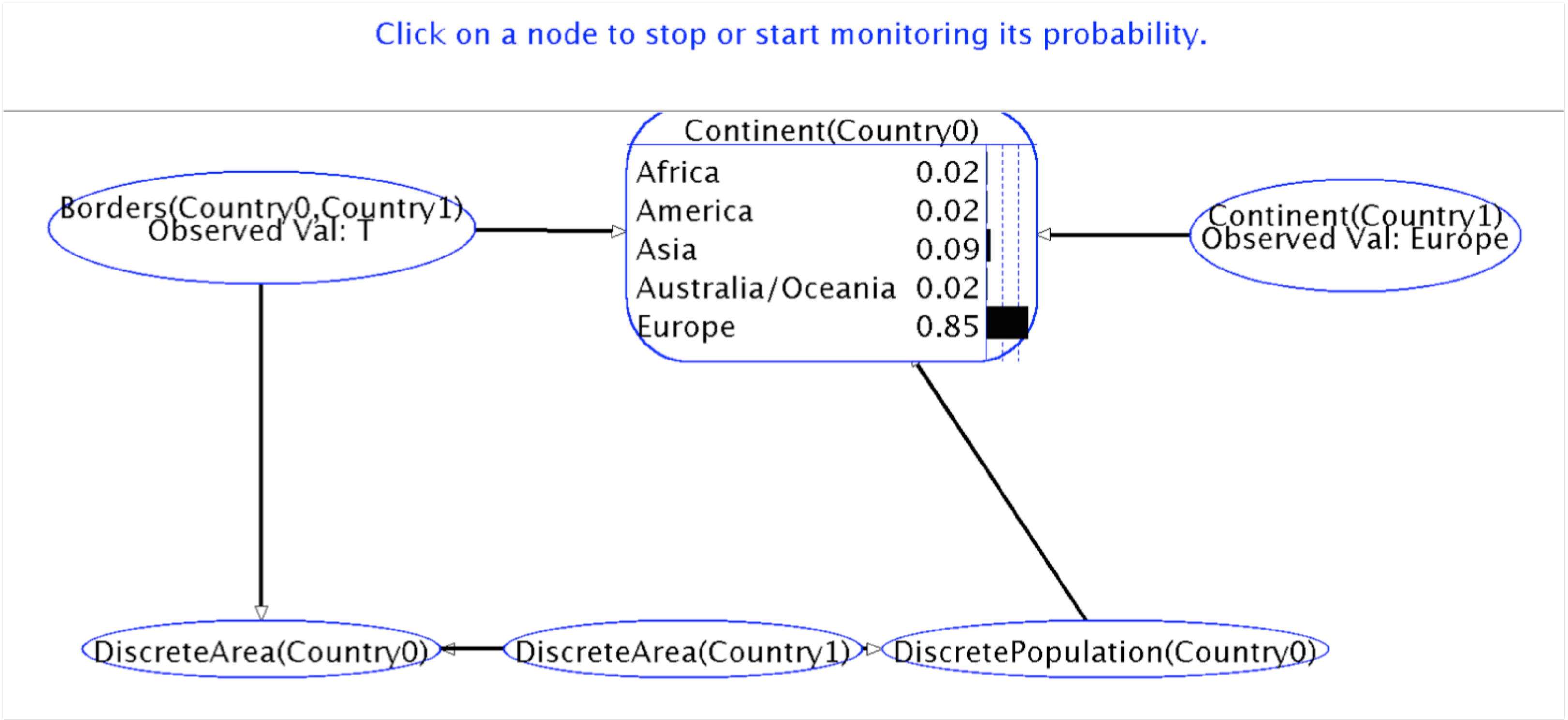
Mondial Data Format



Learned Bayes Net for Mondial



Bayes Net query



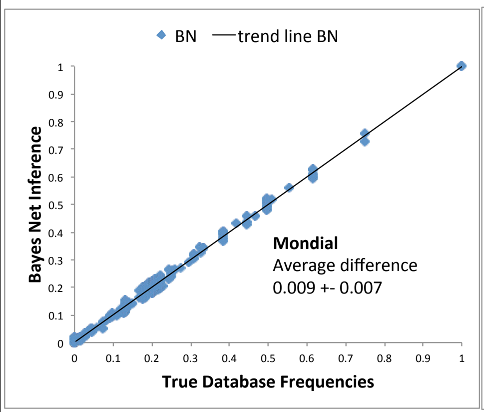
Data Query

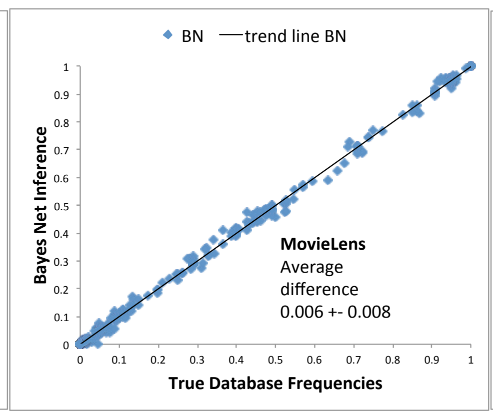
Number of Europe-Europe Borders	**156**
Number of *-Europe Borders	166
P(continent(country1) = Europe | Borders(country1,country2) =T, continent(country2=Europe))	156/166= 93.98%

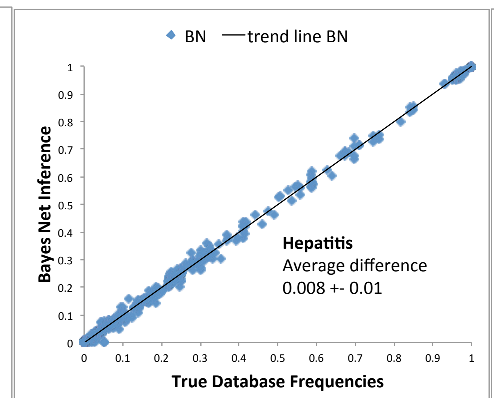
- BN was learned with frequency smoothing (Laplace correction)
- More Examples in spreadsheet on tutorial website

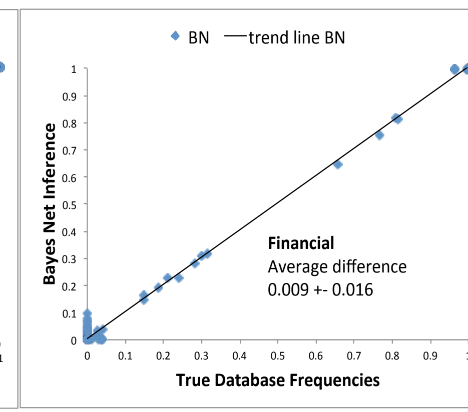
Bayesian Networks are Excellent Estimators of Network Frequencies

- Queries Randomly Generated
- Example: $P(gender(A) = W \mid ActsIn(A,M) = true, Drama(M)=T)$?
- Learn Bayesian network and test on entire database as in Getoor et al. 2001









Schulte, O.; Khosravi, H.; Kirkpatrick, A.; Gao, T. & Zhu, Y. (2014), 'Modelling Relational Statistics With Bayes Nets', Machine Learning 94, 105-125.

Getoor, L.; Taskar, B. & Koller, D. (2001), 'Selectivity estimation using probabilistic models', *ACM SIGMOD Record* *30(2), 461—472.*

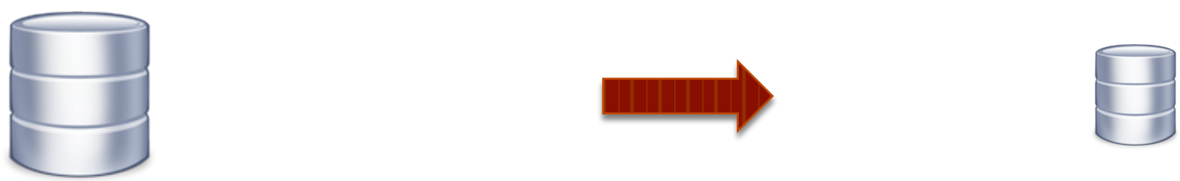
Relational Exception Mining

Random Individuals vs. Specific Individuals

Profile-Based Outlier Detection for Relational Data

Population Database
e.g. IMDB

Individual Database
Profile, Interpretation, egonet
e.g. Brad Pitt's movies



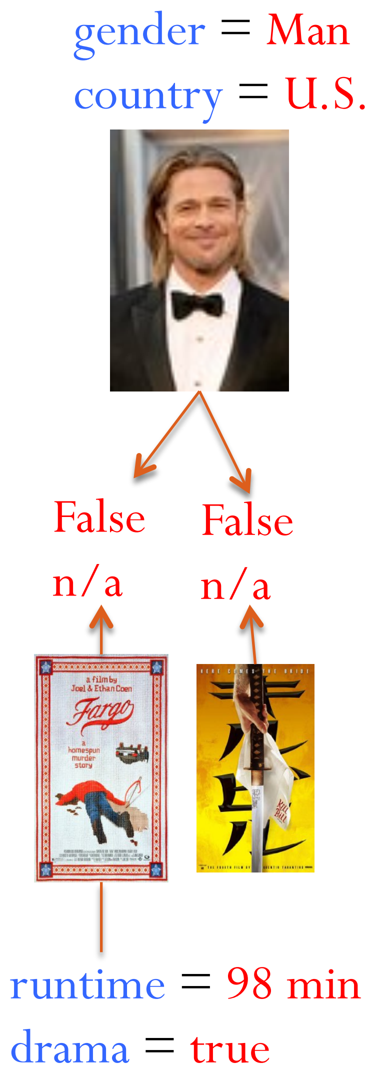
Goal: Identify <u>exceptional individual databases</u>

Akoglu, L.; Tong, H. & Koutra, D. (2015), 'Graph based anomaly detection and description: a survey', *Data Mining and Knowledge Discovery* *29(3), 626--688.*

Maervoet, J.; Vens, C.; Vanden Berghe, G.; Blockeel, H. & De Causmaecker, P. (2012), 'Outlier Detection in Relational Data: A Case Study in Geographical Information Systems', *Expert Systems With Applications* *39(5), 4718—4728.*

Example: population data

Example: individual data



Compare Random Individual to Target Individual

Population Database

Class Bayesian network
(for random individual)

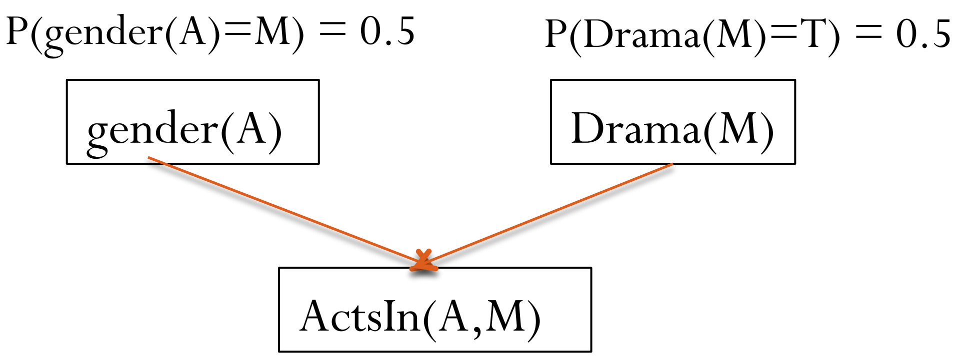
Individual Database

Individual Bayesian network

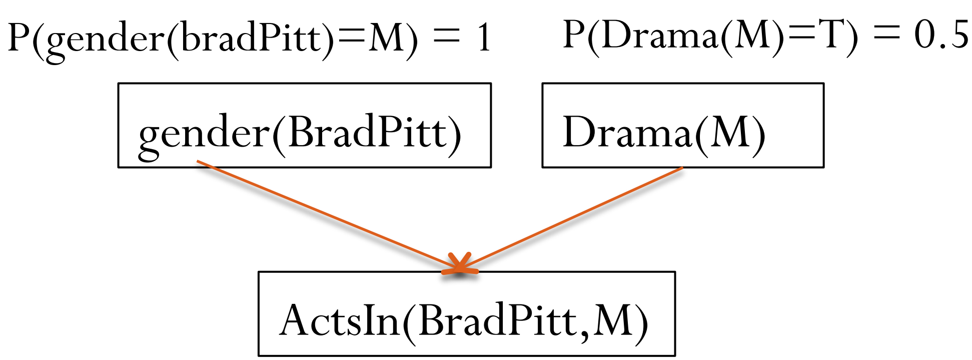
Outlierness Metric (quality measure) =
Measure of dissimilarity between class and individual BN
e.g. KLD, ELD (new)

“Model-based Outlier Detection for Object-Relational Data”. Riahi and Schulte (2015). IEEE SSCI.

Example: class and individual Bayesian network parameters



Gender (A)	Drama(M)	Cond. Prob. of ActsIn(A,M)=T
M	T	1/2
M	F	0
W	T	0
W	F	1



Gender (bradPitt)	Drama (M)	Cond. Prob. of ActsIn(A,M)=T
M	T	0
M	F	0

Case Study: Strikers and Movies

Data are from Premier League Season 2011-2012.

Player Name	Position	KLD Rank	KLD Max Node	Feature Max Value	Individual Probability	Class Probability
Edin Dzeko	Striker	1	Dribble Efficiency	DE = Low	0.16	0.50
Paul Robinson	Goalie	2	SavesMade	SM = Medium	0.30	0.04

Striker = Normal

MovieTitle	Genre	KLD Rank	KLD Max Node	Feature Max Value	Individual Probability	Class Probability
Brave Heart	Drama	1	Actor_Quality	a_quality=4	0.93	0.42
Austin Powers	Comedy	2	Cast_position	cast_num=3	0.78	0.49
Blue Brothers	Comedy	3	Cast_position	cast_num=3	0.88	0.49

How is Relational Learning Different From IID Learning?

Challenges and Solutions

IID Data vs. Relational Data

Traditional Data Matrix represents independent and identically distributed data points (i.i.d.)

- special case of relational data with 0 relationships
- unary functors

gender = Man
country = U.S.

gender = Man
country = U.S.

gender = Woman
country = U.S.

gender = Woman
country = U.S.

Relational Data >1 arity functors

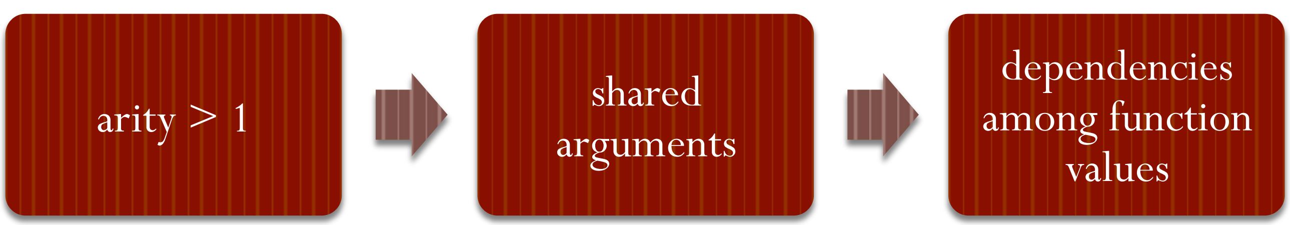
i.i.d. data = single-table data = unary functors

Nickel, M.; Murphy, K.; Tresp, V. & Gabrilovich, E. (2016), 'A review of relational machine learning for knowledge graphs', Proceedings of the IEEE 104(1), 11--33.

Relational Data Are Not Independent

Name	Title	Salary (M$)
Lucy_Liu	Kill_Bill	2
Uma_Thurman	Kill_Bill	5
Uma_Thurman	Be_Cool	9

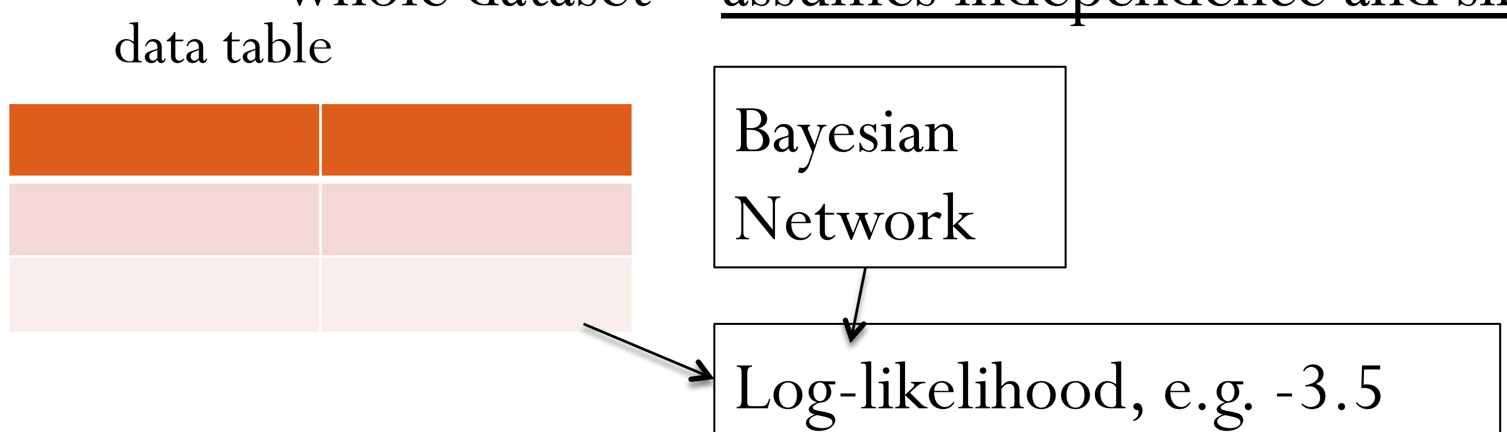
- Uma Thurman's salary in Kill Bill carries information about her salary in Be Cool
- Also carries information about Lucy Liu's salary in Kill Bill



Difficulty #1: Likelihood Function

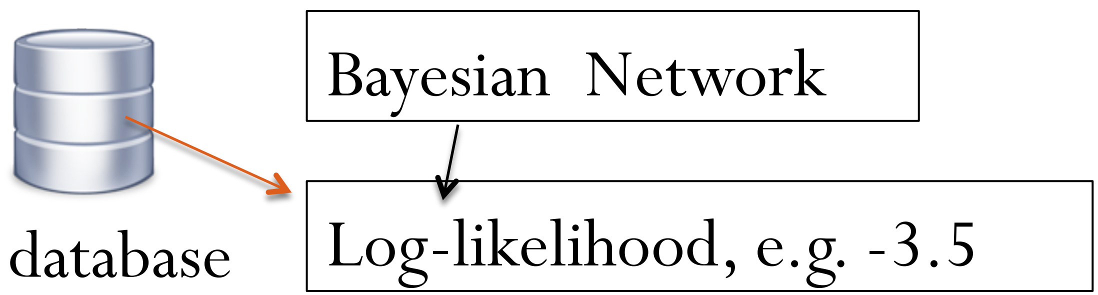
- Most Bayesian network learning methods are based on a **score function**
- Key component: the likelihood function P(data|model)

1. measure how how likely each datapoint is according to the Bayesian network
2. <u>Multiply</u> datapoint probabilities to define likelihood for whole dataset – <u>assumes independence and single table</u>



Solution #1: The Random Selection Likelihood Score

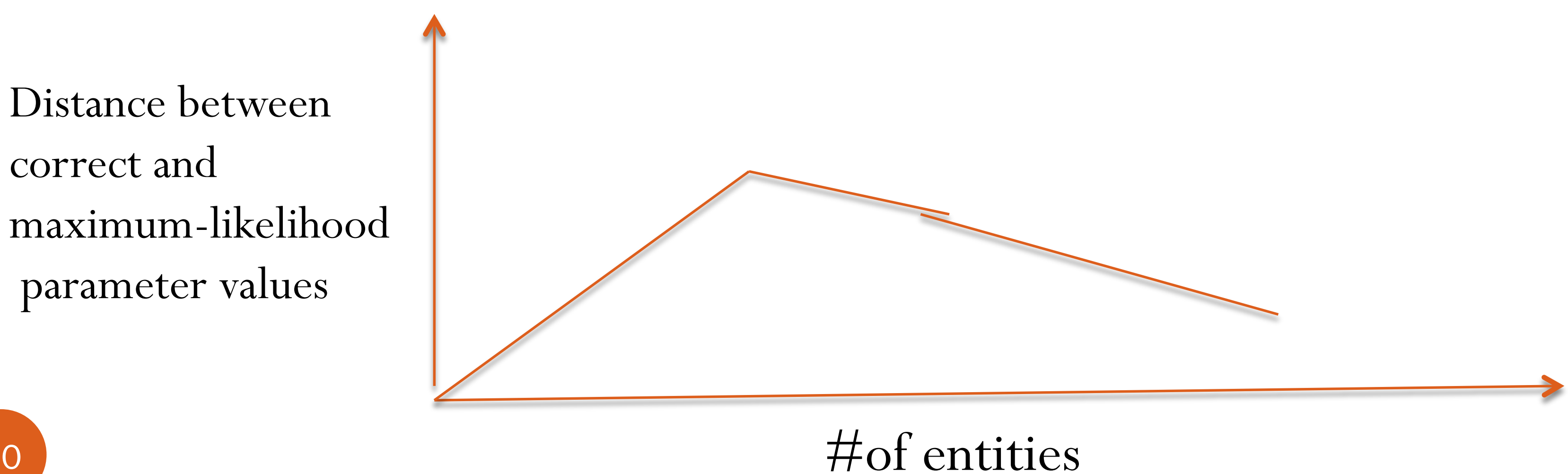
1. Randomly select a grounding/instantiation for **all** first-order variables in the first-order Bayesian network
2. Compute the log-likelihood for the attributes of the selected grounding
3. Log-likelihood score =
   expected log-likelihood for a random grounding



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

Theoretical Validation #1

- **Proposition** (Schulte 2011) The random selection log-likelihood score is maximized by setting the conditional probabilities to the *frequencies observed in the network*.
- **Theorem** (Xiang and Neville 2011) The random selection log-likelihood score is *consistent* (asymptotically correct).



Difficulty #2: No global sample size

- What is the sample size - #Users, #Movies, # Ratings?
- Typical model selection scores are of the form score(model,data) = log-likelihood(data|model)- f(#model parameters, sample size)

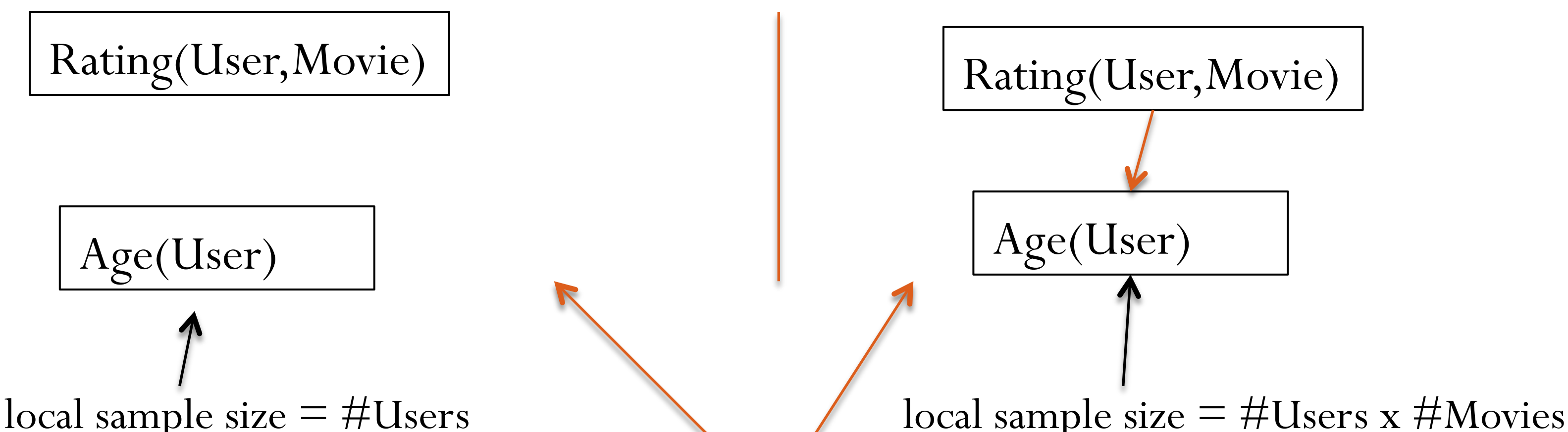
already discussed

penalize complex models

- e.g. for BIC we have f = log(N)/2 x #parameters

Solution #2

- Use local sample sizes = number of possible child-parent instantiations
- When comparing two models, normalize <u>both</u> penalty terms by the larger local sample size.



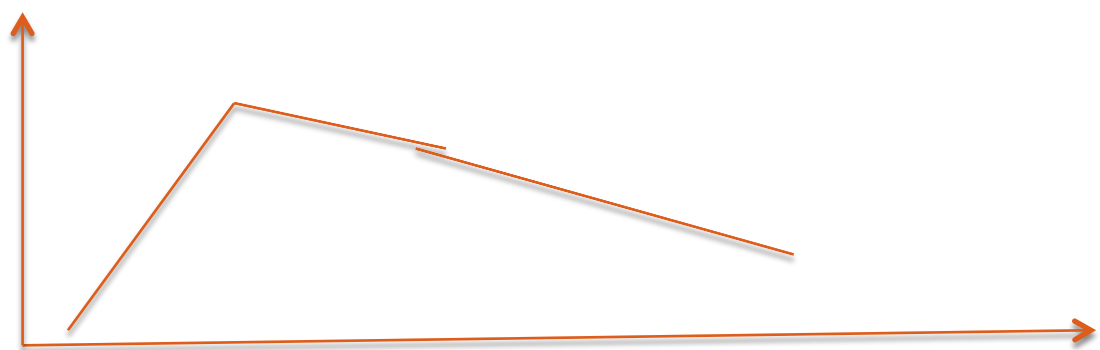
f = [log(#Usersx#Movies)/2 x #parameters]/#Users x #Movies

Schulte, O. & Gholami, S. (2017), Locally Consistent Bayesian Network Scores for Multi-Relational Data, *IJCAI 2017*

Theoretical Validation #2

- **Theorem** (Schulte and Gholami 2017) If a score is consistent for i.i.d. data, then the normalized score is consistent for relational data:
  - converges to a model of the network frequencies
  - with a minimum number of edges

Distance between
network frequencies
and
FOBN joint probabilities



#of entities

Schulte, O. & Gholami, S. (2017), Locally Consistent Bayesian Network Scores for Multi-Relational Data, *IJCAI 2017*

Summary: Information Networks

- Heterogeneous information networks are ubiquitous, go by several names:
  - relational database
  - first-order model
  - matrixes/tensors
- Unifying logic and statistics:
  - Relational random variable = first-order term
  - First-order Bayesian network = BN whose nodes are first-order terms

Summary: Applications of FOBNs

- Modelling correlations and frequencies in relational data
  - applies classic random selection semantics for probabilistic logic
- Exception Mining and Anomaly Detection

Summary: Learning Challenges

- Network nodes and links are *not* independent
- ➢ Difficult to define likelihood for entire network
- ➢ Solution: apply random selection semantics to define *expected log-likelihood* from random instances
- There is no global sample size *N*
- ➢ Difficult to define model selection score
- Normalize score by (max) local sample size
- Theoretical and extensive empirical validation

There's More (In Tutorial)

- https://oschulte.github.io/srl-tutorial-slides/
- Scalable Algorithms:
  - for counting relational frequencies
  - for relational model structure search
- Latent variable models for clustering, community detection, matrix factorization, relational deep learning
- Applications:
  - link-based classification
  - link prediction
  - feature extraction

References

- Github https://github.com/sfu-cl-lab
  - Code and names of collaborators (thank you thank you!)

- Russell, S. (2015), 'Unifying logic and probability', *Communications of the ACM* ***58(7), 88--97.***
- Nickel, M.; Murphy, K.; Tresp, V. & Gabrilovich, E. (2016), 'A review of relational machine learning for knowledge graphs', *Proceedings of the IEEE* ***104(1), 11--33.***
- Domingos, P. & Lowd, D. (2009), *Markov Logic: An Interface Layer for Artificial Intelligence, Morgan and Claypool Publishers.*
- Kimmig, A.; Mihalkova, L. & Getoor, L. (2014), 'Lifted graphical models: a survey', *Machine Learning, 1—45.*

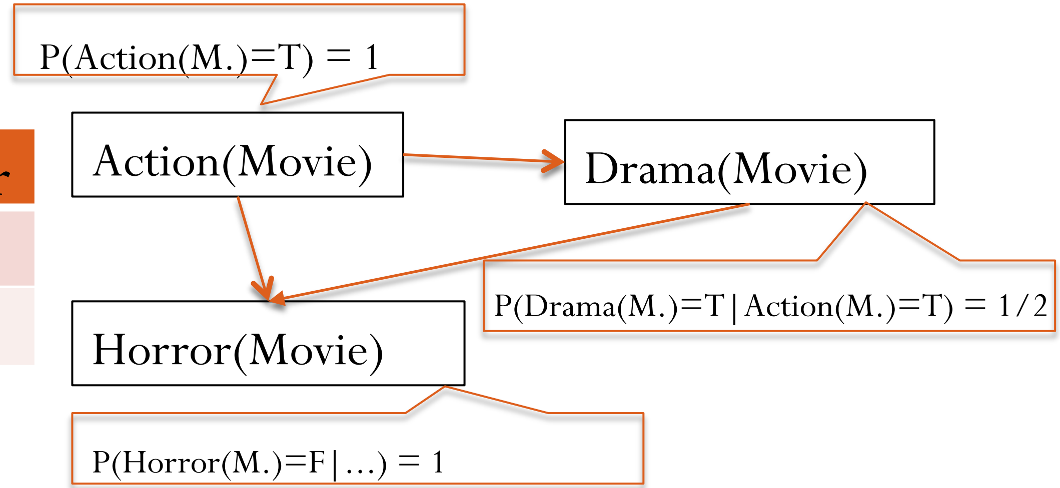
The Bayes Net Likelihood Function for IID data

1. For each row, compute the log-likelihood for the attribute values in the row.
2. Log-likelihood for table = sum of log-likelihoods for rows.

<u>Assumes independence of rows (data points)</u>

IID Example

Title	Drama	Action	Horror
Fargo	T	T	F
Kill_Bill	F	T	F

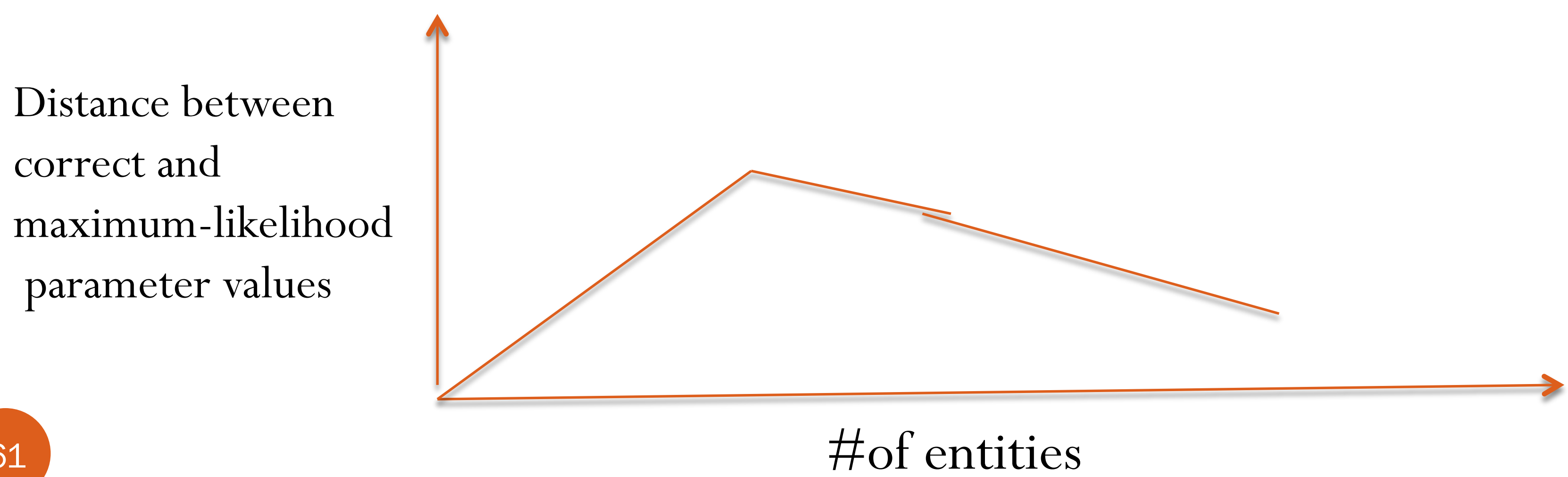


Title	Drama	Action	Horror	P_B	$\ln(P_B)$
Fargo	T	T	F	1x1/2x1 = 1/2	-0.69
Kill_Bill	F	T	F	1x1/2x1 = 1/2	-0.69

Total Log-likelihood Score for Table = -1.38

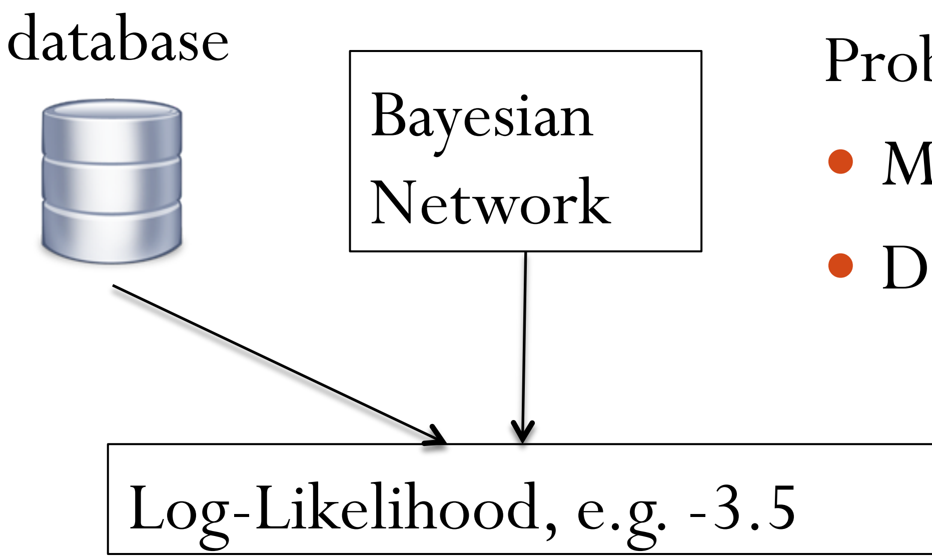
Theoretical Validation #1

- **Proposition** (Schulte 2011) The random selection log-likelihood score is maximized by setting the conditional probabilities to the *frequencies observed in the network*.
- **Theorem** (Xiang and Neville 2011) The random selection log-likelihood score is *consistent* (asymptotically correct).



Likelihood Function for Relational Data

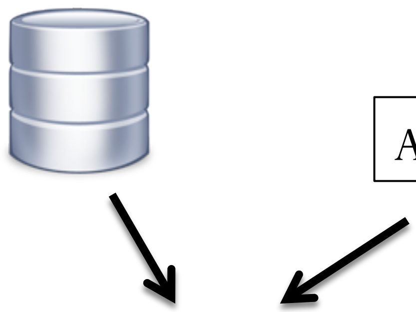
Wanted: a likelihood score for relational data

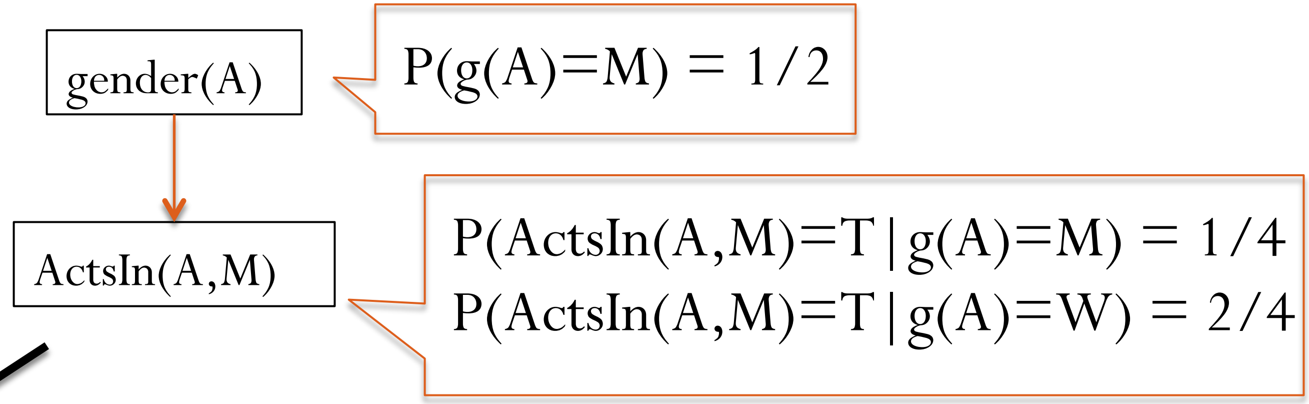


Problems

- Multiple Tables.
- Dependent data points

Example

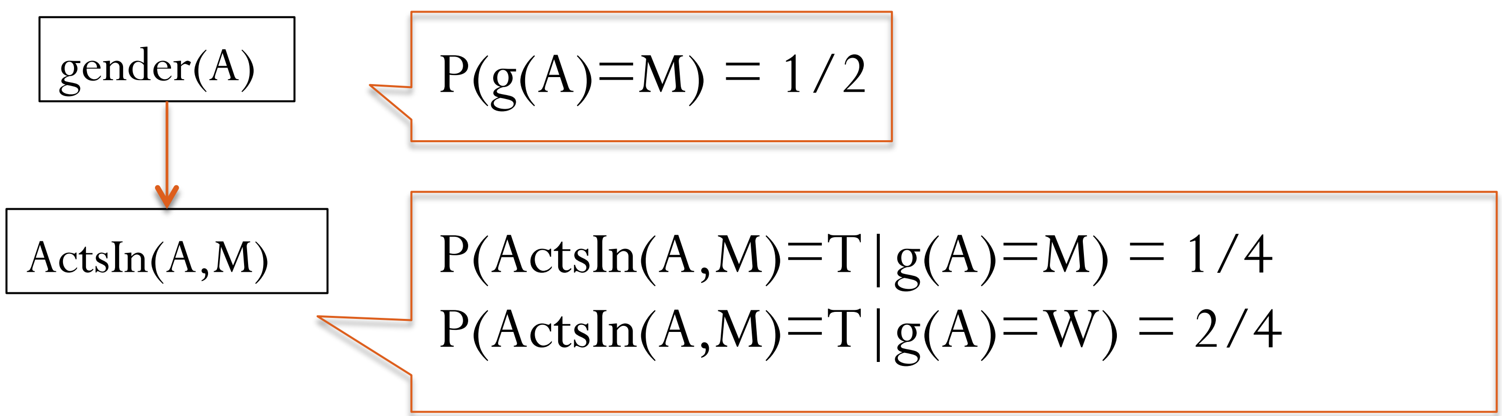




Prob	A	M	gender(A)	ActsIn(A,M)	P_B	$\ln(P_B)$
1/8	Brad_Pitt	Fargo	M	F	3/8	-0.98
1/8	Brad_Pitt	Kill_Bill	M	F	3/8	-0.98
1/8	Lucy_Liu	Fargo	W	F	2/8	-1.39
1/8	Lucy_Liu	Kill_Bill	W	T	2/8	-1.39
1/8	Steve_Buscemi	Fargo	M	T	1/8	-2.08
1/8	Steve_Buscemi	Kill_Bill	M	F	3/8	-0.98
1/8	Uma_Thurman	Fargo	W	F	2/8	-1.39
1/8	Uma_Thurman	Kill_Bill	W	T	2/8	-1.39
					0.27 geo	-1.32 arith

Observed Frequencies Maximize Random Selection Likelihood

Proposition The random selection log-likelihood score is maximized by setting the Bayesian network parameters to the observed conditional frequencies



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