# Relational Latent Class Models

Volker Tresp

Siemens, Corporate Research and Technology

Munich, Germany

Collaborators: Zhao Xu (1,2), Stefan Reckow (1,2), Achim Rettinger (2,4), Matthias Nickles (4,5), Kai Yu (3), Shipeng Yu (2) and Hans-Peter Kriegel (1)

1: University of Munich, Germany

2: Siemens AG

3: NEC Laboratories America

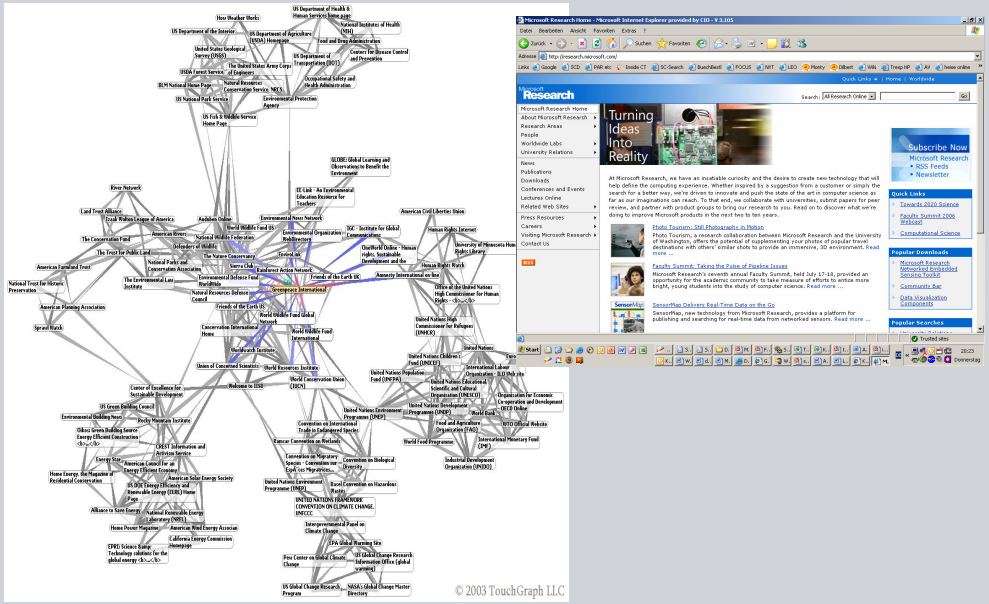
4: Technical University of Munich

5: University of Bath

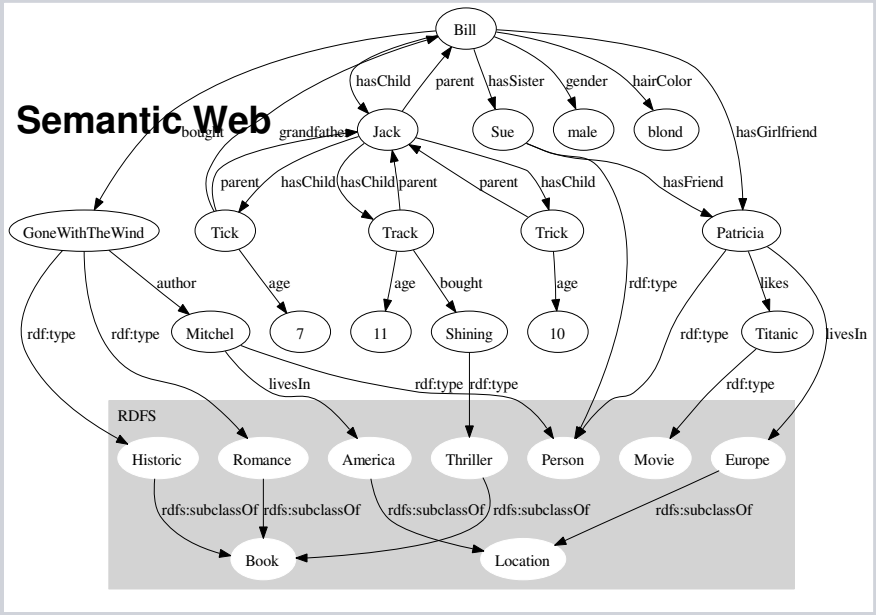
## Overview

- Relational domains are quite important and interesting
- Bayesian methods can be quite effective in modeling relational domains

# Relational Problems are About Networks



# Relational Problems Might Involve Multiple Classes of Entities and Relations

**John Donne, 1572 – 1631**
**a Jacobean poet and preacher**

***No man is an island,*** **entire of itself; every man is a piece of the continent, a part of the main.** If a clod be washed away by the sea, Europe is the less, as well as if a promontory were, as well as if a manor of thy friend's or of thine own were. **Any man's death diminishes me because I am involved in mankind; and therefore never send to know for whom the bell tolls; it tolls for thee.**

(Also: Epigraph of Hemingway's 1940 novel, For Whom the Bell Tolls)

**John Donne (1572 – 1631)**

## Overview: Learning with Relations (incomplete)

**Social Network Analysis:**
- Descriptive, deterministic (network structure analysis)
- Increasing focus on statistical inference
- Driven by solving the problem at hand; often one type of actor and relation

**Specialized Algorithms:**
- Page Rank, most collaborative filtering algorithms …

**Inductive Logic Programming (ILP):**
- FOL based; focus on generality; lost in generality?
- Learning of rules for prediction of predicates (relationships, attributes)
- Mostly deterministic; but recent extensions: Stochastic Logic Programs (Muggleton), Bayesian Logic programs (Kersting et al.))

**Statistical Relational Learning**
- Principled probabilistic approaches from machine learning and AI
- Focus on uncertainty in relational domains;
- Analysis of dependencies; prediction of attributes and relations

## Statistical Relational Learning

- **Probabilistic relational models (PRM)** (Koller, Friedman, Getoor, Pfeffer, …)
  - Combines a relational description with components from frame-based systems and Bayesian networks
- **Directed acyclic probabilistic entity-relationship (DAPER) model** (Heckermann, Meek, Koller)
  - ER (entity relationship) models with Bayesian networks
- **Relational Dependency Networks** (Jennifer Neville, David Jensen)
- **Relational Markov Networks** (Taskar, …)
- **Markov Logic Networks** (Richardson, Domingos, …)

## This Work

What do we do:

- In this work we apply nonparametric hierarchical Bayesian modeling to relational learning and achieve nonparametric relational Bayes in form of an infinite hidden relational model (IHRM) and touch on related approaches

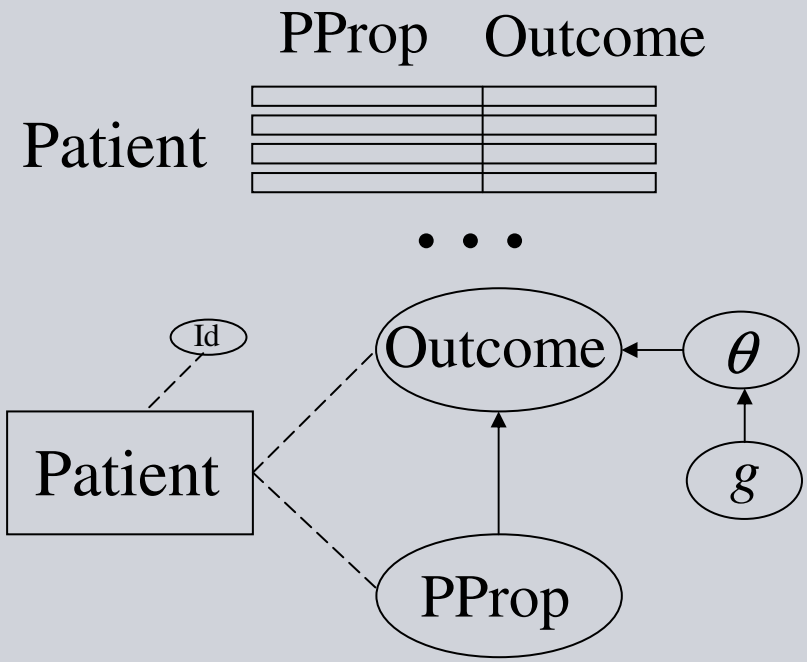
Advantages of the IHRM

- Straightforward to apply without any extensive structural learning
- Attributes, relationships and identities of entities can have predictive power
- Clustering in relational domain (multi-relational clustering)
    - Identify roles of actors

# II. Before Relational Learning

# IID Learning: The Matrix

- Traditionally, the relational structure is ignored and a flat representation is applied
- The standard assumption is that data points are sampled independently

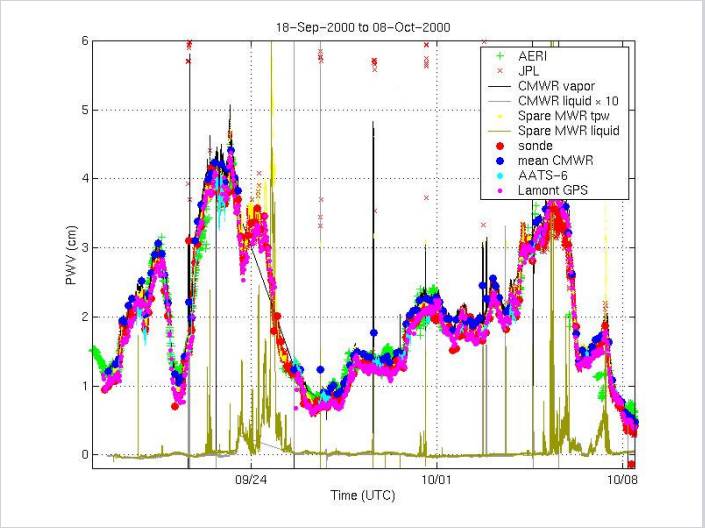


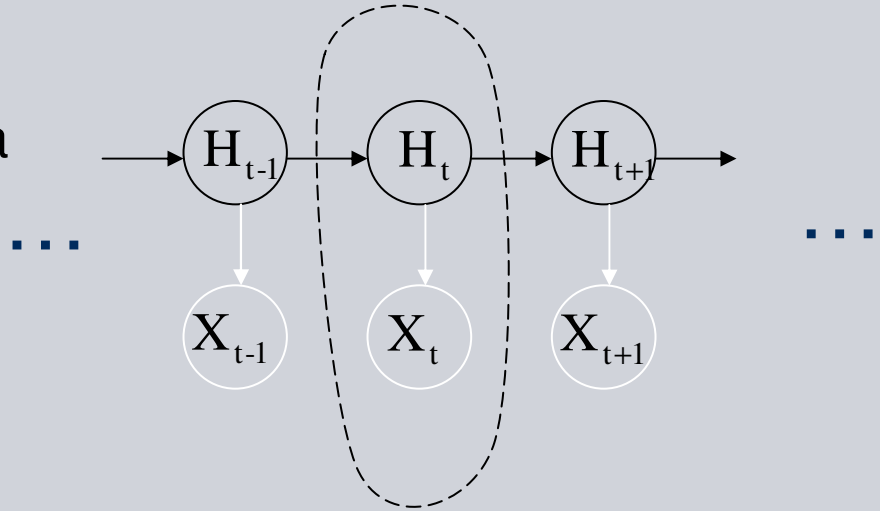
## Towards Relational Learning: Time Series Models

In a standard time-series model, the data are still displayed as a matrix but the temporal ordering of the rows is important

Often, we use a simple template to define the probabilistic model

Although the model itself might factorize, the complete data (one particular time series) is "one data point": For example, if all H are latent variables then all measurements of X influence the probability of H



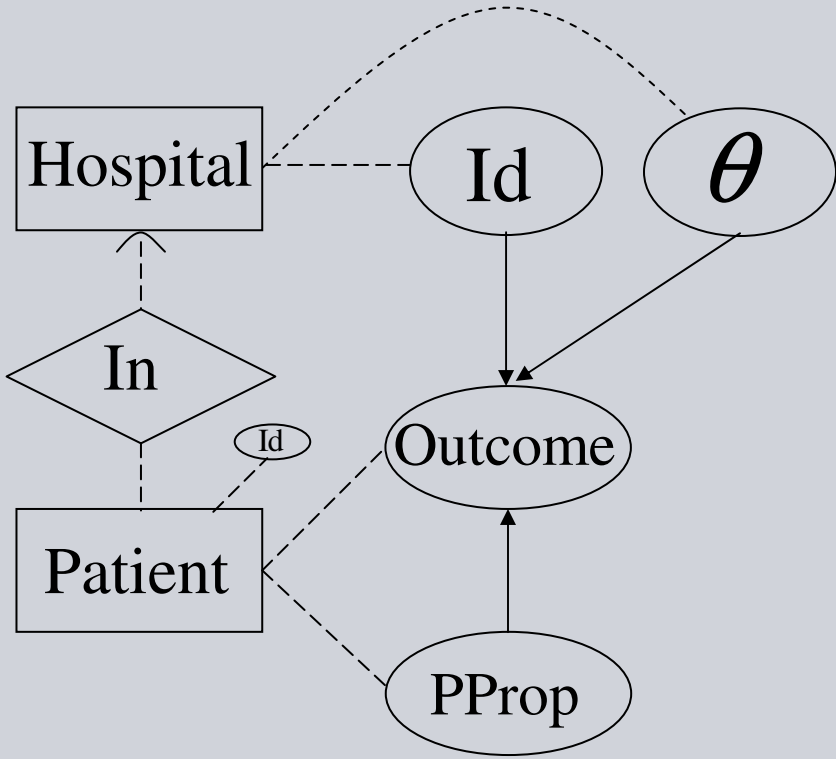



# Towards Relational Learning: Hierarchical Bayesian Modeling

Stochastic sampling in a ornithological hierarchical Bayesian model

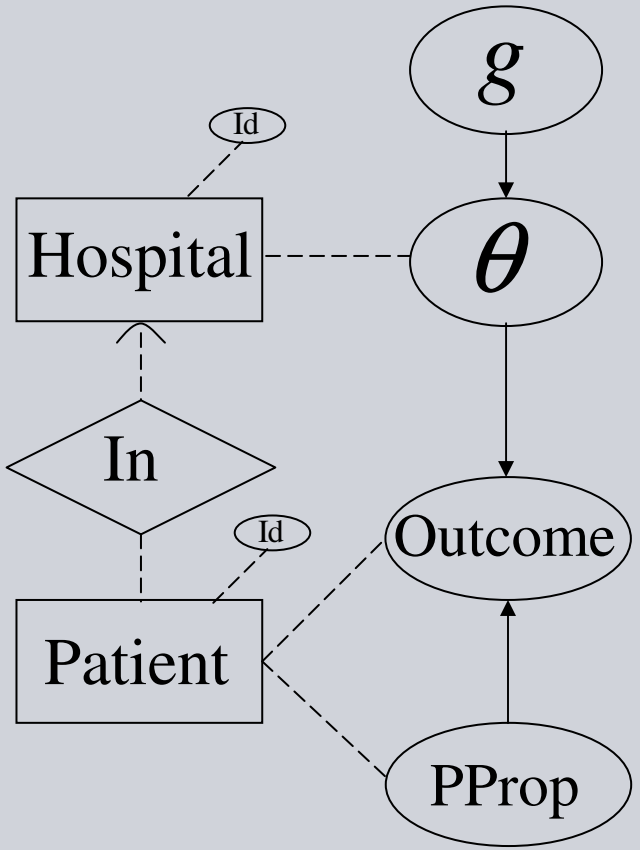
## Learning with Related Tasks

- In many applications different situations might be related but are not identical:
  - Patients are in different hospitals
  - The outcome might depend on unknown attributes of the hospital
- Somehow the Id of the hospital should influence the outcome
- Simply taking the Id as input leads easily to over fitting and models with bad generalization



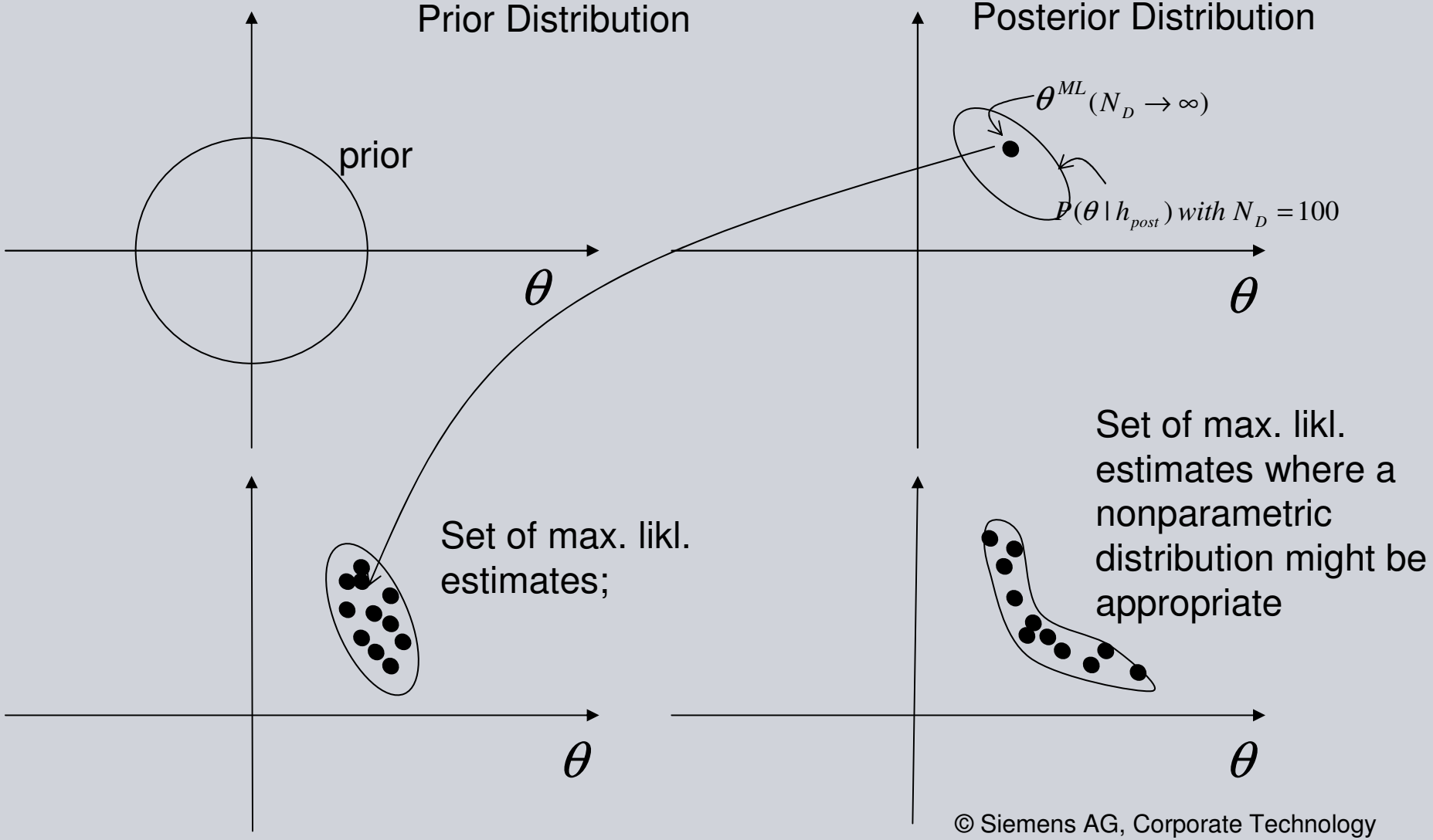
## A Hierarchical Bayesian Model

- In hierarchical Bayesian (HB) modeling, it is assumed that the *parameters* for the outcome prediction in different hospitals are generated from the same prior distribution, but otherwise are independent
- The *hyperparameters g in the prior* distribution are learned (by adapting g); we achieve an informative prior; thus knowledge can be shared between hospitals and can be transferred to a new hospital
- Great flexibility is assumed if we use a nonparametric prior distribution (generated from a Dirichlet process)



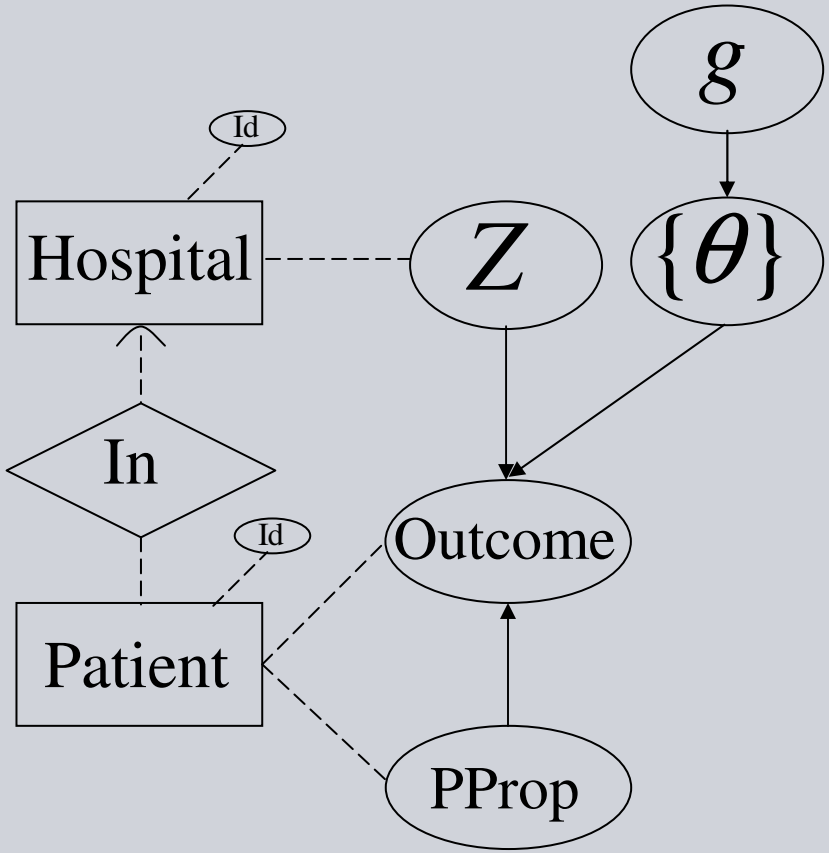

$$G \sim \mathrm{DP}(G_0, \alpha_0)$$

## Parametric HB is too Stiff!



## A Mixture Model

- Alternatively one might assume that the hospital belongs to particular hospital cluster and the cluster influences the Outcome
- We can let the numbers of clusters go to infinity and obtain the Dirichlet process mixture (DPM) model
- ***The DPM is a nonparametric hierarchical Bayesian approach with a Dirichlet process prior!***
- In the Gibbs sampling process (e.g., Chinese restaurant process), the number of (true?) clusters is determined automatically
- Large cluster / individual cluster



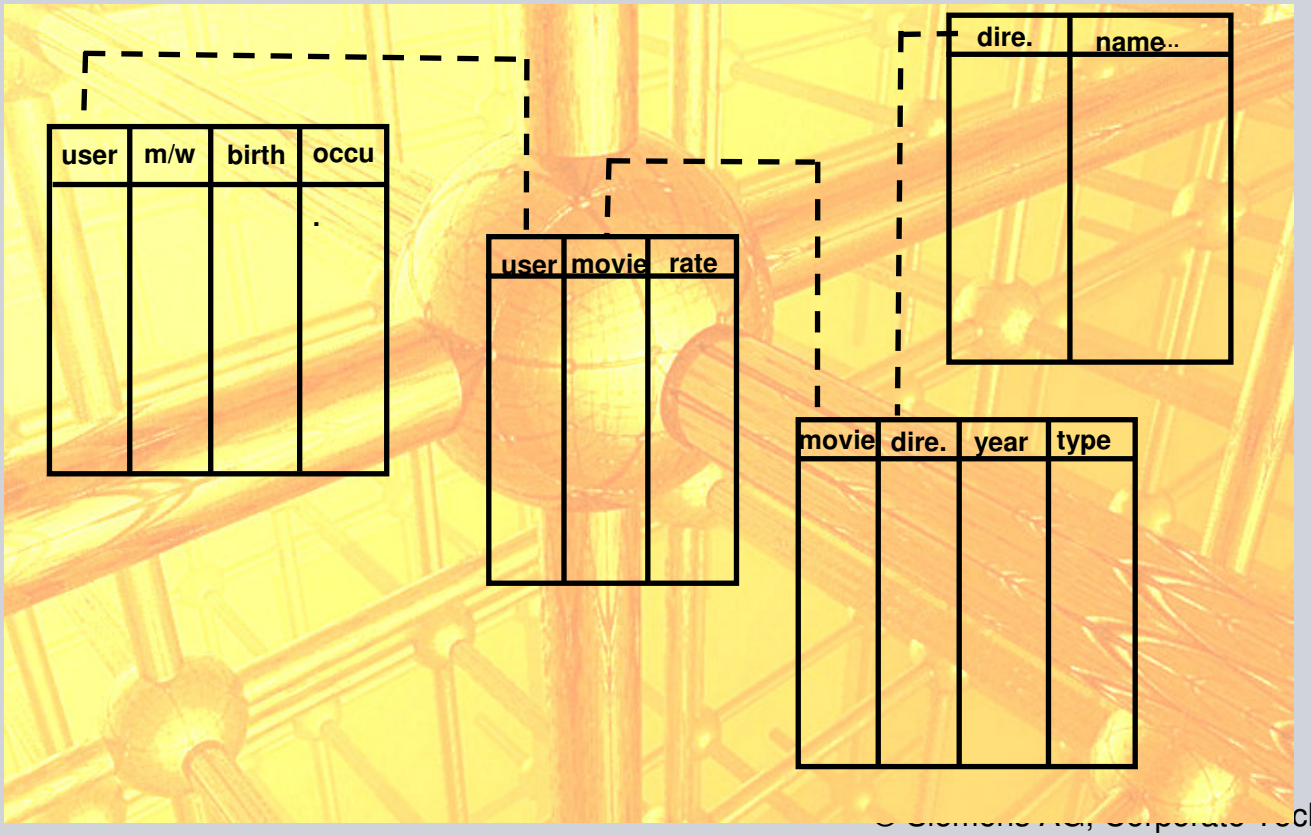
# III Relational Modeling and Learning

*Statistical machine learning is in the midst of a "relational revolution"*

*T. Dietterich*

# Learning with Relational Data

**Not surprisingly, relational data are often stored in a relational data base: both the relational model and the entity relationship model ER are useful description of the structure of a database (DB)**

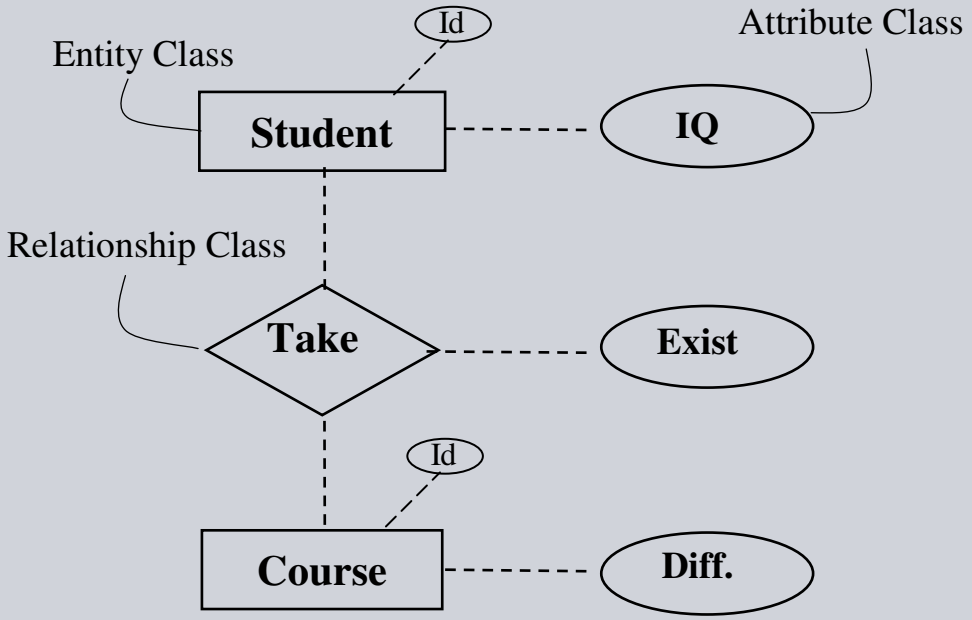


# Entity Relationship Model

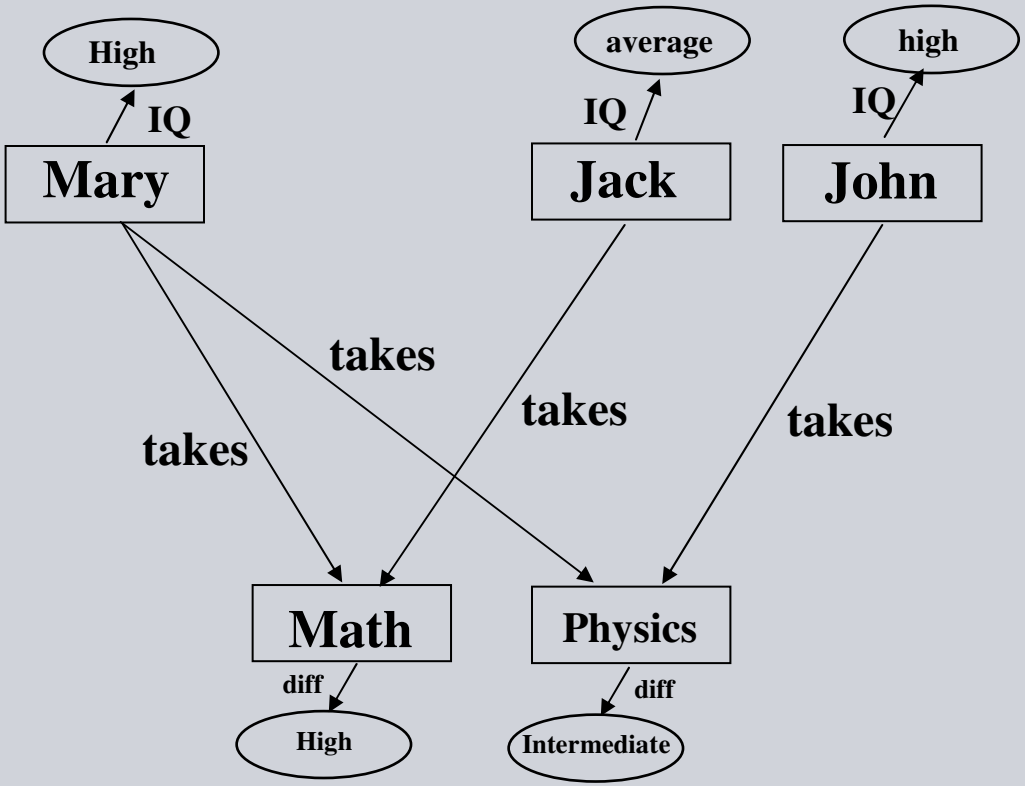
- The ER model is a concise description of a data base schema
- Very general and powerful

Main Components:

- Entity class
- Relationship class
- Attribute class



# Representing Ground Facts



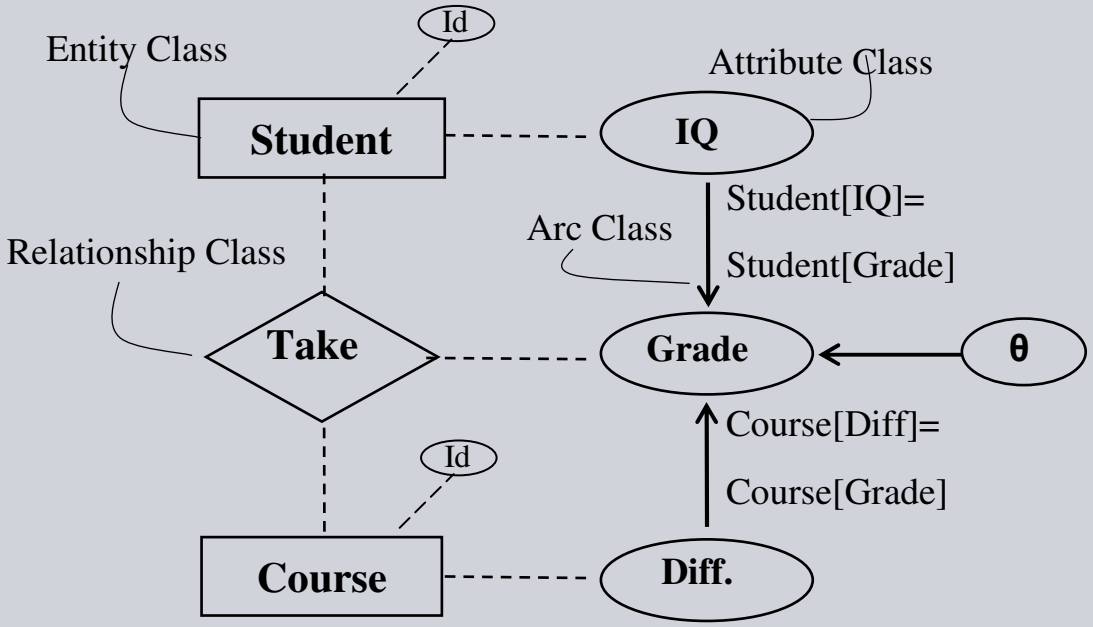

For binary and unary relations, the ground facts can graphically be described as

- Resource Description Framework (RDF)-graph used in the Semantic Web
- Sociogram used in social network analysis

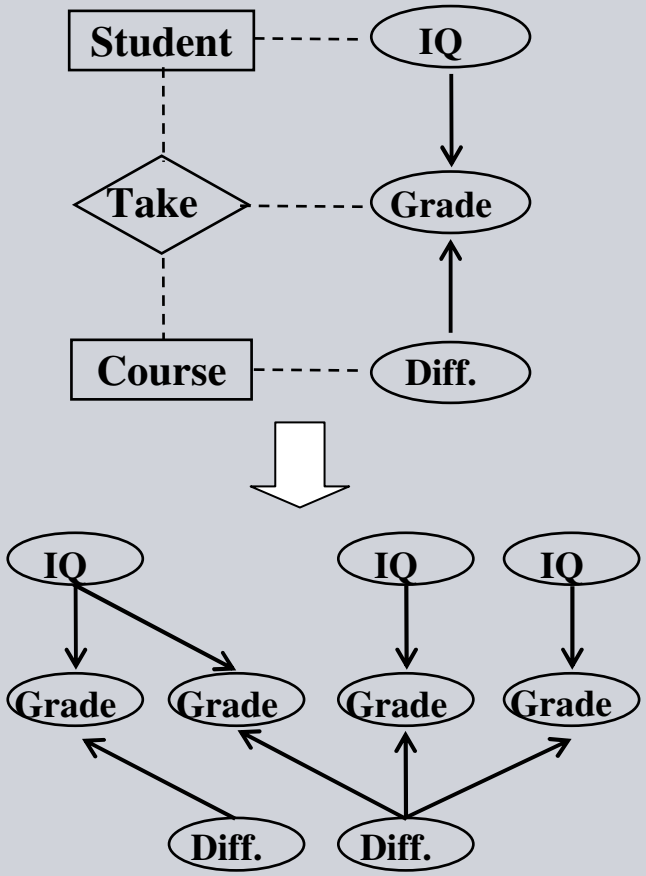
# Directed Acyclic Probabilistic Entity Relationship (DAPER) Model

**In the DAPER model [Heckerman, et al, 2004], probabilistic constraints are formulated at the level of an ER model (class level) and act as a template for forming the ground DAG**

- Entity class
- Relationship class
- Attribute class
- **Arc class**
- **Local distribution class**
- **Constraint class (constraints among attributes)**



# DAPER and Ground Networks



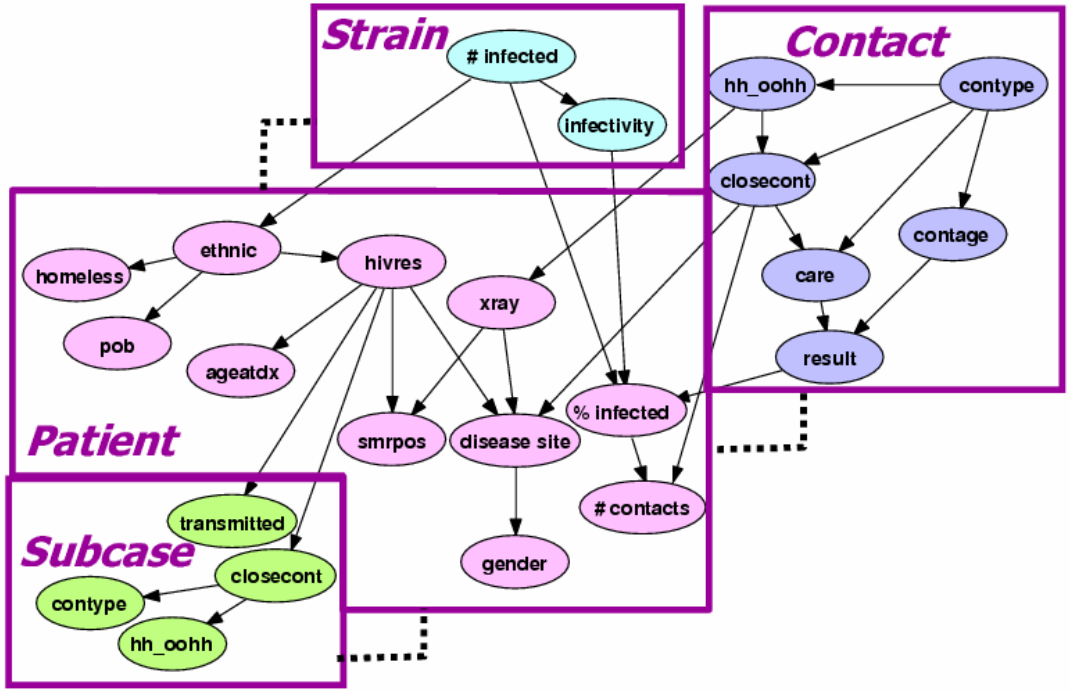

**DAPER describes a template**

**Ground Bayesian Network**

# Structural Learning in Relational Modeling

In many applications it is unreasonable to assume that the probabilistic dependency structure is known

Considerable work in PRM modeling has been devoted to structural learning




Structural learning in relational models is more involved than on non-relational Bayesian networks, due to the explosion in possible attributes candidates as parents

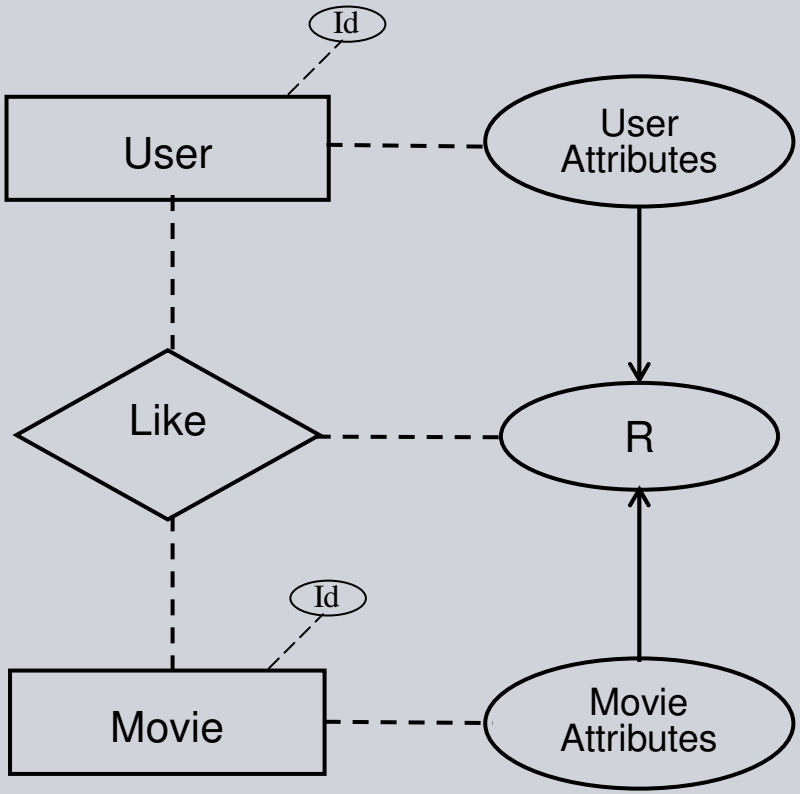
Typically a Bayesian score is optimized using some reasonable search strategy

# IV Infinite Hidden Relational Modeling: Combining Relational Learning with nonparametric Hierarchical Bayes

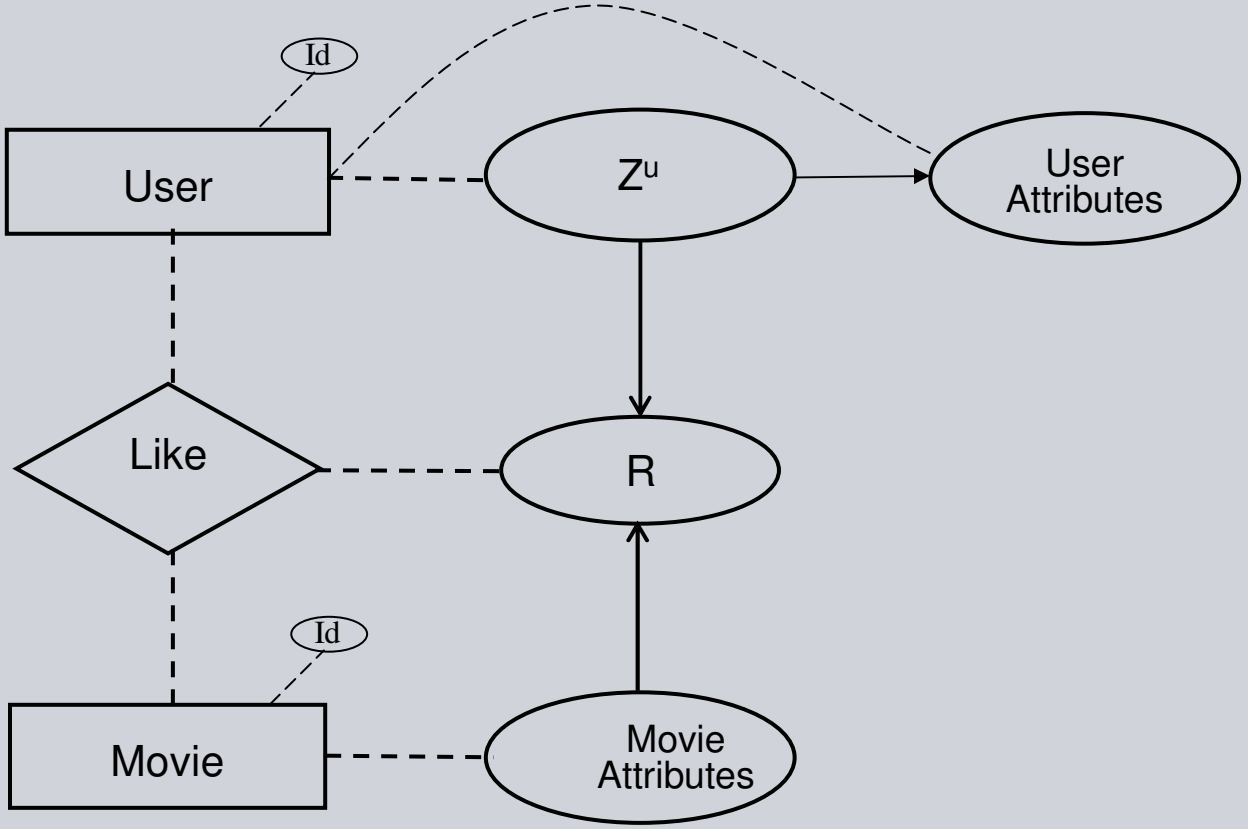
## Hierarchical Bayes and Relational Learning

- Probabilistic relational models (PRM, DAPER) provide templates leading to parameter sharing in the ground BN
- This might be too stiff for many applications
- We have seen how hierarchical Bayesian modeling allowed parameters to be personalized in a sensible way: patient outcome could have some hospital specific effects
- *Thus the natural question is how to generalize HB to relational modeling*
- If the parameter dependency is relational, a parametric HB approach is quite difficult to conceive: one would have to define a prior distribution whose hyperparameters depend on two or more entities
- Fortunately, a nonparametric HB approach is much easier to generalize!

# Relationship Prediction with Strong Attributes

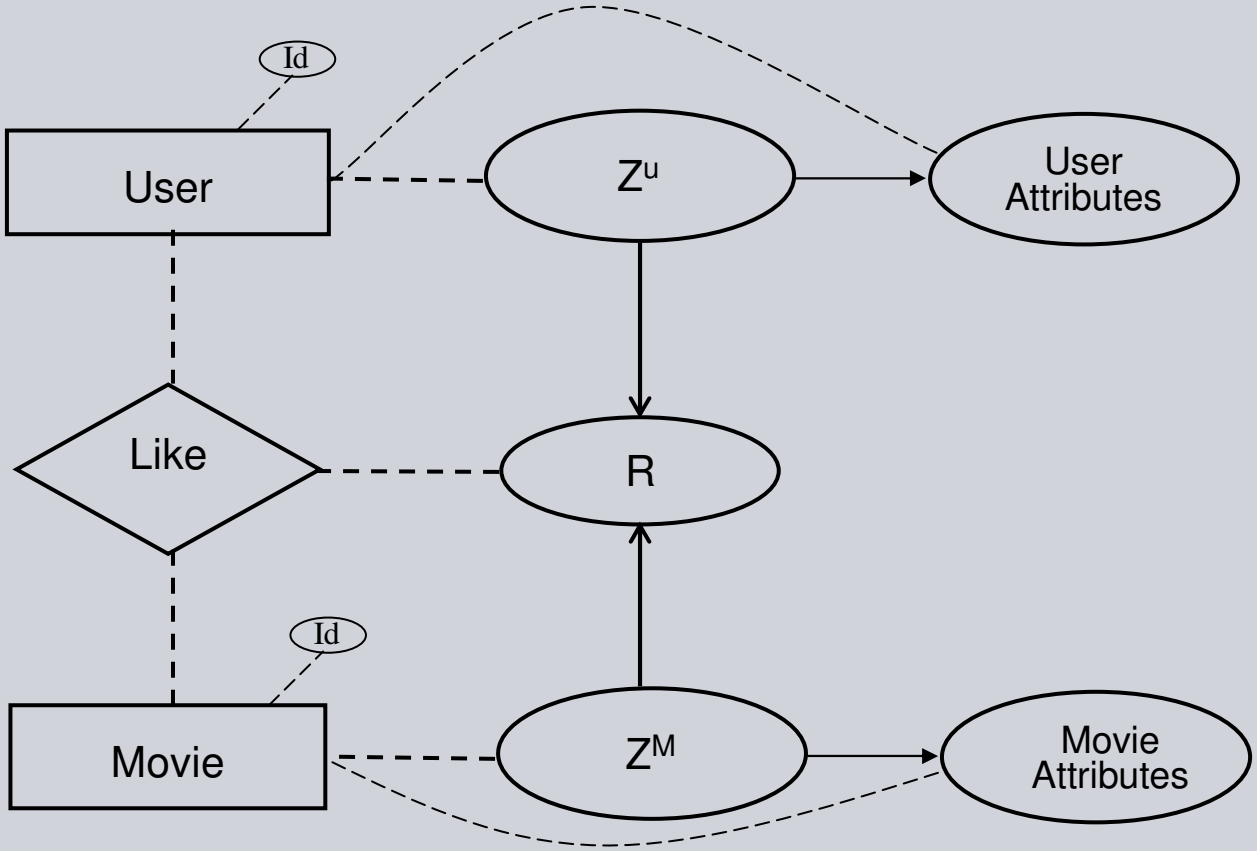


# Relationship Prediction with Weak (or no) User Attributes: nonparametric Hierarchical Bayes



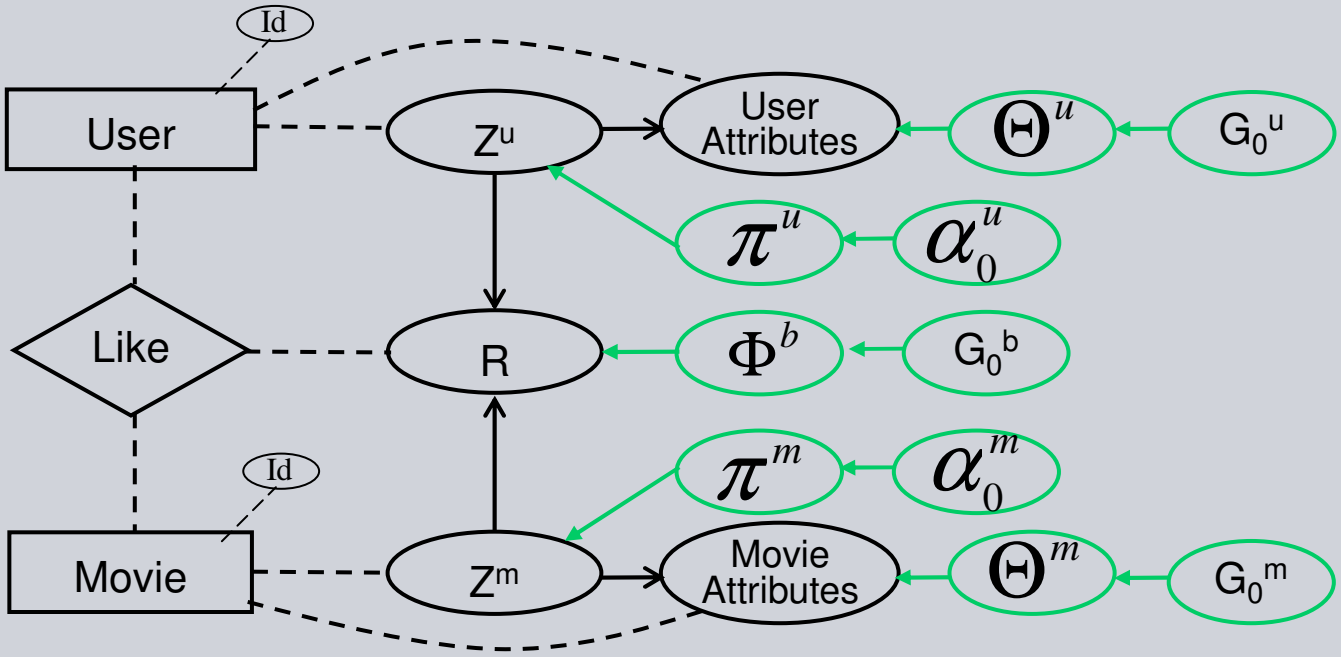
# Nonparametric Relational Bayes: Infinite Hidden Relational Model

**Key-Slide of the Talk!**




**Interacting DPM**

## IHRM with Parameters



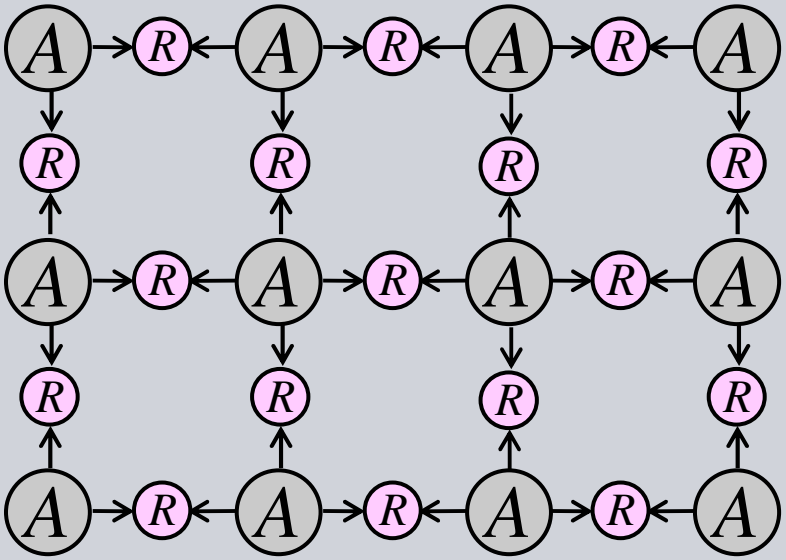

$$R^b \sim \mathrm{Mult}(\phi^b(f Z_i^{c_i} g_{i=1}^{M_b}))$$

## The Recipe

- To each entity an infinite latent variable, specific to each entity class, is assigned
- This latent variable is the parent of the (remaining) attributes of the entity
- The parents of the attributes of a relationship are the latent variables of the associated entities

- But isn't this too limited? The model implies local dependencies following the relationship structure
- Not necessarily: information can propagate through the network of latent variables

## Ground Network With an Image Structure



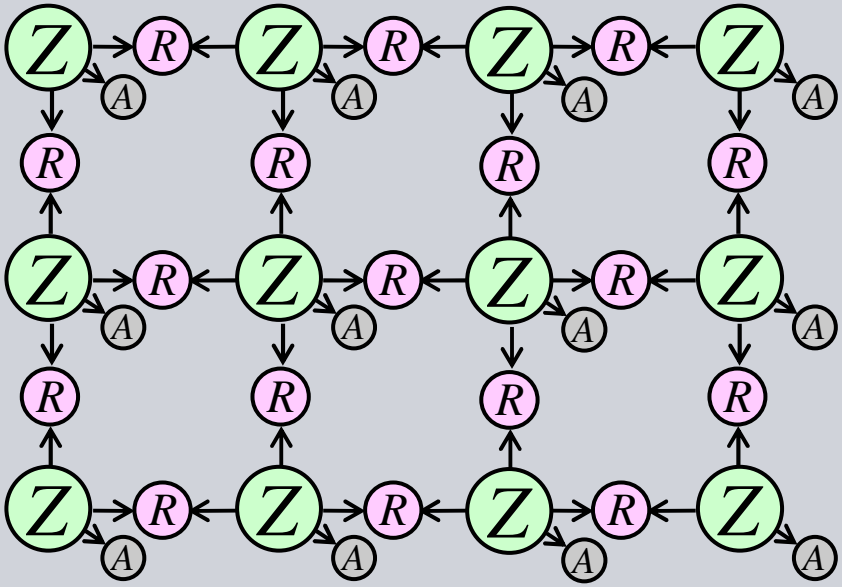

Ground Network

- A: entity attributes
- R: relational attributes (e.g., exist, not exist)

Limitations

- Attributes locally predict the probability of a relational attribute
- Given the parent attributes, all relational attributes are independent

# Ground Network With an Image Structure and Latent Variables: The IHRM




Latent class membership (roles) for two entities tends to be the same if the two entities have comparable relationships to entities with comparable latent class memberships (roles) and if attributes are similar

Thus the John Donne principle “everything depends on everything” is a consequence of the “we never know it all” principle

## Work on Latent Class Relational Learning

**IRM** ∞

C. Kemp, T. Griffiths, and J. R. Tenenbaum (2004). Discovering Latent Classes in Relational Data (Technical Report AI Memo 2004-019)

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. AAAI 2006

**IHRM** ∞

Z. Xu, V. Tresp, K. Yu, S. Yu, and H.-P. Kriegel (2005). Dirichlet enhanced relational learning. In Proc. 22nd ICML, 1004-1011. ACM Press

Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel. Infinite hidden relational models. In Proc. 22nd UAI, 2006

**"HRM"**

K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. JASA 96(455), 2001

Jake M. Hofman, Chris H. Wiggins . A Bayesian Approach to Network Modularity (NIPS*2007)

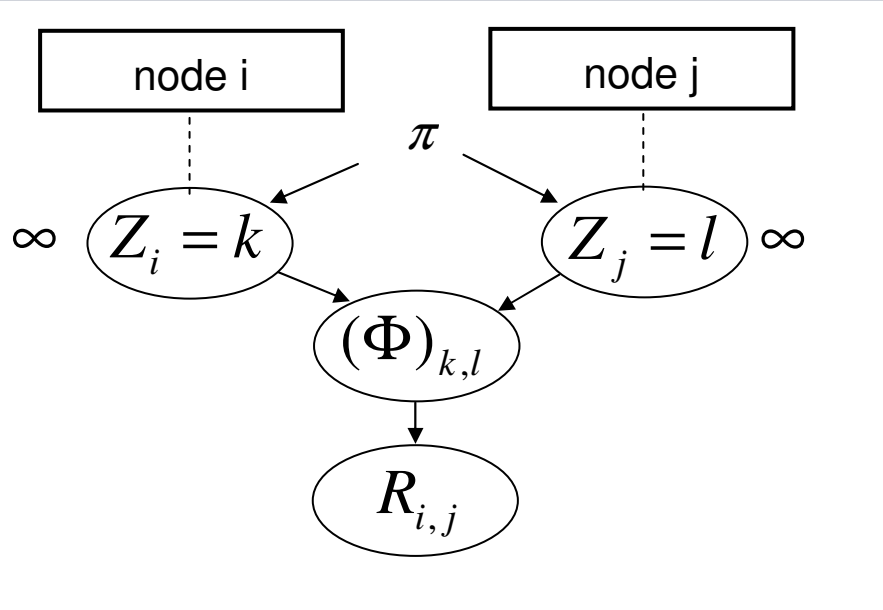
**MMSP**

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2006) Mixed-membership stochastic block models for relational data with application to protein-protein interaction. In Proceed. of the Intern. Biometrics Society Annual Meeting.

∞

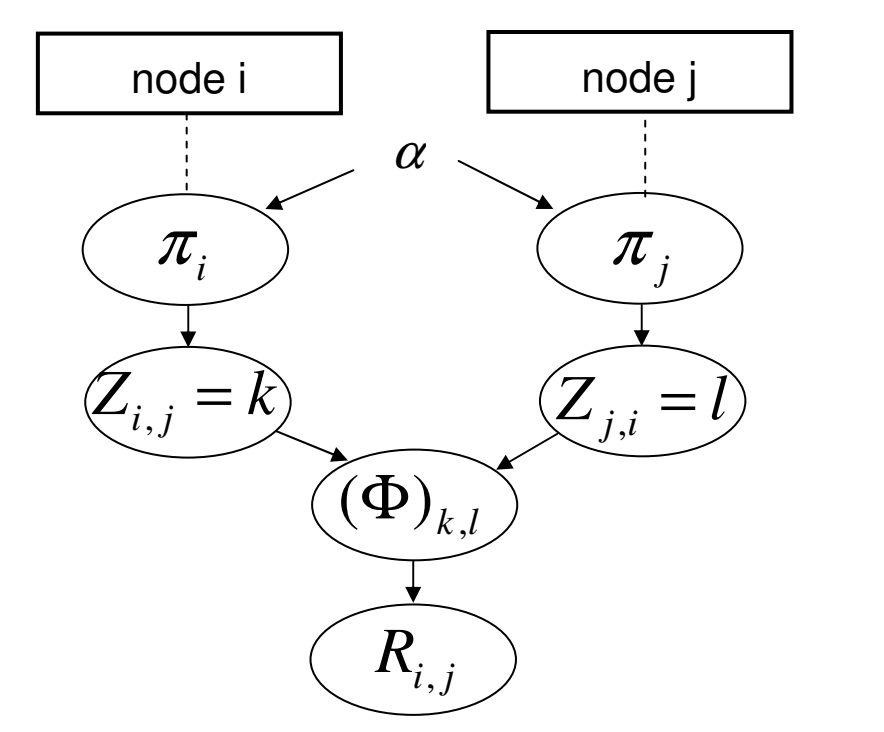
J. Sinkkonen, J. Aukia, and S. Kaski. Inferring vertex properties from topology in large networks. MLG 2007.

## The Generative Model (IHRM)
## Single Entity class; one relation class




- The ground truth is that each node belongs to exactly one class
- The states of the latent variables determines which Bernoulli parameter is selected
- Class membership of both relations determine the probability the existence of a relation

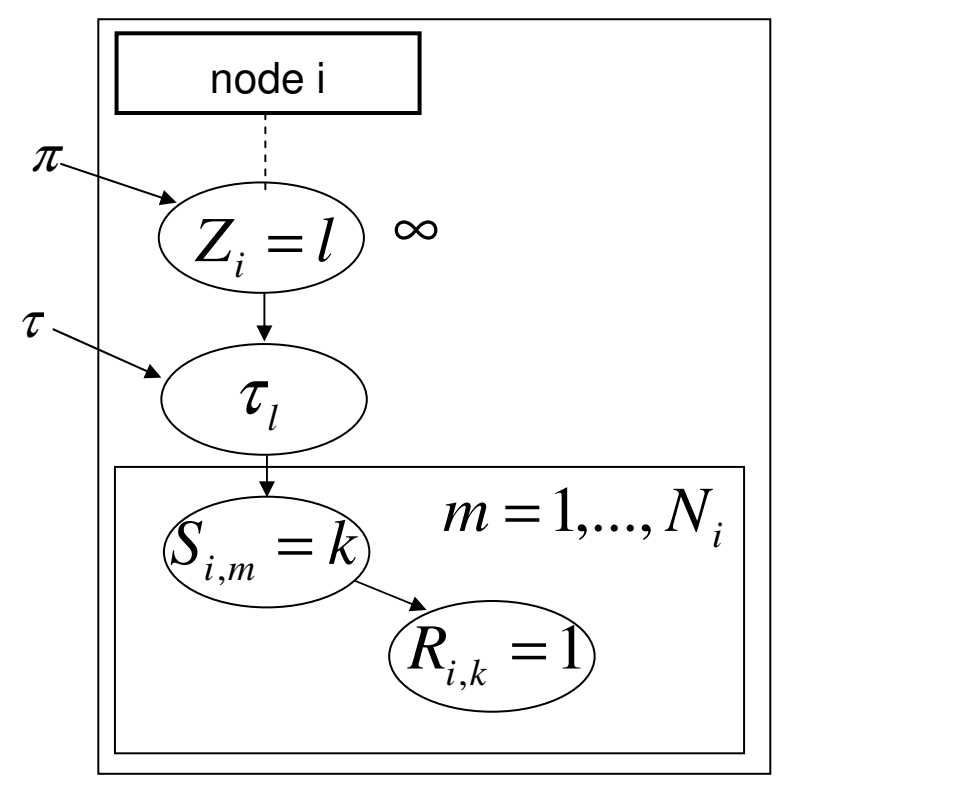
## The Generative Model (MMSB)




- Associated to each node i is a multinomial parameter vector $\pi_i$
- For each link to be formed two multinomial variables are sampled
- The states of the latent variables determines which Bernoulli parameter is selected
- This parameter determines the probability for forming a link
- Note that the ground truth is that each node belongs to several classes (topics)

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2006) Mixed-membership stochastic block models for relational data with application to protein-protein interaction. In Proceed. of the Intern. Biometrics Society Annual Meeting.

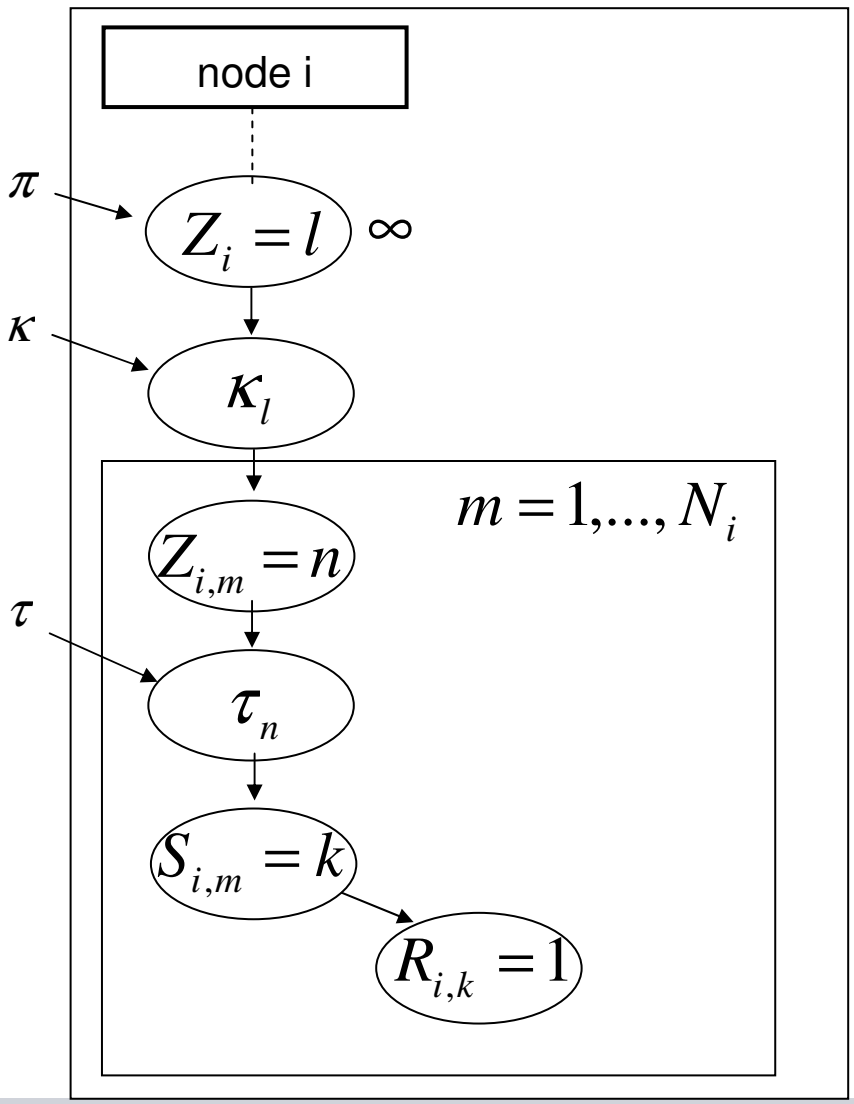
# The Generative Model (DERL)




- A node i belongs to one class l
- A multinomal parameter vector $\tau_l$ is selected
- $\tau_l$ determines the pd of the repeatedly sampled state of the multinomial selection variable $S_{i,m}$
- The state of $S_{i,m}$ determines to which node a link is formed (here: k)

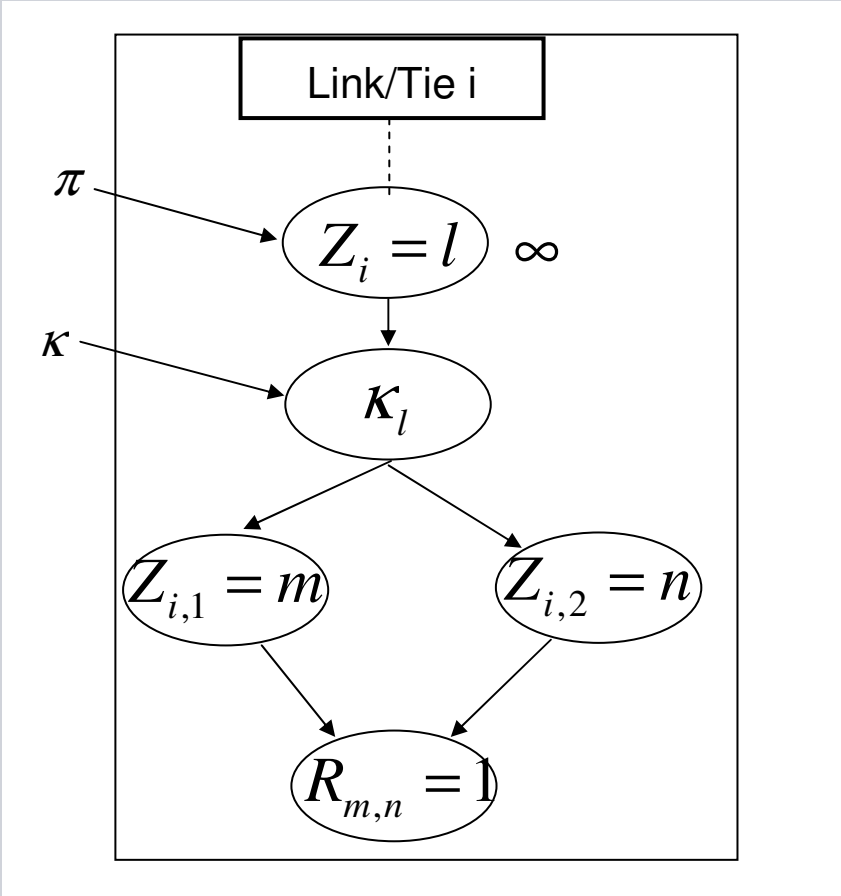
Z. Xu, V. Tresp, K. Yu, S. Yu, and H.-P. Kriegel (2005). Dirichlet enhanced relational learning. In Proc. 22nd ICML, 1004-1011. ACM Press

# The Generative Model (Mixed Membership DERL)




- This is a mixed membership model such that the multinomial parameter vector $\tau_n$ might vary for each link to be formed

## The Generative Model (Sinkkonen et al.)




- Each link I belongs to exactly one class l
- The multinomial parameter vector $\kappa_l$ is selected which determines the probabilities of the two latent variables
- The state of those latent variables determines which two nodes the link I is joining (here: m and n)
- Closely related to the PLSA model

J. Sinkkonen, J. Aukia, and S. Kaski. Inferring vertex properties from topology in large networks. MLG 2007.(this workshop)

# V Making it all work

## Inference in the IHRM

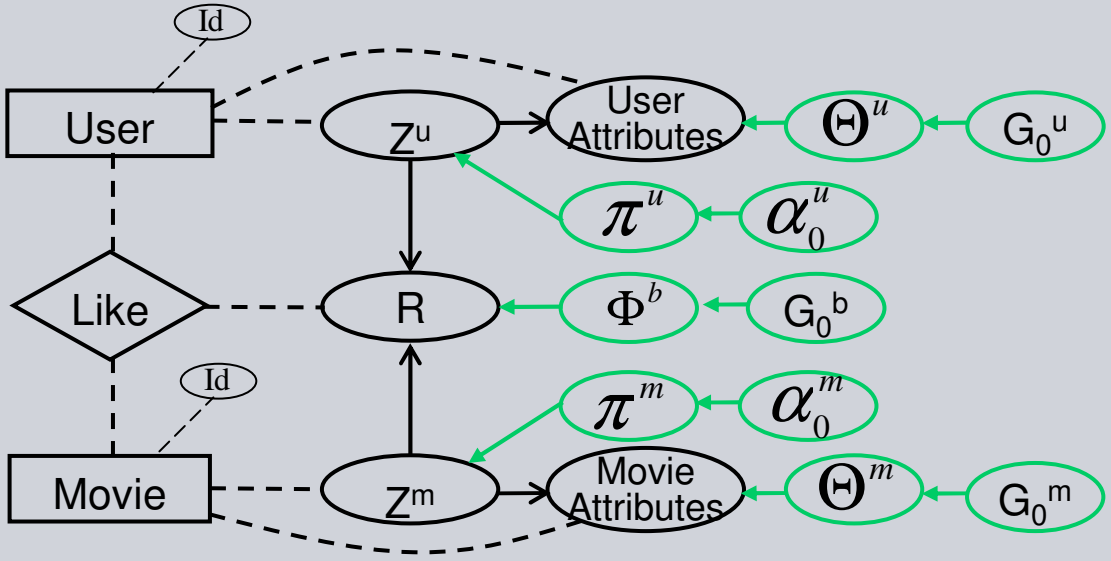
We derived and compared various inference and learning approaches

- Gibbs sampler derived from the Chinese restaurant process representation (Kemp et al. 2004, 2006, Xu et al. 2006);
- Gibbs sampler derived finite approximations to the stick breaking representation
  - Dirichlet multinomial allocation (DMA)
  - Truncated Dirichlet process (TDP)
- Two mean field approximations based on those two approximations
- A memory-based empirical approximation (EA)

## Experiment 1:
## Experimental Analysis on Movie Recommendation

Task description

- To predict whether a user likes a movie given attributes of users and movies, as well as known ratings of users.
- Data set: MovieLens



## MovieLens Attributes

| | | |
|---|---|---|
| User | Age (6) | >61; 60~46; 45~27; 26~19; 18~13; 12~4 |
| | Gender (2) | Female; Male |
| | Occupation (21) | Administrator; Artist; Doctor; Educator; Engineer; Entertainment; Executive; Healthcare; Homemaker; Lawyer; Librarian; marketing; None; Other; Programmer; Retired; Salesman; Scientist; Student; Technician; Writer; |
| Movie | Genre (18) | Action; Adventure; Animation; Children's; Comedy; Crime; Documentary; Drama; Fantasy; Film-Noir; Horror; Musical; Mystery; Romance; Sci-Fi; Thriller; War; Western |
| | Year (2) | 1998~1995; 1994~1990; 1989~1980; after 1979 |

## Experimental Analysis on Movie Recommendation

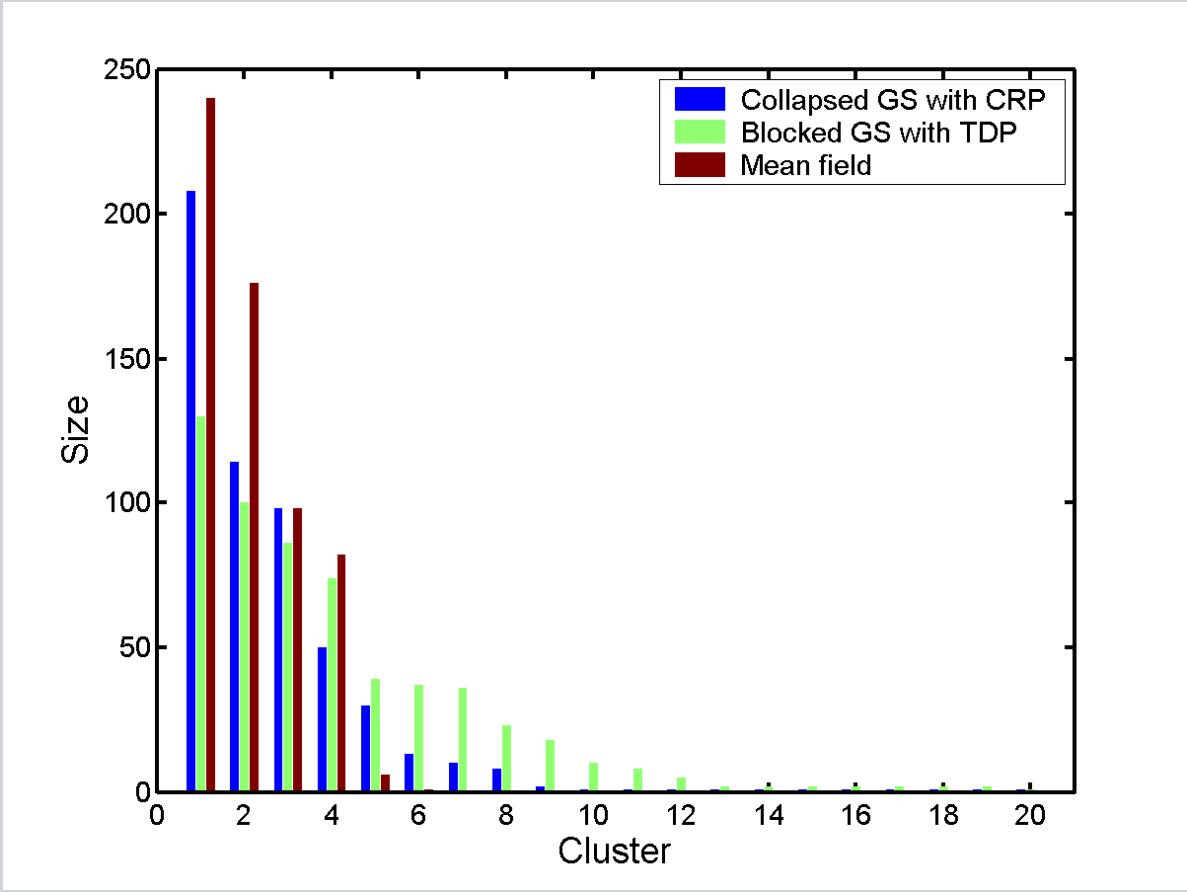
| Method | Prediction Accuracy (%) | | | | Time (s) | #Comp$^u$ | #Comp$^m$ |
|---|---|---|---|---|---|---|---|
| | given5 | given10 | given15 | given20 | | | |
| GS-CRP | 65.13 | 65.71 | 66.73 | 68.53 | 164993 | 47 | 77 |
| GS-TDP | 65.51 | 66.35 | 67.82 | 68.27 | 33770 | 59 | 44 |
| GS-DMA | 65.64 | 65.96 | 67.69 | 68.33 | 25295 | 52 | 34 |
| MF-TDP | 65.26 | 65.83 | 66.54 | 67.63 | 2892 | 9 | 6 |
| MF-DMA | 64.23 | 65.00 | 66.54 | 66.86 | 2893 | 8 | 12 |
| EA | 63.91 | 64.10 | 64.55 | 64.55 | 386 | --- | --- |

- Sampling based on the stick-breaking representation is faster than CRP-based Gibbs sampling since Z can be updated in a block; it also gave comparable performance
- Gibbs sampling finds many more components than mean field but only less than 10 have significant weight

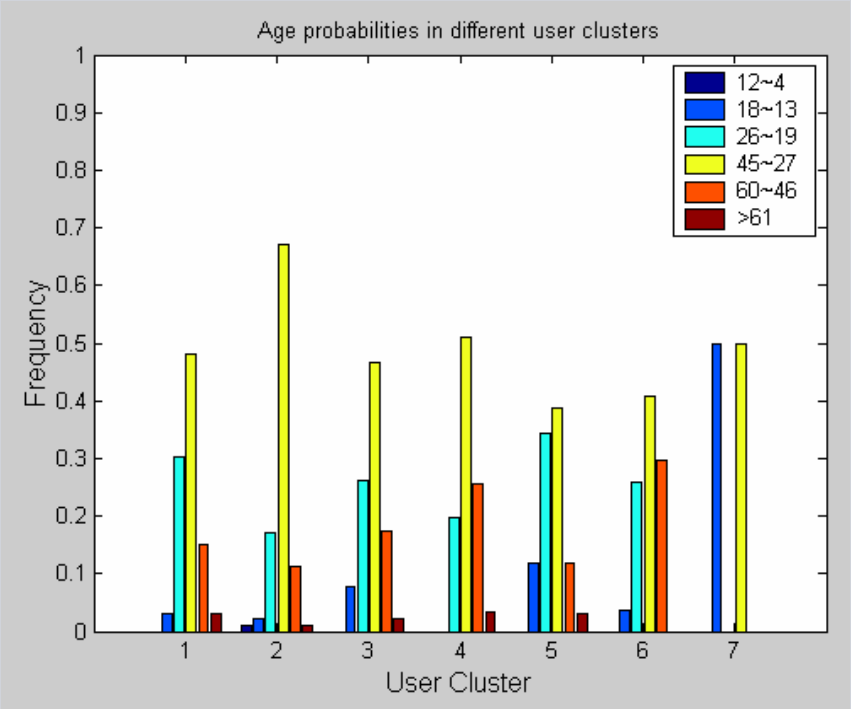
# Movie cluster analysis
# Gibbs sampling with CRP

| Cluster 1 (161/207)<br>very new and popular | Cluster 2 (76/113)<br>old, non US, drama | Cluster 3 (49/98)<br>comedy | Cluster 4 (32/51)<br>children |
|---|---|---|---|
| **My Best Friend's Wedding (1997)**<br>**G.I. Jane (1997)**<br>**The Truth About Cats & Dogs (1996)**<br>**Phenomenon (1996)**<br>**Up Close & Personal (1996)**<br>**Tin Cup (1996)**<br>**Bed of Roses (1996)**<br>**Sabrina (1995)**<br>**Clueless (1995)……** | **Big Night (1996)**<br>**Antonia's Line (1995)**<br>**Three Colors: Red (1994)**<br>**Three Colors: White (1994)**<br>**Cinema Paradiso(1989)**<br>**Henry V (1989)**<br>**Jean de Florette (1986)**<br>**A Clockwork Orange (1971)**<br>**Citizen Kane (1941)**<br>**Mr. Smith Goes to Washington (1939) ……** | **Swingers (1996)**<br>**Get Shorty (1995)**<br>**Mighty Aphrodite (1995)**<br>**Welcome to the Dollhouse (1995)**<br>**Clerks (1994)**<br>**Ed Wood (1994)**<br>**The Hudsucker Proxy (1994)**<br>**What's Eating Gilbert Grape (1993)**<br>**Groundhog Day (1993)……** | **Event Horizon (1997)**<br>**Batman & Robin (1997)**<br>**Escape from L.A. (1996)**<br>**Batman Forever (1995)**<br>**Batman Returns (1992)**<br>**101 Dalmatians (1996)**<br>**The First Wives Club (1996)**<br>**Nine Months (1995)**<br>**Casper (1995)**<br>**……** |
| Cluster 5 (16/27)<br>new action | Cluster 6 (9/15)<br>old action | Cluster 7 (8/13)<br>old drama | Cluster 8 (3/6)<br>H. Ford, Star Wars |
| **Conspiracy Theory (1997)**<br>**The Game (1997)**<br>**Air Force One (1997)**<br>**Ransom (1996)**<br>**The Rock (1996)**<br>**Primal Fear (1996)**<br>**Crimson Tide (1995)**<br>**In the Line of Fire (1993)**<br>**The Abyss (1989)**<br>**……** | **Brave Heart (1995)**<br>**Forrest Gump (1994)**<br>**Fugitive (1993)**<br>**Terminator 2: Judgment Day (1991)**<br>**Indiana Jones and the Last Crusade (1989)**<br>**Die Hard (1988)**<br>**Aliens (1986)**<br>**Terminator (1984)**<br>**Return of the Jedi (1983)** | **Shawshank Redemption (1994)**<br>**Wrong Trousers (1993)**<br>**Schindler's List (1993)**<br>**Silence of the Lambs (1991)**<br>**One Flew Over the Cuckoo's Nest (1975)**<br>**Godfather (1972)**<br>**Rear Window (1954)**<br>**Casablanca (1942)** | **Star Wars (1977)**<br>**Star Wars: The Empire Strikes Back (1980)**<br>**Raiders of the Lost Ark (1981)** |

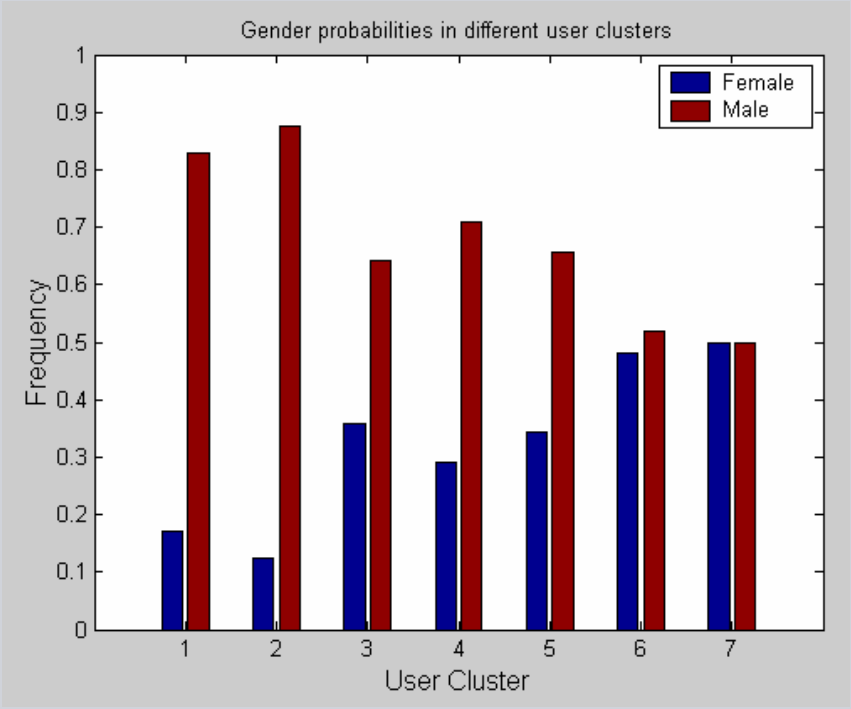
# Movie cluster analysis
# Gibbs sampling with CRP (2)



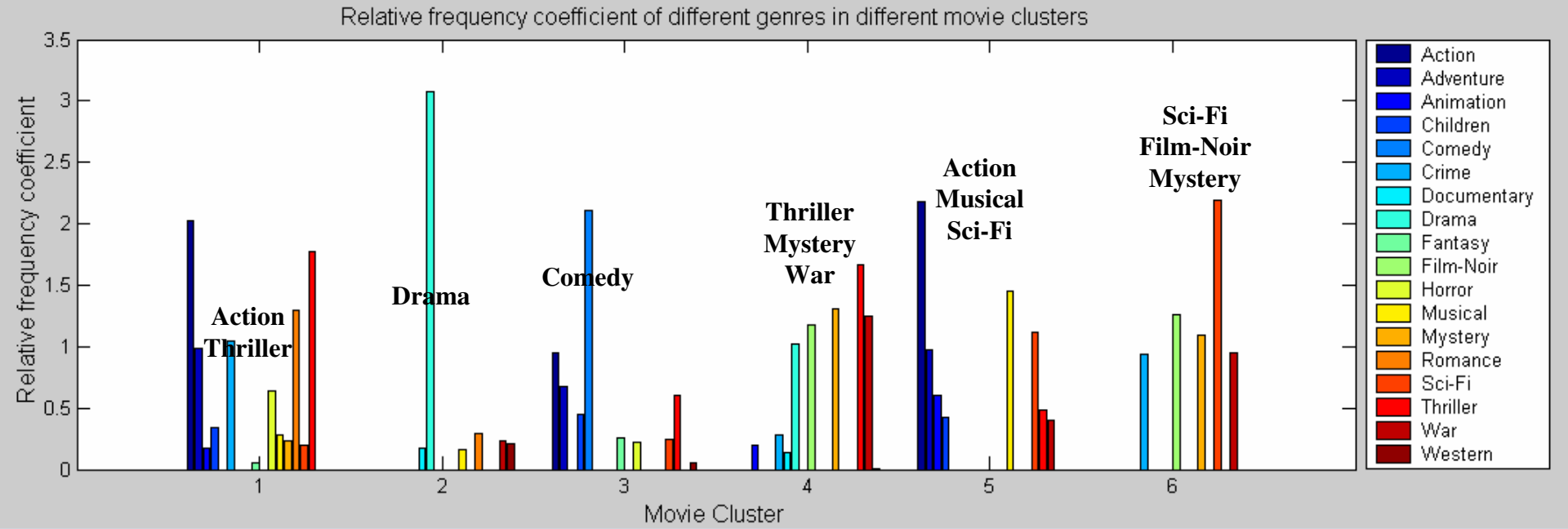
# User Attributes and User Clusters



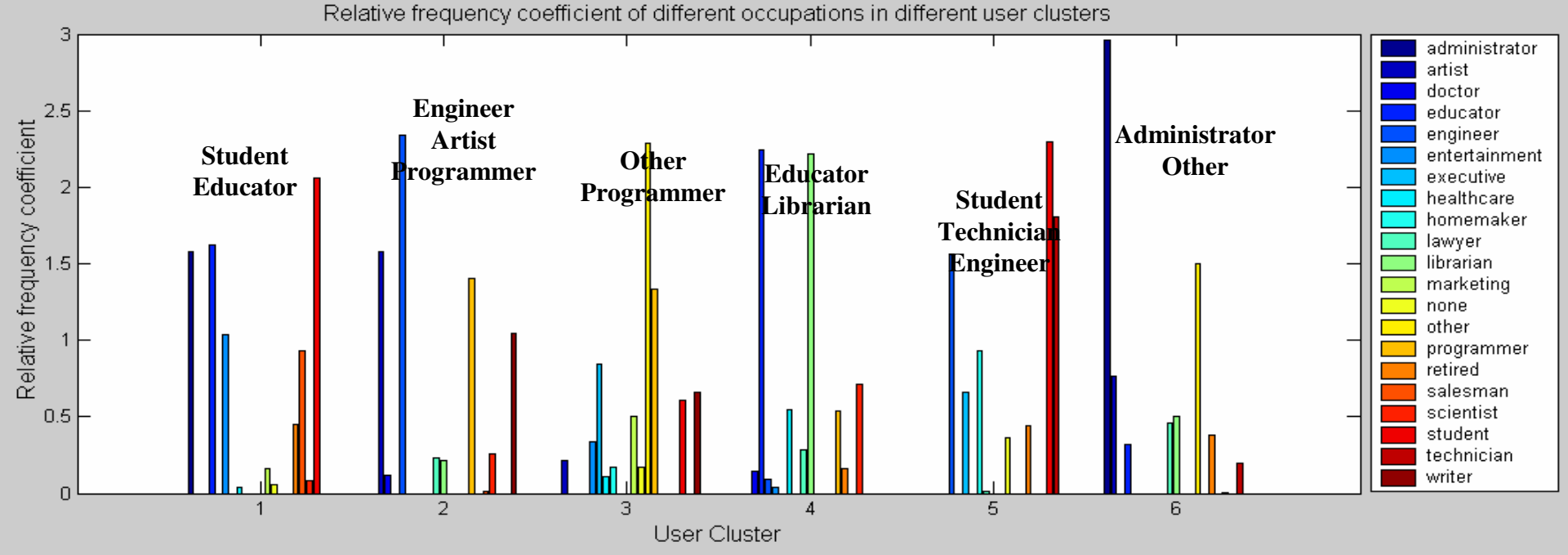

**Age frequency**



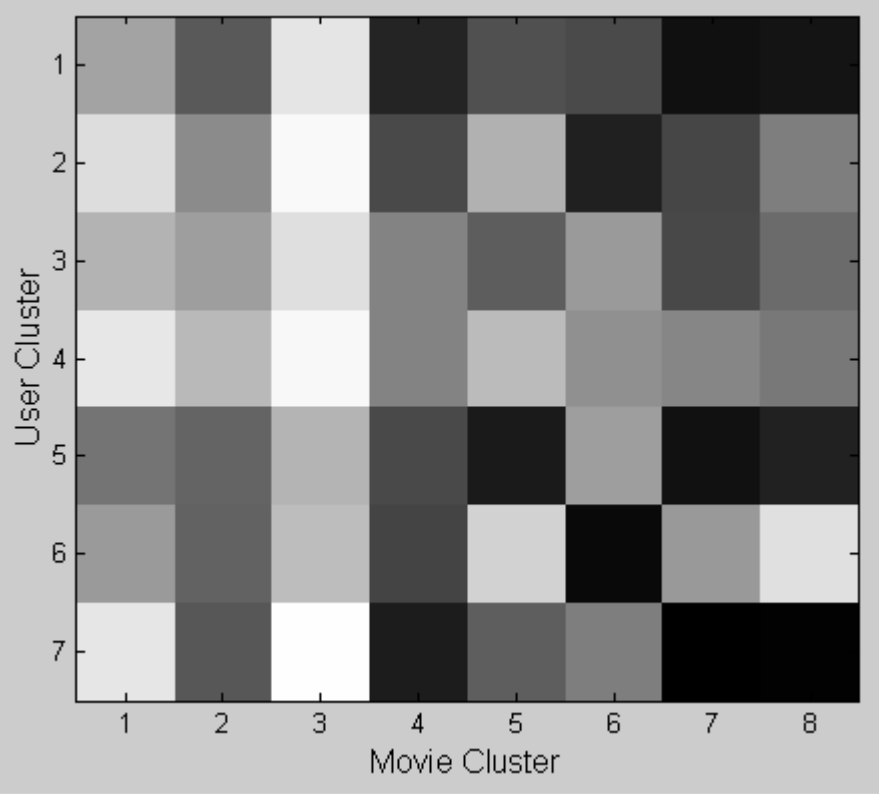

**Gender frequency**



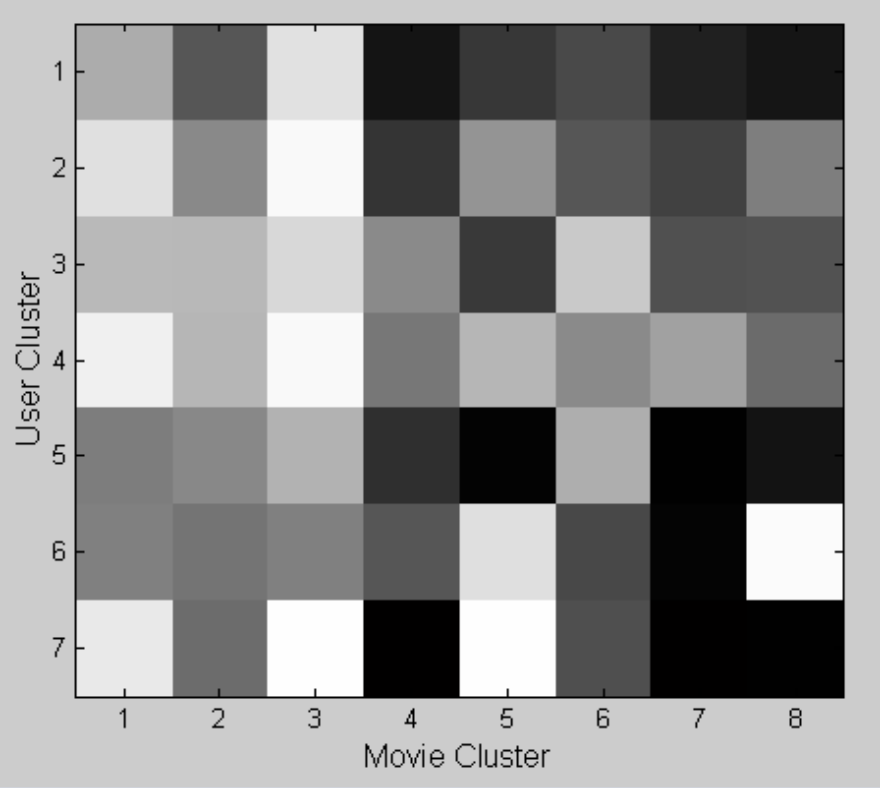

Difference to mean distribution



# User Clusters versus Movie Clusters




**All attributes and relations**




**Only relations**

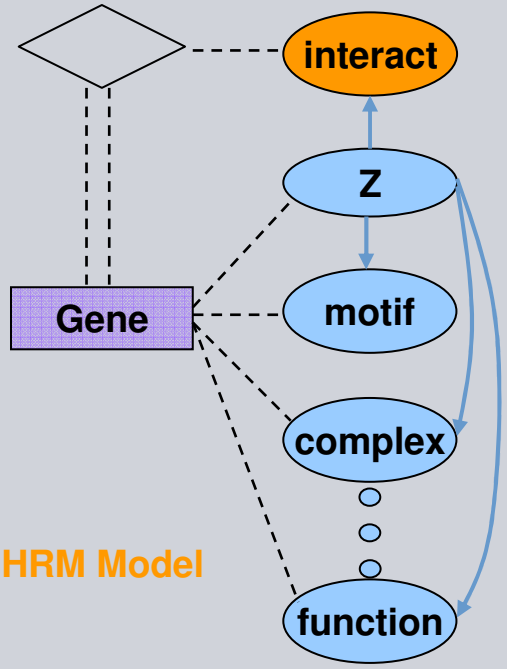
## Experiment 2:
## Gene Interaction and Gene Function

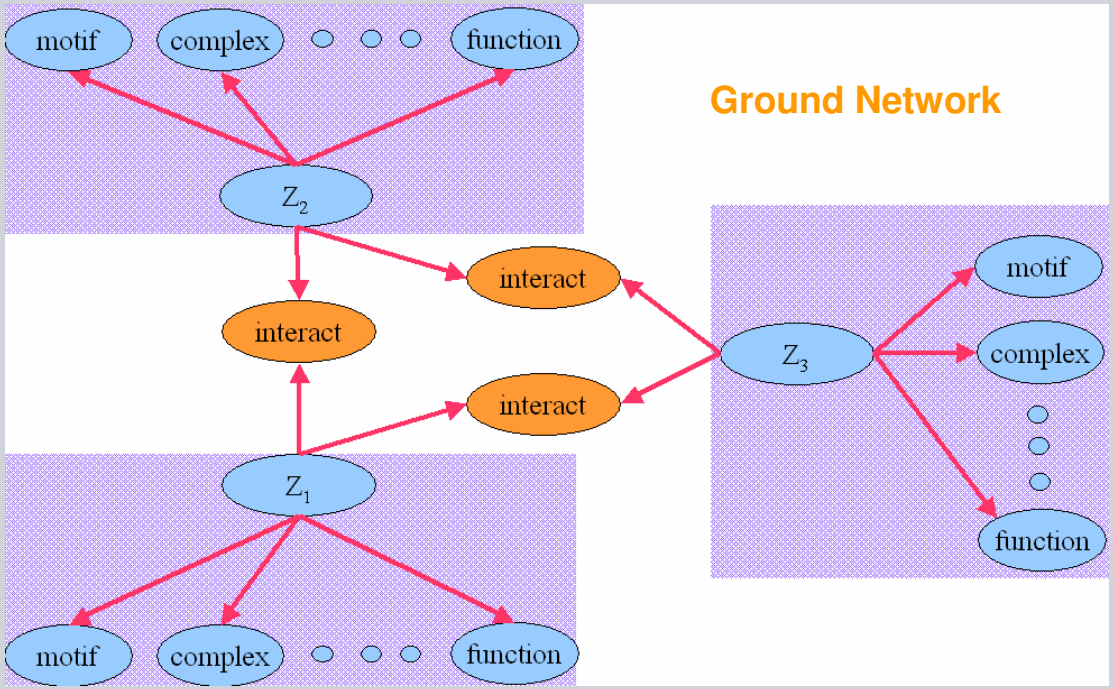
**Task**
- Cluster analysis
- Prediction of gene functions given the information on the gene level and the protein level, as well as the interaction between the genes.

**Attribute data: CYGD** (Comprehensive Yeast Genome Database) from MIPS (Munich Information Center for Protein Sequences)
- 1000 Genes
- Attributes: Chromosome, Motif, Essential, Class, Phenotype, Complex, Function

**Interaction data:** DIP (data base of interacting proteins)

# IHRM Model




Genes (1243) have one or more **functions** (14)[1-4] (cell growth, cell organization, transport, … ) to be predicted; 862 for genes for training, 381 for testing

Genes might **interact** with one another

For a gene one or more **phenotypes** (11)[1-6] are observed in the organism

How the expression of the gene can **complex** with others to form a larger protein (56)[1-3]

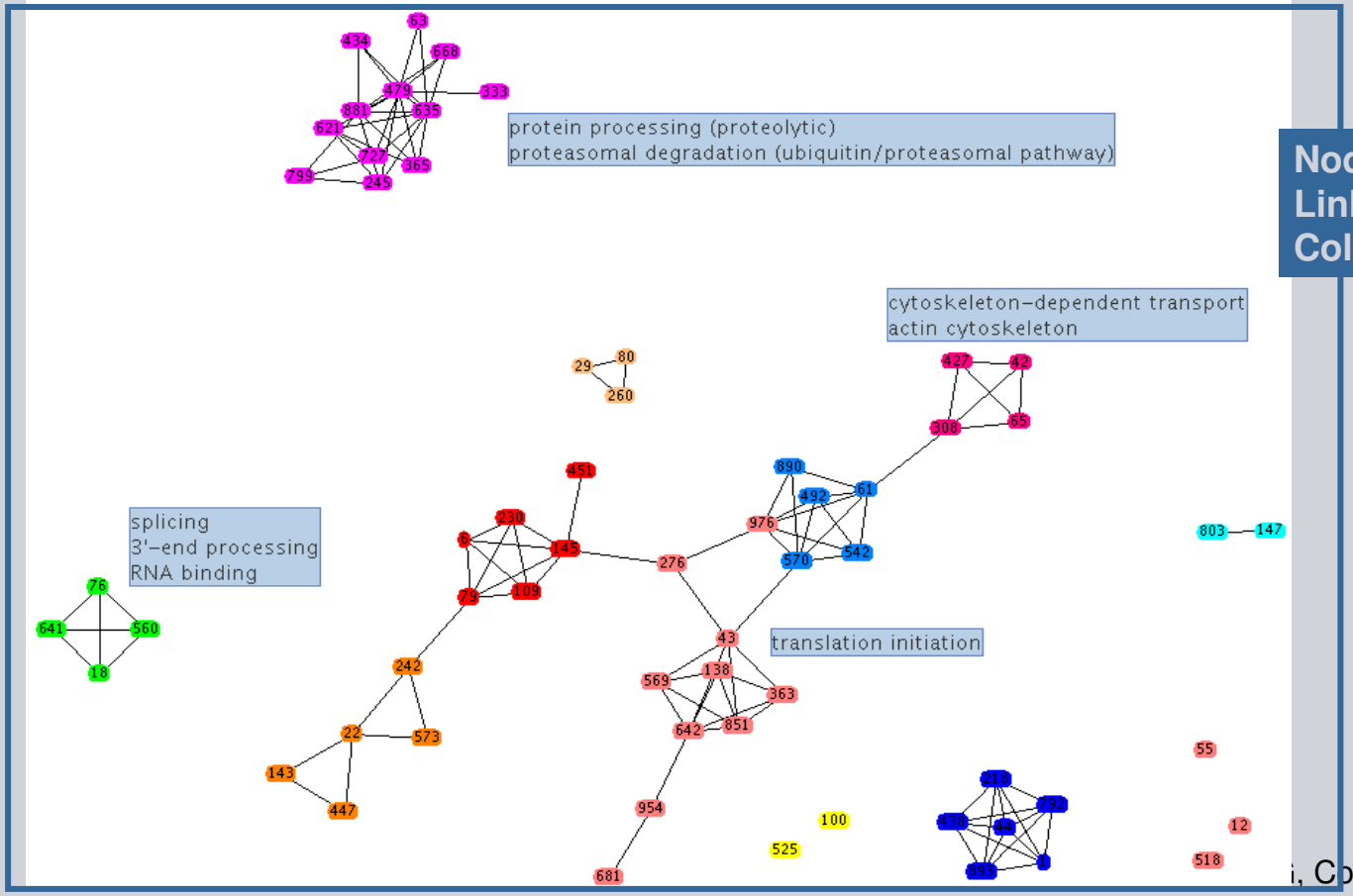
The protein coded by the gene might belong to one or more **structural categories** (24) [1-2]

A gene might contain one or more characteristic **motifs** (351) [1-6] (information about the amino acid sequence of the protein)

Gene **attributes** are: essential (an organism with a mutation can survive?), which chromosome

# Cluster Structure

Some gene clusters: the genes in the same cluster have dense interactions; but the genes in the different clusters have rare interactions.



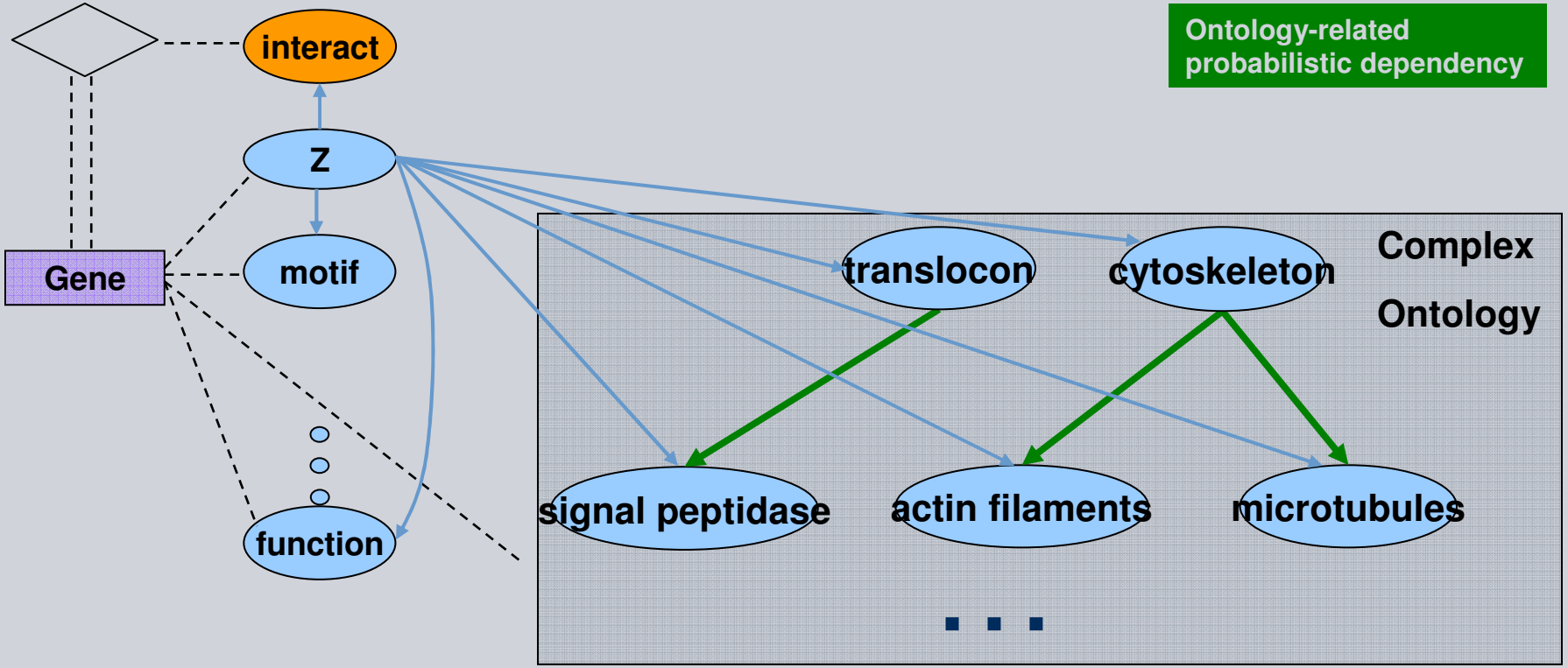
# Relevance of Attributes and Relationships

**The importance of a variety of relationships in function prediction of genes**

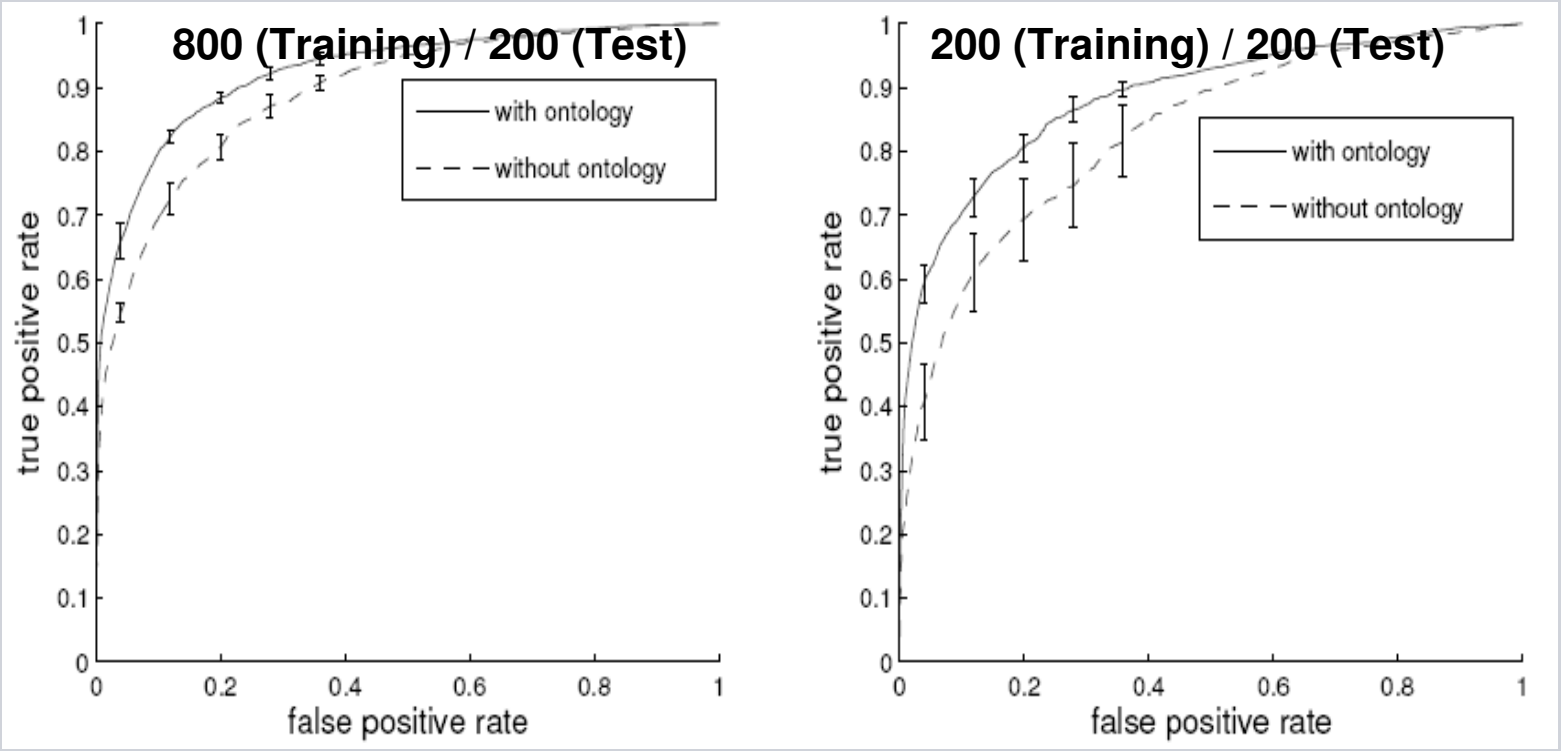
| Relationships | Prediction Accuracy (%) (without the relationship) | Importance |
|---|---|---|
| Complex | 91.13 | 197 |
| Interaction | 92.14 | 100 |
| Structural Category | 92.61 | 55 |
| Phenotype | 92.71 | 45 |
| Attributes of Gene | 93.08 | 10 |
| Motif | 93.12 | 6 |

# Ongoing Work: Integrate Ontology into IHRM (1)

**Ontologies are a valuable source of prior information**



# Ongoing Work: Integrate Ontology into IHRM (2)




| AUC | | |
|---|---|---|
| | Without Ontology: 0.89 | Without Ontology: 0.83 |
| | With Ontology: 0.93 | With Ontology: 0.89 |

## Experiment 3:
## Clinical Decision Support

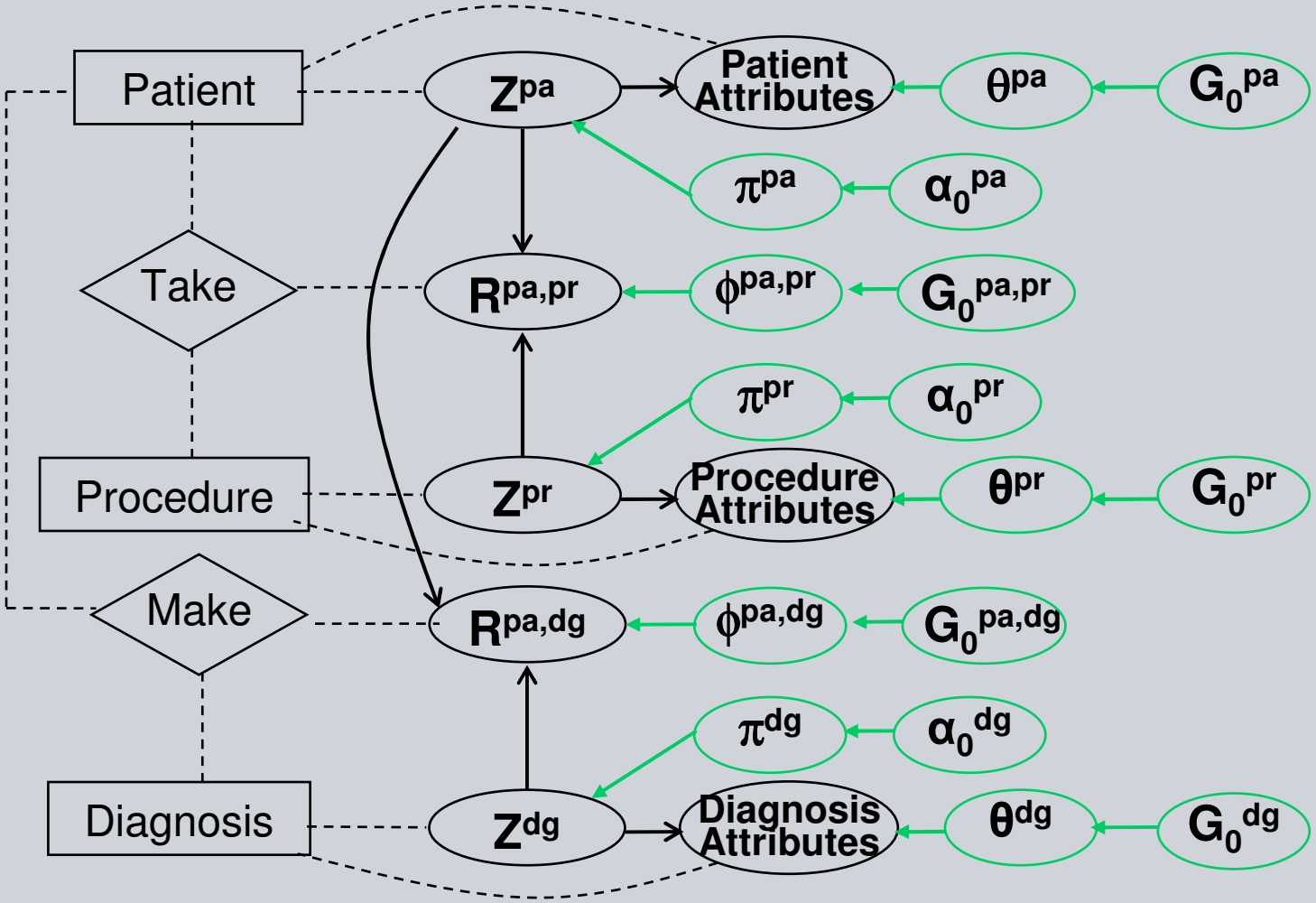
**Task description**

- To predict future procedures for patients given attributes of patients and procedures, as well as prescribed procedures and diagnosis of patients.

**Model**

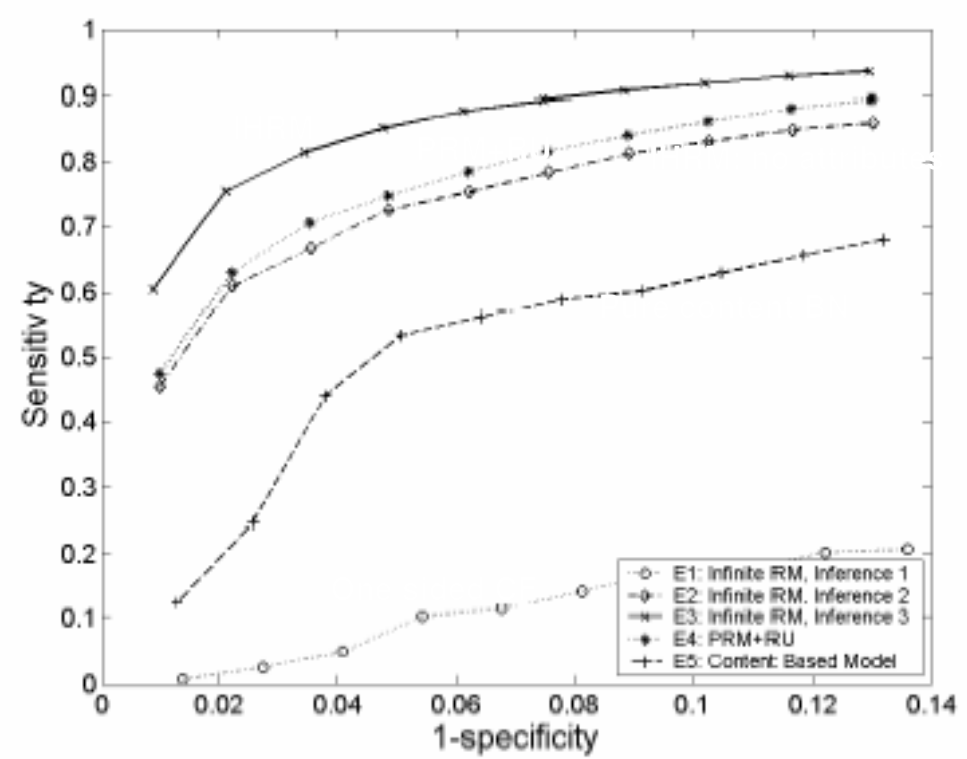
- Entity classes: Patient (14062), Diagnosis (704), Procedure (367)
- Relationship classes: Make (a diagnosis), Take (a procedure)
- A patient has typically multiple diagnosis and procedures
- Patient attributes: Age, Gender, Primary Complaint
- Diagnostic attributes: classes in ICD-9,
- Procedures: class as specified CPT4 code

# IHRM Model for Clinical Decision Support

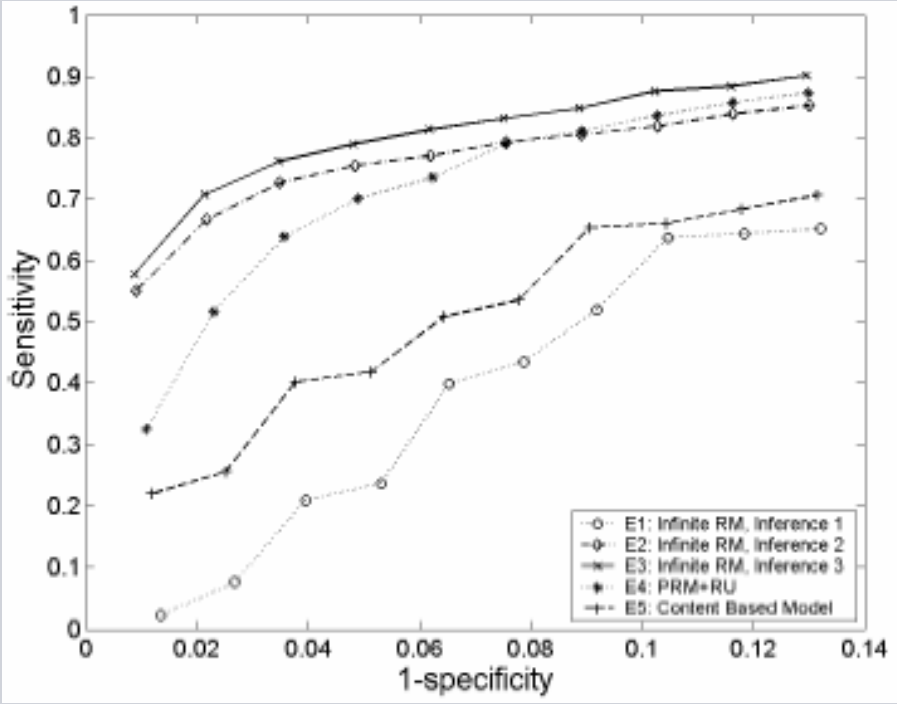


# Procedure Prediction: Given First Procedure

**Average on all patients**




**Only patients with prime complaint**
***circulatory problem***




**E3 (best): full IHRM [1]**
**E2: Same but without attributes [3]**
**E1: One-sided "collaborative" [5]**
**E5: Pure content based [4]**
**E4: PRM+RU [2]**

## Experiment 4:
## Context-Dependent Statistical Trust Learning: Who do you trust? When? [Rettinger, Nickles, Tresp; AAMAS 2008]

- The need for an **evaluation of trustworthiness of agents in future encounters** is getting increasingly important in distributed systems since contemporary developments such as the Semantic Web, Service Oriented Architectures, Pervasive Computing, Ubiquitous Computing and Grid Computing are applied mainly to open and dynamic systems with interacting autonomous agents
- Most existing statistical trust models do not perform well when there is no long history of interactions in a predefined and consistent environment

- We implement and learn **context sensitive trust** from past experience using a probabilistic relational model
  - A seller might be trustworthy if offering a specific product, but not another product.

- Being the most popular online auction and shopping website, fraud on eBay is a serious and well-known issue.
- eBay users leave feedback about their experiences

# Infinite Hidden Relational Trust Model

$ATT^a$
- % of positive ratings[2]
- eBays feedback score [5]
  - More than x number of positive ratings
- Member since

$ATT^s$
- Top eBayCategory[47]
- Condition [new/used]

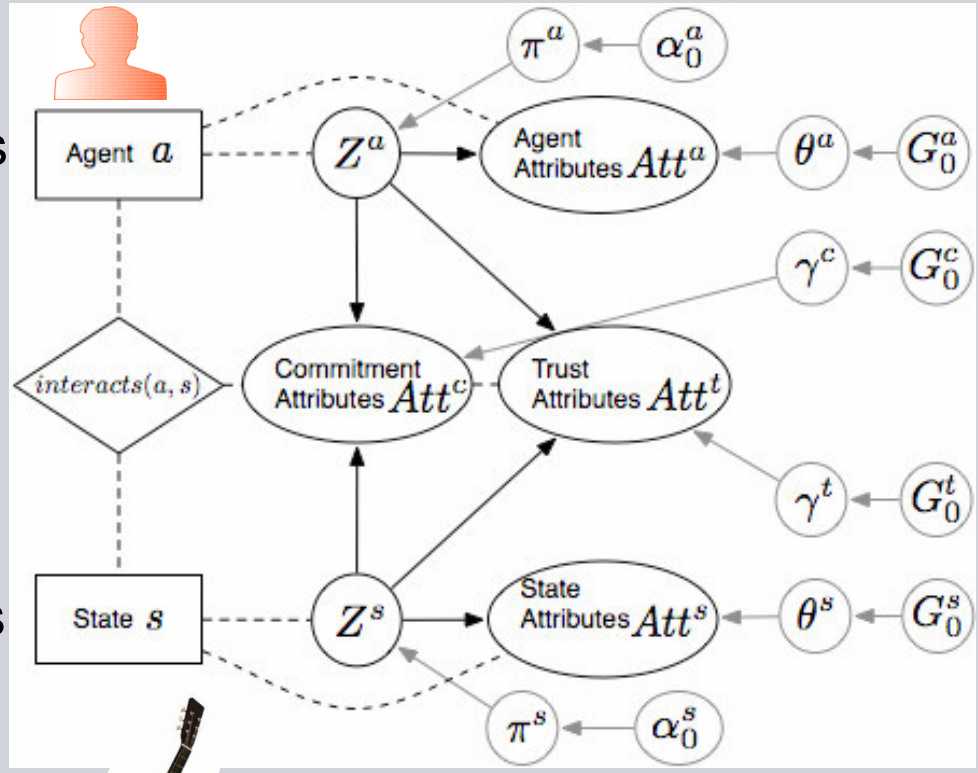
$ATT^c$
- Final price
- # of bids

$ATT^t$
- Feedback [2]

- Task:
  - Predict $ATT^t$ for new situation

sellers

items

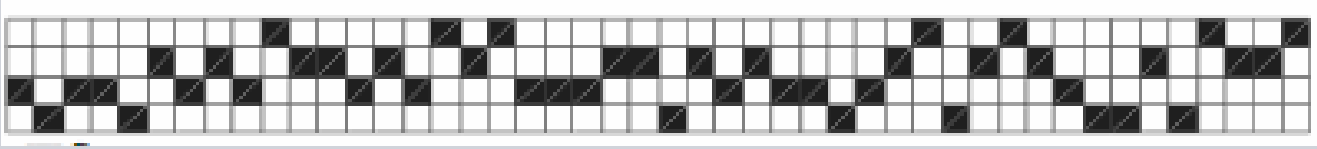



## eBay Data

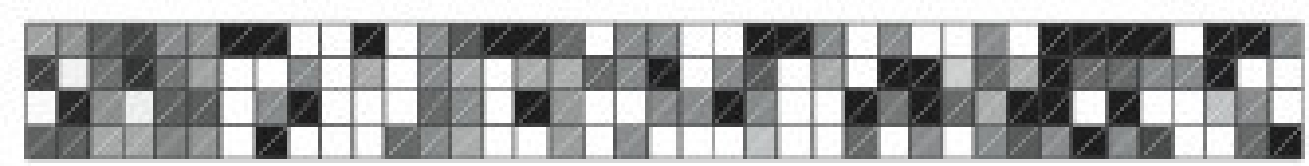
- 47 sellers (agents)
- 631 different items (states)

- 1818 rated sales (47x631 possible sales)

**47 agents in 4 agent clusters**



**4 agent clusters versus 40 item clusters (black: trustworth)**



## Predictive Performance

- Predicting Ratings:
  - 95% confidence interval, 5-fold cross-validation
- Ratio: Baseline
- SVM: Support Vector Machine, DecTree: Decision Tree
- +ID: Different way of propositionalizing by adding an ID-number for every entry

| | Accuracy | ROC Area |
|---|---|---|
| Ratio | 48.5334 (± 3.2407) | - |
| SVM | 54.1689 (± 3.5047) | 0.512 (± 0.0372) |
| DecTree | 54.6804 (± 5.3826) | 0.539 (± 0.0502) |
| SVM+ID | 56.1998 (± 3.5671) | 0.5610 (± 0.0362) |
| DecTree+ID | 60.7901 (± 4.9936) | 0.6066 (± 0.0473) |
| IHRM | 71.4196 (± 5.5063) | 0.7996 (± 0.0526) |

## Conclusion

- We have introduced the IHRM to realize nonparametric relational Bayes and suggest that it might be an interesting model for a number of relational problems
- Advantages
  - Reducing the need for extensive structural learning
  - Expressive ability via coupling between heterogeneous relationships
  - The model decides itself about the optimal number of states for the latent variables
  - Clusters can be analyzed
- Many interesting extensions:
  - The approach can be generalized to cluster relations (Kemp at al.)
    - applies(Jack, Mary, loves), applies(.,.,likes), applies(.,.,hates), ….
  - Interplay with ontologies