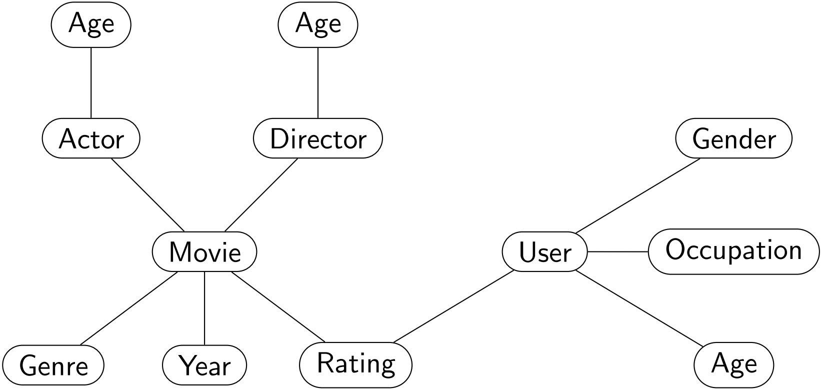
Evaluating Query Result Significance in Databases via Randomizations

Markus Ojala, Gemma Garriga, Aristides Gionis, Heikki Mannila

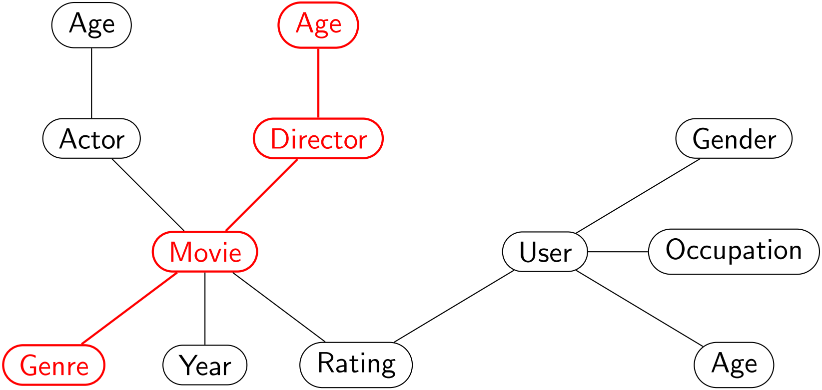
Aalto University
School of Science
and Technology

Introduction



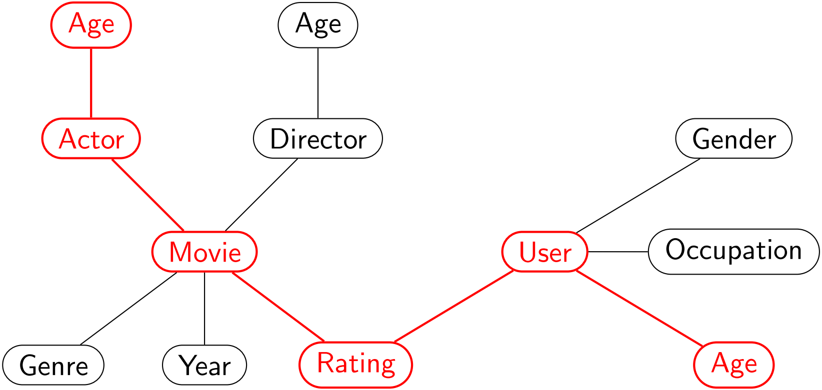
- Database with interrelated tables
- Queries are used to answer questions

1st Example on Movie Database



Hypothesis: **Are the directors of drama movies older than the directors of action movies?**

2nd Example on Movie Database



Hypothesis: **Do old people like old actors?**

Problem

Are the results of the queries statistically significant?

Problem

- *Database $\mathcal{D}$ with multiple binary relations*
- *Query $q(\mathcal{D})$ of interest*
- *Statistic $f(q(\mathcal{D})) \in \mathbb{R}$ of the result $q(\mathcal{D})$,*
- *Is the value of $f(q(\mathcal{D}))$ significant (in some sense)?*

Example

Is the average age of drama directors surprising?

$$\text{GM} = \text{Genre–Movie}, \text{MD} = \text{Movie–Director}, \text{DA} = \text{Director–Age}$$

$$q = \text{ages of directors for drama movies}$$

$$f = \text{average age}$$

Examples of Binary Relations

GM

Genre	Movie
Romance	m_1
Romance	m_2
Drama	m_3
Drama	m_4
Drama	m_5
Drama	m_6
Drama	m_7
History	m_6
History	m_7

MD

Movie	Director
m_1	C. Waitt
m_2	C. Waitt
m_3	C. Waitt
m_4	C. Waitt
m_5	C. Waitt
m_6	T. George
m_7	T. George

DA

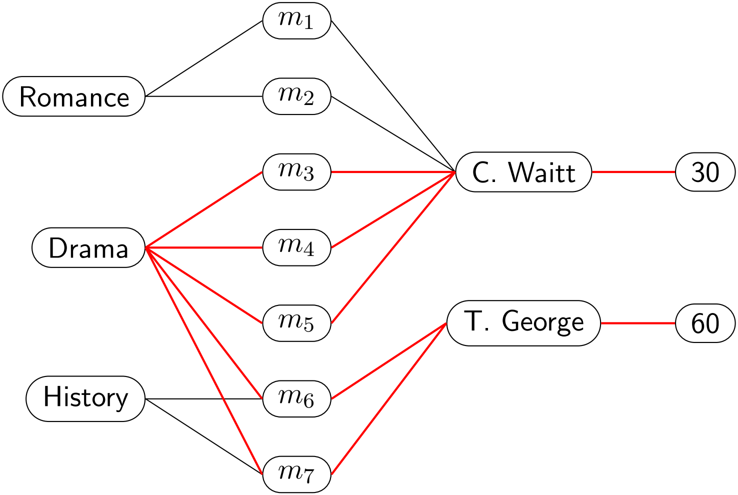
Director	Age
C. Waitt	30
T. George	60

$q =$ ages of directors for drama movies

$= \{(m_3, 30), (m_4, 30), (m_5, 30), (m_6, 60), (m_7, 60)\}$

$f = 42$

The Same Example as Bipartite Graphs



Randomization Approach

Basic approach

- Original database $\mathcal{D}$
- Produce k randomized databases $\widehat{\mathcal{D}}_1, \ldots, \widehat{\mathcal{D}}_k$
- Empirical p-value gives the significance of $f(q(\mathcal{D}))$

$$p = \frac{|\{i : f(q(\widehat{\mathcal{D}}_i)) \leq f(q(\mathcal{D}))\}| + 1}{k + 1}$$

Where and how to randomize

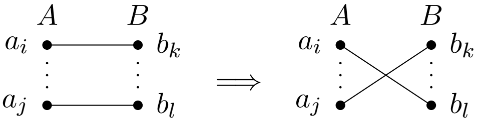
- Each relation separately
- Connections between relations

Randomization Methods

Randomizations for a single binary relation AB

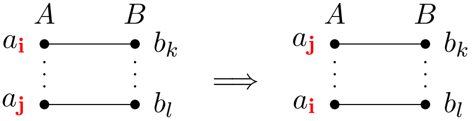
1. *Swap randomization* of AB, sw(AB) (Gionis *et al.*, 2007):

- performs swaps
- preserves degrees



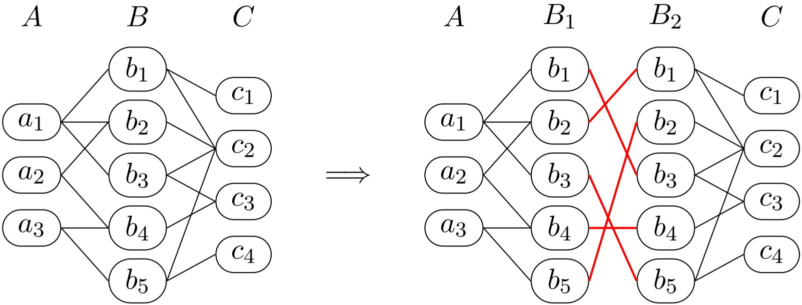
2. *Label permutation* of AB

- permutes the labels in one attribute



Connection between randomizations

Permuting labels in B



- Permuting labels in B = swap randomizing identity relation I_B

Different randomizations for database $\mathcal{D} = \{AB, BC\}$

1. sw(AB), 2. sw(I_B), 3. sw(BC)

Experiments: Synthetic data

Three structured relations plus structureless versions

$$\begin{aligned} \mathsf{SU} &= \text{Gender–User} \quad (2 \times 50) \\ \mathsf{UM} &= \text{User–Movie} \quad (50 \times 100) \\ \mathsf{MG} &= \text{Movie–Genre} \quad (100 \times 6) \\ \mathsf{rXX} &= \text{structureless version of XX} \end{aligned}$$

Hypothesis

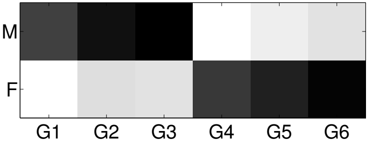
Men watch different types of movies than women.

Statistic

L_1 *distance between the distribution of genres of the movies that men and women have watched.*

Experiments: Results for synthetic data

$\mathcal{D} = \{\mathsf{SU}, \mathsf{UM}, \mathsf{MG}\}$

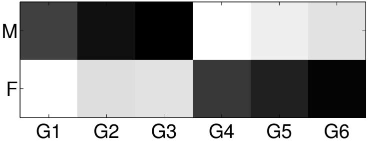


L_1 distance $= 1.23$
Significant with all randomizations

black $= 30\%$, white $= 4.5\%$

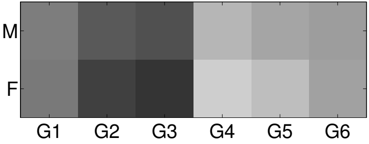
Experiments: Results for synthetic data

$\mathcal{D} = \{\text{SU}, \text{UM}, \text{MG}\}$



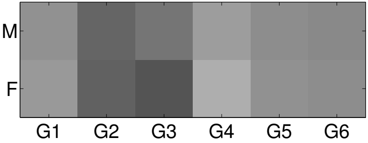
L_1 distance $= 1.23$
Significant with all randomizations

$\mathcal{D} = \{\color{red}{\text{rSU}}, \text{UM}, \text{MG}\}$



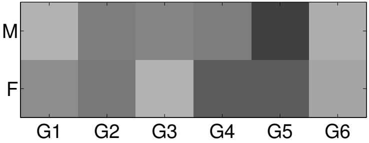
L_1 distance $= 0.10$
Nonsign: sw($\color{red}{\text{rSU}}$), sw($I_{\color{red}{\text{U}}}$)

$\mathcal{D} = \{\text{SU}, \color{red}{\text{rUM}}, \text{MG}\}$



L_1 distance $= 0.08$
Nonsign: sw($I_{\color{red}{\text{U}}}$), sw($\color{red}{\text{rUM}}$), sw($I_{\color{red}{\text{M}}}$)

$\mathcal{D} = \{\text{SU}, \text{UM}, \color{red}{\text{rMG}}\}$



L_1 distance $= 0.15$
Nonsign: sw($I_{\color{red}{\text{M}}}$), sw($\color{red}{\text{rMG}}$)

Experiments: MovieLens dataset

MovieLens: 100,000 ratings from 942 users on 1680 movies

Relation	$\#A$	$\#B$	$\|AB\|/\#A$
User – Movie	943	1680	106
Movie – Genre	1680	18	1.7
User – Occupation	943	21	1
User – Gender	943	2	1
Movie – Age	1680	1680	1
Movie – Rating	943	943	1
User – Age	943	943	1
User – Rating	943	943	1

Experiments: MovieLens dataset

Hypothesis

Men watch different types of movies than women.

Statistic

L_1 *distance between the distribution of genres of the movies that men and women have watched.*

Randomization	Statistic	p-value
Original result	0.16	
sw(Gender–User)	0.03	0.001
sw(I_{User})	0.03	0.001
sw(User–Movie)	0.01	0.001
sw(I_{Movie})	0.03	0.001
sw(Movie–Genre)	0.02	0.001

Experiments: MovieLens dataset

Hypothesis

Men watch genre G more (or less) than women.

Statistic

The difference between the %-proportions of the movies from genre G among all the movies men and women have watched.

Results — equal with all randomizations

More:	Action (2.5), Science fiction (1.5), Thriller (1.1)
Equal:	Documentary (0.0), Fantasy (-0.1)
Less:	Comedy (-1.3), Drama (-2.3), Romance (-2.3)

Experiments: MovieLens dataset

Hypothesis

Old people watch old movies.

Statistic

Correlation between the age of the movies and the age of the users who have watched the movie.

Randomization	Statistic	p-value
Original result	0.16	
sw(Age–User)	0.00	0.001
sw(User–Movie)	0.00	0.001
sw(Movie–Age)	0.00	0.033

Conclusions

Summary

- Assessing queries on multirelational databases
- Randomize relations: sw(AB), sw(I_B), sw(BC)
- Empirical p-values give the structural impact of each relation
- First steps for understanding how the structure hidden in the data affects the significance of the results

Future

- What to do with all the p-values?
- How to conclude the correct inference?
- More study on combinatorial properties and its connection to the significance of queries and patterns