Towards Privacy-Preserving Relational Data Synthesis via Probabilistic Relational Models

KI 2024 – Würzburg, Germany

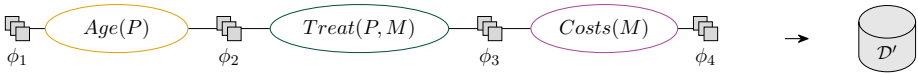
Malte Luttermann[1], Ralf Möller[2], and Mattis Hartwig[1,3]

[1]German Research Center for Artificial Intelligence (DFKI), Lübeck
[2]Institute for Humanities-Centered Artificial Intelligence (CHAI), University of Hamburg
[3]singularIT GmbH, Leipzig

September 25, 2024

Motivation and Problem Setup

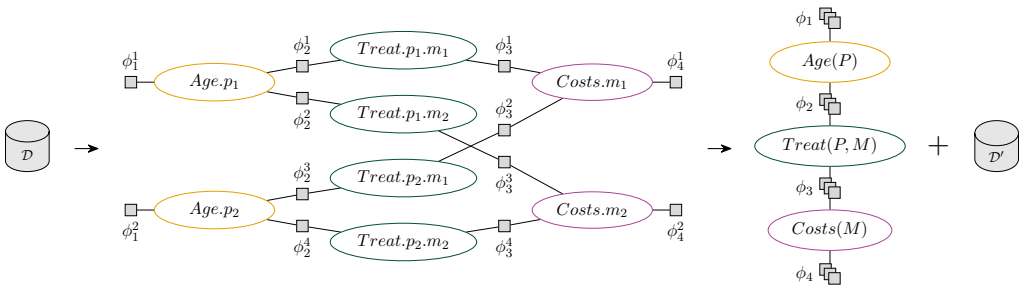
- Many machine learning tasks require access to relational training data
- Collecting real-world data is often challenging
- Synthetic data might help to overcome these challenges
- Idea: Use probabilistic relational models to generate synthetic data



Our Contributions

Pipeline to generate synthetic relational data via probabilistic relational models:

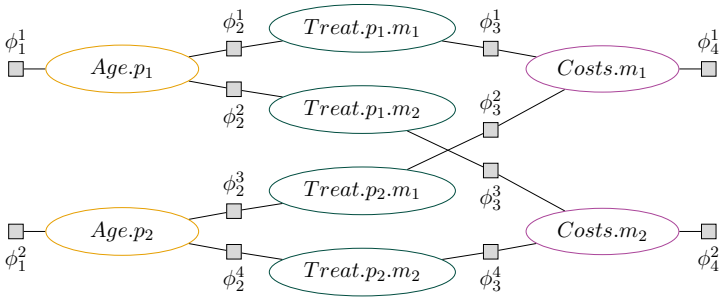
1. Construction of a factor graph
2. Transformation of the factor graph into a parametric factor graph
3. Sampling the parametric factor graph to generate synthetic data samples



Background I

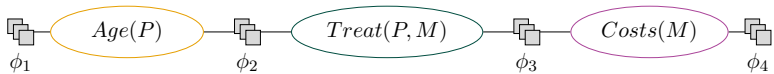
- A factor graph G compactly encodes a full joint probability distribution
- Semantics of G over a set of factors $\mathbf{\Phi} = \{\phi_1, \ldots, \phi_m\}$ is given by

$$P_G = \frac{1}{Z} \prod_{j=1}^{m} \phi_j$$



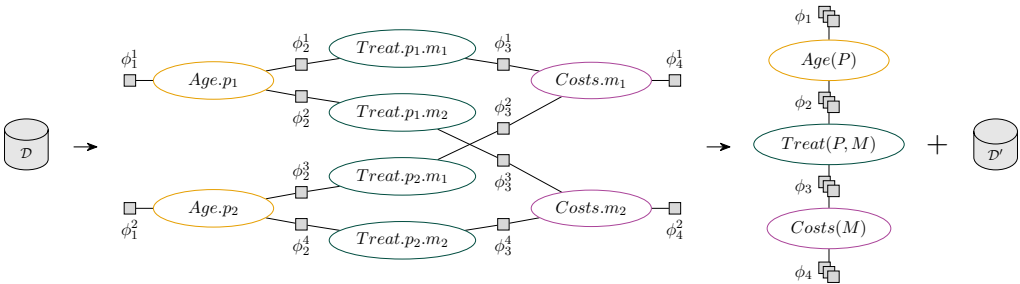
Background II

- A parametric factor graph exploits symmetries in a factor graph
- Introduction of logical variables to represent groups of random variables
  - $\mathrm{dom}(P) = \{p_1, p_2\}$
  - $\mathrm{dom}(M) = \{m_1, m_2\}$



Framework Overview

1. Construction of a factor graph
2. Transformation of the factor graph into a parametric factor graph
3. Sampling the parametric factor graph to generate synthetic data samples



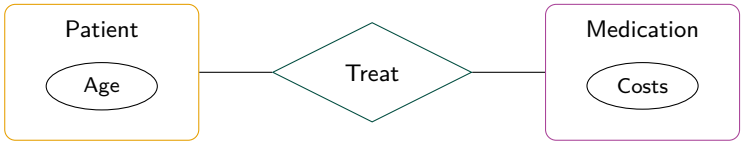
Learning a Parametric Factor Graph from Relational Data

Overview

1. Cluster each entity in the database
2. Add a node for every attribute and every relation to the graph
3. Use conditional independence tests to add edges between nodes
4. Count values in the database to obtain potential values for the factors
5. Detect and exploit symmetries in the resulting factor graph

Learning a Parametric Factor Graph from Relational Data

Example Database



PatientId	Age
alice	≥ 18
bob	≥ 18
charlie	≥ 18
dave	< 18
eve	< 18

PatientId	MedicationId
alice	*myalept*
alice	*danyelza*
bob	*paracetamol*
charlie	*ibuprofen*
eve	*eliquis*

MedicationId	Costs
myalept	*high*
danyelza	*high*
paracetamol	*low*
ibuprofen	*low*
eliquis	*high*

Learning a Parametric Factor Graph from Relational Data

Step 1: Cluster Entities

- Compute clusters for all entities using an arbitrary clustering algorithm
- Patient: $p_1 = \{alice, eve\}$, $p_2 = \{bob, charlie, dave\}$
- Medication: $m_1 = \{myalept, danyelza, eliquis\}$, $m_2 = \{paracetamol, ibuprofen\}$

PatientId	Age	C
$alice$	≥ 18	p_1
bob	≥ 18	p_2
$charlie$	≥ 18	p_2
$dave$	< 18	p_2
eve	< 18	p_1

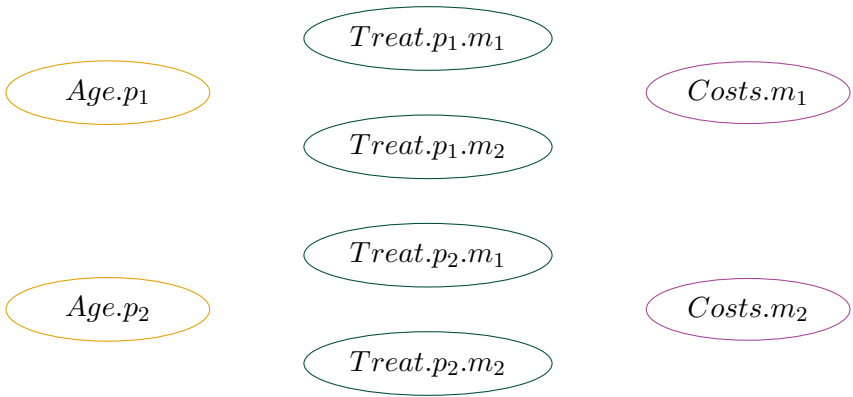
PatientId	MedicationId
$alice$	$myalept$
$alice$	$danyelza$
bob	$paracetamol$
$charlie$	$ibuprofen$
eve	$eliquis$

MedicationId	Costs	C
$myalept$	$high$	m_1
$danyelza$	$high$	m_1
$paracetamol$	low	m_2
$ibuprofen$	low	m_2
$eliquis$	$high$	m_1

Learning a Parametric Factor Graph from Relational Data

Steps 2+3: Construct the Graph Structure

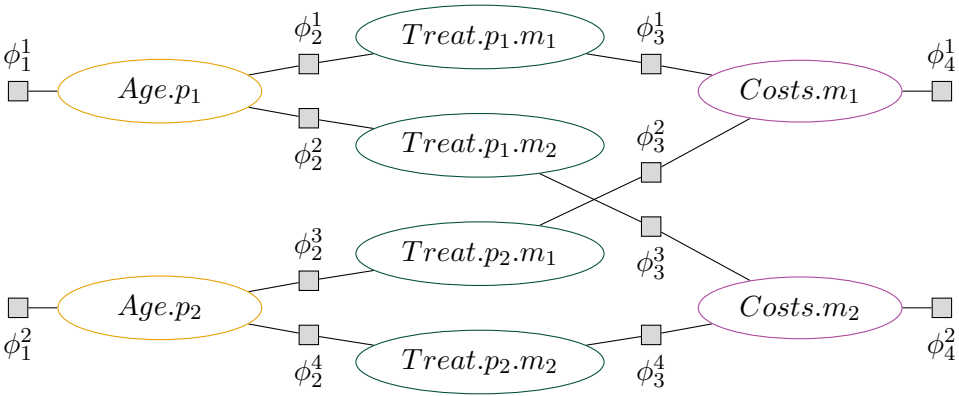
- Use clusters to add nodes for every attribute and every relation



Learning a Parametric Factor Graph from Relational Data

Steps 2+3: Construct the Graph Structure

- Use clusters to add nodes for every attribute and every relation
- Use conditional independence tests to add edges between nodes



Learning a Parametric Factor Graph from Relational Data

Step 4: Learn the Graph Parameters

- For each factor, count occurrences of values by cluster
- E.g., count $Age \geq 18$ and $Age < 18$ for p_1, p_2

$Age.p_1$	ϕ_1^1
≥ 18	$\varphi_1 \in \mathbb{R}^+$
< 18	$\varphi_2 \in \mathbb{R}^+$

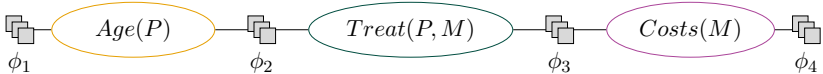
$Age.p_2$	ϕ_1^2
≥ 18	$\varphi_3 \in \mathbb{R}^+$
< 18	$\varphi_4 \in \mathbb{R}^+$

- Analogously for remaining factors

Learning a Parametric Factor Graph from Relational Data

Step 5: Detect and Exploit Symmetries

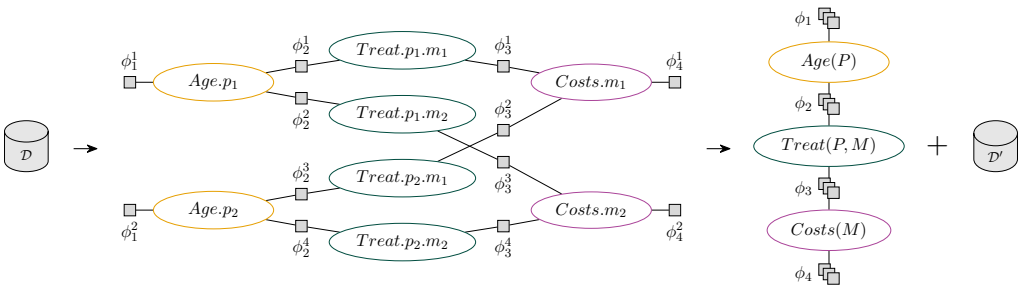
- Group together indistinguishable clusters
  - Advanced Colour Passing algorithm (Luttermann et al., 2024)
- E.g., assume p_1, p_2 and m_1, m_2 are indistinguishable:
  - Logical variable P with $\mathrm{dom}(P) = \{p_1, p_2\}$
  - Logical variable M with $\mathrm{dom}(M) = \{m_1, m_2\}$



Malte Luttermann et al. (2024). »Colour Passing Revisited: Lifted Model Construction with Commutative Factors«. *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*. AAAI Press, pp. 20500–20507.

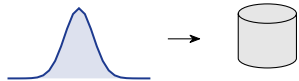
Framework Overview

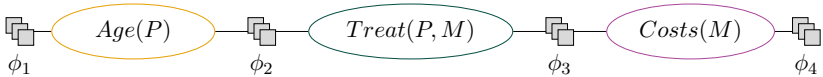
1. Construction of a factor graph
2. Transformation of the factor graph into a parametric factor graph
3. Sampling the parametric factor graph to generate synthetic data samples



Sampling the Parametric Factor Graph

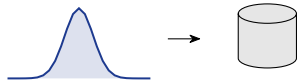
- Sample the underlying probability distribution
- Data samples for each combination of clusters

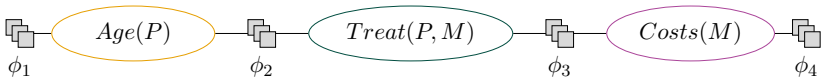




Sampling the Parametric Factor Graph

- Sample the underlying probability distribution
- Data samples for each combination of clusters





- E.g., $Age.p_1 < 18$, $Age.p_2 \geq 18$, $Treat.p_1.m_1 = false$, $Treat.p_1.m_2 = false$, $Treat.p_2.m_1 = true$, $Treat.p_2.m_2 = false$, $Costs.m_1 = low$, $Costs.m_2 = high$

Summary

- We propose an algorithm to learn a parametric factor graph from data
- The parametric factor graph can be sampled to generate synthetic data
- The parametric factor graph can be used for probabilistic and causal inference