

**KU LEUVEN**

**ARENBERG DOCTORAL SCHOOL**  
Faculty of Engineering Science

**DRAFT**

To remove, add 'final' to class options

# Modeling Relational Data Mining

**Sergey Paramonov**

Supervisors:  
Prof. dr. Luc De Raedt  
Prof. dr. Marc Denecker

Dissertation presented in partial  
fulfillment of the requirements for the  
degree of Doctor of Engineering  
Science (PhD): Computer Science

August 2017



# **Modeling Relational Data Mining**

**Sergey PARAMONOV**

Examination committee:

Prof. dr. ir. The Chairman, chair

Prof. dr. Luc De Raedt, supervisor

Prof. dr. Marc Denecker, supervisor

Prof. dr. Gerda Janssens

Dr. Matthijs van Leeuwen

Helmut Simonis

(University College Cork, Insight Centre for  
Data Analytics)

Prof. dr. Christian Bessiere

(CNRS, U. Montpellier, LIRMM)

Dissertation presented in partial  
fulfillment of the requirements for  
the degree of Doctor of Engineering  
Science (PhD): Computer Science

August 2017

© 2017 KU Leuven – Faculty of Engineering Science  
Uitgegeven in eigen beheer, Sergey Paramonov, Celestijnenlaan 200A, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

# Preface

...

## **Instructies van de faculteit:**

In het voorwoord wordt de algemene doelstelling van het werk samengevat in enkele regels en worden personen, diensten of firma's bedankt voor hun medewerking bij het tot stand komen van het werk.

De naam van firma's en personen uit deze firma's mogen slechts worden vermeld mits hun uitdrukkelijke toelating én na overleg met de supervisor(en)! Steeds wordt de supervisor(en) vermeld, de verantwoordelijke en eventueel de personen die rechtstreeks geholpen hebben bv. door het ter beschikking stelling van meetresultaten, faciliteiten. Ook de instantie die eventueel een doctoraatsbeurs heeft toegekend wordt bedankt (bv. FWO, IWT, ...).



# Abstract

## < What data mining is >

Data Mining is the process of discovering new knowledge from the data. A significant attention in the research community is devoted to data analysis in the presence of the data independence assumption, i.e., when the collected data points are independent from each other.

## < What relational means >

However, it is often the case that the objects are connected and related to each other by means of relations. This setting is called *relational* and the associated task is referred as *relational reasoning*.

## < Relational problems >

Electronic tables, spreadsheets, and databases are all examples of relational data that is in a wide use today. The objects are connected by means of relations and constraints. In spreadsheets these are tables and formulae, and in databases schema relations and integrity constraints.

## < Key issue >

We argue that a general approach for modeling and solving data mining problems in the relational setting is missing. The goal of this thesis is to fill in this gap.

## < Contribution >

Firstly, we demonstrate how the problem of learning in relation setting is different from the classical machine learning approach and propose a system named TaCLe as the first working in this setting.

Secondly, we demonstrate how the relational approach generalizes a classical problem of boolean matrix factorization into the Relational Data Factorization, which allows to model a spectrum of classical data mining problems and introduces new ones as well.

Thirdly, we demonstrate how existing relational reasoning formalisms, such as Answer Set Programming, can be enhanced by relational learning techniques

iv \_\_\_\_\_ ABSTRACT

known as sketching.

Last but not least, we demonstrate how relational approach can be used to mine relational patterns, known as structured pattern mining in data mining community.



# Beknopte samenvatting

...

**Instructies van de faculteit:**

In een beknopte tekst van maximum 2 pagina's worden de belangrijkste doelstellingen en besluiten geformuleerd, zowel in het Nederlands als in het Engels. Zulke samenvattingen kunnen worden gebruikt in wetenschappelijke verslagen van het departement of de faculteit. Het Engels moet vlekkeloos zijn.



# List of Abbreviations

**MD** molecular dynamics. 3



# List of Symbols

$\Theta$       A nice symbol



# Contents

<b>Abstract</b>	<b>iii</b>
<b>List of Abbreviations</b>	<b>vii</b>
<b>List of Symbols</b>	<b>ix</b>
<b>Contents</b>	<b>xi</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xv</b>
<b>1 This is introduction</b>	<b>3</b>
1.1 First Order Logic . . . . .	5
1.2 Inductive Logic Programming . . . . .	5
1.3 Answer Set Programming . . . . .	5
1.4 $\text{FO}(\cdot)$ and the IDP system . . . . .	5
<b>2 This is conclusion</b>	<b>6</b>
<b>A This is myappendix</b>	<b>7</b>
<b>Bibliography</b>	<b>9</b>

xii \_\_\_\_\_ CONTENTS

**This is curriculum 11**



# List of Figures

1.1	Illustration of how to include a figure. . . . .	4
-----	--	---



# List of Tables



LIST OF TABLES .....	1
----------------------	---

### **Instructies van de faculteit:**

De hoofdstukken: Elk hoofdstuk is ingelast met een bepaald doel voor ogen. Dit doel wordt vermeld in de eerste paragraaf van elk hoofdstuk. Naargelang de aard van de tekst (experiment, uitvoering, theoretische ontwikkeling, ...) volgen de paragrafen elkaar op. Beweringen worden altijd gestaafd, hetzij door eigen experimenten, hetzij door een theoretische afleiding, hetzij door verwijzingen naar de literatuur. Elk hoofdstuk eindigt met een kort samenvattend besluit waarbij nagegaan wordt in hoeverre de doelstelling van het betrokken hoofdstuk verwezenlijkt is. De deelbesluiten moeten de lezer automatisch leiden naar het algemeen besluit aan het einde van het werk.



# Chapter 1

## This is introduction

### Instructies van de faculteit:

De inleiding situeert de problematiek, beschrijft de stand van de huidige kennis terzake, omschrijft de voornaamste doelstellingen van het werk, samen met de beperkende randvoorwaarden en de ter beschikking gestelde middelen en poneert de belangrijkste stellingen.

Illustration of how to include citations [2] and [3]. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

And yet another citation [1].

Introducing some symbol:  $\Theta$ .

Introducing an acronym: MD.

4 \_\_\_\_\_ THIS IS INTRODUCTION

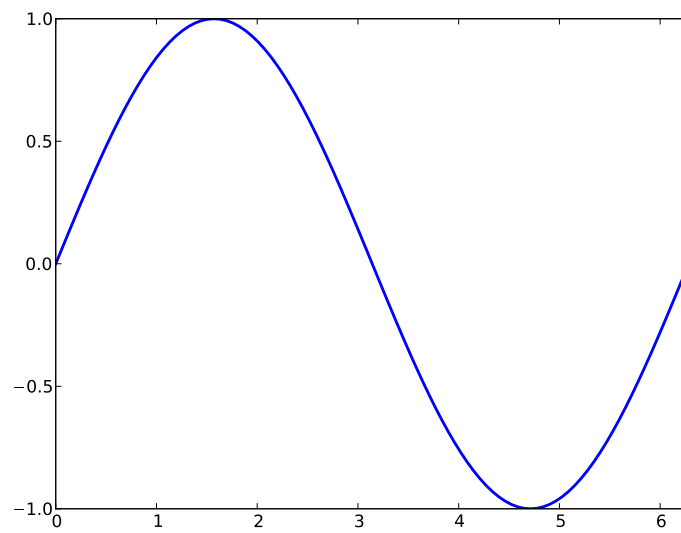


Figure 1.1: Illustration of how to include a figure.



FIRST ORDER LOGIC \_\_\_\_\_ 5

From now on you shall be called  
Brian that is called Brian.

---

The Life of Brian

## **1.1 First Order Logic**

text

## **1.2 Inductive Logic Programming**

## **1.3 Answer Set Programming**

## **1.4 $FO(\cdot)$ and the IDP system**

## Chapter 2

# This is conclusion

...

### **Instructies van de faculteit:**

Algemene besluiten: Verwijzend naar de inleiding en naar de besluiten van de afzonderlijke hoofdstukken worden op het einde van het proefschrift de voornaamste besluiten gebundeld. Hier wordt de nadruk gelegd op de eigen inbreng, de verworven resultaten, de ‘stellingen’ van het proefschrift en de originele bijdragen tot het onderzoeksdomein. De onopgeloste problemen worden aangestipt en suggesties voor eventueel verder onderzoek worden gemaakt.

## Appendix A

# This is myappendix

...

### Instructies van de faculteit:

De appendices: ze omvatten alle gedeelten uit de tekst die weliswaar essentieel zijn voor het proefschrift, maar waarvan de inlassing in de tekst de leesbaarheid ervan nadelig zouden beïnvloeden bv. omwille van hun lengte. Zo kunnen bv. de brute meetresultaten of een computerprogramma met zijn bron, commentaar en voorbeelden beter thuishoren in een appendix dan in de tekst zelf. De appendices kunnen desgevallend worden gebundeld in een apart boekdeel.



# Bibliography

- [1] FREDERIX, Y., AND ROOSE, D. A drift-filtered approach to diffusion estimation for multiscale processes. In *Coping with complexity: model reduction and data analysis* (2010), vol. 75 of *Lecture Notes in Computational Science and Engineering*, Springer-Verlag.
- [2] MEERT, W. *Inference and Learning for Directed Probabilistic Logic Models*. PhD thesis, Informatics Section, Department of Computer Science, Faculty of Engineering, Mar. 2011. Blockeel, Hendrik (supervisor).
- [3] VAN DEN BROECK, G., TAGHIPOUR, N., MEERT, W., DAVIS, J., AND DE RAEDT, L. Lifted probabilistic inference by first-order knowledge compilation. In *Proceedings of the 22th International Joint Conference on Artificial Intelligence (IJCAI)* (2011).

## Instructies van de faculteit:

De bibliografie. Departementale richtlijnen terzake te volgen.



# This is curriculum

...

**Instructies van de faculteit:**

Beknopt CV van de doctorandus.





# List of publications

Input file chapters/publications/publications.tex does not exist. Make sure its starts with “\chapter{List of publications}”. To not include this chapter in the table of contents, use the starred version of the \chapter command. . .

**Instructies van de faculteit:**

Lijst van de publicaties door de doctorandus/a (auteur of co-auteur).





FACULTY OF ENGINEERING SCIENCE  
DEPARTMENT OF COMPUTER SCIENCE  
DTAI  
Celestijnenlaan 200A  
B-3001 Leuven  
sergey.paramonov@kuleuven.be  
sergey-paramonov.com

