N° d'ordre NNT :2019LYSEI086

Thèse de doctorat de l'Université de Lyon
opérée au sein de
L'INSA de Lyon

Ecole Doctorale N° 512
Mathématiques et Informatique (InfoMaths)

Spécialité / discipline de doctorat : Informatique

À soutenir publiquement par
Adnene Belfodil

Exceptional Model Mining for Behavioral Data Analysis

Devant le jury composé de:

Sihem Amer-Yahia	Directrice de recherche, CNRS	Rapporteure
Arno Siebes	Professeur des Universités, Université d'Utrecht	Rapporteur
Arno Knobbe	Maître de conférences, Université de Leiden	Examinateur
Ioana Manolescu	Directrice de recherche, INRIA	Examinatrice
Amedeo Napoli	Directeur de recherche, CNRS	Examinateur
Philippe Lamarre	Professeur des Universités, INSA-Lyon	Directeur de thèse
Sylvie Cazalens	Maître de conférences, INSA-Lyon	Co-directrice de thèse
Marc Plantevit	Maître de conférences, Université Claude Bernard Lyon 1	Co-directeur de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales – Quinquennal 2016-2020

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage **secretariat@edchimie-lyon.fr** INSA : R. GOURDON	M. Stéphane DANIELE Institut de recherches sur la catalyse et l'environnement de Lyon IRCELYON-UMR 5256 Équipe CDFA 2 Avenue Albert EINSTEIN 69 626 Villeurbanne CEDEX **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec. : M.C. HAVGOUDOUKIAN **ecole-doctorale.eea@ec-lyon.fr**	M. Gérard SCORLETTI École Centrale de Lyon 36 Avenue Guy DE COLLONGUE 69 134 Écully Tél : 04.72.18.60.97 Fax 04.78.43.37.17 **gerard.scorletti@ec-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : H. CHARLES **secretariat.e2m2@univ-lyon1.fr**	M. Philippe NORMAND UMR 5557 Lab. d'Ecologie Microbienne Université Claude Bernard Lyon 1 Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://www.ediss-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 INSA : M. LAGARDE **secretariat.ediss@univ-lyon1.fr**	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 Avenue Jean CAPELLE INSA de Lyon 69 621 Villeurbanne Tél : 04.72.68.49.09 Fax : 04.72.68.49.16 **emmanuelle.canet@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 **infomaths@univ-lyon1.fr**	M. Luca ZAMBONI Bât. Braconnier 43 Boulevard du 11 novembre 1918 69 622 Villeurbanne CEDEX Tél : 04.26.23.45.52 **zamboni@maths.univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **ed.materiaux@insa-lyon.fr**	M. Jean-Yves BUFFIÈRE INSA de Lyon MATEIS - Bât. Saint-Exupéry 7 Avenue Jean CAPELLE 69 621 Villeurbanne CEDEX Tél : 04.72.43.71.70 Fax : 04.72.43.85.28 **jean-yves.buffiere@insa-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bât. Direction **mega@insa-lyon.fr**	M. Jocelyn BONJOUR INSA de Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69 621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **http://ed483.univ-lyon2.fr** Sec. : Véronique GUICHARD INSA : J.Y. TOUSSAINT Tél : 04.78.69.72.76 **veronique.cervantes@univ-lyon2.fr**	M. Christian MONTES Université Lyon 2 86 Rue Pasteur 69 365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Publications

Publications presented in the thesis

The contributions presented in this thesis appear in the following publications:

Peer-Reviewed International Journals

- Adnene Belfodil, Sylvie Cazalens, Philippe Lamarre and Marc Plantevit. Identifying exceptional (dis) agreement between groups. *Accepted in Data Mining and Knowledge Discovery.*

Peer-Reviewed International Conferences

- Adnene Belfodil, Wouter Duivesteijn, Marc Plantevit, Sylvie Cazalens and Philippe Lamarre. DEvIANT: Discovering Significant Exceptional (Dis-) Agreement Within Groups. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), 2019.*
- Adnene Belfodil, Sylvie Cazalens, Philippe Lamarre and Marc Plantevit. Flash Points: Discovering Exceptional Pairwise Behaviors in Vote or Rating Data. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pages 442-458, 2017.*

Peer-Reviewed National Conferences

- Charles de Lacombe, Antoine Morel, Adnene Belfodil, François Portet, Cyril Labbé, Sylvie Cazalens, Marc Plantevit and Philippe Lamarre. Analyse de comportements relatifs exceptionnels expliquée par des textes[1]. *In Extraction et Gestion des connaissances - Démo Track (EGC), Pages 437-440, 2019*

[1]Rewarded by the EGC'2019 award committee as the best demo paper of the year.

CONFERENCE PAPERS NOT COVERED IN THIS DISSERTATION

- Adnene Belfodil, Aimene Belfodil, Anes Bendimerad, Philippe Lamarre, Celine Robardet, Mehdi Kaytoue and Marc Plantevit. FSSD - A Fast and Efficient Algorithm for Subgroup Set Discovery. *In The 6th IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2019.*
- Aimene Belfodil, Adnene Belfodil and Mehdi Kaytoue. Mining Formal Concepts using Implications between Items. *In International Conference on Formal Concept Analysis (ICFCA), pages 173-190, 2019.*
- Aimene Belfodil, Adnene Belfodil and Mehdi Kaytoue. Anytime Subgroup Discovery in Numerical Domains with Guarantees[2]. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), pages 500-516, 2018.*

[2]Rewarded by the ECML/PKDD'2018 award committee as the best student data mining paper of the year.

"Science involves confronting our absolute stupidity "

- Schwartz, 2008 -

Abstract

With the rapid proliferation of data platforms collecting and curating data related to various domains such as governments data, education data, environment data or product ratings, more and more data are available online. This offers an unparalleled opportunity to study the behavior of individuals and the interactions between them. In the political sphere, being able to query datasets of voting records provides interesting insights for data journalists and political analysts. In particular, such data can be leveraged for the investigation of exceptionally consensual/controversial topics.

Consider data describing the voting behavior in the European Parliament (EP). Such a dataset records the votes of each member (MEP) in voting sessions held in the parliament, as well as information on the parliamentarians (e.g., gender, national party, European party alliance) and the sessions (e.g., topic, date). This dataset offers opportunities to study the agreement or disagreement of coherent subgroups, especially to highlight unexpected behavior. It is to be expected that on the majority of voting sessions, MEPs will vote along the lines of their European party alliance. However, when matters are of interest to a specific nation within Europe, alignments may change and agreements can be formed or dissolved. For instance, when a legislative procedure on fishing rights is put before the MEPs, the island nation of the UK can be expected to agree on a specific course of action regardless of their party alliance, fostering an exceptional agreement where strong polarization exists otherwise. In this thesis, we aim to discover such exceptional (dis)agreement patterns not only in voting data but also in more generic data, called behavioral data, which involves individuals performing observable actions on entities. We devise two novel methods which offer complementary angles of exceptional (dis)agreement in behavioral data: within and between groups. These two approaches called Debunk and Deviant, ideally, enables the implementation of a sufficiently comprehensive tool to highlight, summarize and analyze exceptional comportments in behavioral data. We thoroughly investigate the qualitative and quantitative performances of the devised methods. Furthermore, we motivate their usage in the context of computational journalism.

Keywords: Subgroup Discovery, Exceptional Model Mining, Behavioral Data Analysis, Computational Journalism.

Résumé

Avec la prolifération rapide des plateformes de données qui récoltent des données relatives à plusieurs domaines tels que les données de gouvernements, d'éducation, d'environnement ou les données de notations de produits, plus de données sont disponibles en ligne. Ceci représente une opportunité sans égal pour étudier le comportement des individus et les interactions entre eux. Sur le plan politique, le fait de pouvoir interroger des ensembles de données de votes peut fournir des informations intéressantes pour les journalistes et les analystes politiques. En particulier, ce type de données peut être exploité pour l'investigation des sujet exceptionnellement conflictuels ou consensuels.

Considérons des données décrivant les sessions de votes dans le parlement Européen (PE). Un tel ensemble de données enregistre les votes de chaque député (MPE) dans l'hémicycle en plus des informations relatives aux parlementaires (e.g., genre, parti national, parti européen) et des sessions (e.g., sujet, date). Ces données offrent la possibilité d'étudier les accords et désaccords de sous-groupes cohérents, en particulier pour mettre en évidence des comportements inattendus. Par exemple, il est attendu que sur la majorité des sessions, les députés votent selon la ligne politique de leurs partis politiques respectifs. Cependant, lorsque les sujets sont plutôt d'intérêt d'un pays particulier dans l'Europe, des coalitions peuvent se former ou se dissoudre. À titre d'exemple, quand une procédure législative concernant la pêche est proposée devant les MPE dans l'hémicycle, les MPE des nations insulaires du Royaume-Uni peuvent voter en accord sans être influencés par la différence entre les lignes politiques de leurs alliances respectives, cela peut suggérer un accord exceptionnel comparé à la polarisation observée habituellement. Dans cette thèse, nous nous intéressons à ce type de motifs décrivant des (dés)accords exceptionnels, pas uniquement sur les données de votes mais également sur des données similaires appelées données comportementales. Nous élaborons deux méthodes complémentaires appelées Debunk et Deviant. La première permet la découverte de (dés)accords exceptionnels entre groupes tandis que la seconde permet de mettre en évidence les comportements exceptionnels qui peuvent au sein d'un même groupe. Idéalement, ces deux méthodes ont pour objective de donner un aperçu complet et concis des comportements exceptionnels dans les données comportementales. Dans l'esprit d'évaluer la capacité des deux méthodes à réaliser cet objectif, nous évaluons les performances quantitatives et qualitatives sur plusieurs jeux de données réelles. De plus, nous motivons l'utilisation des méthodes proposées dans le contexte du journalisme computationnel.

Titre: Fouille de Modèles Exceptionnels dans les Données Comportementales.

Mots-Clés: Découverte de Sous-Groupes, Fouille de Modèles Exceptionnels, Analyse de Données Comportementales, Journalisme Computationnel.

Contents

List of Figures

List of Tables

List of Definitions

List of Algorithms

1

Introduction

"Journalism is the activity of gathering, assessing, creating, and presenting news and information"[1]. The primary objective of journalism is to "provide citizens with the information they need to be free and self-governing." as argue Kovach and Rosenstiel, 2014 in the *Elements of Journalism*. In this book, the authors underline ten enduring values of journalism. We highlight in the following two values that represent the main motivations behind the project **ContentCheck**[2,3] within which this thesis is conducted:

1. "Journalism's first obligation is to the **truth**."
2. "Its essence is a discipline of **verification**."

■ ***Truth****, **Accuracy** and **Verifiability** are the backbone of a Trustworthy Journalism.* ■

The digital era and the advent of social media platforms brought sweeping changes to how information is published and consumed (Alejandro, 2010). This affected the whole process of traditional journalism and undermined its credibility and quality with the rise of misinformation (Ireton and Posetti, 2018). Despite the undeniable potential of social media in improving the life of citizens (e.g. organizing efforts in the aftermath of natural disasters (Palen and Hughes, 2018)), its weaponisation[4] impacted profoundly the landscape of journalism (Kucharski, 2016). For instance, according to a recent survey on Internet Security and Trust[5], 85% of the respondents said they had fallen for fake news at least once, with 44% saying they sometimes or frequently did. In this context, journalists around the world gathered in a joint-initiative to fight the scourge of misinformation. For instance,

[1] https://www.americanpressinstitute.org/journalism-essentials
[2] https://contentcheck.inria.fr/
[3] ContentCheck is funded by the French National Research Agency (ANR) under the project code: ANR-15-CE23-0025 - https://anr.fr/Projet-ANR-15-CE23-0025
[4] https://www.rappler.com/nation/148007-propaganda-war-weaponizing-internet
[5] https://www.cigionline.org/internet-survey-2019

The International Fact-Checking Network[6] (IFCN) was launched in 2015 to support Fact-Checking in the world. More than 65 established news organizations, such as The Washington Post Fact Checker[7] and Le Monde - Les Décodeurs[8], are signatory of the IFCN *code of principles*[9] which is a series of commitments organizations abide by to promote excellence in transparent fact-checking. In the same spirit, by July 2019, there were 188 fact-checking projects active in more than 50 countries according to Duke Reporters' Lab[10].

Within this ecosystem, we have been collaborating with journalists from Le Monde (Les Décodeurs Team) since 2015[11] in a research and development project (ContentCheck (Manolescu, 2017)). The goal is to assist and empower journalists by *content management technologies* in order to improve fact-checking work. Content management technologies are cross-fertilization of methods (Cazalens et al., 2018) pertaining to various data-driven computer science disciplines including: data and knowledge management, data mining, information retrieval and natural language processing. This extends the capabilities of a nascent inter-disciplinary field known as *Computational Journalism* (Cohen et al., 2011; Hamilton and Turner, 2009). This field represent the application scope of the tools devised in this thesis.

During the last decade, the field of computational journalism has witnessed growing efforts to meet journalists' needs in many facets of their work (Caswell and Dörr, 2018; Cazalens et al., 2018; Cohen et al., 2011; Flew et al., 2012; Young and Hermida, 2015) . Essential to this work are data, of any kind and on any topic, which have to be collected, understood and analyzed (Coddington, 2015). Sources complying with the open data movement offer good quality information in many domains such as science, government, health, etc. (Charalabidis, Alexopoulos, and Loukis, 2016). In particular, parliamentary institutions make voting data available for transparency. For instance, Voteview[12] offers access to every congressional roll-call votes in American history. Similarly, Parltrack[13] publishes on a daily basis vote results in the European Parliament. Such data can be leveraged to objectively analyze several aspects of the democratic process (Hix, Noury, and Roland, 2007; Poole and Rosenthal, 2000). This kind of data has been the main driver of the research conducted in this thesis.

In this context, fine-grained analysis of voting behaviors is necessary, as it would help in holding politicians accountable for their actions and voting behavior. Such investigation can make use of simple queries to obtain basic information like whether a given parliamentarian has voted for or against in a given voting session. Deeper analyses may take advantage of other methods such as query perturbation (Yang et al., 2018) or data mining techniques in general (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). These last years, descriptive data mining algorithms (such as *Subgroup Discovey* (Klösgen, 1996; Wrobel, 1997)) have

[6] https://www.poynter.org/ifcn/
[7] https://www.washingtonpost.com/news/fact-checker
[8] https://www.lemonde.fr/les-decodeurs/
[9] https://ifcncodeofprinciples.poynter.org/
[10] https://reporterslab.org/fact-checking/
[11] More precisely, this joint-initiative between Le Monde and Four Research Laboratory in France started in December 2015. This thesis work started in October 2016.
[12] https://voteview.com/data
[13] https://parltrack.org/

proved to be helpful to explore such datasets (Etter et al., 2014; Grosskreutz, Boley, and Krause-Traudes, 2010) and to point out interesting relationships between elements in specific data regions, in any application domain (Duivesteijn, Feelders, and Knobbe, 2016; Herrera et al., 2011). Such tools are particularly compelling in the context of fact-checking as they can rapidly uncover useful insights to put claims into perspective and evaluate their veracity. Furthermore, what makes descriptive data mining techniques particularly appealing in the context of computational journalism in general, is the fact that they involve discovering hypotheses from data. For instance, Subgroup discovery (a descriptive data mining technique) has been explained as “a convenient hypothesis generator for further analysis” (Wrobel, 2001). This perfectly fits one of the main endeavors of computational journalism which, as argued by Cohen et al., 2011, is not only about finding answers but finding interesting questions to ask starting from the data of interest (e.g. voting data).

Considering parliamentary institutions and their votes which constitutes our data of interest, to understand the political positions, it is of major interest to find the contexts hardening or softening oppositions. Accordingly, the problem we focus on is to find peculiar behavior of groups of individuals (e.g. parliamentarians) in some context (e.g. judicial legislative procedures) when compared to the behavior of groups observed in overall terms. For instance, In the European Parliament, despite the fact that the votes of the French MEPs (Members of the European Parliament) reflect a strong disagreement between “Rassemblement National” and the “Front de Gauche” in overall terms, there is a strong agreement when voted legislative procedures concerns external relations of the EU. Such elements of information can provide valuable insights for both political analysts and journalists, as it allows, amongst others, (i) to help discover ideological idiosyncrasies when comparing parliamentarians against their peers, (ii) determining red lines between political groups and (iii) exhibiting contexts where nations’ representatives coalesce against others in critical matters.

The main endeavor of this thesis is to expand the portfolio of tools of computational journalism for the analysis of voting data and “similar data”, called next **Behavioral Data**. In this thesis, we are primarily interested in:

■ *Discovering and characterizing* ***Exceptional Behaviors***
between *and* ***within sub-populations*** *in* ***Behavioral data****.* ■

The statement above brings to the fore three important questions whose answers define the scope of this thesis:

- What are **behavioral data**?
- What is **behavioral data analysis**?
- What kind of **exceptional behaviors** are we looking for?

This chapter aims to provide answers to the these questions. First, it briefly defines the research background of this thesis by introducing **behavioral data** (Section 1.1), **behavioral data analysis** and its related works (Section 1.2). Subsequently, the chapter formulates the research questions we address from the view point of behavioral data analysis and characterizes what kind of **exceptional behaviors** we are interested in (Section 1.3). Finally, an overview of the contributions of this thesis (Section 1.4) and its general structure (Section 1.5) are given.

1.1 Behavioral Data

The data we are interested in consist of a set of individuals (e.g. social network users, parliamentarians, patients) who express outcomes (e.g. opinions, ratings, votes, purchases) on entities (e.g. legislative procedures, movies, restaurants, drugs). We call data of this type: **Behavioral data**. Similarly structured data have been considered in several previous works Bendimerad et al., 2017b; Das et al., 2011; Lemmerich et al., 2016; Omidvar-Tehrani and Amer-Yahia, 2018; Omidvar-Tehrani and Amer-Yahia, 2019; Omidvar-Tehrani, Amer-Yahia, and Borromeo, 2019. UGA (User Group Analytics) (Omidvar-Tehrani and Amer-Yahia, 2019; Omidvar-Tehrani, Amer-Yahia, and Borromeo, 2019) is the most generic and mature framework for behavioral data analysis whose main objective is to "breakdown users into groups to gain a more focused understanding of their behavior" (Omidvar-Tehrani and Amer-Yahia, 2019). In UGA, **Behavioral Data** are called **User Data**. Behavioral Data/User Data can be seen as bipartite graphs having individuals on one side and entities on the other side. An edge linking an individual to an entity indicates that the corresponding individual expressed an outcome on the referred entity. Hence, each edge carries information about the expressed outcome (cf. Figure 1.1).

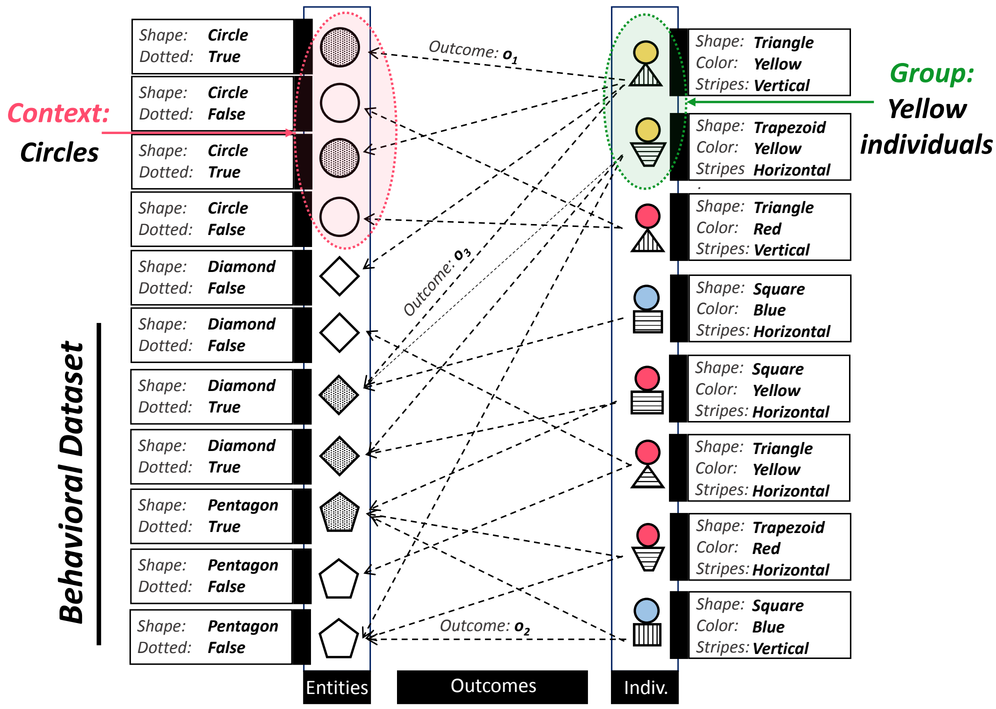


Figure 1.1: Behavioral data as an attributed bipartite graph

While, behavioral data and user data are practically the same in terms of their structure, we choose the term behavioral data to refer to our data of interest. This choice is mainly motivated by the fact that the term "behavioral data" covers, in our view, a broader range of collections of data describing individuals (social network users, parliamentarians, patients) who express outcomes on entities. In contrast, the term "user data", in turn, suggests a more restrictive collection of data where only social network users are considered. Below, we give the definition of a behavioral dataset (Definition 1.1.1).

Definition 1.1.1 — Behavioral Dataset. A behavioral dataset $\mathscr{B} = \langle G_I,\ G_E,\ O,\ o\rangle$ is defined by (i) a collection of Individuals G_I, (ii) a collection of Entities G_E, (iii) a domain of possible Outcomes O, and (iv) a function $o : G_I \times G_E \to O$ that gives the outcome of an individual i over an entity e, if applicable.

The two sets G_I and G_E are collections of records defined over a set of descriptive attributes. We denote such collection of records by G, reintroducing the subscripts only in case of possible confusion. We assume $\mathscr{A} = \{a_1, ..., a_m\}$ to be the set of attributes constituting the schema of G. Each attribute a_j has a domain of interpretation, noted $\mathrm{dom}(a_j)$, which corresponds to all its possible values. We denote $\mathrm{dom}(\mathscr{A}) = \mathrm{dom}(a_1) \times ... \times \mathrm{dom}(a_m)$. Hence, each record $r \in G$ can be seen as a tuple $r = (a^r_1, ..., a^r_m) \in \mathrm{dom}(\mathscr{A})$ where a^r_j corresponds to the value of $a^r_j \in \mathrm{dom}(a_j)$ in the record r. Finally, the domain of possible outcomes O can include, but not limited to, numerical outcomes (e.g. ratings), ordinal outcomes (e.g. preference), categorical outcomes (e.g. votes), texts (e.g. opinions).

Several real-world datasets can be modeled as behavioral datasets. For instance, The European Parliament Voting Dataset[14] (cf. Table 1.1) features parliamentarians who cast votes on legislative procedures in the European parliament. In turn, Movielens[15] (cf. Table 1.2) corresponds to a movie review dataset featuring users who rate movies on a 5-star scale.

ide	themes	date
e_1	1.20 Citizen's rights	20/04/16
e_2	2.10 Free Movement of goods	16/05/16
e_3	1.20 Citizen's rights; 7.30 Judicial Coop	04/06/16
e_4	7 Security and Justice	11/06/16
e_5	7.30 Judicial Coop	03/07/16
e_6	7.30 Judicial Coop	29/07/16

(a) Entities (Voting sessions)

idi	country	group	age
i_1	France	S&D	26
i_2	France	PPE	30
i_3	Germany	S&D	40
i_4	Germany	ALDE	45

(b) Individuals (Parliamentarians)

idi	ide	outcome
i_1	e_1	For
i_1	e_2	Against
i_1	e_5	For
i_1	e_6	Against
i_2	e_1	For
i_2	e_3	Against
i_2	e_4	For
i_2	e_5	For
i_3	e_1	For
i_3	e_2	Against
i_3	e_3	For
i_3	e_5	Against
i_4	e_1	For
i_4	e_4	For
i_4	e_6	Against

(c) Outcomes

Table 1.1: Example of a behavioral dataset - European Parliament Voting dataset. Individuals are described by categorical attributes (country, group) and a numerical attribute (age). Entities are described by a categorical attribute augmented with a taxonomy (themes) and a date percieved as a numerical attribute (date). Outcomes are categorical (not ordered)

[14] http://parltrack.euwiki.org/
[15] https://grouplens.org/datasets/movielens/100k/

ide	genres	releaseDate
e_1	Comedy	1987
e_2	Crime; Drama; SciFi	1992
e_3	Action; Adventure; Crime	1996
e_4	Animation; Comedy	1996
e_5	Action; Romance; War	1992
e_6	Comedy	1997

(a) Entities (Movies)

idi	gender	age	occupation
i_1	M	30	programmer
i_2	F	53	healthcare
i_3	F	48	educator
i_4	M	55	marketing

(b) Individuals (Users)

idi	ide	outcome
i_1	e_1	4
i_1	e_2	2
i_1	e_4	5
i_1	e_5	3
i_2	e_2	1
i_2	e_3	2
i_2	e_5	2
i_2	e_6	5
i_3	e_1	5
i_3	e_2	3
i_3	e_4	5
i_3	e_6	5
i_4	e_1	4
i_4	e_3	1
i_4	e_4	5

(c) Outcomes

Table 1.2: Example of a behavioral dataset - Movielens Dataset. Individuals are described by categorical attributes (gender,occupation) and a numerical attribute (age). Entities are described by a categorical attribute augmented with a taxonomy (genres) and a date percieved as a numerical attribute (releaseDate). Outcomes are numerical (totally ordered).

In this thesis, we are interested in **characterizing** exceptional behaviors in behavioral datasets. For now, we do not introduce what kind of exceptional behaviors we are looking for, though we introduce the concepts required to characterize such peculiarities. For this, two concepts are central and recurrent through this thesis: **Groups** and **Contexts** whose generic definitions are given below (Definition 1.1.2 and Definition 1.1.3). In short, **Groups** characterize subsets of individuals in G_I and **Contexts** characterize subsets of entities in G_E.

Definition 1.1.2 — Group. A group u is a selection predicate which, when applied over a behavioral dataset $\mathscr{B}$, returns a subset of individuals $G_I^u \subseteq G_I$ for which the selection predicate holds:

$$G_I^u = \{i \in G_I \,|u(\mathscr{B},i) = \text{True}\} \text{ with } \mathscr{B} = \langle G_I, G_E, O, o\rangle$$

We depict a group in a behavioral dataset in example 1.1.

■ **Example 1.1** Given the behavioral dataset $\mathscr{B} = \langle G_I,\ G_E,\ O,\ o\rangle$ depicted in Table 1.2, the following group description:

$$u = \langle\, (\texttt{gender},\texttt{F}), (\texttt{age},[25,55]), (\texttt{genre},\texttt{comedy})\rangle$$

Covers all female individuals whose age is in between 25 and 55 in the G_I who reviewed comedy movies, i.e. $G_I^u = \{i_2, i_3\}$.

■

Definition 1.1.3 — Context. A context c is as a selection predicate which, when applied over a behavioral dataset $\mathscr{B}$, returns a subset of entities $G_E^c \subseteq G_E$ for which the selection predicate holds:

$$G_E^c = \{e \in G_E \,|c(\mathscr{B}, e) = \text{True}\} \text{ with } \mathscr{B} = \langle G_I, G_E, O, o\rangle$$

Several description languages can serve to characterize subsets of individuals or subset of entities. For attribute-value data, the most common and easy-to-interpret languages are propositional languages, where subsets of data are characterized by conjunctions of predicates of the corresponding attributes (Duivesteijn, Feelders, and Knobbe, 2016; Kloesgen, 2000; Lemmerich et al., 2016; Omidvar-Tehrani, Amer-Yahia, and Borromeo, 2019). For now, we confine ourselves to such generic definition. Having in mind what are behavioral data, we give in the following section a brief overview of the state-of-the-art of behavioral data analysis.

1.2 Behavioral Data Analysis

With the advent of platforms collecting and curating data related to various domains such as governments data, education data, environment data, product ratings, social network data, outpatient data, more and more behavioral data are available online. This offers an unparalleled opportunity to study the behavior of individuals and the interactions between them. This attracted the interest of both researchers and practitioners from various disciplines such as, social network analysis (Wasserman and Faust, 1994), biology and medicine (De Nooy, Mrvar, and Batagelj, 2018; Zitnik et al., 2019), political analysis (Clinton, Jackman, and Rivers, 2004), psychology (Smith and Osborn, 2004), journalism (Cohen et al., 2011), education (Romero and Ventura, 2013; Romero et al., 2010), marekting (Erevelles, Fukawa, and Swayne, 2016), commerce (Kohavi, 2001), etc.

One of the appealing possibilities that behavioral data analysis can deliver, is the study of how groups of individuals sharing the same characteristics (e.g. young students, smoking patients, left-wing parliamentarians) behave with regards to entities of interest (e.g. horror movies, chemotherapy, European integration related matters). Pieces of information uncovered from such data can help both novice and seasonal analysts to generate hypotheses on group behaviors and to investigate them in keeping with exploratory data analysis (Behrens, 1997; Tukey, 1977). In this spirit, User Group Analytics (UGA) (Omidvar-Tehrani and Amer-Yahia, 2019; Omidvar-Tehrani, Amer-Yahia, and Borromeo, 2019) brings under its umbrella a broad range of literature approaches that address the task of discovery, exploration and visualization of user group behaviors. In a nutshell, UGA can be performed along three principled components summarized below (cf. (Omidvar-Tehrani and Amer-Yahia, 2019)):

Discovery: it concerns the set of approaches that strive to discover a collection of *interesting* groups $S \subseteq 2^{G_I}$ given a behavioral data $\mathscr{B}$ with regards to some property of interest φ and multiple optimization criteria. Typically, this class of methods can be divided into two complementary categories:

Global Behavior Model: this category encompasses techniques whose aim is to provide a comprehensive and global characterization of the behavior of the whole population of interest. The most typical methods are: community detection (Fortunato, 2010; Pool, Bonchi, and Leeuwen, 2014; Rossetti and Cazabet, 2018) and clustering (Xu and II, 2005). For example, one can build a similarity graph where each vertex represents an individual from the underlying population G_I and each edge represents the similarity between two individuals. Using this data and by applying, for instance, Louvain algorithm (Blondel et al., 2008), one can extract groups where similar behaving individuals are put together. In this spirit, Amelio and Pizzuti, 2012 study the voting behavior in the Italian parliament based on, amongst other techniques, community detection. Similarly, Jakulin et al., 2009 propose to study the US Senators voting behavior using agglomerative hierarchical clustering algorithm (Murtagh and Contreras, 2012). In a related effort to analyze political related data, Garimella et al., 2018 investigate how to characterize controversy on social media (e.g. in Twitter) given a topic of interest. In a nutshell, the proposed approach start by building a conversation graph where vertices represent users and edges represent interactions between them. Next a graph partitioning technique (Karypis and Kumar, 1995) is used to produce two disjoints partitions (aka. the two sides of the debate) on the conversational graph. Last, a controversy measure is used to evaluate how controversial the topic is, by using, for instance, betweenness centrality (Freeman, 1977).

Local Behavior Model: this category refers to the set of methods that attempt to characterize the behavior of sub-populations rather than the whole population. description-oriented community detection (Atzmueller, 2017), multi-objective group discovery (Das et al., 2011; Omidvar-Tehrani et al., 2016), patients cohorts discovery (Li et al., 2005; Mullins et al., 2006), subgroup discovery (Grosskreutz, Boley, and Krause-Traudes, 2010), etcetera. For instance, Li et al., 2005 propose the task of identifying risk patterns in medical data where each patient is labeled by a target class: abnormal (disease, identified risk) or normal. In summary, the aim is to identify from such data a group of patients (cohort) characterized by demographic and inpatient attributes where a high risk is observed. Similarly, one can leverage educational data to identify influencing factors on students' success rate. In this perspective, Lemmerich, Ifl, and Puppe, 2011 discuss how subgroup discovery can be utilized to mine for groups of students where the drop-out is relatively high compared to the rest of students. In contrast to the community detection approaches that aim to characterize the global behavior model mentioned in the former category, the goal of COMODO (Atzmueller, Doerfel, and Mitzlaff, 2016) is to identify top-k communities from a given behavioral data (seen as an attributed graph) using some adapted interestingness measure (e.g. Newman's modularity (Newman, 2004)). Each uncovered community is characterized by the set of descriptive attributes augmenting the behavioral data in question.

Exploration: it concerns the set of approaches that provide an in-depth understanding of groups by navigating the space of groups S (that may be provided by the *discovery* step). In this category, the end-user is an active part of the process of exploration. This process can be seen as a sequence of interactive steps (Dzyuba, 2017; Omidvar-Tehrani, Amer-Yahia, and Termier, 2015; Van Leeuwen, 2014). In short, each step requires an input group $u \in S$ provided by the end-user. An exploration phase consists in looking within the provided group (Sozio and Gionis, 2010) or around it (Omidvar-Tehrani, Amer-Yahia, and Lakshmanan, 2018), returning a collection of groups in the powerset 2^S. The exploration process restarts by considering the output of the previous step as the input collection of groups, from which the end-user picks her next group of interest. For instance, Omidvar-Tehrani et Al. propose GNavigate (Omidvar-Tehrani, Amer-Yahia, and Borromeo, 2019; Omidvar-Tehrani, Amer-Yahia, and Termier, 2015) an interactive tool that enables to navigate among groups of individuals which are as diverse as possible while covering some seed group given upfront by the end-user. Interactive database exploration (Dimitriadou, Papaemmanouil, and Diao, 2014; Huang et al., 2018) makes it possible to select some individuals to form group of interest, which evolves and converges after successive iterations toward the terminal relevant group. This is done by actively integrating the end-user feedback in the underlying classification model (e.g. decision tree (Breiman et al., 1984)). In the same vein, Siren (Galbrun and Miettinen, 2018) enables to interactively explore redescriptions (Galbrun and Miettinen, 2017) of some starting sub-population which can evolve through multiple interactions of the end-user with the system by modifying either the characterization (description) of some returned subset of individuals or by updating the subset itself.

Visualization: it concerns approaches that transform a collection of groups (may consists in a single group) $S \subseteq 2^{G_I}$ to visual variables through visual views. Several techniques in the literature fall into this category. Graph visualization can be applied when social links between individuals are available (Herman, Melançon, and Marshall, 2000). For instance, Vizster (Heer and Boyd, 2005) enables to visualize social networks users and community structures. Also g-Miner (Cao et al., 2015) allows to visualize multivariate graphs. Multidimensional scaling (MDS) (Cox and Cox, 2000) can be employed to graphically represent groups of individuals by leveraging a pairwise distance based on their outcomes. For example, Jakulin et al., 2009 employ Rajski's distance (Rajski, 1961) to illustrate dissimilarities between parliamentarians based on their votes. Similarly, in the political sphere, Poole and Rosenthal, 1985 propose Nominate, a MDS technique tailored specifically for the analysis and visualization of legislative roll-call voting behavior. Time-based visualization can also be crucial to understand trends of groups behavior. For example, Silva, Spritzer, and Freitas, 2018 propose a tool which provide the big picture of groups cohesiveness over time by leveraging similarity between actions expressed by the individuals comprising the group of interest. Furthermore, one can utilize off-the shelf softwares such as Gephi[16] or Tableau[17] to visualize either raw behavioral data or results obtained by pre-processing such as graphs of similarities between actions of individuals.

[16] https://gephi.org/
[17] https://www.tableau.com/

Above, we gave a brief overview of UGA framework components which revealed the rich array of methods available for analyzing behavioral data and the behavior of groups from various perspectives. The scope of this thesis falls within the perimeter of the **discovery component**. Recall that the objective in such a component is to transform an input raw behavioral data to a concise collection of "interesting" patterns (e.g. groups) with regards to some property of interest. More precisely, we are interested in locally characterizing exceptional behavior of groups by using the descriptive attributes to unveil easily-interpretable insights. This pertains to the second category "**local behavior model**" in the discovery component. We review below some existing approaches specifically tailored for **attribute-based discovery** of interesting groups in behavioral data.

Representative Groups' Discovery: the goal here is to extract groups of individuals $S \subseteq 2^{G_I}$ that best represent a selected group or selected distribution of ratings. For instance, Das et al., 2011 aim is to identify, given a probe group (e.g. users who rated Toy Story), subgroups of raters that substantially agree or disagree while using the average rating within the group as an interestingness measure. Extensions have been proposed to enable multi-objective groups identification, thanks to more complex statistical measure: rating distributions (Amer-Yahia et al., 2017; Omidvar-Tehrani, Amer-Yahia, and Borromeo, 2019; Omidvar-Tehrani, Amer-Yahia, and Termier, 2015; Omidvar-Tehrani et al., 2016). These approaches take into account several criteria as diversity, coverage, size or proximity with a desired opinions distribution (e.g. polarized opinions, homogeneous opinions, etc.). Groups discovered can be abstracted in a smaller number of groups using the descriptive attributes to reduce information overload for the end users (Omidvar-Tehrani and Amer-Yahia, 2017).

Subgroup Discovery: given an input behavioral dataset $\mathscr{B}$ and an interestingness measure φ defined according to the aim of study, the objective is to uncover a collection of interesting groups $S \subseteq 2^{G_I}$ essentially with regards to φ (e.g. top-k groups). Standard Subgroup Discovery (Atzmueller, 2015; Herrera et al., 2011; Klösgen, 1996; Wrobel, 1997) (detailed in Chapter 2) encompasses several techniques of this type. Here, we give two examples where the methods were specifically designed for the category of data we are interested in (e.g. votes). For instance, Grosskreutz, Boley, and Krause-Traudes, 2010 investigate election result data to study what socio-economic variables, determining a subpopulation, characterize a voting behavior that substantially differ from the global voting behavior of the whole population. For this task, the authors compute for each subpopulation: the mean vector representing the share of votes of each party and compare it against the mean vector representing the share of votes of each party for the whole population. For this comparison, they propose an interestingness measure φ which evaluates the weighted difference between the two vectors where the weight is equal to the size of the sub-population (group) in question. In the same spirit, Du, Duivesteijn, and Pechenizkiy, 2018 propose ELBA to mine for subgroups of students that have significantly high dropout rate compared to the whole population of study. The authors utilize, among others, the Weighted Relative Accuracy (WRAcc) (Lavrač, Flach, and Zupan, 1999) as the interestingness measure φ to evaluate to what extent the dropout rate changes for some subpopulation.

Preference Mining: such approaches are interested in finding socio-demographic factors that substantially impact the preferences of subpopulations. For instance, consider a political survey where respondants emit their vote preferences for particular national parties (e.g. p_1, p_2, p_3). Each individual i in G_I is associated to her personal details (e.g. age, family income) along with her vote preference (e.g. $p_2 \succ p_1 \succ p_3$). One can obtain the overall preference of the entire population by aggregating individual preferences, then look for subgroups where the aggregated preference relation between subsets of the parties significantly differ from the aggregated overall preference. It is the goal of Exceptional Preference Mining (EPM) (Sá et al., 2016; Sá et al., 2018) which is grounded on the Exceptional Model Mining framework (Duivesteijn, Feelders, and Knobbe, 2016). EPM aims to uncover exceptional subgroups where preference between some labels significantly differs from the overall preference. In EPM, the authors propose several interestingness measures to gauge the exceptionality of a subgroup. For example, the labelwise measure aims to identify subgroups where only a single label behaves differently, disregarding the interaction between the other labels.

Transition Behavior Mining: the objective here is to find exceptional transition behavior of groups of individuals. In this case, the behavioral dataset $\mathscr{B}$ given as input describes a collection of individuals $i \in G_I$ and their transitions (e.g. from location a to location b at a time t) between entities $e \in G_E$ (e.g. locations, web pages, etc.). To discover and extract hypotheses about human navigation, Lemmerich et al., 2016 propose to model the transition behavior of a group by a first-order markov chain (Norris, 1998). In order to extract the exceptional transition behaviors, the proposed algorithm mines for subgroups whose fitted markov transition' matrix significantly differs (using an adapted manhattan distance) from the one computed over the entire population. Similarly, HypTrails (Singer et al., 2015) extended to MixedTrails (Becker et al., 2017) operationalizes bayesian model comparsion on simple markov chains (HypTrails) and hetergeneous mixed comparison markov chains (MixedTrails). Although, these methods do not consider descriptive attributes to extract groups but rather evaluate the transition behavior of an input group. Similarly as the work of Lemmerich et al., 2016, Kaytoue et al., 2017 and Bendimerad et al., 2017a strive to find exceptional transition of groups of individuals between areas in a city. To this aim, the behavioral data is modeled as an attributed graph where vertices depict places and edges represent the trips. This enables the enumeration of contextual subgraphs where each represents a subset of places characterized by means of the descriptive attributes. Each contextual subgraph may suggest an exceptional transition behavior according to the used interestingness measures φ. Several interestingness measures have been investigated to measure to what extent the number of transitions in a subgraph is high compared to the expected number of transitions. The latter being estimated either by considering a simple contingency matrix (Bendimerad et al., 2017a) or more sophisticated models (Kaytoue et al., 2017) as the gravity model (Zipf, 1946) and the radiation model (Simini et al., 2012).

1.3 Research Questions

While each of the aforecited methods aims to uncover various insights from behavioral data, they share in common the fact that entities attributes $\mathscr{A}_E$ and individuals attributes $\mathscr{A}_I$ are confounded. Both collections of attributes serve to characterize groups of individuals (cf. definition 1.1.2). For example, in **representative group discovery**' methods, a group in a movies review dataset can be described by $\langle$ (`gender`,`female`), (`location`,`DC`), (`genre`,`comedy`)$\rangle$ which contains individual who are all females living in Washington D.C. and who expressed at least one rating over a comedy movie. In contrast, in **preference mining**, only individual attributes are used to characterize groups with exceptional preferences, leaving the labels (perceived as entities) without characterization. By merging $\mathscr{A}_E$ and $\mathscr{A}_I$, some insights of crucial importance cannot be unveiled. For instance, consider the European parliament voting dataset (an excerpt is given in table 1.1) featuring parliamentarians voting for legislative procedures. Each parliamentarian is associated to his national party, country and political group and each voting session is characterized by its date and the topics of interest. Using this dataset, a data journalist can be interested in answering the following question:

What are the controversial topics between French parties representatives in the European Parliament?

At first sight, the question seems simple, yet finding an answer is a daunting task if the journalist investigates manually all possible topics treated in the European parliament between all possible combinations of French national parties. If we take a deep look in the question, it can be brought down to three elements highlighted below:

*What are the **[controversial] [topics] [between French parties]** in the European Parliament?*

In this configuration, the french parties are the **groups** (cf. definition 1.1.2) of interest, the topics represent the **contexts** (cf. definition 1.1.3) containing the voted legislative procedures. Both contexts and groups must be characterized and enumerated independently by their corresponding attributes: groups by using descriptors from $\mathscr{A}_I$ and contexts by using descriptors from $\mathscr{A}_E$. Finding controversial topics now requires the definition of proper interestingness measures that objectively capture such an information by analyzing, for instance, the inter-group agreement observed in each context (e.g. agriculture, judicial matterns, Citizen's rights) related legislative procedures.

This is a challenging task as it requires to handle both complex search spaces induced by the set of all possible groups and all possible contexts that one can characterize using attributes from $\mathscr{A}_I$ and $\mathscr{A}_E$ respectively. Moreover, one needs to only return the most relevant combinations of groups and context to avoid overwhelming the end-user with too many options. As discussed earlier, the state-of-art does not offer an off-the-shelf method that enables providing a ready answer to the aforementioned question or to similar ones (e.g. what are the contexts that divides groups sharing naturally the same political line ?). Having these elements in mind, we formulate in the following the two main complementary **research questions** for which we endeavor to provide answers in this thesis:

Research Question. 1 How to characterize, discover, summarize and present **exceptional (dis) agreement <u>between</u> groups** (sub-populations) in **behavioral data** ?

Research Question. 2 How to characterize, discover, summarize and present **exceptional (dis) agreement <u>within</u> groups** (sub-populations) in **behavioral data** ?

These are the challenges we are addressing in this thesis. Interestingly, these challenges pertain to the scope of the generic framework of **Subgroup discovery** (Klösgen, 1996; Wrobel, 1997), a popular task in the data mining research field (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). Subgroup discovery has been extended recently to **Exceptional Model Mining** (Leman, Feelders, and Knobbe, 2008) (both detailed in Chapter 2). The techniques falling in Subgroup Discovery (**SD**) or Exceptional Model Mining (**EMM**) frameworks aim to discover interpretable patterns in the data that stand out w.r.t. some property of interest.

1.4 Contributions

This thesis bring three main contributions that aim to provide solutions in response to **R.Q. 1** and **R.Q. 2**. These contributions are summarized in what follows:

1.4.1 Contribution 1: From Behavioral Data to Exceptional Inter-group (dis)agreements

In response to **Research Question 1**, we propose to define the task of "*Discovering Exceptional (Dis)Agreement between Groups*" grounded on SD/EMM. The solution of such a task is a list of triples (patterns) of the form (c, u_1, u_2); where c is a context (cf. Definition 1.1.3) and (u_1, u_2) are two groups (cf. Definition 1.1.3), (c, u_1, u_2) reads as follows:

■ *There is an **exceptional disagreement** (or **agreement**) between group u_1 and group u_2 in the context c compared to the overall **inter-group agreement*** ■

To tackle this task and retrieve such patterns, we first define the underlying **search space** corresponding to all the characterizable[18] contexts and groups. Subsequently, we define an **inter-group agreement measure** to evaluate to what extent two groups are in agreement with regards some subset of entities. This enables the definition of **interestingness measure** which assesses how **exceptional** the inter-group agreement observed in a context is, compared to the one observed for the whole collection of entities. Once these elements are defined, we propose two algorithms: DEBuNk and Quick-DEBuNk.

- DEBuNk is an exhaustive branch and bound algorithm which guarantees the retrieval of all the desired patterns. To make this possible, we propose several optimizations to avoid enumerating uninteresting patterns. For instance, optimistic estimates are used to safely prune as soon as possible unpromising areas of the search space.

[18]Characterizable subset means a subset of the data that can be retrieved by a conjunctive query on the attributes value domains. This notion will be properly formalized in Chapter 2 and appropriately instantiated afterward for both tasks associated to: **Contribution 1** (Chapter 3) and **Contribution 2** (Chapter 4).

- Quick-DEBuNk is a stochastic algorithm which heuristically approximates the exact solution of the task of finding exceptional inter-group agreement. This algorithm is proposed to render the solving of such a task tractable, since an exhaustive traversal of the search space is computationally expensive even when optimizations are used.

In order to evaluate both the usefulness of exceptional inter-group agreement patterns and the efficiency of the proposed algorithms (DEBuNk and Quick-DEBuNk), a thorough experimental study is performed over four real-world datasets relevant to three different application domains: political analysis, rating data analysis and healthcare surveillance.

A preliminary version of this contribution has appeared in the proceedings of the The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'2017) (Belfodil et al., 2017a). An extended version has been accepted for publication in Data Min. Knowl. Disc. Jounral (Belfodil et al., 2019c).

1.4.2 Contribution 2: From Behavioral Data to Exceptional Intra-group (dis)agreements

In response to **Research Question 2**, we propose the task of "*Discovering statistically significant exceptional (Dis)agreement within Groups*" grounded on SD/EMM. The solution of such a task is a list of pairs (u,c) where u is a group and c a context; (u,c) reads as follows:

■ *There is a **systematic exceptional disagreement** (or **agreement**) among members of the group u in the context c compared to what is **expected** in overall terms.* ■

Along the same lines as in the previous contribution, we first model the underlying **search space** by identifying and formally characterizing all candidate patterns (groups and contexts). We propose an adequate **intra-group agreement measure** to capture how consensual/conflictual the situation is between members of a group when a subset of entities is selected. Particularly, the proposed measure needs to handle the sparsity encountered in behavioral data. Subsequently, we formally define an **interestingnesss measure** which rates how exceptional a contextual intra-group agreement is. Once these elements are defined, we devise an algorithmic solution, named DEvIANT, to solve efficiently and optimally the search of such patterns. To do so, several optimizations are integrated into the algorithm to avoid enumerating uninteresting patterns. Finally, we study the efficiency of the proposed algorithm DEvIANT and show the interpretability of such patterns via two application domains: political analysis and rating data analysis.

This contribution has appeared in the proceedings of the the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'2019) (Belfodil et al., 2019a).

1.4.3 Contribution 3: A web platform for exceptional voting behaviors analysis

The third contribution pertains to the two fundamental research questions that we ask in this thesis (**R.Q. 1** and **R.Q. 2**) and comes as a use case which links the two precedent contributions. Hence, providing two complementary angles of exceptional (dis)agreement in behavioral data: between and within groups. With this aim in mind, we propose ANCORE, a web-platform[19] which is tailored specifically for the analysis of exceptional behaviors in

[19] The web platform is available online on https://contentcheck.liris.cnrs.fr.

voting data (e.g. European Parliament Voting Data and United States Congresses).

ANCORE allows, via a user-interface, to query an input voting dataset perceived as a behavioral dataset (cf. Definition 1.1.1) for exceptional (dis)agreement within and between groups. Moreover, for a better understanding and interpretation for each pattern, the platform offers a fine-grained visualization tool which offers the possibility to retrieve the data that were used to assess the exceptionality of the findings. In order to evaluate the usefulness of such a tool, we give two exemplary applications which are relevant to computational journalism field: **Fact-checking** and **Lead-finding**. For fact-checking, we paints several portraits of how ANCORE can be used to provide contextual counter-arguments, if possible, for some given claim on the voting behavior of parliamentarians. Furthermore, for lead-finding which consists on finding interesting information nuggets from the data that can raise further investigations or stories around them, we discuss several scenarii using voting data.

This contribution extends the work that appeared in the proceedings of Extraction et Gestion des connaissances - Demo Track (EGC'2019) (Lacombe et al., 2019).

1.5 Thesis Outline

This chapter presented the context of this thesis: first, by depicting the data we are interested in, i.e. behavioral data, and by briefly introducing behavioral data analysis. Upon this background, the chapter draws particular attention to the main research questions motivating the contributions of this thesis. The remainder of this thesis is organized as follows:

- **Chapter 2** is devoted to the presentation of the theoretical background of the three aforementioned contributions, namely: Subgroup Discovery (SD) and Exceptional Model Mining (EMM). The chapter reviews the state-of-the-art works and outlines the main building blocks of both frameworks. This building blocks are required to formally define and optimally solve the underlying mining tasks.
- **Chapter 3** details **Contribution 1** by introducing the problem of discovering exceptional (dis)agreement between groups in behavioral data. The chapter expands the possibilities of SD/EMM framework by instantiating its building blocks according to the problem statement. The chapter discusses an algorithmic solution (DEBuNk and Quick-DEBuNk) to the problem and evaluate its efficiency through a comprehensive experimental evaluation.
- **Chapter 4** concerns **Contribution 2**, it introduces the problem of discovering exceptional (dis)agreement within groups in behavioral data. In the same spirit as the precedent chapter, it instantiates SD/EMM building blocks for such a task in order to propose an adequate and efficient algorithm (DEvIANT) for solving the problem of finding the desired patterns. A thorough experimental evaluation is conducted to evaluate the effectiveness and efficiency of the proposed algorithm.
- **Chapter 5** details **Contribution 3** by presenting ANCORE. This tool consolidates the results of the two precedent chapters by illustrating how they can be used in the context of a Computational Journalism process.
- **Chapter 6** concludes this thesis by summarizing its contributions and by discussing opportunities for future work.

2

Subgroup Discovery and Exceptional Model Mining

Subgroup Discovery and its extension Exceptional Model Mining provide generic frameworks that enable to define descriptive data mining tasks and to efficiently solve them. This chapter addresses the formalization of these two frameworks. Furthermore, it reviews state-of-the-art works that are relevant in the scope of this thesis. However, it is not the sole aim. Our endeavor via this chapter, is to create the theoretical foundations on which this thesis is grounded. Moreover, our aim is to consolidate the concepts required for the understanding of the main contributions of this work discussed in details in the following Chapters (Chapter 3 and Chapter 4).

2.1 Introduction

Subgroup Discovery and Exceptional Model Mining provide generic frameworks that can be used to model several mining tasks while handling appropriately the complexity of both the underlying search space and the interestigness measures. The aim of this chapter is to review the work that has been done in the state-of-the-art and to build a theoretical background of the algorithms proposed in Chapter 3 and Chapter 4.

Scientists have always seen Exploratory Data Analysis (EDA) as an important research area since its introduction (Tukey, 1977). Among the various EDA techniques that aim to maximize insight into datasets and uncover underlying structures, Subgroup Discovery (SD) (Atzmueller, 2015; Herrera et al., 2011; Klösgen, 1996; Wrobel, 1997) is a generic data mining task concerned with finding regions in the data that stand out with respect to a given target[1]. Many other data mining tasks have similar goals as SD, e.g., emerging patterns (Dong and Li, 1999), significant rules (Terada et al., 2013), contrast sets (Bay and Pazzani, 2001) or classification association rules (Liu, Hsu, and Ma, 1998). However, among these different tasks, SD is known as the most generic one, especially SD is agnostic of the data and the pattern domain. For instance, subgroups can be defined with a conjunction of conditions on symbolic (Lavrač et al., 2004) or numeric attributes (Atzmüller and Puppe, 2006; Grosskreutz and Rüping, 2009) as well as sequences (Grosskreutz, Lang, and Trabold, 2013). Furthermore, the single target can be discrete or numeric (Lemmerich, Atzmueller, and Puppe, 2016). Exceptional Model Mining (EMM) (Leman, Feelders, and Knobbe, 2008), while sharing exactly the same exploration space (i.e., the description space), extends SD by offering the possibility to handle complex targets, e.g., several discrete attributes (Duivesteijn, Feelders, and Knobbe, 2016; Duivesteijn et al., 2010; Leeuwen and Knobbe, 2012) or graphs (Bendimerad, Plantevit, and Robardet, 2016; Bendimerad et al., 2017b; Kaytoue et al., 2017).

Roadmap. The remainder of this section is organized as follows. We first introduce the generic framework of Subgroup discovery in section 2.2 and discuss the related works. Subsequently, in Section 2.3 we introduce its generalization called Exceptional Model Mining and review its literature. Section 2.4 summarizes the concepts introduced in both precedent sections. Moreover, it presents two guideline algorithms which serves as backbone for the algorithms presented in the next chapters. Section 2.5 concludes the chapter by discussing the potential and limitations of state-of-the-art SD/EMM techniques for behavioral data analysis.

2.2 Subgroup Discovery

Subgroup discovery as a research field, although called *Data Surveying*, dates back to the seminal paper of Siebes, 1995 where it is described as “the discovery of interesting subgroups”. The term *Subgroup Discovery* was coined by Klösgen, 1996 and Wrobel, 1997. It is defined as the problem of finding statistically unusual subgroups in a given database (Wrobel, 1997). Below, we give a generic definition of SD, first introduced in (Wrobel, 2001) and pointed out in a recent survey (Herrera et al., 2011).

[1]Subgroup discovery definitions corresponds here more to supervised descriptive rule discovery (Carmona, Jesus, and Herrera, 2018; Herrera et al., 2011; Kralj Novak, Lavrač, and Webb, 2009) than the original definition (Klösgen, 1996; Siebes, 1995; Wrobel, 1997)

Definition 2.2.1 — Subgroup Discovery. (***Generic Definition***) In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically “most interesting", i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

For example, consider a patient dataset where each patient is associated with her demographic attributes (e.g. age) along with her inpatient data (e.g. average heart beat rate per minute). More over each patient is classified to a variable which states whether or not she has developed lung cancer. One interesting investigation that can be conducted over such a dataset is the search of subgroups whose lung cancer is substantially higher than the average. Figure 2.1 illustrates such a dataset.

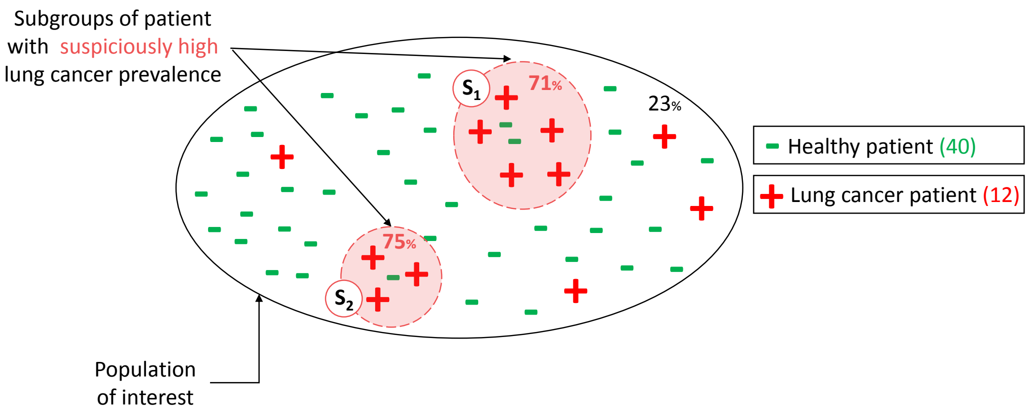


Figure 2.1: A patient dataset describing individuals and whether they have a lung cancer.

Figure 2.1 shows an example where in the overall terms, only 23% of the patients are diagnosed with a lung cancer. In this Figure, two subgroups are highlighted S_1 and S_2 which bring to the fore two subgroups where the cancer prevalence is substantially higher than in the rest of the dataset. Considering such subgroups of patients, a medical researcher can be interested in finding an answer to the following question:

■ *What are the common characteristics shared by patients of subgroup S_1 (or S_2) ?* ■

This is the prime objective of Subgroup Discovery, finding interpretable links between different characteristics (descriptive variables) and the property of those individuals we are interested in (e.g. cancer incidence in this example) as argued by Siebes, 1995: “Clearly, the result of a data mining session should never be a listing of the members of such a subgroup. Rather, it should result in a (characteristic) description of the subgroup".

Consider a collection G of records g and its underlying schema $\{a_1, a_2, ..., a_m, t\}$ (the schema as previously presented is extended with a new attribute t). Each attribute a_j has a domain of interpretation, noted $\text{dom}(a_j)$, which corresponds to all its possible values. Attributes a_1, a_2,..., a_m are called **descriptive attributes** and are denoted $\mathscr{A}$. We have $\text{dom}(\mathscr{A}) = \text{dom}(a_1) \times ... \times \text{dom}(a_m)$. t is an attribute called **target attribute** and represent the property of interest. The target attribute has also a domain of its possible values denoted $\mathrm{dom}(t)$. Hence, each record $r \in G$ can be seen as a tuple $g = (a_1^g, \ldots, a_m^g, t^g) \in \mathrm{dom}(\mathscr{A}) \times \mathrm{dom}(t)$ where a_j^g corresponds to the value of $a_j \in \mathrm{dom}(a_j)$ in the record g and $t^g \in \mathrm{dom}(t)$ the associated target value for g. Tables 2.1 and 2.2 give two standard SD dataset extracted respectively from the behavioral datasets depicted in tables 1.1 and 1.2. The two datasets will serves for running example through this section. The two datasets differ mainly on the domain of interpretation of the target attribute. Table 2.1 describes a dataset with a categorical target. Table 2.2 describes a dataset with a numerical target.

idi	country	group	national party	age	**Vote**
i_1	France	S&D	PS	26	For
i_2	France	PPE	LR	30	For
i_3	France	PPE	LR	40	Against
i_4	France	ENF	RN	45	Against
i_5	Germany	ENF	BP	26	For
i_6	Germany	PPE	CDU	30	For
i_7	Germany	S&D	SPD	40	Against
i_8	Germany	PPE	CSU	45	Against

Table 2.1: Example of behavioral dataset - European Parliament dataset depicting the votes of parlementarians for a single voting session (session 72229) concerning the second amendment of Social dumping in the European Union. The **descriptive attributes** characterizing parlementarians are: **country**, **group**, **national party** and **age**. The **target attribute** is **Vote** (**categorical attribute**) representing the voting outcome of the parliamentarians.

idi	gender	age	occupation	**Rating**
i_1	M	30	programmer	4
i_2	F	53	healthcare	5
i_3	F	48	marketing	1
i_4	M	21	healthcare	5
i_5	M	25	educator	3
i_6	F	19	educator	5
i_7	F	61	educator	4
i_8	M	55	marketing	1

Table 2.2: Example of behavioral dataset - Movielens dataset depicting the ratings of users for a single movie (Pulp Fiction). The **descriptive attributes** characterizing users are: **gender**, **age** and **occupation**. The **target attribute** is **Rating** (a **numerical attribute**) representing the rating outcome of the users for Pulp Fiction.

The aim of subgroup discovery is to find **characteristic descriptions** for **interesting** subgroups. Two important concepts are highlighted here: the notion of “**description**" and the notion of “**interestingness**". Let us first consider the descriptions.

Descriptions (also called selectors (Kloesgen, 2000)) represent by intent a subgroup which is by extent a subset of individuals $S \subseteq G$. In its most generic definition, a description d of a subgroup S is a statement in a **subgroup description language**, noted hereafter $\mathscr{D}$, that specifies the properties that must be satisfied by the subgroup records (Kloesgen, 2000). It can be seen as a selection query on the underlying database (Siebes, 1995) using the descriptive attributes. The literature abounds of possible descriptions language: itemsets (Agrawal, Imielinski, and Swami, 1993), hyper-rectangles (Grosskreutz and Rüping, 2009; Kaytoue, Kuznetsov, and Napoli, 2011; Kaytoue et al., 2011; Mampaey et al., 2012) , polygones (Belfodil et al., 2017b), sequences (Agrawal and Srikant, 1995; Grosskreutz, Lang, and Trabold, 2013; Mathonat et al., 2019), graphs (Kaytoue et al., 2017; Yan and Han, 2002) which define the space (set) of possible descriptions defining, by extent, the set of possible subsets of records that one can consider in the analysis task. In the scope of this thesis, we confine ourselves to propositional languages which are the most commonly used languages for attribute-value data (Kralj Novak, Lavrač, and Webb, 2009). In this case, descriptions are formalized as conjunction of conditions (restrictions), each corresponding to a single attribute. The subset of elements of G supporting the description is the subset of elements for which the conjunction of conditions hold. For example:

■ **Example 2.1** Given the collection G depicted in table 2.1 and the following description:

$$d = \langle \texttt{Country} \in \{\ \texttt{France}\ \} \text{ and } \texttt{age} \in [20, 39] \rangle$$

Subset $S \subseteq G$ supporting the description d is $S = \{i_1, i_2\}$. ■

Below, we give the generic definition of a description, also called the intent or pattern, in the conjunctive descriptions language $\mathscr{D}$.

Definition 2.2.2 — Description. Let G be a collection of records with $\mathscr{A} = \{a_1, ..., a_m\}$ the descriptive attributes. A **description** $d \in \mathscr{D}$ is a conjunction of **conditions** of the form $d = \langle r_1, ..., r_m \rangle$ where r_j is a membership test in a subset χ_j of the value domain $\text{dom}(a_j)$ of the attribute a_j. A description d is hence given by:

$$d = r_1 \wedge r_2 \wedge ... \wedge r_m \text{ where } r_j : a_j \in \chi_j \text{ with } \chi_j \subseteq \text{dom}(a_j).$$

Note that, if $\chi_j = \text{dom}(a_j)$, the condition r_j can be removed from d or replaced in d by the wildcard $*$ which means that the condition do not restrict the domain of possible values of the attribute a_j.

A description d characterizes a subset of records, also called the *extent*, the *support* or the *cover* of d.

Definition 2.2.3 — Extent. Let G be a collection of records with $\mathscr{A} = \{a_1, ..., a_m\}$ the descriptive attributes. Let $d = \langle r_1, ..., r_m \rangle \in \mathscr{D}$ a description $d \in \mathscr{D}$ with $r_j : a_j \in \chi_j$. The **extent** of d denoted G^d is the subset of records $g \in G$ fulfilling the conditions of d, hence:

$$G^d = \{g = (a_1^g, ..., a_m^g, t^g) \in G \text{ s.t. } \forall j \in 1..m \ : \ a_j^g \in \chi_j\}.$$

Note that, we also denote the extent of a description d by $\text{ext}(d)$ where $ext : \mathscr{D} \to 2^G$ with $\text{ext}(d) = G^d$.

Through this thesis, if no confusion can arise, the term *subgroup* is interchangeably used to express a description d or its extent G^d. Note also that the term support is used in the literature both to express the extent G^d or its cardinality $|G^d|$. To avoid confusion, we will use *cardinality* or *size* of the subgroup to refer to the number of records in G fulfilling the conditions of d.

Descriptions are partially ordered in $\mathscr{D}$ by a *specialization relationship* defined as follows.

Definition 2.2.4 — Specialization $\sqsubseteq$. Let d and d' be two descriptions from $\mathscr{D}$. d' is said to be a *specialization* of d, denoted $d \sqsubseteq d'$, iff $d' \Rightarrow d$.

As a consequence, if $d \sqsubseteq d'$ then $G^{d'} \subseteq G^d$, since each record supporting d' supports by definition d. For example:

■ **Example 2.2** Given the collection G depicted in table 2.1 and the two following descriptions:

$$d = \langle \texttt{Country} \in \{\ \texttt{France}\ \} \text{ and } \texttt{age} \in [20,39] \rangle$$
$$d' = \langle \texttt{Country} \in \{\ \texttt{France}\ \} \text{ and } \texttt{age} \in [20,39] \text{ and } \texttt{National Party} = \texttt{PS}\ \rangle$$

We have $d \sqsubseteq d'$ since $d' \Rightarrow d$. We have: $G^d = \{i_1, i_2\}$ and $G^{d'} = \{i_1\}$, thus $G^{d'} \subseteq G^d$. ■

In most standard subgroup discovery enumeration algorithms (Atzmüller and Puppe, 2006; Leeuwen and Knobbe, 2012), the search space induced by $(\mathscr{D}, \sqsubseteq)$ is explored in a top-down fashion starting from the most general description, it proceeds by atomic refinements to progress, step by step, toward more specific descriptions with regard to $\sqsubseteq$. Intuitively, an atomic refinement of a description d produces a more specific description d' by reinforcing the condition of one attribute only. Furthermore, such refinement is minimal. Such descriptions are provided by a refinement operator η.

Definition 2.2.5 — Refinement operator η. A refinement operator is function $\eta : \mathscr{D} \longrightarrow 2^{\mathscr{D}}$ that maps each description $d \in \mathscr{D}$ to its **neighbors** in $\mathscr{D}$, i.e.

$$\eta(d) = \{d' \in \mathscr{D} \text{ s.t. } d \sqsubset d' \ \wedge \ \nexists e \in \mathscr{D} \ : \ d \sqsubset e \sqsubset d'\}$$

■ **Example 2.3** Resuming the example 2.2, where the two following descriptions are:

$$d = \langle \texttt{Country} = \texttt{France} \text{ and } \texttt{age} \in [20,39] \rangle$$
$$d' = \langle \texttt{Country} = \texttt{France} \text{ and } \texttt{age} \in [20,39] \text{ and } \texttt{National Party} = \texttt{PS}\ \rangle$$

We have $d \sqsubseteq d'$, moreover $d' \in \eta(d)$ as, colloquially, d' contains only a new atomic condition `National Party = PS`. ■

For now, we confine ourselves to this high definition of the refinement operator, we shall return to this point later in section 2.2.1.

Recall that the aim of subgroup discovery is to find the collection of "**interesting**" subgroups. For this, an objective characterization of interestingness measurement is required. For this purpose, a quality measure is generally defined to evaluate the interestingness of a subgroup transforming it to a quantity in a totally ordered set, most usually $\mathbb{R}$ (Wrobel, 1997).

Definition 2.2.6 — Quality measure. A quality measure is a function $\varphi : \mathscr{D} \to \mathbb{R}$ which assigns to each description $d \in \mathscr{D}$ a real number $\varphi(d) \in \mathbb{R}$.

Whilst some of the work in the literature use the description (syntax) to compute the interestingness of a subgroup (Bie, 2011a; Lijffijt et al., 2018; Siebes, Vreeken, and Leeuwen, 2006; Vreeken, Leeuwen, and Siebes, 2011), we emphasize in the scope of this thesis, on **extent-based quality measures** that are computed exclusively using the extent of a description and mostly relies on the target value. Therefore, we will occasionally use $\varphi(G^d)$ to denote the quality of a subgroup whose description is d. Hence, by abuse of notation and to avoid overloading notations, φ is also defined on the powerset of 2^G, i.e. $\varphi : 2^G \to \mathbb{R}$. It follows that, two equivalent descriptions $d, d' \in \mathscr{D}$ in terms of their respective extent ($G^d = G^{d'}$) have the same quality, i.e. $\varphi(d) = \varphi(G^d) = \varphi(G^{d'}) = \varphi(d')$.

The notions of **subgroups** (descriptions d and extents G^d) along with the **specialization** $\sqsubseteq$ and the **quality measure** φ allows to define the task of subgroup discovery. In short, the task of subgroup discovery consists of exploring the search space defined by the description language $\mathscr{D}$ and structured with the partial order $\sqsubseteq$ in order to find a succinct list of subgroups $L = \{d_1, d_2, ...d_k\}$ ($k \in \mathbb{N}$) where each subgroups d_i observe a high interestingness score $\varphi(d)$. Additionally, a set of constraints $\mathscr{C}$ can be given by the end-user to limit the collection of valid subgroups. Usually these constraints encompasses a cardinality constraint and a minimal quality constraint. Cardinality constraints imposes that subgroups are of sufficient size. This translates to a minimum size threshold σ_G that must be satisfied by subgroups in L, i.e. $\forall d \in L : |G^d| \geq \sigma_G$. Minimal quality constraint requires that the subgroups in L are above a minimal quality threshold $\sigma_\varphi \in \mathbb{R}$, i.e. $\forall d \in L : \varphi(d) \geq \sigma_\varphi$. We give in the following, a typical subgroup discovery task which consists of finding the top-k subgroups with regard to the defined quality measure (solving the problem 2.2.1). The task is simlarly formalized as the generic problem defined in (Duivesteijn, Feelders, and Knobbe, 2016) .

Problem 2.2.1 (*Top-k* Subgroup Discovery Problem).

Given a collection G, a description language $\mathscr{D}$, a quality measure φ, a quality threshold σ_φ and a minimum support threshold σ_G, the problem is to find the list $L = \{d_1, ..., d_k\} \subseteq \mathscr{D}$ such that:

[*Validity*] $\forall d \in L : d$ *valid*, that is $|G^d| \geq \sigma_G$ and $\varphi(d) \geq \sigma_\varphi$.

[*Top-k*] $(\forall d' \in (\mathscr{D} \setminus L))\,(\forall d \in L) : \varphi(d) \geq \varphi(d')$.

To solve this problem, we need to devise an efficient **algorithm** which explores the search space $\mathscr{D}$ by smartly leveraging both its **structure** induced by the partial $\sqsubseteq$ and the properties of the **quality measure** φ. Figure 2.2 summarizes the building blocks of a subgroup discovery task. In summary, when it comes to defining a subgroup discovery task and solve it, one need to answer the three following questions:

Language: what is the description space $\mathscr{D}$ of candidate subgroups?

Interestingness: how to assess the interestingness of a subgroups (quality measure)?

Algorithm: how to explore the search space of candidate subgroups?

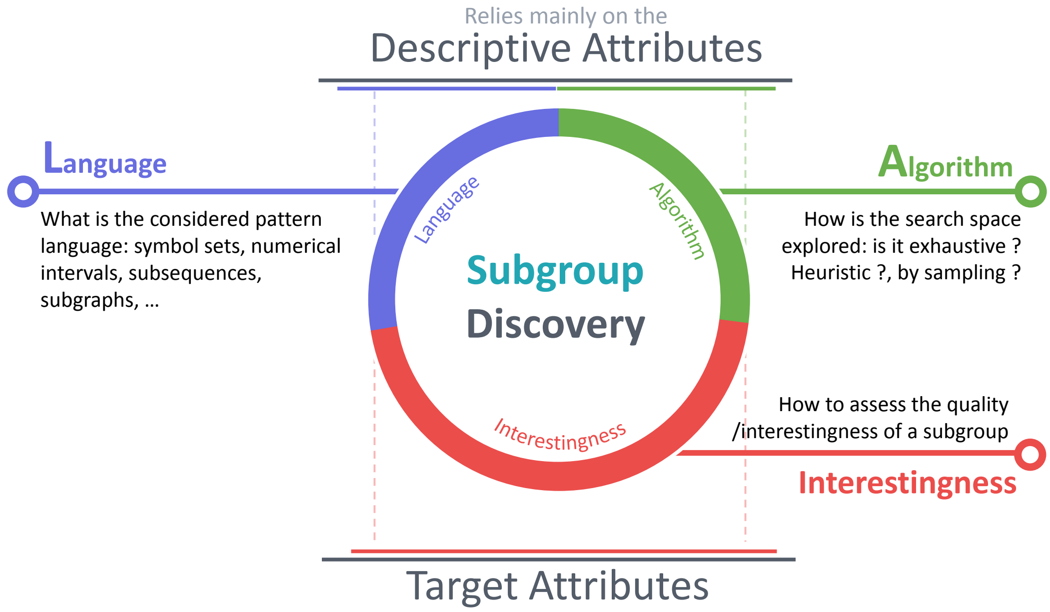


Figure 2.2: Building blocks of a subgroup discovery task (Summary)

The remaining of this section is organized as follows. We give in Section 2.2.1 an overview of how description spaces are structured and the main properties that one can leverage for an efficient enumeration of candidate subgroups. Next, we enumerate in Section 2.2.2 examples of noteworthy quality measures that can be used to assess the interestingness of subgroups. Eventually, we discuss in Section 2.2.3 several search strategies developed in the literature to explore the potentially exponential search space.

2.2.1 On Description Languages

In the scope of this thesis, we only consider descriptions that are conjunction of conditions restricting the domain of values of the descriptive attributes (cf. definition 2.2.2. These descriptions are members of a description language denoted $\mathscr{D}$ and are partially ordered by the operator $\sqsubseteq$ (cf. definition 2.2.4) which roughly translates to a logical implication. This induces a **partially ordered set** (**poset**) that is denoted henceforth $(\mathscr{D}, \sqsubseteq)$. In the search space related to a subgroup discovery task, we have on one hand descriptions from the **description language**[2] $(\mathscr{D}, \sqsubseteq)$ and in the other hand objects from the collection G. These two collections are closely related, hence a mapping linking G with $\mathscr{D}$ is essential to manipulate objects and descriptions when it comes to look for interesting subgroups. Below, we define a mapping function δ:

$$\delta : G \longrightarrow \mathscr{D}$$

δ maps each record $g \in G$ to the tightest (maximum) description $\delta(g)$ in $\mathscr{D}$ with regard to $\sqsubseteq$. Given this mapping, a record $g \in G$ supports a description d in $\mathscr{D}$ if and only if $d \sqsubseteq \delta(g)$.

[2]From now on, **description space** refers to $\mathscr{D}$, **search space** refers to $(\mathscr{D}, \sqsubseteq)$ and, if no confusion can arise, **description language** interchangeably refers to both $\mathscr{D}$ and $(\mathscr{D}, \sqsubseteq)$.

It follows that the extent of a description d can be formalized as such:

$$\text{ext} : \mathscr{D} \longrightarrow 2^G \quad, d \longmapsto \text{ext}(d) = \{g \in G \mid d \sqsubseteq \delta(g)\} = G^d \tag{2.1}$$

For the ease of presentation, we will consider for now G as a finite collection of single attributed records. Table 2.3 is extracted from Table 2.2 and gives an example of such a dataset and the mapping operator δ.

idi	age	δ(age)
i_1	30	[30,30]
i_2	53	[53,53]
i_3	48	[48,48]
i_4	21	[21,21]
i_5	25	[25,25]
i_6	19	[19,19]
i_7	61	[61,61]
i_8	55	[55,55]

Table 2.3: Example dataset G with a single numerical attribute **age**. The mapping describes δ the transformation of an attribute value to its corresponding description in $\mathscr{D}$. For numerical attributes, the most commonly used and easy to interpret language is interval language where $\mathscr{D}$ contains all intervals that one can form using the values in dom(age).

When grouped, these concepts form a **pattern setup** $(G, (\mathscr{D}, \sqsubseteq), \delta)$ (Lumpe and Schmidt, 2015) which builds upon Formal Concept Analysis (FCA) (Ganter and Wille, 1999; Wille, 1982). Although, several structures can be induced from pattern setups (Belfodil, Kuznetsov, and Kaytoue, 2018; Belfodil, Kuznetsov, and Kaytoue, 2019), we emphasize on **pattern structures** (Ganter and Kuznetsov, 2001) as they provide a sufficient framework to manipulate datasets with various complex attributes (numerical, categorical, etc.).

Definition 2.2.7 — Pattern Structure. A pattern structure is essentially a Pattern Setup: $(G, (\mathscr{D}, \sqsubseteq), \delta)$ where G is a collection of records, $(\mathscr{D}, \sqsubseteq)$ is a poset (a description space $\mathscr{D}$ partially ordered with $\sqsubseteq$). δ is a mapping function $\delta : G \longrightarrow \mathscr{D}$ which maps each record $g \in G$ to the tightest (maximum) description $\delta(g)$ in $\mathscr{D}$ with regard to $\sqsubseteq$. $(G, (\mathscr{D}, \sqsubseteq), \delta)$ is a **pattern structure** if and only if the poset $(\mathscr{D}, \sqsubseteq)$ is a meet-semilattice.

Below, we give the definition of a meet-semilattice and the important surrounding concepts. For more details about lattices and order, we invite the reader to consult (Davey and Priestley, 2002; Roman, 2008).

In what follows, $(\mathscr{D}, \sqsubseteq)$ is a poset and $S \subseteq \mathscr{D}$ an arbitrary subset.

Definition 2.2.8 — Lower bound and Upper bound of S. The lower bound (resp. upper bound) of S denoted S^l (resp. S^u) is the subset of elements in $\mathscr{D}$ that are below (resp. above) all elements in S, Formally:

$$S^l = \{d \in \mathscr{D} \mid (\forall s \in S)\ d \sqsubseteq s\} \qquad S^u = \{d \in \mathscr{D} \mid (\forall s \in S)\ s \sqsubseteq d\}$$

The lower bound concept allow to, among other things, to formalize the collection of common descriptions between records in G. For instance:

■ **Example 2.4** in Table 2.3, the description language $\mathscr{D}$ considers all possible intervals. Hence, the common descriptions between individuals i_4 and i_8 are all possible intervals $d \in \mathscr{D}$ that contains simultaneously 21 and 55. That is $\{\delta(21), \delta(55)\}^l = \{[21,21],[55,55]\}^l$. If we restricts the domain to the values appearing in the dataset (i.e. $\{19,21,25,30,48,53,55,61\}$), we have $\{\delta(21), \delta(55)\}^l$ contains all intervals whose left endpoints are lower or equal to 21 and whose right endpoints are higher than 55, that is $\{\delta(21), \delta(55)\}^l = \{[21,55],[19,55],$ $[21,61],[19,61]\}$. ■

Definition 2.2.9 — Meet and Join. The meet also called infimum or minimum (resp. join also called supremum or maximum) of a subset S denoted $\bigwedge S$ (resp. $\bigvee S$) is the greatest lower bound (resp. least upper bound) in $\mathscr{D}$ that is above (resp. below) all elements in S^l (S^u), Formally:

$$\forall d' \in S^l \text{ we have } d' \sqsubseteq \bigwedge S \text{ also } \forall d \in S \text{ we have } \bigwedge S \sqsubseteq d$$
$$\forall d' \in S^u \text{ we have } \bigvee S \sqsubseteq d' \text{ also } \forall d \in S \text{ we have } d \sqsubseteq \bigvee S$$

The meet concept allows in turn to characterize the maximum common description between records in G. An example is given below:

■ **Example 2.5** We resume the example 2.4. Given the two individuals i_4 and i_8, the infimum of the two descriptions corresponding to i_4 and i_8 is the maximum element of the set of lower bounds $\{\delta(21), \delta(55)\}^l = \{[21,55],[19,55],[21,61],[19,61]\}$. That is: $\bigwedge\{i_4, i_8\}$ $[21,55]$. ■

The definition of meets and joins makes $\wedge$ a binary operations which given two descriptions d_1, d_2 returns the maximum common description between them. i.e. $d_1 \wedge d_2 = \bigwedge\{d_1, d_2\}$. The same goes for the join operation $\vee$. By definition, the two operations are idempotents, commutatives and associatives.

Below, we give a definition of lattices from a partial order theory point of view:

Definition 2.2.10 — Meet-semilattice, Join-semilattice and Lattice. A description space with the specialization operator $(\mathscr{D}, \sqsubseteq)$ form:

1. a **meet-semilattice** if and only if every finite non-empty subset $S \subseteq \mathscr{D}$ has a meet $\bigwedge S$ in $\mathscr{D}$.
2. a **join-semilattice** if and only if every finite non-empty subset $S \subseteq \mathscr{D}$ has a join $\bigvee S$ in $\mathscr{D}$.
3. a **lattice** if and only if it is a both a meet-semilattice and a join-semilattice.

In the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$, one can extend by abuse of notation, the operator δ to map subsets F of G to their maximum common description.

$$\delta : 2^G \longrightarrow \mathscr{D} \;, F \longmapsto \delta(F) = \bigwedge F \tag{2.2}$$

Considering the two operations: ext (cf. equation 2.1) and δ (cf. equation 2.2), we can go back and forth between the description space $\mathscr{D}$ and the collection of records G.

Interestingly, these two operations form a Galois connection between the power set 2^G and $(\mathscr{D}, \sqsubseteq)$. Hence, the composite operator $\text{clo} = \delta \circ \text{ext} : \mathscr{D} \to \mathscr{D}$ is a **closure operator** (Ganter and Kuznetsov, 2001; Ganter and Wille, 1999). Using this operator one can compute the closed descriptions (also called closed patterns (Pasquier et al., 1999)) which are useful to reduce redundancy when it comes to generate all characterizable subsets in a collection G. In summary, using the aforementioned concepts, in a pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$, for every description $d \in \mathscr{D}$, $\text{clo}(d) = \delta(\text{ext}(d)) = \delta(G^d)$ is a closed description.

As stated above, the closed descriptions serves to summarize the characterizable subsets in the underlying description language $(\mathscr{D}, \sqsubseteq)$ (i.e. $\text{ext}[\mathscr{D}] = \{G^d \mid d \in \mathscr{D}\}$). These come from the fact that many descriptions in $\mathscr{D}$ may have the same extent in G, such descriptions are said to be equivalents. That is:

Definition 2.2.11 — Equivalence relationship. Let $(G, (\mathscr{D}, \sqsubseteq), \delta)$ be a pattern structure and let d, d' be two descriptions from $\mathscr{D}$. d and d' are said to be equivalents if and only if $G^d = G^{d'}$. Hence, the equivalence class of a description d is denoted $\dot{d} = \{d' \in \mathscr{D} \mid G^d = G^{d'}\}$.

Hence, the collection of closed descriptions $\text{clo}[D] = \{d \in \mathscr{D} | d = \text{clo}(d)\}$ contain a unique representative description per equivalence class of $\mathscr{D}$, each corresponding to a characterizable subset in $\text{ext}[\mathscr{D}]$. This is closely related to pattern concepts (Ganter and Kuznetsov, 2001) (linked to formal concepts (Ganter and Wille, 1999)) where each pair $(F, d) \mid F = \text{ext}(d)$ and $d = \delta(F)$ contain a closed description from $\text{clo}[D]$ and its characterizable subset in $\text{ext}[\mathscr{D}]$. These pattern concept form what the so called **concept lattice** (Ganter and Kuznetsov, 2001) which contain the smallest possible lattice representing the whole information (from the extents point of view) of the original pair lattice (d, G^d) as it provides a one to one correspondance between descriptions in $\mathscr{D}$ and characterizable subsets of G.

In subgroup discovery and considering the fact that we are interested only on extent-based quality measures, candidate subgroups can be generated solely from the concept lattice (e.g. to solve problem 2.2.1). This enables to avoid redundancy in the resulting list of interesting subgroups. Several algorithms enables to efficiently traverse the concept lattice in order to generate all candidate subgroups and their associated closed descriptions (Ganter et al., 2016; Kuznetsov and Obiedkov, 2002). We shall return to this point later in this chapter. Figure 2.3 summarizes the concepts presented so far in this section.

So far, we introduced the pattern structure in an abstract way without instantiating it to formalize the complex search space dealing with multiple heterogeneous attributes (itemsets, numerical, categorical, etc.) (see Table 2.1 and Table 2.2). Interestingly, in order to handle such description language, given to a certain extent in definition 2.2.2, one can build the lattice of descriptions on each attribute independently, and then perform a Cartesian product between these lattices to obtain the full one dealing with the whole attribute set $\mathscr{A} = \{a_1, a_2, ..., a_m\}$ (Roman, 2008).

A description in our full setting is seen as a tuple $d = \langle r_1, r_2, ..., r_m \rangle \in \mathscr{D}$ where each **condition** r_j is a restriction on the domain of values of the corresponding attribute a_j (cf. definition 2.2.2). In a such configuration, we can build a **condition space** $(\mathscr{D}^j, \sqsubseteq)$, along with its meet operation $\wedge_j$,the mapping function $\delta_j : G \to \mathscr{D}^j$ and the induced refinement

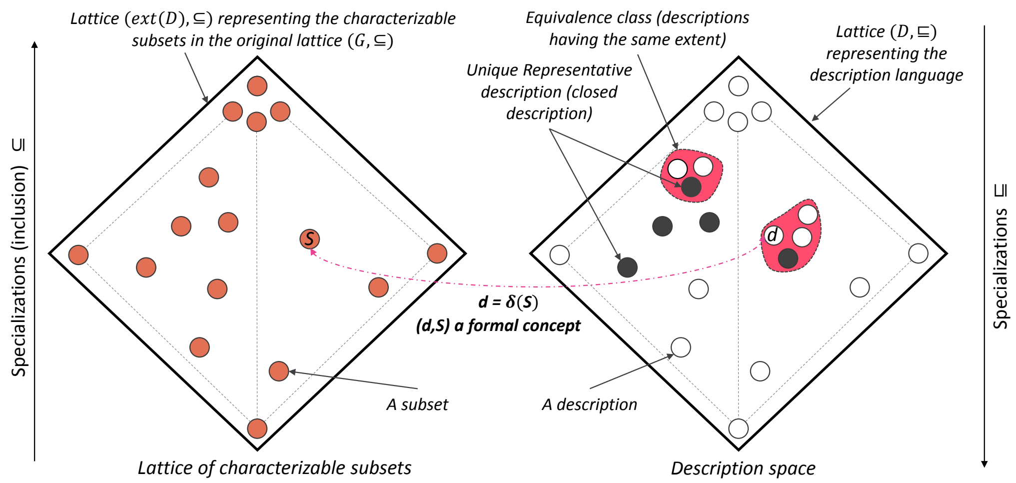


Figure 2.3: Pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ represented by its associated description langauge $(\mathscr{D}, \sqsubseteq)$ and the collection of characterizable subsets of the powerset 2^G. Note that, the collection of characterizable subsets $\text{ext}[\mathscr{D}]$ with the inclusion form a lattice. Also, the derivation operators ext and δ are order reversing, i.e. $d \sqsubseteq d' \Rightarrow G^{d'} \subseteq G^d$ and $F \subseteq F' \Rightarrow \delta(F') \sqsubseteq \delta(F)$.

operator η_j (cf. definition 2.2.5). All this being done to consider the specificity of the corresponding attribute a_j. This, will build the corresponding pattern structure for each attribute a_j, i.e. $(G, (\mathscr{D}^j, \sqsubseteq), \delta_j)$. From this point of view, it follows that:

$$\mathscr{D} = \mathscr{D}^1 \times \mathscr{D}^2 \ldots \times \mathscr{D}^m \tag{2.3}$$

$$(G, (\mathscr{D}, \sqsubseteq), \delta) = (G, (\mathscr{D}^1 \times \ldots \times \mathscr{D}^m, \sqsubseteq), \langle \delta_1(\square), \ldots, \delta_m(\square) \rangle) \tag{2.4}$$

$$\delta(a_1^g, a_2^g, \ldots, a_m^g) = (\delta_1(a_1^g), \delta_2(a_2^g), \ldots, \delta_m(a_m^g)) \tag{2.5}$$

$$\langle r_1, r_2, \ldots, r_m \rangle \sqsubseteq \langle r'_1, r'_2, \ldots, r'_m \rangle \Leftrightarrow \forall j \in 1..m \,|\, r_j \sqsubseteq r'_j \tag{2.6}$$

$$\langle r_1, r_2, \ldots, r_m \rangle \wedge \langle r'_1, r'_2, \ldots, r'_m \rangle = \langle r_1 \wedge_1 r'_1, r_2 \wedge_2 r'_2, \ldots, r_m \wedge_m r'_m \rangle \tag{2.7}$$

$$\eta(d) = \begin{array}{r} \{\langle r'_1, \ldots, r'_m \rangle \in \mathscr{D} : \exists! i \in 1..m \mid r'_i = \eta_i(r_i) \\ \text{and } (\forall j \in 1..m) j \neq i \Rightarrow r'_j = r_j\} \end{array} \tag{2.8}$$

What remains now, is to build properly the description language associated to each attribute a_j by defining the mapping function δ_j and the meet operation $\wedge_j$. The associated partial order is induced by the latter meet operation, this comes from the fact that for any two restriction r_j and r'_j from $(\mathscr{D}^j, \sqsubseteq)$, we have as usual $r_j \sqsubseteq r'_j \Longleftrightarrow r_j \wedge_j r'_j = r_j$. Recall that we use the wildcard $*$ to say that the condition is always valid and can be omitted from the description d. in the following a_j^g is the value of an arbitrary attribute a_j in a an arbitrary record $g \in G$:

Categorical attribute: if a_j is categorical, the domain $\text{dom}(a_j)$ is a collection of unordered values v (Wrobel, 1997).

Condition: it can be seen as an equality test, i.e. $a_j = v$ with $v \in \text{dom}(a_j)$ which roughly translate to $a_j \in \{v\}$.

Condition space: it is the domain of all singletons augmented with the $*$, i.e. $\mathscr{D}^j = \{*\} \cup \{\{v\} \mid v \in \mathrm{dom}(a_j)\}$ with $* = \mathrm{dom}(a_j)$.

Partial Order: Correspond to an inclusion between sets i.e. $r_j \sqsubseteq r'_j \equiv r'_j \subseteq r_j$.

Mapping: returns the singleton corresponding to the value a_j^g, i.e. $\delta^j : G \rightarrow \mathscr{D}^j$, $g \mapsto \delta^j(g) = \{a_j^g\}$

Meet operator: $r_j \wedge r'_j = \begin{cases} r_j & \text{if } r_j = r'_j \\ * & \text{else} \end{cases}$

Refinement operator: the atomic refinement of a condition $*$ gives a condition of the form $a_j = v \in \mathrm{dom}(a_j)$. Otherwise, a condition of the form $a_j = v$ does not admit any refinement, i.e. $\eta_j(*) = \{\{v\} | v \in \mathrm{dom}(a_j)\}$ and $\eta_j(v) = \emptyset$.

Numerical attribute: if a_j is numerical, the domain $\mathrm{dom}(a_j)$ is a list of totally ordered values (some total order $\leq$). This has been formalized by pattern structure tools by Kaytoue et Al. (Kaytoue, Kuznetsov, and Napoli, 2011; Kaytoue et al., 2011).

Condition: it can be seen as a membership test in an interval, i.e. $a_j \in [v, w]$ with $v, w \in \mathrm{dom}(a_j)$, this roughly translate to $a_j \in \{x \in \mathrm{dom}(a_j) \mid v \leq x \leq w\}$.

Condition space: it is the domain of all closed intervals, i.e. $\mathscr{D}^j = \{[v,w] | v, w \in \mathrm{dom}(a_j) \text{ and } v \leq w\}$, we have $* = [\min(\mathrm{dom}(a_j)), \max(\mathrm{dom}(a_j))]$.

Partial Order: Correspond to inclusion between intervals i.e. $r_j \sqsubseteq r'_j \equiv r'_j \subseteq r_j$. Given $r_j = [v, w]$ and $r'_j = [v', w']$, we have $r_j \subseteq r'_j \equiv v \leq v' \leq w' \leq w$.

Mapping: returns the degenerate interval corresponding to the value a_j^g, i.e. $\delta^j : G \rightarrow \mathscr{D}^j$, $g \mapsto \delta^j(g) = [a_j^g, a_j^g]$

Meet operator: Given $r_j = [v, w]$ and $r'_j = [v', w']$, we have:

$$r_j \wedge r'_j = [v, w] \wedge [v', w'] = [min(v, v'), max(w, w')]$$

Refinement operator: the atomic refinement of an interval $r_j = [v, w]$ returns two intervals, one resulting on minimal left change and the second resulting on a right minimal change, i.e. $\eta_j([v,w]) = \{[\underline{v}, w], [v, \overline{w}]\}$. With $\underline{v}$ (resp. $\overline{w}$) the predecessor of v (resp. successor of w) in the totally ordered domain $\mathrm{dom}(a_j)$.

Itemset attribute: if a_j is itemset, the domain $\mathrm{dom}(a_j) = 2^Z$ (the powerset of Z) with $Z = \{v_1, ..., v_l\}$ the possible items. Recall that in itemset language (Agrawal, Imielinski, and Swami, 1993) each record is associated to a set of items.

Condition: it can be seen as a superset test of the form $a_j \supseteq S$ with $S \in \mathrm{dom}(a_j)$. This roughly translates to: $a_j \in \{X \in \mathrm{dom}(a_j) \mid S \subseteq X\}$.

Condition space: it is the domain of all subsets of Z, i.e. $\mathscr{D}^j = 2^Z = \mathrm{dom}(a_j)$. We have $* = \emptyset$.

Partial Order: Correspond to inclusion between sets, i.e. $r_j \sqsubseteq r'_j \Leftrightarrow r_j \subseteq r'_j$.

Mapping: the mapping is straightforward as the condition space and the domain are equal: $\delta^j : G \rightarrow \mathscr{D}^j$, $g \mapsto \delta^j(g) = a_j^g$.

Meet operator: the meet $\wedge_j$ between two itemset conditions correspond to a simple intersection, i.e. $r_j \wedge_j r'_j = r_j \cap r'_j$.

Refinement operator: the atomic refinement of a condition (itemset) r_j correspond to adding a new item from Z in r_j, i.e. $\eta_j(r_j) = \{r_j \cup \{x\} \mid x \in Z\}$.

Clearly, one can augment the collection of attributes presented above as long as all the required components are appropriately stated, we will see in Chapter 3 a new descriptions language dealing with nominal attributes (or itemsets) augmented with a taxonomy. In a nutshell, the definition of a description given in Definition 2.2.2 can be summarized along the handled types of attributes in the following definition:

Definition 2.2.12 — Description (instantiated attributes). Let G be a collection defined over the schema $\mathscr{A} = \{a_1, ..., a_m\}$, a **description** $d \in \mathscr{D}$ is a conjunction of **conditions** of the form $d = \langle r_1, ..., r_m \rangle$ (cf. Definition 2.2.2) where r_j depends on the type of attribute a_j:

- If a_j is a categorical attribute then **condition** r_j is an equality test of the form $a_j = v$ with $v \in \text{dom}(a_j)$;
- If a_j is a numerical attribute then **condition** r_j is a membership test of the form $a_j \in [v..w]$ with $v, w \in \text{dom}(a_j)$.
- If a_j is an itemset attribute then **condition** r_j is a superset test of the form $a_j \supseteq S$ with S an itemset $\in \text{dom}(a_j)$.

A description d characterizes a subgroup $G^d = \{g \in G \text{ s.t. } d \sqsubseteq \delta(g)\}$.

With these instances of condition spaces along with the equations (2.3 — 2.8), we can use algorithms that enumerates efficiently candidate subgroups by traversing the concept lattice induced by the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ (Boley et al., 2010; Ganter et al., 2016; Kuznetsov and Obiedkov, 2002). A standard enumeration algorithm will be discussed later in Section 2.4.

Summary: in this section, we have discussed the component "**description language**" of a subgroup discovery task, by providing a deeper understanding of the search space induced by an attribute-value datasets. We discussed particularly descriptions that are conjunction of conditions over multiple and different types of attributes. We explained how to transform it to a pattern structure and defined the closure operator which will be useful to reduce substantially the number of enumerated descriptions when it comes to generate candidate subgroups. Now that candidate subgroups are characterized, we discuss in the following section 2.2.2 how to evaluate their interestingness, which represents the second component **"interestingness"** of an SD task (cf. Figure 2.2).

2.2.2 On Subgroup Interestingness Evaluation

Up to now, we have treated aspects about candidate subgroups, mostly how they are characterized by a description language and how we can enumerate them exhaustively. In subgroup discovery, one need to evaluate in an objective manner the interestingness of subgroups. This in order to return a list or the most interesting one with regard to the conducted analysis. In this section, we will discuss some examples of interestingness measures (also called quality measures). Several surveys has been proposed in the litterature to address, mostly, the discriminative power of descriptions (**patterns** (Tan, Kumar, and Srivastava, 2004), **descriptive rules** (Kralj-Novak, Lavrac, and Webb, 2009), **subgroups** (Wrobel, 1997), **association rules** (Lenca et al., 2008), **classification rules** (Todorovski, Flach, and Lavrač, 2000), etc.) with regard to a target categorical class (Geng and Hamilton, 2006; Hébert and Crémilleux, 2007; Janssen and Fürnkranz, 2006; Kirchgessner et al., 2016; Lavrac, Flach, and Zupan, 1999; Lenca et al., 2008; Tan, Kumar, and Srivastava, 2004).

In the remaining of this subsection, we have in mind a pattern structure $(G,(\mathscr{D},\sqsubseteq),\delta)$ with G a set of records associated to a single target label t (cf. Table 2.1 and Table 2.2), $(\mathscr{D},\sqsubseteq)$ the lattice of descriptions and δ a mapping between records g in G and their maximum description δ in $\mathscr{D}$ with regard to $\sqsubseteq$.

Recall that we are interested in the scope of this thesis on *extent-based interestingness measures*. That is, measures whose computations depends solely on the extent of a descriptions, i.e.

$$\forall d,d' \in \mathscr{D} \ : \ G^d = G^{d'} \implies \varphi(d) = \varphi(d') \tag{2.9}$$

Recall that we extend the definition of a quality measure over the powerset 2^G, we have: $\forall d \in \mathscr{D} \ : \ \varphi(d) = \varphi(G^d)$ (cf. Definition 2.2.6) .

In what follows, we give a quick overview of measures that are used in SD and divide them onto three categories: quality measures that are agnostic of the target label, quality measures that address categorical target label and quality measures that address numerical target label. Note that, in this section, we do not seek to provide an extensive review of interestingness measures in the litterature. For such, we invite the reader to consult, as a starting point Tan, Kumar, and Srivastava, 2004 Survey and Geng and Hamilton, 2006 Survey.

2.2.2.1 Quality measures that are agnostic of a target label:

In subgroup discovery, the support size is the most basic quality measure that, given a description $d \in \mathscr{D}$, counts the number of records supporting d, i.e. the cardinality $|G^d|$. This measure is used as the main interestingness measure for a frequent pattern mining task (Agrawal, Imielinski, and Swami, 1993) and play usually the role of a constraint in SD. We have:

Support size: also called cardinality:

$$\text{supsize} : \mathscr{D} \to \mathbb{R}^+ \ : \ d \mapsto \text{supsize}(d) = |G^d| \tag{2.10}$$

Frequency: it is the relative support size, i.e.

$$\text{freq} : \mathscr{D} \to [0,1] \ : \ d \mapsto \text{freq}(d) = \frac{|G^d|}{|G|} \tag{2.11}$$

■ **Example 2.6** Consider Table 2.1, for the description:

$$d = \langle \texttt{Country} = \texttt{France} \text{ and } \texttt{age} \in [20,39] \rangle$$

We have $\text{supsize}(d) = |\{i_1,i_2\}| = 2$, hence $\text{freq}(d) = \dfrac{1}{4}$.

■

2.2.2.2 Quality measures for categorical target labels:

Given a target class t with $\text{dom}(t) = \{c_1, c_2, ...c_k\}$ with k labels. The aim of such measures is to evaluate the descriminative power of a description d to some target values $c \subseteq \text{dom}(t)$ selected upfront. In this category, usually the domain of possible labels is partitioned to two sets $\text{dom}(t)^+$ and $\text{dom}(t)^-$ shortly and respectively denoted $+$ and $-$. Where $+$ is the property of interest. This consequently partitions the collection G into $G_+ = \{g \in G \,|\, t^g \in \text{dom}(t)^+\}$ and $G_- = \{g \in G \,|\, t^g \in \text{dom}(t)^-\}$. For example, in Table 2.1, if some analyst is interested in explaining **For** votes with the parliamentarians attributes, $+ = \{\text{For}\}$ and $- = \{\text{Against}\}$. Having this in mind, most measures are defined in terms of the frequency counts tabulated in 2×2 contingency table (Tan, Kumar, and Srivastava, 2004) where the descriptions can be seen as if-then rules $d \to +$ with $d \in \mathscr{D}$ and $+ \subseteq \text{dom}(t)$:

	+	-	
d	$\frac{\lvert G_+^d \rvert}{\lvert G_+ \rvert}$	$\frac{\lvert G_-^d \rvert}{\lvert G_- \rvert}$	$\frac{\lvert G^d \rvert}{\lvert G \rvert} = \text{freq}(d)$
$\overline{d}$	$\frac{\lvert \overline{G_+^d} \rvert}{\lvert G_+ \rvert}$	$\frac{\lvert \overline{G_-^d} \rvert}{\lvert G_- \rvert}$	$\frac{\lvert \overline{G^d} \rvert}{\lvert G \rvert} = 1 - \text{freq}(d)$
	$\frac{\lvert G_+ \rvert}{\lvert G \rvert} = \alpha^+$	$\frac{\lvert G_- \rvert}{\lvert G \rvert} = \alpha^-$	1

Table 2.4: A 2×2 contingency table for $d \to +$ with d a description characterizing the subset G^d and $\overline{d}$ is an abuse of notation characterizing the complement set of G^d. Thus, we have $\overline{S} = G \setminus S$ with $S \subseteq G$.

From Table 2.4, we call positive prevalence denoted α^+, the proportion of records in G labeled by the positive target class $+$. Dually, α^- is the negative prevalence and is defined as $1 - \alpha^+$. The two most basic interestingness measures that are present in the contingency table are the true positive rate (**tpr**) and the false positive rate (**fpr**) which are usually used to express several other interestingness measures (Fürnkranz and Flach, 2005).

True Positive Rate: it is the relative support size of a description d in G_+:

$$\text{tpr} : \mathscr{D} \to [0,1] \; : \; d \mapsto \text{tpr}(d) = \frac{|G_+^d|}{|G_+|} \tag{2.12}$$

False Positive Rate: it is the relative support size of a description d in G_-:

$$\text{fpr} : \mathscr{D} \to [0,1] \; ; \; d \mapsto \text{fpr}(d) = \frac{|G_-^d|}{|G_-|} \tag{2.13}$$

One of the most standard measures that is used in discriminative tasks in SD is the weighted relative accuracy (WRAcc)(Lavrac, Flach, and Zupan, 1999) which is closely related to Piatetsky-Shapiro Measure (Piatetsky-Shapiro, 1991) (cf. (Kralj Novak, Lavrač, and Webb, 2009)). The measure aims to discover subgroups which fosters the presence of positive instances while disadvantaging the presence of negative instances:

$$\text{WRAcc} : \mathscr{D} \to [-0.25, 0.25] \; ; \; d \mapsto \text{WRAcc}(d) = \alpha^+ \alpha^- \left(\text{tpr}(d) - \text{fpr}(d)\right) \tag{2.14}$$

■ **Example 2.7** In Table 2.1, if we consider the vote **for** as the positive class, we have $\alpha^+ = 0.5$. For the description:

$$d = \langle \texttt{Country} = \texttt{France} \text{ and } \texttt{age} \in [20,39] \rangle$$

We have $\text{tpr}(d) = 0.5$ and $\text{fpr}(d) = 0$, thus WRAcc $= 0.125$. Note that the best subgroup maximizing the WRAcc is obviously the one covering all positive instances while not containing any negative instance. The best subgroup w.r.t. the for votes as the property of interest and the *WRAcc* measure is:

$$d = \langle \texttt{age} \in [20,39] \rangle, \text{ as } \text{WRAcc}(d) = 0.25.$$

■

In the same spirit, other SD interestingness measures rely on the contingency table 2.4 such as: Accuracy, Precision, Laplace Correction, Linear Correlation coefficient, Cohen's Kappa, FMeasure, Cosine, etc... (Fürnkranz and Flach, 2005; Geng and Hamilton, 2006; Tan, Kumar, and Srivastava, 2004).

2.2.2.3 Quality measures for numerical target labels:

For this category of measures, the underlying dataset has a totally ordered domain of the target class t that is usually a subset $\text{dom}(t) \subseteq \mathbb{R}$ (e.g. Table 2.2). For a comprehensive state of the art of measures dealing with continuous target label, we invite the reader to consult (Lemmerich, Atzmueller, and Puppe, 2016). For a brief overview of such measures, we advise the reader to refer to (Pieters, Knobbe, and Dzeroski, 2010) and (Atzmueller, 2015). We explore some of these measures in what follows:

Mean: the simplest way to determine the interestingness of a subgroup in numerical target dataset is to use the deviation between the mean value observed in the subgroup d and the mean observed in the whole dataset, i.e.

$$\varphi_{\text{mean}} : \mathscr{D} \to \mathbb{R} \, ; \, d \mapsto \varphi_{\text{mean}}(d) = \frac{1}{|G^d|} \sum_{g \in G^d} t^g - \frac{1}{|G|} \sum_{g \in G} t^g \tag{2.15}$$

Mean-Test: this measures was first proposed by Klösgen (Klösgen, 1996) and was used in a dedicated subgroup discovery task by Grosskreutz (Grosskreutz, 2008). Mean test measure is a weighted version of the mean quality. It uses the root of the support size of the subgroup in question to ponderate the mean deviation, i.e.:

$$\varphi_{\text{mean}-\text{test}} : \mathscr{D} \to \mathbb{R} \, ; \, d \mapsto \varphi_{\text{mean}-\text{test}}(d) = \sqrt{|G^d|} \cdot \varphi_{\text{mean}} \tag{2.16}$$

The mean-test can be divided by the standard deviation $\text{std}(G)$ of the target value over the entire dataset to obtain a standardized **Z-Score** (Trajkovski, Lavrač, and Tolar, 2008). i.e. $\varphi_{z-\text{score}} = \frac{1}{\text{std}(G)} \varphi_{\text{mean}-\text{test}}$. Note that this measure is compatible (Fürnkranz and Flach, 2005) with the mean-test as the two measures equivalently order the subgroups w.r.t. their interestingness. Similarly, the **t-statistic** (Klösgen, 2002; Pieters, Knobbe, and Dzeroski, 2010) can be obtained by weighting the mean-test measure by the standard deviation $\text{std}(G^d)$ of the subgroup d instead of the standard deviation of the whole population.

Other measures exist which deal with the specificities of numerical target variables, such as Median χ^2 statistic (Pieters, Knobbe, and Dzeroski, 2010) which uses the median to calculate the difference in distributions. The median-based measures has been extended recently by Boley et al., 2017 to take into account the dispersion of the target values in the subgroup while providing an efficient branch and bound algorithm to handle the complexity of the measure.

■ **Example 2.8** Consider the dataset given in Table 2.2, for the description:

$$d = \langle \texttt{occupation} = \texttt{marketing} \rangle$$

We have $\varphi_{\text{mean}}(d) = 1 - 3.5 = -2.5$ which reads: the users in the group of individuals whose occupation is marketing strongly dislikes Pulp Fiction compared to the whole population. ■

Summary: in this section, we gave a brief overview of interestingness measures that can be used to evaluate the quality of candidate subgroups generated by some enumeration algorithm in order to return the most relevant subgroups to the end-user. Of course, the literature abounds of interestingness measures and the choice depends tightly on the desired objective of an SD task. In this section, we discussed what is dubbed objective interestingness measures. Other measures are addressed in the state-of-the-art and takes into account the prior knowledge of the end-user (formalized as a set of constraints) and are called subjective interestingness measures (Bie, 2011a; Bie, 2011b; Lijffijt et al., 2018), although these measures are not extent-based quality measures. Moreover, some of the work address the statistical significance of the deviation between some quantity observed in the subgroup and the one expected over the whole dataset (Hämäläinen, 2010b; Hämäläinen, 2012; Hämäläinen and Webb, 2019; Webb, 2007).

2.2.3 On Search Space Exploration

Up until now, we have presented two out of three building bricks of a subgroup discovery task, namely: **Language** and **Interestingness**. The third component revolves around **Algorithms** and links between the two aforementioned components to enable solving a subgroup discovery task (e.g. Problem 2.2.1) and return a collection of interesting subgroups. In what follow, we have in mind a pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ and some given interestingness measures φ.

Many SD algorithms exists in the literature covering multiple search space exploration paradigms: **exhaustive search algorithms** (Atzmüller and Puppe, 2006; Grosskreutz and Rüping, 2009; Kavšek and Lavrač, 2006; Klösgen, 1996; Lemmerich, Atzmueller, and Puppe, 2016; Lemmerich, Rohlfs, and Atzmueller, 2010; Spyropoulou, De Bie, and Boley, 2014; Wrobel, 1997; Zimmermann and Raedt, 2009), **heuristic search algorithms** (Boley, Gärtner, and Grosskreutz, 2010; Carmona et al., 2014; Jesús et al., 2007; Klösgen and May, 2002; Lavrac et al., 2004; Leeuwen and Knobbe, 2011; Leeuwen and Knobbe, 2012; Lucas, Vimieiro, and Ludermir, 2018; Luna et al., 2013; Mampaey et al., 2012; Moens and Goethals, 2013), **sampling algorithms** (Al Hasan and Zaki, 2009; Boley, Moens, and Gärtner, 2012; Boley et al., 2011; Diop et al., 2018; Dzyuba, Leeuwen, and De Raedt, 2017; Giacometti and Soulet, 2018; Li and Zaki, 2016) and **anytime algorithms** (Belfodil, Belfodil, and Kaytoue, 2018; Bosc et al., 2018). Furthermore, many off-the-shelf tools and softwares have been proposed for a plug-and-play subgroup discovery: Orange (Demsar et al., 2013), Cortana (Meeng and Knobbe, 2011), Vikamine (Atzmueller and Lemmerich, 2012), PySubgroup (Lemmerich and Becker, 2018).

In the following, we give an overview about the search space exploration paradigms presented above. Particular attention is drawn to exhaustive search algorithms and branch-and-bound like algorithms, since the core algorithms proposed in this thesis mostly follow a branch-and-bound scheme (Land and Doig, 1960; Little et al., 1963; Narendra and Fukunaga, 1977) to solve subgroup discovery like tasks (e.g. top-k interesting subgroups as stated in Problem 2.2.1). In a simple formalization, Algorithms (hereafter denoted Solve) can be formalized under constraint-based pattern mining framework (Boulicaut and Jeudy, 2009; Nijssen and Zimmermann, 2014). Consider an input dataset G and its associated description space $\mathcal{D}$, a quality measure φ and a collection of constraints $\mathcal{C}$ that need to hold for the subgroups resulting in the list L. i.e. $\text{Solve}(G,\mathcal{D},\mathcal{C}) \to L = \{d \in \mathcal{D} | \mathcal{C}(d,G) = \text{True}\}$. $\mathcal{C}$ can be roughly translated to an operator which combines all constraints and transform them to a boolean value. The constraints in $\mathcal{C}$ in most standard SD algorithms consider a threshold on the quality measure σ_φ ($\varphi(d) \geq \sigma_\varphi$), a threshold σ_G on the frequency of the descriptions ($|\text{freq}d| \geq \sigma_G$), a plethora of constraints exists in the literature that can be tailored for different tasks of SD, we refer the interested reader to (Bonchi et al., 2009).

Exhaustive Algorithms: an exhaustive search algorithm $\text{Solve}_{\text{exh}}$ explores the whole search space defined by $\mathcal{D}$ to return the subgroups of interest. Th most straightforward approach is to perform a Brute-Force search algorithm by enumerating every possible descriptions d in the description language $(\mathcal{D},\sqsubseteq)$. This is clearly unfeasible in most settings since the number of possible descriptions is exponential to the number of attributes. e.g. if we have an itemset attribute with 100 items, the number of possible descriptions is equal to $2^{100} \simeq 10^{30}$. To enable an exhaustive search one need to exploit efficient pruning properties and data-structures (e.g. FP-Trees (Han, Pei, and Yin, 2000) or vertical representations (Zaki, 2000)). For instance, given a frequency threshold σ_G, one can exploit the monotonicity of the constraint $|G^d| \geq \sigma_G$ (e.g. Apriori (Agrawal and Srikant, 1994) and Apriori-SD (Kavšek and Lavrač, 2006)) to prune the sub-search space of a description d whenever its frequency is below some given threshold. In general, one can push any monotonous constraints (Bonchi et al., 2003; Jeudy and Boulicaut, 2002) to prune the sub-search space related to some given descriptions when traversing the search space in a top-down (bottom-up) fashion. More sophisticated constraints can also be exploited to reduce substantially the size of the search space while guaranteeing the completeness of the algorithm (i.e. $\forall d \in \mathcal{D}\, \mathcal{C}(d,G) = \text{True} \Rightarrow d \in L$ with L the returned subgroups) (Bonchi et al., 2009). Moreover, interestingness measures φ properties can be leveraged to avoid generating subgroups in unpromising areas of the search space. For instance, by defining proper optimistic estimates (bounds) (Grosskreutz, 2008; Grosskreutz, Rüping, and Wrobel, 2008). In practice, two standard algorithms are provided in the state-of-the-art to mine for interesting subgroups while ensuring completeness: (i) SD-Map (Atzmüller and Lemmerich, 2009; Atzmüller and Puppe, 2006) with its extension for numerical target concepts NumBSD (Lemmerich, Atzmueller, and Puppe, 2016) and RMiner (Spyropoulou, De Bie, and Boley, 2014). Both exploit efficient data structures and pruning properties over constraints while leveraging optimistic estimates for several interestingness measures. The main feature of R-Miner is the fact that it exploits closure operators to reduce the number of generated candidates when the quality-measure is extent-based. It relies on Divide-and-Conquer algorithm (Boley et al., 2010). In this thesis and since we focus particularly on extent-based interestingness measures, the core algorithms resemble - in terms of the search space explored not the functioning - to RMiner (Spyropoulou, De Bie, and Boley, 2014).

Heuristic Algorithms: most typical SD algorithms are heuristic algorithms. A heuristic algorithm $\text{Solve}_{\text{heur}}$ abandons the completeness property of exhaustive search algorithms in favor of runtime, hence tractable. Standard heuristic algorithms in SD rely on a beam-search strategy (Lowerre, 1976). In a nutshell, a simple beam search SD algorithm perform a level-wise search (similarly to a breadth-first-search (BFS)). At each level, a **breadth-width** number of *valid* subgroups are chosen with regard to the constraints $\mathscr{C}$. The choice is usually made by considering the top subgroups with regard the interestingness measure φ. Other techniques, CN2-SD (Lavrac et al., 2004) and DSSD (Leeuwen and Knobbe, 2011; Leeuwen and Knobbe, 2012) among others, diversify the beam so as to have a better trade-off between exploration and exploitation. Once the beam is selected, only its description are used to generate the next level. The stop-condition is commonly fixed by a **depth-level** which specifies how far in-depth the algorithms goes in the search space. The depth-level corresponds usually to the descriptions length, i.e. the maximum number of conditions allowed in a description $d \in \mathscr{D}$. Since beam-search grounded algorithms are enumeration algorithms, they can take into consideration most of the properties of constraints and optimistic estimates to avoid having in some current beam uninteresting subgroups (e.g. (Mampaey et al., 2012)). Other techniques follow an evolutionary scheme (genetic algorithms (Whitley, 1994)) (Carmona et al., 2014). Having the fitness operator which corresponds to the interestingness measures φ. Genetic algorithm additionally requires the definition of proper generation selection, mutation and crossover operators. For instance, SSDP+ (Lucas, Vimieiro, and Ludermir, 2018) select a diversified beam on each generation by considering only the non-dominated subgroups (Relevance theory (Garriga, Kralj, and Lavrač, 2008)) and a diversification criterion (e.g. Jaccard index). The mutation consists of an itemset language to remove, update or insert a random item. The crossover consists of a uniform crossover between two descriptions where the output have the same number of items by randomly taking items from both input descriptions.

Sampling Algorithms: Similarly to heuristic algorithm, a sampling algorithm $\text{Solve}_{\text{samp}}$ abandon the completeness condition. Several techniques had been proposed in the literature to provide guarantees while relying on a small sample of drawn descriptions of the whole description space $\mathscr{D}$. For instance, algorithms proposed in (Al Hasan and Zaki, 2009; Boley, Gärtner, and Grosskreutz, 2010) rely on Markov Chain Monte Carlo (MCMC) to generate descriptions according to some desired probability distribution (e.g. a subgroup $d \in \mathscr{D}$ chance to be returned in the resulting set proportional to its quality $\varphi(d)$). Interestingly (Boley, Gärtner, and Grosskreutz, 2010) implements a Metropolis–Hastings algorithm to generate only closed descriptions (formal concepts -

see Section 2.2.1) while guaranteeing the former property. Despite the generic nature and the interesting guarantees that MCMC algorithms provide, it requires a number of steps that grows exponentially in the input size to generate a single pattern which usually hinders their usage. To overcome this issue, some techniques abandon the genericity of MCMC techniques while maintaining hard theoretical guarantees. This is done by devising direct-output sampling algorithms (Boley, Moens, and Gärtner, 2012; Boley et al., 2011; Diop et al., 2018; Giacometti and Soulet, 2018) tailored for specific quality measures. Direct-output sampling technique (Boley et al., 2011) are non-enumerative methods which sample subgroups directly from the full search space. They enable to produce a collection of subgroups, each of which is generated following exactly some distribution (e.g. frequency, discriminativity, etc.). Other algorithms tackles the sampling by combining the advantages of exhaustive search algorithms and sampling (Dzyuba, Leeuwen, and De Raedt, 2017; Riondato and Vandin, 2018) while ensuring guarantees on the quality of the returned subgroups. For instance, MiSoSouP (Riondato and Vandin, 2018) sample the input dataset G and perform an exhaustive search afterwards. It derive bounds to the sample size sufficient to ensure that an ε-approximation of the top-k subgroups hold with a sufficiently high probability. Conversely, Flexics (Dzyuba, Leeuwen, and De Raedt, 2017) maintains the full collection of records G and proposes to sample the description space $\mathcal{D}$ beforehand. This, followed by an exhaustive search on the sampled description space.

Anytime Algorithms: this category of algorithms combines the properties of the three first categories in the aim of providing tractable algorithms ensuring a completeness guarantee. Anytime pattern mining algorithms Solve$_{\text{any}}$ (Belfodil, Belfodil, and Kaytoue, 2018; Bosc et al., 2018; Hu and Imielinski, 2017) are enumerative methods which exhibit the anytime feature (Zilberstein, 1996), a solution is always available whose quality improves gradually over time and which converges to an exhaustive search if given enough time, hence ensuring completeness. While MCTS4DM (Bosc et al., 2018) ensures interruptibility (i.e. the execution can be interrupted anytime) and an exhaustive exploration if given enough time and memory budget, it does not ensures any theoretical guarantees on the distance from optimality and on the diversity. In contrast, RefineAndMine (Belfodil, Belfodil, and Kaytoue, 2018) is an anytime algorithm tailored specifically for subgroup discovery in numerical attributed dataset which ensures hard guarantees on the quality of the found solutions upon interruption.

Summary: in this section, we have discussed the component "**Algorithms**" of a subgroup discovery task which comes in between the "**Description Language**" and "**Interestingness**" components. In short, an algorithm enumerates candidate subgroups from the search space defined upon the description space, measures their interestingness and returns the most important ones. Although, as discussed in this section, multiple paradigms exists to handle this task, we will focus on the exhaustive search paradigm where the aim is to **guarantee** that the best patterns are found and returned to the end-user given a SD problem (e.g. Problem 2.2.1). Since the enumeration algorithms for SD are roughly equivalents to the ones used for Exceptional Model Mining (EMM), we will first introduce and discuss EMM in the next section.

2.3 Exceptional Model Mining

Exceptional model mining (EMM) is a framework (Duivesteijn, Feelders, and Knobbe, 2016; Leman, Feelders, and Knobbe, 2008) can be seen as a multi-target generalization of Subgroup discovery (SD) framework (standard SD seen as supervised descriptive rule discovery (Kralj-Novak, Lavrac, and Webb, 2009)). In this perspective, EMM deals with similarly structured dataset as SD where the schema of attributes is partitioned to two parts: descriptive attributes and target attributes (rather than a single target attribute). Hence, the underlying dataset is a collection G of records g with its schema $\{a_1,a_2,...,a_m,t_1,...,t_l\}$. Recall that each attribute has a domain of interpretation which corresponds to all its possible values. Attributes $\mathscr{A} = \{a_1,a_2,...,a_m\}$ are called **descriptive attributes** which are used to characterize subsets (subgroups) of data in the same way as in SD. t_1, t_2,..., t_l are called **target attributes** and are used to build **models** and evaluate **interestingness** of subgroups. First of all, we give an example of a standard EMM dataset that can be seen as an excerpt of the behavioral dataset given in Table 1.2.

ide	genres	releaseDate	**RatingDate**	**RatingAvg**
e_1[3]	Comedy	1974	1998	2.5
e_1	Comedy	1974	2001	3.5
e_1	Comedy	1974	2007	4
e_2	Crime; Drama; SciFi	1992	1999	4
e_2	Crime; Drama; SciFi	1992	2002	4.5
e_3	Action; Adventure; Crime	1996	2002	3
e_4	Animation; Comedy	1996	2003	4
e_5	Action; Romance; War	1992	1999	2
e_6	Comedy	1997	2005	1.5

Table 2.5: Example of behavioral dataset - Movielens dataset depicting the average ratings per year given by users on movies. The **descriptive attributes** characterizing movies are: **genres** and **releaseDate**. The **target attributes** are **RatingDate** and **RatingAvg**. **RatingAvg** represents the average rating of users given in a **RatingDate**.

Considering this dataset given in Table 2.5, one can ask the following question: "Do movies get better or worse ratings over time?"[4]. In order to give elements of answer to this question, one can use linear regression to explain **RatingAvg** with **RatingDate** to provide an initial answer. Furthermore, this analysis can be refined to produce additional details on subgroups of movies, particularly those who do not follow the norm. This is the main objective of Exceptional Model Mining:

■ ***finding subgroups** where an **unusual interaction** between the targets is observed* ■

[3] Moviename = The Return of The Pink Panther

[4] an example of such a question can be found in https://www.quora.com/Does-the-IMDb-rating-for-a-movie-change-over-time

In EMM, and having in mind the pattern structure $(G,(\mathscr{D},\sqsubseteq),\delta)$[5], the **interaction** between targets is captured by what is called a **model**. The **unusualness** is captured by an interestingness (quality) measure φ. This interestingness measure relies on an objective function which compare the **model fitted on the targets in the subgroup** G^d $(d \in \mathscr{D})$ with the **model induced over the whole dataset** G. **Subgroups** are characterized as in SD by descriptive attributes where the syntax is defined by the description language $(\mathscr{D},\sqsubseteq)$ and **finding** interesting subgroups consists in enumerating candidate subgroups in $(G,(\mathscr{D},\sqsubseteq),\delta)$ by some exploration algorithm. We will summarize this concepts in Figure 2.4 after briefly introducing a standard EMM task.

A typical task of EMM is to find the top-k exceptional subgroups as formalized in (Duivesteijn, Feelders, and Knobbe, 2016). It is similar to the task of finding top-k interesting subgroups formerly introduced in Problem 2.2.1. Clearly, the main difference resides in how interestingness measure is evaluated for some given candidate subgroup. In SD[6], the interestingness of a subgroup is evaluated using directly the target values of records as detailed in Section 2.2.2. Conversely, in EMM, the interestingness evaluation of a subgroup in EMM combines:

1. The computation of a model class $M : \mathscr{D} \to \Omega$ over the targets of some given subgroup.
2. The evaluation of a designed distance measure $\Delta : \Omega \times \Omega \to \mathbb{R}$ comparing between the model induced on a subgroup characterized by a description d and the model induced on the description $*$ corresponding to the whole dataset (sometimes the complement of the subgroup $G \setminus G^d$ is used instead). Intuitively, this distance captures how significant the model fitted on the subgroup deviates from the norm, i.e. the model fitted on the whole dataset.

This roughly translates to:

$$\varphi(d) = \Delta(M(d), M(*)) \text{ with } d \text{ an arbitrary description in } \mathscr{D}$$

Similarly as in subgroup discovery, we confine ourselves to extent-based interestingness measures. That is, the model is computed by relying solely on the extent of a description d.

Having this concepts in mind, we give in the following the standard top-k exceptional model mining problem.

Problem 2.3.1 (*Top-k* Exceptional Model Mining Problem).
Given a collection G, a description language $\mathscr{D}$, a model class $M : \mathscr{D} \to \Omega$, a quality measure $\varphi : \mathscr{D} \to \mathbb{R}$, a quality threshold σ_φ and a minimum support threshold σ_G, the problem is to find the list $L = \{d_1,\ldots,d_k\} \subseteq \mathscr{D}$ such that:

[*Validity*] $\forall d \in L \ :\ d$ *valid*, that is $|G^d| \geq \sigma_G$ and $\varphi(d) \geq \sigma_\varphi$.

[*Top-k*] $(\forall d' \in (\mathscr{D} \setminus L))(\forall d \in L) \ :\ \varphi(d) \geq \varphi(d')$.

[5] Of course, any collection of records G with a schema $\mathscr{A}$ inducing a description space $\mathscr{D}$ can be considered as an input to an Exceptional Model Mining task. As discussed in the former section 2.2, we confine ourselves to descriptions languages that induces a lattice structure. It follows that, the input can be formalized as a pattern structure $(G,(\mathscr{D},\sqsubseteq),\delta)$.

[6] from the perspective of Supervised Descriptive Rule Discovery (Kralj-Novak, Lavrac, and Webb, 2009)

In short, when it comes to define an exceptional model mining task and solve it, one need to answer to the four following questions which are summarized in Figure 2.4:

($\mathbf{Q}_1$) **Language**: what is the description space $\mathscr{D}$ of candidate subgroups?

($\mathbf{Q}_2$) **Model**: what is the model class used to capture interaction between target attributes?

($\mathbf{Q}_3$) **Interestingness**: how to assess the interestingness of a subgroups (quality measure) - how to compare between the model fitted on the targets in the subgroup and the model fitted on the targets in the whole dataset?

($\mathbf{Q}_4$) **Algorithm**: How to explore the search space of candidate subgroups?

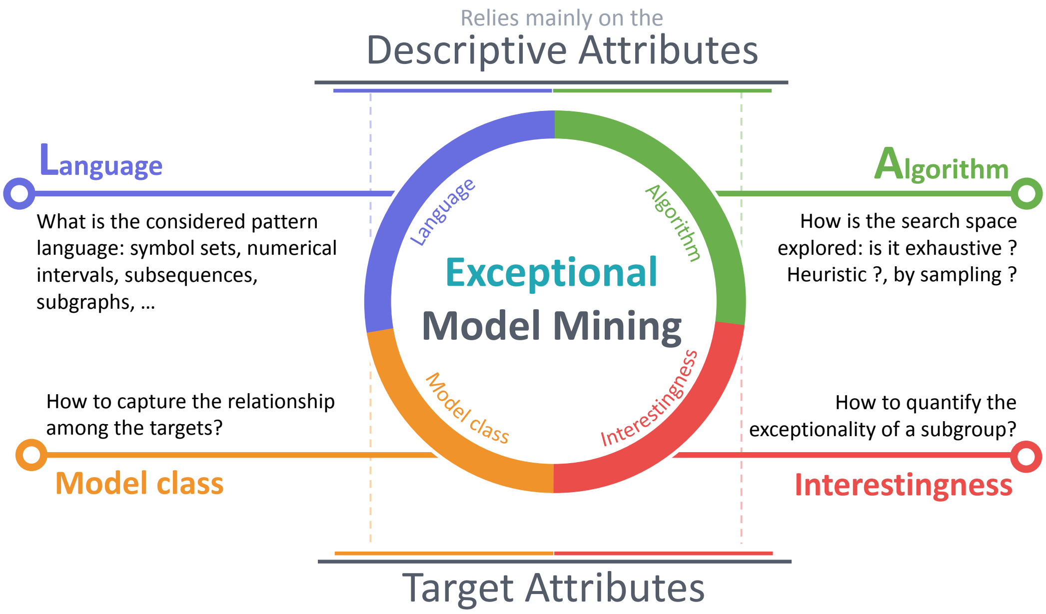


Figure 2.4: Building blocks of an exceptional model mining task (Summary)

2.3.1 On Description Languages and On Search Space Exploration

The top aspects in Figure 2.4, namely, the description language and the algorithms have been already discussed in the former section 2.2 for Subgroup Discovery. Almost nothing changes for these two building blocks, subgroups are characterized in most configurations by conjunction of conditions on the attributes values as discussed in Section 2.2.1. Since the considered interestingness measures are extent-based, one can use any algorithm that exhaustively generate all candidate subgroups by traversing the concept lattice induced from the underlying pattern structure $(G,(\mathscr{D},\sqsubseteq),\delta)$. For this task, algorithms like Close-By-One (Kuznetsov and Obiedkov, 2002) or Divide-and-Conquer (Boley et al., 2010) can be used. We will elaborate a standard enumeration algorithm later in Section 2.4). In the literature, some algorithms have been specifically designed for generic EMM tasks (Krak and Feelders, 2015; Lemmerich, Becker, and Atzmueller, 2012; Moens and Boley, 2014). As an example, GP-growth (Lemmerich, Becker, and Atzmueller, 2012) is an exhaustive search algorithm for EMM. It extends the well-known FP-tree (Han, Pei, and Yin, 2000) data structures to GP-tree data structures to efficiently handle the computation of models, which is usually the most computationally extensive part in an EMM algorithm. Hence, ($\mathbf{Q}_1$) and ($\mathbf{Q}_4$) are already covered in Section 2.2.1 and Section 2.2.3. In the following, we emphasize on the evaluation of interestingness of candidate subgroups, that is: the models that had been proposed in the literature ($\mathbf{Q}_2$) and the associated interstingness measures ($\mathbf{Q}_4$) as they are tightly linked.

2.3.2 On Model Classes and Interestingness Measures

Several model classes and interestingness measures have been proposed in the state-of-the-art since the seminal paper (Leman, Feelders, and Knobbe, 2008). We briefly review in the following some of these models. Recall that, we have in mind a pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ as an input, a subgroup is denoted $d \in \mathscr{D}$ and its extent is denoted G^d. For a comprehensive survey, we draw the reader's attention to Duivesteijn, Feelders, and Knobbe, 2016 work and Ventura et Al. book (Ventura and Luna, 2018, Chapter 6).

Correlation and Rank Correlations Models: given two target attributes t_1 and t_2, the simplest correlation model was defined in the original EMM model classes (Leman, Feelders, and Knobbe, 2008) by measuring the linear association between the two pearson's correlation coefficient in the subgroup G^d and its complement in the whole dataset $G \setminus G^d$. Several interestingness measures had been proposed to capture how significant is the differences between the two correlation models. For instance, the simplest measure that had been proposed is the absolute difference between the two correlation coefficients (Duivesteijn, Feelders, and Knobbe, 2016). This models suffered from several pitfalls, mainly, the high sensitivity to outliers and the (hard) assumption targets normality. To mitigate these problems, Downar and Duivesteijn (Downar and Duivesteijn, 2015; Downar and Duivesteijn, 2017) proposed to use rank correlation models. In short, rather than measuring the interaction between the two targets t_1 and t_2 with pearson's correlation coefficients, it uses rank correlation coefficients like the well-known Spearman's rank correlation coefficient and Kendall's rank correlation coefficient (Kendall, 1948). The authors propose several interestingness measures to capture the difference between the rank correlation model computed over the subgroup and its complement. This work have been extended to evaluate rank correlation between more than two target attributes in (Hammal et al., 2019).

Classification Models: in this category, methods (Duivesteijn, Feelders, and Knobbe, 2016; Leman, Feelders, and Knobbe, 2008) are given a set of target attributes t_1,...,t_{l-1}, t_l along with the descriptive attributes $a_1, \ldots, a_m$. The model used to capture interaction between the target attributes is a classifier on the discrete target value t_l (boolean or categorical) using $t_1, \ldots, t_{l-1}$. To judge whether the effect of the these descriptive target attributes is substantially different in a subgroup d, the methods builds a classifier over both the collection of records G^d and its complement $G \setminus G^d$ and measure the difference between the two classifier with an adapted interestingness measure. In this spirit two classification models had been proposed in the original EMM framework (Leman, Feelders, and Knobbe, 2008): logistic regression (Neter et al., 1996) and simple decision tables (Decision Table Majority - DTM) (Kohavi, 1995) with adapted interestingness measures. For instance, in DTM model classes, Leman, Feelders, and Knobbe, 2008 propose to use Hellinger Distance (Le Cam and Yang, 2012) to evaluate how exceptional the subgroup d is. It measures the conditional distribution of t_l in the subgroup and its complement for each possible combination of $t_1, \dots, t_{l-1}$ which are summed to obtain an overall distance (Duivesteijn, Feelders, and Knobbe, 2016).

Regression Models: Conversely to the precedent category, regression models, proposed in (Duivesteijn, Feelders, and Knobbe, 2016; Leman, Feelders, and Knobbe, 2008), are used to characterize interaction between target attributes $t_1, \dots, t_{l-1}, t_l$ when the target of interest t_l is numerical. Simply put, the regression model class in the case of two target attributes t_1, t_2 in EMM is used as follows: a linear regression is fitted on the subgroup d to explain t_2 with t_1 and is compared to the linear regression fitted on the whole dataset G or the complement of the subgroup. To measure how significant the difference is between the two regressions model in hand, several interestingness measures φ had been proposed (Duivesteijn, Feelders, and Knobbe, 2012; Duivesteijn, Feelders, and Knobbe, 2016; Leman, Feelders, and Knobbe, 2008). The simplest one consists in comparing the two slopes β_d (subgroup) and $\overline{\beta_d}$ (complement) and measure the p-value resulting from the following hypothesis testing: $H_0 : \beta_d = \overline{\beta_d}$ against $H_1 : \beta_d \neq \overline{\beta_d}$. An illustration of such a model class is given in Figure 2.5 where the input dataset is the one given in Table 2.5. In order to compare between multiple linear regression models where the target of interest t_l is explained with multiple descriptive target attributes, i.e. $t_1, \dots, t_{l-1}$, Duivesteijn, Feelders, and Knobbe, 2012 propose to use Cook's distance.

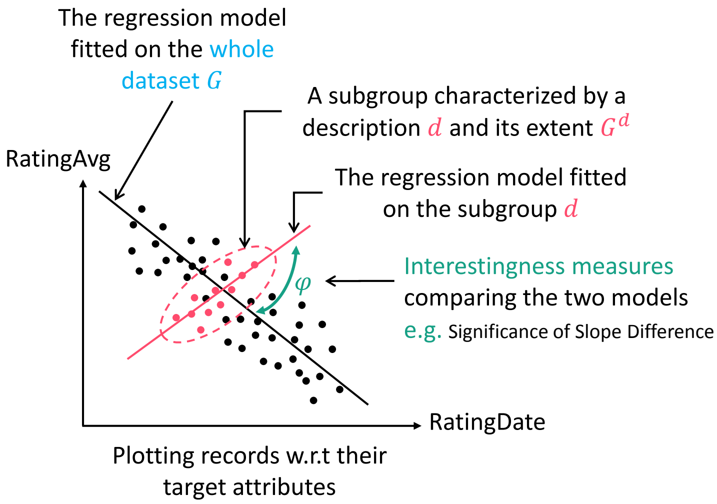


Figure 2.5: Exceptional model mining - Regression model as a model class to capture relationship between two numerical targets values. This illustration considers the example given in Table 2.5 where the aim is to analyze how movies ratings evolve with time .
Note: this figure was largely inspired by a figure in Bendimerad, 2019 thesis manuscript.

Bayesian Networks Models: Duivesteijn et al., 2010 propose to analyze exceptionality of inter-dependencies between discrete target variables $t_1, \dots t_l$ in subgroups by using Bayesian Networks. In short, the task of analyzing exceptional subgroups based on

Bayesian networks consists in finding subgroups whose fitted Bayesian network significantly differ from the one computed on the whole dataset. The adapted interestingness measure is grounded on edit distance (Shapiro and Haralick, 1985). Such a model can be useful for several tasks where an exceptional inter-dependencies between variables can highlight spurious correlations. For instance, one of the interesting findings in this work (Duivesteijn et al., 2010) resuts from analyzing the emotions dataset (Trohidis et al., 2008) (songs associated with rhythmic and timbre descriptive attributes and emotions (e.g. *sad-lonely*) as targets attributes. In this dataset, the emotion *sad-lonely* is correlated with all the other emotions (e.g. happy pleased) in overall terms. When the dataset is restricted on songs having bright sounds (the subgroup), the *sad-lonely* emotion becomes not correlated with none of the other emotions.

Markov Chains Models: Lemmerich et al., 2016 proposed to mine for exceptional transition behaviors by utilizing first-order Markov Chains as the model class (Norris, 1998). The dataset given as input represents sequential data where the records are transitions characterized by multiple descriptive attribute $a_1,...,a_m$ (e.g. weekday) and three target attributes t_1,t_2,t_3, where t_1,t_2 represent respectively the source state and the target state in the transition, while t_3 represent the number of visits. A records can be roughly translated to: $\#t_3$ individuals went from t_1 to t_2 on Wednesday (a_1) morning (a_2). Considering this target attributes and some given subgroup d, the transition matrix is built and the first-order markov chain is subsequently computed. A subgroup is considered exceptional if its associated markov chain deviates significantly from the markov chain fitted on the whole dataset G. The deviation is captured by an adapted Manhattan distance.

Graph Models: this category pertains to those techniques that model the input dataset as attributed graphs and mine for exceptional sub-graphs with regard to some interestingness measure. For instance, similarly as the work of Lemmerich et al., 2016, Kaytoue et al., 2017 mine for exceptional transition behavior of groups. To this aim, the proposed model consists of contextual sub-graphs which capture the transitions between nodes (e.g. areas of a city). The contextual sub-graph is computed for each subgroup and the number of transition in the subgroup are compared to the overall context in the same sub-graph via a WRAcc-like measure (Lavrac, Flach, and Zupan, 1999). Other attributed graph models have been proposed in the literature (Bendimerad, Plantevit, and Robardet, 2016; Bendimerad, Plantevit, and Robardet, 2018) to capture exceptional characteristics in sub-graphs by looking for significant increase or decrease in some numerical target attributes of interest. This can be used to mine for predominant activities in cities neighborhoods (Bendimerad, Plantevit, and Robardet, 2016) (e.g. there is substantially more bars and restaurants in the subgroup (subgraph) neighborhood compared to the rest of the city) or to extract exceptional activated area in brain (Moranges et al., 2018).

Preferences Models: Sá et Al. (Sá et al., 2016; Sá et al., 2018) proposed exceptional preference mining (EPM) to look for sub-population (subgroup) having exceptional preferences compared to the whole population. To this aim, the input of an EPM approach is a dataset which describe individual and his preferences (partial order) with regard to a collection of discrete targets $(t_1, t_2, ..., t_l)$. To mine for exceptional preference behavior, the model chosen is a preference matrix which aggregate the preferences of the subgroup. Several interestingness measures had been proposed to capture how significant the deviation of preferences is, compared the preferences of the overall population. For instance, the author propose to calculate the Frobenius norm of the distance matrix to measure how unusual the average ranking is for the subgroup. The distance matrix used corresponds to the difference between the preference matrix fitted on the whole population and the one fitted on the subgroup.

Compression Models: Leeuwen and Knobbe, 2012 propose *Krimp* code tables (Vreeken, Leeuwen, and Siebes, 2011) as a new model class. They utilize WKG (Weighted Krimp Gain) to evaluate the interestingness of a subgroup. In short, a subgroup is considered interesting if it can be compressed much better by its own compressor, than by the compressor induced on the overall dataset.

Summary: this section was devoted to exceptional model mining framework and how it generalizes Subgroup discovery to analyze multiple target attributed dataset. In short, a task grounded in exceptional model mining goes in the same line as SD where the aim is to discover exceptional subgroups. Exceptionality is captured by: (1) defining a model characterizing the interaction between the target attributes and (2) comparing the model fitted on the subgroup with the one fitted on the overall population. Once the model and the interestingness measure are properly defined, several algorithms can be used to approach the solution of an EMM task (e.g. Problem 2.3.1). The focus of the next section is to provide a standard Branch and Bound algorithm that can be used for such a task.

2.4 Standard Exploration Algorithms

Section 2.2 and Section 2.3 gave an overview of the theoretical background of Subgroup Discovery (SD) and Exceptional Model Mining (EMM). We have discussed the main building blocks in SD and EMM frameworks required to define and solve a mining task. We briefly recall below these building blocks while bringing to the fore the main concepts that we are going to use to formulate a standard and guideline algorithm for SD/EMM.

Description Language: as discussed in Section 2.2.1, one need to define the syntax used to characterize subgroups. In this thesis, in the same spirit of most past works in SD/EMM (Kloesgen, 2000; Klösgen, 1996; Leman, Feelders, and Knobbe, 2008; Wrobel, 1997), we choose to characterize subgroups by conjunctions of conditions (cf. Definition 2.2.2 and Definition 2.2.12). Although, many formalisms exist in the literature to build the search space induced by such a description language, we choose Pattern structures $(G, (\mathscr{D}, \sqsubseteq), \delta)$ (Ganter and Kuznetsov, 2001) (cf. Definition 2.2.7). $(\mathscr{D}, \sqsubseteq)$ and $(2^G, \subseteq)$ are both lattices (cf. definition 2.2.10). Recall that, in pattern structures, two operators are important and allow to go back and forth between the two lattices: $\delta : 2^G \to \mathscr{D}$ and ext$\mathscr{D} \to 2^G$. δ computes the maximum common description between records belonging to a subset of G and ext computes the extent (support) of a description in G. These two operations form a Galois connection between the power set $(2^G, \subseteq)$ and $(\mathscr{D}, \sqsubseteq)$. Hence, the composite operator clo $= \delta \circ \text{ext} : \mathscr{D} \to \mathscr{D}$ is a closure operator (Ganter and Kuznetsov, 2001; Ganter and Wille, 1999). In this thesis, we are interested in generating candidate subgroups from the collection of closed descriptions $\mathrm{clo}[D] = \{d \in \mathscr{D} \,|\, d = \mathrm{clo}(d)\}$. The latter contain a unique representative description per equivalence class of $\mathscr{D}$ (cf. Definition 2.2.11), each corresponding to a characterizable subset in $\mathrm{ext}[\mathscr{D}]$ (cf. Figure 2.3).

Interestingness Measures (and Model classes): in SD/EMM, the objective is to find "**interesting**" subgroups with regard a property of interest. The latter is usually implemented via a quality measure $\varphi : \mathscr{D} \to \mathbb{R}$ (cf. Definition 2.2.6. In this thesis, we are solely interested by extent-based quality measures ($\forall d \in \mathscr{D} : \varphi(d) = \varphi(G^d) = \varphi(\mathrm{ext}(d))$). As discussed in Section 2.2.2 and Section 2.3.2, several interestingness measures have been proposed in the literature (Duivesteijn, Feelders, and Knobbe, 2016; Fürnkranz and Flach, 2005; Geng and Hamilton, 2006; Kralj-Novak, Lavrac, and Webb, 2009; Lavrac, Flach, and Zupan, 1999; Tan, Kumar, and Srivastava, 2004) depending on the target attributes types (numerical, categorical) and the study objective. In the scope of this thesis, no interesting measure in the literature makes it possible to convey the semantic of the desired patterns (i.e. exceptional (dis)agreement). Hence, one of the main contributions of this thesis is to define proper and interpretable model classes and interestingness measures to capture (dis)agreement between and within groups in behavioral data.

Algorithms: Section 2.2.3 discussed the multitude of possible paradigms that one can follow to explore the search space related to an pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$. We briefly recall these paradigms in here: **Exhaustive** search algorithms (e.g. SD-Map (Atzmüller and Lemmerich, 2009; Atzmüller and Puppe, 2006), NumBSD (Lemmerich, Atzmueller, and Puppe, 2016) and RMiner (Spyropoulou, De Bie, and Boley, 2014)); **Heuristic** search algorithms (e.g. beam-search algorithms, CN2-SD (Lavrac et al., 2004), DSSD (Leeuwen and Knobbe, 2011; Leeuwen and Knobbe, 2012) and FSSD (Belfodil et al., 2019b)); **Sampling** Algorithms (e.g. Direct-output sampling techniques (Boley et al., 2011) and MiSoSouP (Riondato and Vandin, 2018)); and **Anytime** Algorithms (e.g. MCTS4DM Bosc et al., 2018 and Refine&Mine (Belfodil, Belfodil, and Kaytoue, 2018)). In this thesis, we are mainly interested by providing complete solutions for the problem of discovering exceptional behavior in behavioral data. Hence, we emphasize on designing efficient exhaustive search algorithms.

In what follows, and given an input pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$, we will first design in Section 2.4.1 a standard enumeration algorithm, dubbed EnumCC, which generates all candidate subgroups corresponding to closed descriptions $\mathrm{clo}[D] = \{d \in \mathscr{D} \,|\, d = \mathrm{clo}(d)\}$. This choice is motivated by (i) the fact that enumerating only closed descriptions substantially reduces the number of generated candidates and also (ii) the fact that we consider only extent-based quality measures φ. In algorithm EnumCC, the interestingness measure φ is not taken into account. Hence, subgroups quality is not evaluated. For this aim, we devise in Section 2.4.2 a standard branch-and-bound algorithm called B&B4SDEMM. The algorithm perform an exhaustive search to find all interesting subgroups w.r.t. φ in order to solve Top-k SD/EMM problems (see Problem 2.2.1 and Problem 2.3.1). B&B4SDEMM leverages EnumCC and optimistic estimates for an efficient exhaustive traversal of the search space.

2.4.1 A Standard Enumeration Algorithm For SD/EMM

Considering instances of condition spaces in Defintion 2.2.12 along with the equations (2.3 — 2.8), we can use algorithms that enumerates efficiently[7] subgroups corresponding to formal concepts by traversing the concept lattice induced by the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ (Boley et al., 2010; Ganter et al., 2016; Kuznetsov and Obiedkov, 2002). We give in this section an exhaustive algorithm which enumerate all candidate subgroups (closed descriptions) corresponding to $\text{clo}[D] = \{d \in \mathscr{D} \,|\, d = \text{clo}(d)\}$.

A simple yet efficient[7] algorithm to enumerate all formal concepts is Close-By-One (CbO for short) (Kuznetsov, 1993; Kuznetsov, 1999; Kuznetsov and Obiedkov, 2002). The algorithm functioning is similar to Divide-and-Conquer (Boley et al., 2010) which enumerates all closed elements in a closure system given a closure operator clo (e.g. $\text{clo} = \delta \circ \text{ext}$). CbO was defined particularly to handle itemsets, even though the functioning is closely similar, the algorithm that enumerate closed descriptions in the complex search space containing heterogeneous attributes will be dubbed here EnumCC (introduced and formalized first in (Belfodil et al., 2017a)).

Given G a collection of records and its schema $\mathscr{A} = \{a_1, a_2, ..., a_m\}$ (given in an arbitrary order fixed upfront) inducing the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$, Algorithm 1 called EnumCC (**Enume**rate Closed **C**andidate) enumerates once and only once all closed descriptions in $(\mathscr{D}, \sqsubseteq)$ whose associated support fulfill the minimum support constraint $\sigma_G \in \mathbb{N}$. It traverses the search lattice $(\mathscr{D}, \sqsubseteq)$ in a top-down, DFS fashion starting from the most general description $*$ whose extent is the entire collection G. It proceeds by atomic refinements to progress, step by step, toward more specific descriptions. This is enabled by the refinement operator η (cf. definition 2.2.5 and equation 2.8). We override its previous definition (given in equation 2.8) below to specify that only the condition corresponding to the attribute whose index is equal to some given index $k \in [1, m]$ should be refined. For any description $d \in \mathscr{D}$, we have:

$$\eta(d,k) = \{\langle r'_1, ..., r'_m \rangle \in \mathscr{D} : r'_k = \eta_k(r_k) \text{ and } (\forall j \in 1..m)\, j \neq k \Rightarrow r'_j = r_j\} \quad (2.17)$$

$$\eta(d) = \bigcup_{j \in [1,m]} \eta(d,k) \quad (2.18)$$

Starting from a description d, EnumCC first computes its corresponding support G^d. If the size exceeds the threshold (line 1), the closure of d is computed (line 2). Subsequently, a *canonicity test* between *closure_d* and d is assessed (line 3). It allows to determine if a description after closure was already generated and to discard it, if appropriate, without addressing the list of already generated closed descriptions requiring hence no additional storage. The canonicity test relies on an arbitrary order between attributes in $\mathscr{A} = \{a_1, a_2, ..., a_m\}$ indicating that, in the enumeration process, attribute conditions are refined following this

[7] Efficiency here corresponds to the fact that the algorithm in question enumerates all closed descriptions in the concept lattice and which is **polynomial delay**(Johnson, Papadimitriou, and Yannakakis, 1988) and **PSPACE** (Arora and Barak, 2009). Polynomial delay algorithms are algorithms where the delay between the beginning and the first output, two outputs and the final output and the end is polynomial to the input size. PSPACE algorithms are algorithms using a polynomial amount of space w.r.t. the input size. This is valid as long as the computation of closure (i.e. $clo(d) = \delta(\text{ext}(d))$) is polynomial time which is the case in our setting (itemsets, numerical and categorical attributes and also heterogeneous attributes with a mixed schema).

arbitrary order. Let $d = \langle r_1, ..., r_f, ..., r_m \rangle$ a description resulting from the refinement of the f^{th} condition of some preceding description, and $d' = \langle r'_1, ..., r'_f, ..., r'_m \rangle = clo(d)$ the closure of d. Following the arbitrary order between attributes, we expect for d', if it is the first time that it is encountered, that no condition before r'_f (i.e. $r'_1, ..., r'_{f-1}$) is refined; otherwise, $clo(d)$ was already generated after a refinement of preceding conditions and need thus to be discarded. The intuition behind the canonicity test being explained, a canonicity test rests essentially on a lectic order (cf. (Ganter and Wille, 1999, p.66-68)) between d and its closure d' denoted $d \lessdot_f d'$ which is defined as follows: $d \lessdot_f d' \iff \forall i \in [1..f-1] \mid r_i = r'_i \wedge r_f \lessdot r'_f$. The latter condition, $r_f \lessdot r'_f$, corresponds to an analogous canonicity test between conditions and makes sense for multi-valued attributes types only (e.g. itemsets[8] (Ganter and Wille, 1999, p.66-68)). It does not need to be calculated for simple attributes (numerical, categorical). If the canonicity test is successful (line 3), *closure_d* is returned as a valid closed candidate (line 5). The algorithm then generates the neighbors by refining the attributes $\{a_f, ..., a_n\}$ continuing from d on the condition that *cnt_c* is not switched to *False* (lines 6-8). Flag f determines the index of the last attribute that was refined in the description d (operator η). Boolean cnt_c can be modified externally by some caller algorithm to prune the search space, for instance, when using optimistic estimates on the quality measures. Eventually, a recursive call is done to explore the sub search space related to d (lines 9-10). Hence, to enable the full exploration of search space related to the pattern structure $(G, (\mathcal{D}, \sqsubseteq), \delta)$, the algorithm is called with this initial parameters EnumCC$(G, *, \sigma, 1, true)$. Recall that $*$ is the description $\langle *, *, \ldots * \rangle$ having the complete collection G as its support.

Algorithm 1: EnumCC(G, d, σ_G, f, cnt)

Inputs : G is the collection of records, each encompassing m attributes,
d is a description from $\mathcal{D}$,
σ_G is a minimum support threshold,
$f \in [1, m]$ is a refinement flag,
`cnt` is a Boolean.

Output: yields all closed descriptions, i.e. $clo[\mathcal{D}] = \{clo(d) \text{ s.t. } d \in \mathcal{D}\}$

1 **if** $|G^d| \geq \sigma$ **then**
2   `closure_d` $\leftarrow$ clo$(d) = \delta(G^d)$
3   **if** $d \lessdot_f$ `closure_d` **then**
4   `cnt_c` $\leftarrow$ `copy(cnt);` `// can be modified by a caller algorithm`
5   ***yield*** (`closure_d`, $G^{\texttt{closure\_d}}$, `cnt_c`); `// yield results and wait`
6   **if** `cnt_c` **then**
7   **foreach** $j \in [f, m]$ **do**
8   **foreach** $d' \in \eta$(`closure_d`, j) **do**
9   **foreach** (`nc`, $G^{\texttt{nc}}$, `cnt_nc`) $\in$ EnumCC($G, d', \sigma_G, j,$ `cnt_c`) **do**
10   ***yield*** (`nc`, $G^{\texttt{nc}}$, `cnt_nc`)

Figure 2.6 illustrates the area and the elements of the search space explored by EnumCC, its depiction rely on the figure 2.3.

[8]Let $r_j = \{v_1, ..., v_q\}$ be an itemset condition and its closure $r'_j = clo(r_j) = \{v'_1, ..., v'_q, ..., v'_s\}$ with $Z = \{v_1, ..., v_l\}$ the set of possible items, all r_j, r'_j and Z are ordered using some arbitrary total order $\lessdot$ defined on Z. To assess the *canonicity test* between r_j and r'_j, and considering that r_j is generated after a refinement of its previous f^{th} item, the lectic order is defined as: $r_j \lessdot_f r'_j \Leftrightarrow \forall i \in [1..f-1] : v_i = v'_i \wedge t_f \lessdot u_f$

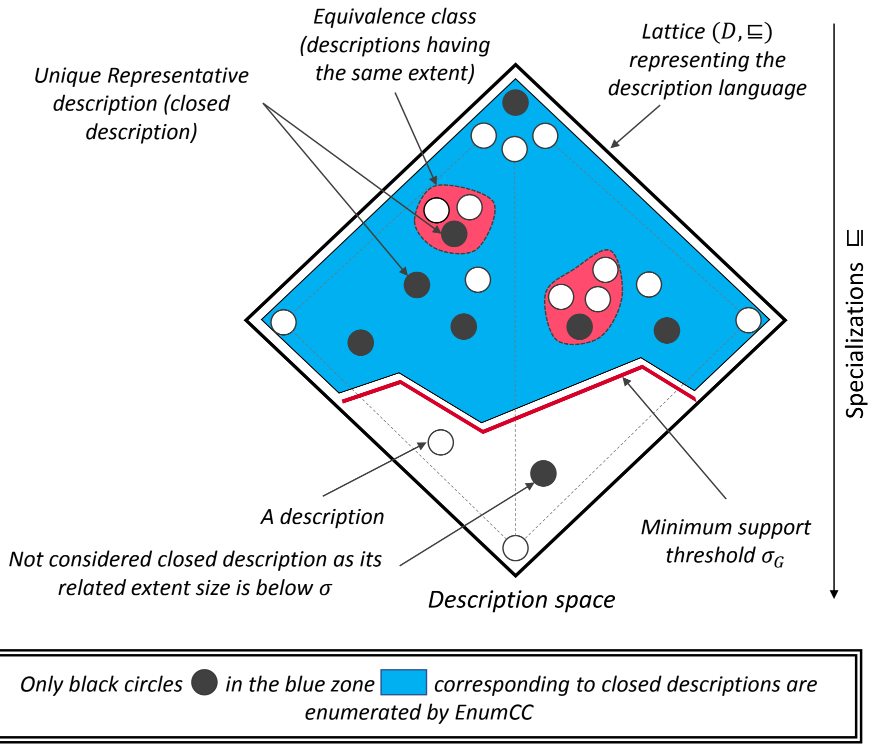


Figure 2.6: Illustration of the area and elements (formal concepts, closed descriptions) enumerated by EnumCC in some given pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ represented by its associated description langauage $(\mathscr{D}, \sqsubseteq)$ and the collection of characterizable subsets of the powerset 2^G (c.f. Figure 2.3).

2.4.2 A Standard Branch and Bound Algorithm For SD/EMM

Having in mind the three building components of subgroup discovery (cf. Figure 2.2) and algorithm EnumCC (cf. Algorithm 1). We explain below the standard scheme of a branch and bound algorithm which efficiently leverages the properties of the description language $(\mathscr{D}, \sqsubseteq)$ in the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ and extent-based interesting measures φ. Since, most interestingness measures are not monotonous, one needs to define proper optimistic estimates (upper bounds) on the quality measure (Grosskreutz, Rüping, and Wrobel, 2008) to quickly discard unpromising parts of the search space. In general, an optimistic estimate oe for a quality measure is defined as follows:

Definition 2.4.1 — Optimistic Estimate. An optimistic estimate oe for a given quality measure φ is a function such that:

$$\forall\ d' \in \mathscr{D} \,.\, d \sqsubseteq d' \Rightarrow \varphi(d') \leq \mathrm{oe}(d)$$

Intuitively, the definition above states that: given a description d from the lattice $(\mathscr{D}, \sqsubseteq)$, an optimistic estimates ensure that every description d' subsumed by d has its quality $\varphi(d')$ bounded by the quantity $\mathrm{oe}(d)$.

Several optimistic estimates had been proposed in the literature to allow an efficient exhaustive search for specific quality measures. For instance, Webb, 2001 proposed an optimistic estimate for the impact interestingness measure used for numerical target attributed datasets, i.e. $\varphi_{\text{impact}}(d) = |G^d|\varphi_{\text{mean}}(d)$. For a survey on optimistic estimates on quality measures for numerical target labels, we refer the interested reader to (Lemmerich, Atzmueller, and Puppe, 2016). Morishita and Sese, 2000 exploit convexity of interestingness measures in the ROC (coverage) space (Fürnkranz and Flach, 2005) (x-axis defined by **fpr** and y-axis defined by **tpr**, see Section 2.2.2) to provide proper optimistic estimates for correlation measures like χ_2 (Chi-squared) statisic and information gain (see (Abudawood and Flach, 2009; Fürnkranz and Flach, 2005)). The convexity property of interestingness measures in the ROC space has been also leveraged for defining optimistic estimates for other interestingness measures like the well-known Weighted Relative Accuracy (WRAcc), the proof of WRAcc convexity can be found in (Zimmermann and Raedt, 2009). In summary, one need to leverage properties of the underlying interestingness measures in the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ to devise adapted optimistic estimates.

Some optimistic estimates are better than other in terms of their pruning abilities (conservativeness (Grosskreutz, Rüping, and Wrobel, 2008)). Grosskreutz, Rüping, and Wrobel, 2008 defined the concept of *tight optimistic estimates* to refer to optimistic estimates that are as efficient as possible.

Definition 2.4.2 — Tight Optimistic Estimate. An optimistic estimate oe for a given quality measure φ is said to be tight if and only if:

$$\forall d \in \mathscr{D}\ \exists S \subseteq G^d \text{ s.t. } \mathrm{oe}(d) = \varphi(S)$$

Intuitively, an optimistic estimate is said to be tight if there exists a subset S in the extent of some given description whose quality $\varphi(S)$ is equal to the upper bound of the description $\mathrm{oe}(d)$. Note that, the subset does not need to be characterized by a description in $\mathscr{D}$.

Considering the Problem 2.2.1 (or 2.3.1) with the common SD constraints $\mathscr{C}$: minimum support size $|G^d| \geq \sigma_G$, a minimum threshold on the quality of subgroups $\varphi(d) \geq \sigma_\varphi$. A standard SD/EMM branch-and-bound algorithm performs a full traversal of the concept lattice induced from the pattern structure $(G, (\mathscr{D}, \sqsubseteq), \delta)$ to generate candidate subgroups without redundancy (cf. Section 2.2.1). This can be done by relying on the formerly presented algorithm EnumCC (cf. Algorithm 1). Each generated candidate subgroup have its quality evaluated, if it is above the required threshold σ_φ, it need to be kept in the final result set. Otherwise, the optimistic estimate oe is evaluated and the sub-search space of the current candidate can be pruned if the corresponding oe is below the quality threshold σ_φ. The algorithm stops when there is no remaining candidate subgroup. We call this algorithm B&B4SDEMM and we illustrate its pseudo-code in Algorithm 2.

We conclude this section by summarizing the concepts that have been introduced through this section in Figure 2.7. We augment Figure 2.6 depicting the search space explored by EnumCC. The figure depicts the subgroups that are traversed by the algorithm B&B4SDEMM. In short, only the closed descriptions are considered since we consider extent-based interestingness measures. Moreover, if applicable, an optimistic estimate is leveraged by the algorithm so as to prune as soon as possible unpromising areas of the search space. It is to be noted that most algorithms proposed in this thesis follow the same scheme defined by the algorithm B&B4SDEMM.

Algorithm 2: B&B4SDEMM$((G,(\mathscr{D},\sqsubseteq),\delta),\sigma_G,\varphi,\text{oe},\sigma_\varphi,k)$

Inputs : $(G,(\mathscr{D},\sqsubseteq),\delta)$ a pattern structure;
σ_G minimum support threshold of a description;
φ the quality measure;
oe the optimistic estimate;
σ_φ quality threshold on the quality; k of the top-k.

Output: L is the list of interesting subgroups.

$L \leftarrow \{\}$
$\sigma_\varphi^{\text{current}} \leftarrow \sigma_\varphi$
foreach $(d, G^d, \mathtt{cont}) \in \text{EnumCC}(G,*,\sigma_G,0,\text{True})$ **do**
  if $\text{oe}(d) < \sigma_\varphi^{\text{current}}$ **then**
    $\mathtt{cont} \leftarrow \text{False}$;   // Prune the sub-search space under d
  else if $\varphi(d) \geq \sigma_\varphi^{\text{current}}$ **then**
    $L \leftarrow (L \cup d)$
    if $|L| > k$ **then**
      $L \leftarrow L \setminus \{r\}$ with $r \in \{d \in L \mid \varphi(d) = \min(\{\varphi(d) \mid d \in L\})\}$
      $\sigma_\varphi^{\text{current}} \leftarrow \min(\{\varphi(d) \mid d \in L\})$
return L

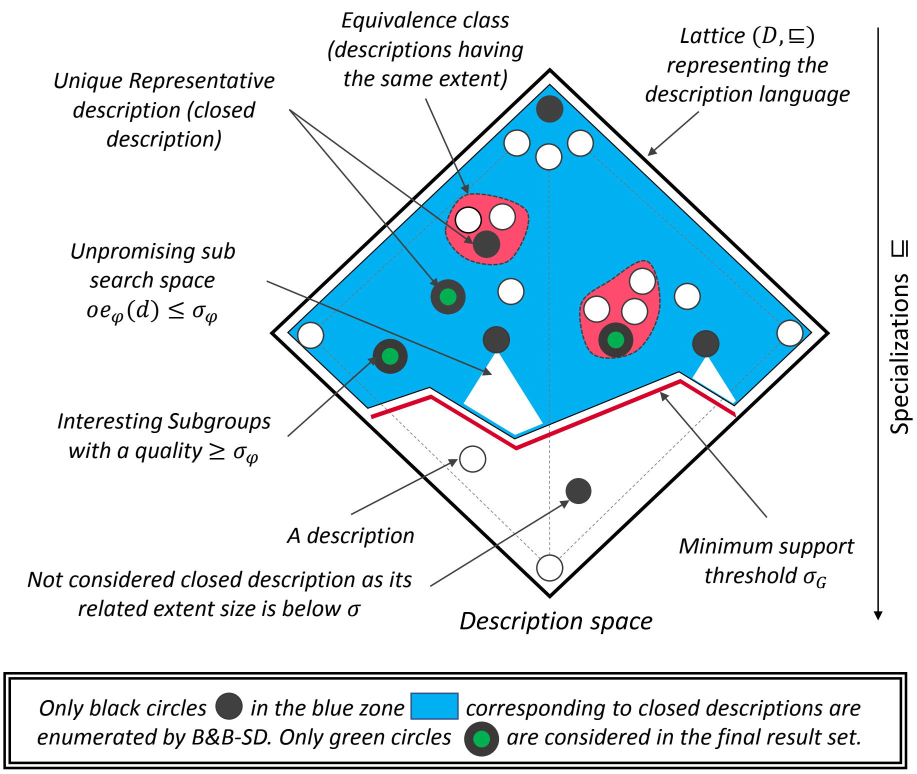


Figure 2.7: Illustration of the area and elements (interesting closed subgroups) enumerated by B&B4SDEMM via EnumCC in some given pattern structure $(G,(\mathscr{D},\sqsubseteq),\delta)$ (updating Figure 2.6). Only the interesting closed subgroups are kept in the final result set while unpromising areas of the search space are pruned by leveraging the optimistic estimates .

2.5 Potentials and Limitations

We conclude this chapter by showing the potential and limitations of SD/EMM[9] frameworks for the discovery of exceptional (dis)agreements in behavioral data (c.f. Definition 1.1.1). Considering the building blocks of SD and EMM (summarized in Figure 2.2 and Figure 2.4) and the algorithms presented in the previous Section 2.4, all we need to do is to instantiate properly these building blocks to enable the efficient discovery of exceptional (dis)agreement between and within groups. Recall that both aforementioned tasks consider contexts (cf. Definition 1.1.3) and groups (cf. Definition 1.1.2) to extract peculiar behavior between/among groups. Since, no model classes and interestingness measures have been proposed in the literature to mine for such patterns, our objective in this thesis is to harness the potentials of SD/EMM. The latter provides a solid theoretical framework to model the desired tasks and to devise efficient exhaustive algorithms for the analysis of exceptional subgroups once the property of interest is appropriately defined.

Before getting into the core of the proposed EMM tasks for behavioral data analysis, let us briefly review the limitations that prevented us from applying straightforwardly the existing EMM models to uncover the desired patterns from behavioral data, i.e. exceptional intra-group agreement and exceptional inter-group agreement.

- In most existing EMM models (cf. Section 2.3.2), the target attributes $(t_1,...,t_l)$ are fixed and given upfront to the task. This is not the case in our setting, a target space (groups) is provided instead of explicit targets. Dynamic EMM (i.e., EMM with a non-fixed model) has been recently investigated for different aims. Bosc et al., 2016 propose a method to handle multi-label data where the number of labels per record is much lower than the total number of labels which prevents the use of usual EMM model. Other dynamic EMM approaches aim to discover exceptional attributed sub-graphs (Bendimerad, Plantevit, and Robardet, 2016; Bendimerad et al., 2017b; Kaytoue et al., 2017) (cf. Section 2.3.2). Although, none of these (dynamic) models are straightforwardly adaptable for the discovery of the desired patterns in behavioral data. This point concerns Chapter 3 and Chapter 4.

- Considering the previous point and since no models have been proposed in the literature to discover exceptional intra/inter-group agreement, it is required to define proper model classes and adapted interestingness measures. Furthermore, correct optimistic estimates need to be devised to make the exhaustive search of the desired patterns possible. This point concerns Chapter 3 and Chapter 4.

- Earlier in this chapter, we discussed several possible interestingness measures and how they are handled both in SD and EMM frameworks. While most of the interestingness measures require a threshold on the quality fixed by the end-user before starting the algorithm (or a number k of desired patterns), evaluating interestingness via statistical

[9] Starting from now, we deliberately confound SD and EMM and we note SD/EMM since: (i) EMM is a generalization of SD when SD is seen from the perspective of supervised descriptive rule discovery (Kralj-Novak, Lavrac, and Webb, 2009) and SD is generalization of EMM when SD is seen from the perspective of Siebes, 1995 (more precisely, SD and EMM can be seen as instances of data surveying) or Wrobel, 1997. Although this choice seems late as it comes in the end of the chapter, it was motivated by the fact that presenting EMM as a generalization of SD is more intuitive and more didactic.

significance (Hämäläinen and Webb, 2019) is an interesting paradigm since: (1) it requires almost no input from the end-user (only the conventional intuitive critical value α), (2) it statistically validates the found patterns, avoiding hence to return spurious findings. However, there is no straightforward adaptations of existing approaches in the literature to handle statistical significance of results in our setting. Moreover, except for works addressing associations rules (Hämäläinen, 2010b; Minato et al., 2014), most of the literature work rely on non-efficient search algorithms (no pruning of uninteresting branches) (Duivesteijn and Knobbe, 2011; Lemmerich et al., 2016) to measure statistical significance of patterns during enumeration. Thus, we need to investigate proper and efficient significance measuring of patterns and associated correct optimistic estimates to render possible an exhaustive search algorithm for the desired patterns. This point concerns Chapter 4.

The next Chapters are devoted to the instanciation of SD/EMM framework for the discovery of exceptional **inter-group** agreement in behavioral data (Chapter 3) and the discovery of exceptional **intra-group** agreement in behavioral data (Chapter 4).

Note: The notations that have been introduced introduced in Chapter 1 and Chapter 2 are listed in Table C.1 in Appendix C.

3

Identifying exceptional (dis)agreement between groups

This chapter addresses the problem of discovering exceptional (dis)agreement patterns between groups in such data, i.e., groups of individuals that exhibit an unexpected mutual agreement under specific contexts compared to what is observed in overall terms. To tackle this problem, we design a generic approach, rooted in the Exceptional Model Mining framework, which enables the discovery of such patterns in two different ways. A branch-and-bound algorithm ensures an efficient exhaustive search of the underlying search space by leveraging closure operators and optimistic estimates on the interestingness measures. A second algorithm abandons the completeness by using a direct sampling paradigm which provides an alternative and tractable algorithm when an exhaustive search approach becomes unfeasible. To illustrate the usefulness of discovering exceptional (dis)agreement patterns, we report a comprehensive experimental study on four real-world datasets relevant to three different application domains: political analysis, rating data analysis and healthcare surveillance.

3.1 Introduction

In the former chapter, we have presented the theoretical background of Subgroup Discovery and Exceptional Model Mining which will serve to model the task of finding exceptional (dis)agreement between groups in behavioral datasets (cf. definition 1.1.1). In a nutshell, the aim of this chapter is to extend the capabilities of SD/EMM in order handle the discovery of such patterns in an efficient way. To this aim, we first need to instantiate the building blocks of EMM for this underlying problem. Also, we need to study the properties of the proposed interestingness measures to propose (tight) optimistic estimates. This enables the discovery of exceptional (dis)agreement in behavioral data in an optimal way.

Consider a behavioral dataset (cf. definition 1.1.1) describing the organization and votes of a parliamentary institution (e.g., European Parliament[1], US Congress[2]). Such datasets record the activity of each member in voting sessions held in the parliament, as well as information on the parliamentarians and the sessions. Table 3.1 provides an example. It reports the outcomes of European parliament members (MEPs) on legislative procedures. These procedures are described by attributes such as themes and dates. MEPs are characterized by their country, parliamentary group and age. The general trends are well known, and easy to check on these data with basics queries on data. For instance, the Franco-German axis is reflected by consensual votes between parliamentarians of both countries as well as the usual opposition between right wing and left wing. An analyst (e.g., a data journalist) is aware of these political positions and expects deeper insights. To this end, it is of major interest to discover groups of individuals that exhibit an unexpected mutual

ide	themes	date
e_1	1.20 Citizen's rights	20/04/16
e_2	2.10 Free Movement of goods	16/05/16
e_3	1.20 Citizen's rights; 7.30 Judicial Coop	04/06/16
e_4	7 Security and Justice	11/06/16
e_5	7.30 Judicial Coop	03/07/16
e_6	7.30 Judicial Coop	29/07/16

(a) Entities (Voting sessions)

idi	country	group	age
i_1	France	S&D	26
i_2	France	PPE	30
i_3	Germany	S&D	40
i_4	Germany	ALDE	45

(b) Individuals (Parliamentarians)

idi	ide	outcome
i_1	e_1	For
i_1	e_2	Against
i_1	e_5	For
i_1	e_6	Against
i_2	e_1	For
i_2	e_3	Against
i_2	e_4	For
i_2	e_5	For
i_3	e_1	For
i_3	e_2	Against
i_3	e_3	For
i_3	e_5	Against
i_4	e_1	For
i_4	e_4	For
i_4	e_6	Against

(c) Outcomes

Table 3.1: Example of behavioral dataset - European Parliament Voting dataset . This dataset is a replica of the dataset presented in Table 1.1

[1] http://parltrack.euwiki.org/
[2] https://voteview.com/data

agreement (or disagreement) under specific conditions (contexts). For example, from Table 3.1, an exceptional inter-group agreement pattern is $p = (c = \langle \text{themes} = 7.30 \text{ Judicial Coop} \rangle$, $u_1 = \langle \text{country} = \text{France} \rangle$, $u_2 = \langle \text{country} = \text{Germany} \rangle)$, which reads: "in overall terms, while German and French parliamentarians are in agreement (comparing majority votes leads to 66%[3] of equal votes), an unexpected strong disagreement between the two groups is observed for *Judicial Cooperation related* legislative procedures (the respective majorities voted the same way only 33% of the time in the corresponding voting sessions, i.e. $\{e_3, e_5, e_6\}$)".

In this chapter, we aim to discover such exceptional inter-group agreement patterns not only in voting data but also in more generic data which involves individuals, entities and outcomes, i.e. behavioral data (cf. definition 1.1.1). From such datasets, we aim to discover exceptional (dis)agreement between groups of individuals on specific contexts. That is to say, an important difference between the groups' behaviors is observed compared to the usual context (i.e., the whole data). This could answer a large variety of questions. For instance, considering political data, an analyst may ask: *what are the controversial subjects in the European parliament in which groups or parliamentarians have divergent points of view?* In collaborative rating analysis, one may ask *what are the controversial items? And which groups are opposed?* In Healthcare surveillance, the analyst may want to know if some medicines are prescribed much more often for one group of individuals than another one.

No model in the SD/EMM framework (cf. Chapter 2) makes it possible to investigate exceptional contextual (dis)agreement between groups. We made a first attempt to discover exceptional inter-group agreement patterns in (Belfodil et al., 2017a). However, the model proposed in (Belfodil et al., 2017a) requires the specification of many non-intuitive parameters that may be the source of misleading interpretation. In this work (Belfodil et al., 2019c), we strive to provide a simpler and more generic framework to analyze behavioral data.

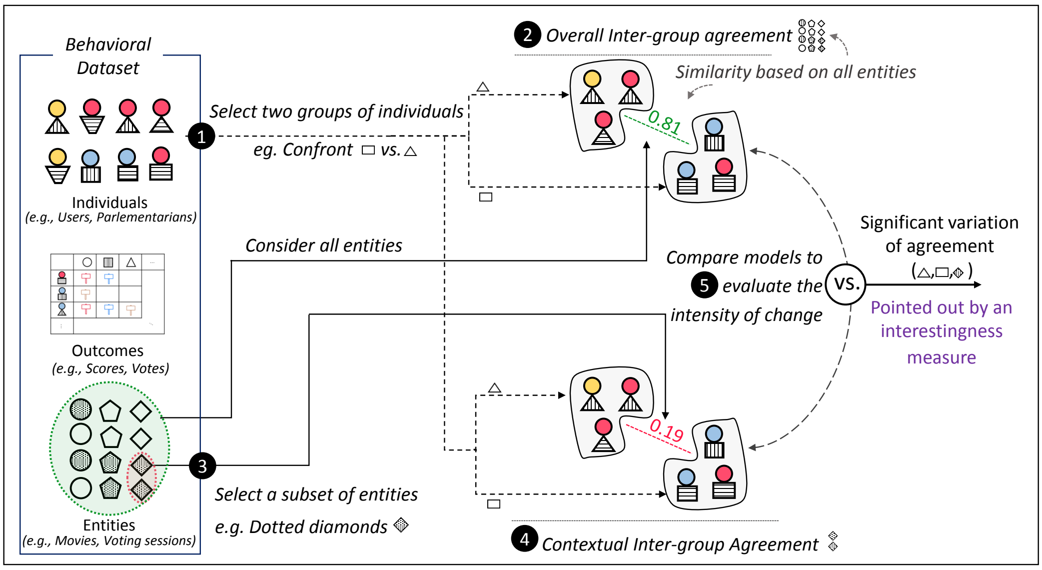


Figure 3.1: Overview of the task of discovering exceptional (dis)agreement between groups

[3] Since the majorities of $\langle u_1, u_2 \rangle$ voted respectively on $\{e_1, e_2, \underline{e_3}, e_4, \underline{e_5}, \underline{e_6}\}$ as follows: $\langle \text{For}, \text{For} \rangle$, $\langle \text{Against}, \text{Against} \rangle$, $\underline{\langle \mathit{Against}, \mathit{For} \rangle}$, $\langle \text{For}, \text{For} \rangle$, $\underline{\langle \mathit{For}, \mathit{Against} \rangle}$, $\underline{\langle \text{Against}, \text{Against} \rangle}$.

Figure 3.1 gives an overview of the approach we devise to discover exceptional agreement/disagreement between groups. At a high level of description, five steps are necessary to discover interesting inter-group agreement patterns. First, two groups of individuals (u_1, u_2) are selected by intents (1) . Then, their usual agreement on all their expressed outcomes is computed in step (2). All characterizable subsets of entities (contexts (c)) are then enumerated (3) and for each selected subset, the agreement between the two groups is measured (4) and compared to their usual agreement (5) to evaluate to what extent the mutual agreement changes, conveyed by an inter-group agreement pattern (c, u_1, u_2). Eventually, all pairs of groups (at least conceptually) are confronted. The discovery of exceptional inter-group agreement patterns requires to explore (simultaneously) both the search space associated to the individuals and the search space related to the entities. Moreover, behavioral datasets may contain several types of attributes (e.g., numerical, categorical attributes potentially organized by a hierarchy), and outcomes. This requires efficient enumeration strategies. Last but not least, different measures to capture agreement may be considered depending on the application domain. Accordingly, the proposed method must be generic.

Contributions. this chapter makes the following contributions:

Problem formulation. We define the novel problem of discovering exceptional (dis)agreement between groups of individuals when considering a particular subset of outcomes compared to the whole set of outcomes.

Algorithms. We propose two algorithms to tackle the problem of discovering exceptional inter-group agreement patterns. DEBuNk[4] is a branch-and-bound algorithm that efficiently returns the complete set of patterns. It takes benefit from both closure operators and optimistic estimates. Quick-DEBuNk is an algorithm that samples the space of inter-group agreement patterns in order to support instant discovery of patterns.

Evaluation. We report an extensive empirical study on both synthetic and real-world datasets. Synthetic datasets with controlled ground truth allows one to make some qualitative comparisons with some existing methods. It gives evidence that existing methods fail to discover inter-agreement patterns. The four real-world datasets are then used to demonstrate the efficiency and the effectiveness of our algorithms as well as the interest of the discovered patterns. Especially, we report three case-studies from different application domains: political analysis, rating data analysis and healthcare surveillance to demonstrate that our approach is generic.

The following content is based on our article on *Flash points* (Belfodil et al., 2017a) and its extension has been accepted in *Data Min. Knowl. Disc. journal* (Belfodil et al., 2019c).
Roadmap. The rest of this chapter is organized as follows. The problem formulation is given in Section 3.2. We present the *agreement* measure and how it is integrated into an interestingness measure to capture changes of inter-group agreement in Section 3.3. DEBuNk algorithm is presented in Section 3.4 while a pattern space sampling version, Quick-DEBuNk, is defined in Section 3.5. We report an empirical study in Section 3.6. Eventually, we discuss the potentials and limitations of the proposed approach in Section 3.7.
Note: Notations used in this chapter are listed in Appendix C and Appendix D.

[4]DEBuNk stands for Discovering Exceptional inter-group Behavior patterNs

3.2 Setup and Problem Formalization

Here, we first define the fundamental concepts that we use throughout the chapter in Section 3.2.1, followed by the formal problem statement in Section 3.2.2. Some definitions and notions that were already introduced in Chapter 2 will be recalled in brief in this section for the convenience of the reader.

3.2.1 Preliminaries

We are interested in discovering exceptional (dis)agreement among groups in *Behavioral Datasets* whose formal definition is given in Definition 1.1.1. Recall that a behavioral dataset is quadruple $\langle G_I,\ G_E,\ O,\ o\rangle$ where G_I is a collection of individuals, G_E is a collection of entities, O is the domain of possible outcomes and $o : G_I \times G_E \to O$ gives the outcome of an individual i over an entity e, if applicable.

In order to define appropriately the form of the sought patterns, we need first to characterize subgroups of data records in G_I and G_E. These two sets are collections of records defined over a set of descriptive attributes (Schema). We denote such *collection of records* by G, reintroducing the subscripts only in case of possible confusion. We assume $\mathscr{A} = (a_1,...,a_m)$ to be the ordered list of attributes constituting the schema of G. Each attribute a_j has a domain of interpretation, noted $\text{dom}(a_j)$, which corresponds to all its possible values. Attributes may be numerical or categorical potentially augmented with a taxonomy referred to by *Hierarchical Multi-Tag* (HMT) attributes (see section 3.4.2). For instance, in Table 3.1, parliamentarians, described by their country (categorical), their political group (categorical) and their age (numerical), decide on some voting sessions outlined by a date (numerical) and themes (HMT attribute). The attributes' domains define a description domain $\mathscr{D}_E$ (resp. $\mathscr{D}_I$) which corresponds to the set of all possible descriptions that one can use to characterize *subgroups* of records in G_E (resp. G_I). Recall that descriptions are conjunction of conditions of the form $d = \langle r_1,...,r_m\rangle$ where r_j depends on the type of the attribute a_j (cf. Definition 2.2.2 and Definition 2.2.12). Descriptions are ordered via a specialization operator $\sqsubseteq$ which roughly translates to: $d \sqsubseteq d'$, iff $d' \Rightarrow d$ (cf. Definition 2.2.4). Formally, the concept of *description* is used to describe both sets of individuals and sets of entities. Yet, for the ease of interpretation, we use two different terms to name them: *group description* and *context*. An example is given below.

Example 3.1 In Table 3.1, the *context* $c = \langle$ date $\in$ [05/06/16..30/07/16] $\rangle$ identifies the set of entities $G_E^c = \{e_4, e_5, e_6\}$. Similarly, the *group description* $u = \langle$ group='S&D' $\rangle$ selects the set of individuals $G_I^u = \{i_1, i_3\}$. ■

In the remaining, we manipulate the two pattern structures (cf. Definition 2.2.7) $(G_E, (\mathscr{D}_E, \sqsubseteq), \delta^E)$ and $(G_I, (\mathscr{D}_I, \sqsubseteq), \delta^I)$. Recall that δ^E (resp. δ^I) is a mapping operator which transforms a record $e \in G_E$ (resp. $i \in G_I$) to its maximal description $c \in \mathscr{D}_E$ (resp. $u \in \mathscr{D}_I$). Thus a subgroup of entities characterized by a *context* $c \in \mathscr{D}_E$ is denoted $G_E^d = \{e \in G_E \mid d \sqsubseteq \delta^E(e)\}$. Similarly a subgroup of individuals by a *group description* $u \in \mathscr{D}_I$ is denoted $G_I^u = \{i \in G_I \mid u \sqsubseteq \delta^I(i)\}$.

Since we are interested in patterns highlighting exceptional (dis)agreement between two groups of individuals described by u_1 and u_2, in a context c compared to the overall context, the sought patterns are defined as follows:

Definition 3.2.1 — Inter-Group Agreement Pattern. A *inter-group agreement pattern* is a triplet $p = (c, u_1, u_2)$ where $c \in \mathscr{D}_E$ is a *context* and $(u_1, u_2) \in \mathscr{D}_I^2$ are two *group descriptions*.

The *extent* of a inter-group agreement pattern p is $ext(p) = (G_E^c, G_I^{u_1}, G_I^{u_2})$ with G_E^c the set of entities satisfying the conditions of context c, and $G_I^{u_1}$ (resp. $G_I^{u_2}$) the set of individuals supporting the description u_1 (resp. u_2). The set of all possible patterns is denoted as $\mathscr{P} = \mathscr{D}_E \times \mathscr{D}_I \times \mathscr{D}_I$. Furthermore, as $\mathscr{P} = \mathscr{D}_E \times \mathscr{D}_I \times \mathscr{D}_I$ is the product of three partially ordered collections, patterns of $\mathscr{P}$ are also partially ordered. Since $(G_E, (\mathscr{D}_E, \sqsubseteq), \delta^E)$ and $(G_I, (\mathscr{D}_I, \sqsubseteq), \delta^I)$ are both pattern structures and the cartesian product of lattices related to that forms a lattice (Roman, 2008), we have $\langle G_E \times G_I \times G_I, (P, \sqsubseteq), \delta = (\delta^E, \delta^I, \delta^I)\rangle$ is a pattern structure (cf. Definition 2.2.7)).

Definition 3.2.2 — Specialization between patterns $\sqsubseteq$. Let p and p' be two patterns from $\mathscr{P}$, p' is a *specialization of a pattern* p, denoted $p \sqsubseteq p'$, iff $c \sqsubseteq c'$, $u_1 \sqsubseteq u_1'$ and $u_2 \sqsubseteq u_2'$.

Notice that, if $p \sqsubseteq p'$ then $\text{ext}(p') \subseteq \text{ext}(p)$, that is $G_E^{c'} \subseteq G_E^c$ and $G_I^{u_1'} \subseteq G_I^{u_1}$ and $G_I^{u_2'} \subseteq G_I^{u_2}$. Some descriptions are considered to be equivalent if they characterize the same subset $S \subseteq G$. i.e. two descriptions $d_1, d_2 \in \mathscr{D}$ are equivalent iff $G^{d_1} = G^{d_2}$ (cf. Definition 2.2.11). Similarly, two patterns $p, p' \in \mathscr{P}$ are equivalent if they share the same extent, i.e. $\text{ext}(p) = \text{ext}(p')$.

To objectively evaluate how interesting an inter-group agreement pattern is, a *quality measure* φ is required as formerly introduced in Definition 2.2.6 and discussed in Section 2.2.2, in the scope of this chapter and since the sought patterns are in $\mathscr{P}$, the quality measure is a function $\varphi : \mathscr{P} \to \mathbb{R}$ which assigns to each pattern $p = (c, u_1, u_2) \in \mathscr{P}$ a real number $\varphi(p) \in \mathbb{R}$.

A quality measure is designed to compare patterns: the quality of one will be compared to the quality of the others, most of the time to choose the best one. Consequently, it must be carefully designed with respect to what the algorithm is expected to produce. Our first objective is to identify particular parts of the data. This naturally leads to quality evaluation functions focusing on the extent of the pattern. Moreover, in this case, any consideration about the syntax of the pattern can only interfere and has to be avoided. Consequently, the quality measures we propose[5] are extent-based quality measures which are of the form: $\varphi(p) = \varphi'(ext(p))$ (see Definition 2.2.6 and its following paragraph). It follows that two patterns characterizing the same data, i.e. with the same extent, share the same quality measure: $\forall p, p' \in \mathscr{P}$, if $ext(p) = ext(p')$ then $\varphi(p) = \varphi(p')$.

3.2.2 Formal Problem Definition

The user will be provided with a collection of patterns that captures exceptional (dis)-agreements in a given behavioral dataset. A first intuitive idea is to provide all patterns of high quality, i.e. with a quality greater than a user-defined threshold σ_φ. However, by construction of the quality measures, different patterns sharing the same extent will

[5]Different quality measures are proposed in Sec. 3.3.

reach the same quality level, leading to multiple descriptions of the same parts of the data. Assuming that the user can be quickly bothered by such duplication, we propose to expose each interesting part of the data only once. More interestingly, the system should provide the user with the best generalizations only, i.e., patterns whose extent is not included in some other found ones. Additionally, some cardinality constraints can be added to avoid patterns of too small extent. Given two minimum support thresholds σ_E and σ_I, these constraints ensure, for a pattern $p = (c, u_1, u_2)$, that the size of the context extent (i.e. $|G_E^c| \geq \sigma_E$) and the size of both groups (i.e. $|G_I^{u_1}| \geq \sigma_I$ and $|G_I^{u_2}| \geq \sigma_I$) are large enough. Now, we introduce formally the core problem we tackle in this chapter.

Problem 3.2.1 (*Discovering Exceptional (Dis)Agreement between Groups*).

Given a behavioral dataset $\langle G_I, G_E, O, o \rangle$, a quality measure φ, a quality threshold σ_φ and a set of cardinality constraints $\mathscr{C}$, the problem is to find the pattern set $P \subseteq \mathscr{P}$ such that the following conditions hold:

1. (*Validity*) $\forall p \in P \ : \ p$ *valid*, that is p *satisfies* $\mathscr{C}$ and $\varphi(p) \geq \sigma_\varphi$.
2. (*Maximality*) $\forall p \in P \ \forall q \in \mathscr{P} : \text{ext}(q) = \text{ext}(p) \Rightarrow q \sqsubseteq p$
3. (*Completeness*) $\forall q \in \mathscr{P} \setminus P : q$ *valid* $\Rightarrow \exists p \in P$ s.t. $\text{ext}(q) \subseteq \text{ext}(p)$
4. (*Generality*) $\forall (p,q) \in P^2 : p \neq q \Rightarrow ext(p) \not\subseteq ext(q)$.

Condition (1) ensures that the patterns in P are of high quality and satisfy the cardinality constraints. Condition (2) retains only one unique representative among patterns sharing the same extent: the maximal one w.r.t. $\sqsubseteq$. Such a pattern exists only if the specialization relation $\sqsubseteq$ over the pattern space induces a lattice structure (Ganter and Kuznetsov, 2001) (we have $\langle G_E \times G_I \times G_I, (P, \sqsubseteq), \delta \rangle$ is a pattern structure). The maximal pattern w.r.t. $\sqsubseteq$ is commonly refered to as the *closed pattern* (Pasquier et al., 1999). We confine ourselves to such pattern spaces. Condition (3) ensures that each valid pattern in $\mathscr{P}$ has a representative in P covering it, while condition (4) ensures that only the most general patterns w.r.t. their extents are in P. In other words, the combination of conditions (3) and (4) guarantees that the solution P is minimal in terms of the number of patterns while having each valid pattern represented in the solution. Considering the generic definition of the quality measure discussed here, this problem extends the top-k problem addressed in (Belfodil et al., 2017a) (see Problem 2.3.1) by introducing conditions (3) and (4). That is, for a sufficiently large k, the method formerly provided in (Belfodil et al., 2017a) solves this problem only limited to the two first conditions providing, hence, a solution with a much larger number of redundant patterns.

3.3 Inter-Group Agreement Measure and Interestingness Evaluation

The previous section has already hinted at the fact that pattern interestingness is assessed with a quality measure φ whose generic definition is given. Here we show how such measure captures the deviation between the *contextual inter-group agreement* and the *usual inter-group agreement*. The inter-group agreement being the **model** (as required by the EMM framework (cf. Figure 2.4)) we choose to capture the inter-group agreement.

3.3.1 Quality Measures

For any pattern $p = (c, u_1, u_2) \in \mathscr{P}$, we denote by p^* the pattern $(*, u_1, u_2)$ which involves all the entities. IAS (p^*) (resp. IAS (p)) represents the usual (resp. contextual) inter-group agreement observed between the two groups u_1, u_2. In order to discover interpretable patterns, we define two quality measures that rely on IAS (p^*) and IAS (p).

- φ_{consent} assesses the strengthening of inter-group agreement in a context c:

$$\varphi_{\text{consent}}(p) = \max\left(\text{IAS}(p) - \text{IAS}(p^*), 0\right) \ .$$

- φ_{dissent} assesses the weakening of inter-group agreement in a context c:

$$\varphi_{\text{dissent}}(p) = \max\left(\text{IAS}(p^*) - \text{IAS}(p), 0\right) \ .$$

For instance, one can use φ_{consent} to answer: "*What are the contexts for which we observe more consensus between groups of individuals than usual?*".

3.3.2 Inter-group Agreement Similarity (IAS)

Several IAS measures can be designed according to the domain in which the data was measured (e.g., votes, ratings) and the user objectives. The evaluation of an IAS measure between two groups of individuals over a context requires the definition of two main operators: the *outcome aggregation operator* (θ) which computes an aggregated outcome of a group of individuals for a given entity, and a *similarity operator* (sim) which captures the similarity between two groups based on their aggregated outcomes over a single entity. These operators are defined in a generic way as following.

Definition 3.3.1 — Outcome Aggregation Operator θ. An aggregation operator is a function $\theta : 2^{G_I} \times G_E \rightarrow \mathbb{D}$ which transforms the outcomes of a group of individuals G_I^u over one entity $e \in G_E$ (i.e. $\{o^a(i,e) \mid i \in G_I^u\}$) into a value in a domain $\mathbb{D}$ (e.g. $\mathbb{R}$, *categorical values*).

[a] $o(i,e)$ returns the outcome expressed by an individual i to an entity e, if given.

Definition 3.3.2 — Similarity between aggregated outcomes sim. $\text{sim} : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}^+$ assigns a real positive value sim(x,y) to any couple of aggregated outcomes (x,y).

Based on these operators, we properly define IAS which assigns to each pattern $p = (c, u_1, u_2) \in \mathscr{P}$ a value IAS$(p) \in \mathbb{R}^+$. This similarity evaluates how the two groups of individuals (u_1, u_2) behave similarly given their outcomes w.r.t. the context c. In the scope of our study, we confine ourselves to IAS measures that can be expressed as weighted averages. The next definition, though limiting, is generic enough to handle a wide range of behavioral data.

Definition 3.3.3 — Inter-group Agreement Similarity Measure IAS. Let w be a function associating a weight to each triple from $(G_E \times 2^{G_I} \times 2^{G_I})$. The IAS of a pattern (c, u_1, u_2) ($IAS : \mathscr{P} \to \mathbb{R}^+$) is the weighted average of the similarities of the aggregated outcomes for each entity e supporting the context c.

$$IAS(c, u_1, u_2) = \frac{\sum_{e \in G_E^c} w(e, G_I^{u_1}, G_I^{u_2}) \times \mathrm{sim}(\theta(G_I^{u_1}, e), \theta(G_I^{u_2}, e))}{\sum_{e \in G_E^c} w(e, G_I^{u_1}, G_I^{u_2})}$$

3.3.3 Examples of IAS Measures

By simply defining sim and θ, we present two instances of IAS measure that address two types of behavioral data with specific aims.

3.3.3.1 Behavioral Data With Numerical Outcomes

Collaborative Rating datasets are a classic example of behavioral data with numerical outcomes. Such datasets describe users who express numerical ratings belonging to some interval $O = [\min, \max]$ (e.g., 1 to 5 stars) over reviewees (*e.g. movies, places*). A simple and adapted measure for aggregating individual ratings over one entity is the weighted mean $\theta_{\text{wavg}} : 2^{G_I} \times G_E \to [min, max]$.

$$\theta_{\text{wavg}}(G_I^u, e) = \frac{1}{\sum_{i \in G_I^u} w(i)} \sum_{i \in G_I^u} w(i) \times o(i, e) \tag{3.1}$$

Weight $w(i)$ corresponds to the importance of ratings given by each individual $i \in G_I$. Such weight may depend on the confidence of the individual or the size of the sample population if fine granularity ratings (*rating of each individual*) are missing. If no weights are given, θ_{wavg} computes a simple average over ratings, denoted θ_{avg}. To measure agreement between two aggregated ratings over a single entity, we define $\mathrm{sim}_{\text{rating}} : [\min, \max] \times [\min, \max] \to [0, 1]$.

$$\mathrm{sim}_{\text{rating}}(x, y) = 1 - \left(\frac{|x - y|}{\max - \min} \right) \tag{3.2}$$

3.3.3.2 Behavioral Data with Categorical Outcomes

A typical example of such datasets are Roll Call Votes (RCVs)[6] datasets where voting members cast categorical votes. The outcome domain O is the set of all possible votes (e.g., $O = \{\text{For}, \text{Against}, \text{Abstain}\}$). To aggregate categorical outcomes we use the majority vote[7] θ_{majority}. We adapt its definition to handle potential ties (i.e., non-unique majority vote). Hence, $\theta_{\text{majority}} : 2^{G_I} \times G_E \to 2^O$ returns all the outcomes that received the majority of votes.

$$\begin{gathered} \theta_{\text{majority}}(G_I^u, e) = \{v \in O \ : \ v = \operatorname*{argmax}_{z \in O} \#\text{votes}(z, G_I^u, e)\} \\ \text{with } \#\text{votes}(z, G_I^u, e) = |\{(i, e) : i \in G_I^u \wedge o(i, e) = z\}| \end{gathered} \tag{3.3}$$

[6] Roll-Call vote is a voting system where the vote of each member is recorded, such as `http://www.europarl.europa.eu` (EU parliament) or `https://voteview.com` (US Congresses).

[7] The same measure is used by ***votewatch*** to observe the voting behaviors of groups of parliamentarians- `http://www.votewatch.eu/blog/guide-to-votewatcheu/`

We use a Jacquard index to assess the similarity between two majority votes x and y. Hence, $\text{sim}_{\text{voting}} : 2^O \times 2^O \to [0,1]$ is defined as follows.

$$\text{sim}_{\text{voting}}(x,y) = \frac{|x \cap y|}{|x \cup y|} . \tag{3.4}$$

3.3.4 Discussion

We introduced above two simple similarity measures that can be used as part of the IAS measure to assess how similar two groups of individuals are. More sophisticated measures can be considered. For instance, in behavioral datasets with categorical outcomes, one can define an outcome aggregation measure which takes into account the empirical distribution of votes and then a similarity measure which builds up on a statistical distance (e.g. Kullback-Leibler divergence (Csisz, 1967; Johnson and Sinanovic, 2001)). Such measures can also be used on behavioral datasets which involves numerical outcomes, for instance *Earth Mover Distance* measure was investigated in similarly structured dataset (rating dataset) in (Amer-Yahia et al., 2017). Several other measures can be relevant to analyze behavioral data with numerical outcomes depending on the aim of the study. In the empirical study, we investigate another similarity measure which relies on a ratio to highlight discrepancies between the medicine consumption rates of two subpopulations.

3.4 Mining Exceptional Inter-Group Agreement Patterns

We address the design of an efficient algorithm for enumerating exceptional inter-group agreement patterns. First, we present how candidates are enumerated without redundancy by relying on the pattern structure formalization (cf. Section 2.2.1). Second, we detail the enumeration process, paying particular attention to the attribute domains depicted by a hierarchy. Next, we propose optimistic estimates for the quality measures to enable pruning uninteresting branches of the search space. Eventually, these elements are used to define an efficient branch-and-bound algorithm which computes the complete set of relevant inter-group agreement patterns.

3.4.1 Enumerating Candidate Subgroups

Exploring the space of inter-group agreement patterns from $\mathscr{D}_E \times \mathscr{D}_I \times \mathscr{D}_I$ is equivalent to enumerating descriptions in $\mathscr{D}_E$ and $\mathscr{D}_I$ concurrently. Given the fact that the quality measures addressed in this work are extent-based quality measures and since $(G_E, (\mathscr{D}_E, \sqsubseteq), \delta^E)$ and $(G_I, (\mathscr{D}_I, \sqsubseteq), \delta^I)$ are two pattern structures (cf. Definition 2.2.7), we enumerate all closed descriptions (closed contexts, and closed groups of descriptions) using Algorithm EnumCC (cf. Algorithm 1). Recall that EnumCC enumerates, in a depth-first search manner, once and only once all the *closed contexts* c (closed group descriptions u) that fulfill the support constraint $|G_E^c| \geq \sigma_E$ (resp. $|G_I^u| \geq \sigma_I$) with σ_E (resp. σ_I) a user defined minimum support threshold on the context (resp. group description) related subgroup.

3.4.2 Hierarchical Multi-Tag Attribute (HMT)

Vote and review datasets often contain multi-tagged records whose tags are part of a hierarchical structure. For instance, voting sessions in the EU parliament can have multiple tags. For example, procedure *Gender mainstreaming in the work of the EU Parliament* is tagged by *4.10.04-Gender equality* and *8.40.01-EU Parliament*. Tag *4.10.04* is identified in a hierarchy as a specialization of tag *4.10* that depicts *Social policy* and which is itself a specialization of tag *4* that covers all the sessions related to *Economic, social and territorial cohesion*. We formally define this type of attribute named HMT.

Definition 3.4.1 — HMT Attribute. Let $T = \{t_1, t_2 \dots t_k\} \cup \{*\}$ be a set of values (also called tags), $<$ be a partial order over T inducing a tree structure $(T,<)$ whose root is '$*$'. $t_i < t_j$ denotes the fact that t_j is a descendant of t_i in T. In addition, the ascendants (resp. descendants) of a tag $t \in T$ is $\uparrow t = \{t' \in T | t' \leq t\}$ (resp. $\downarrow t = \{t' \in T | t' \geq t\}$). If t is a parent of a tag t' according to the tree T, it is denoted by $t = p(t')$.
A HMT attribute a_j takes its values in $dom(a_j) = 2^T$.

As an example, Fig. 3.2b describes G, a set of tag records defined by a unique attribute *tags*. Elements of *tags* are organized through the tree from Fig. 3.2a. We have $* < 1 < 1.20$ and $\uparrow 1.20 = \{1.20, 1, *\}$.

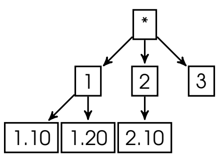
For a HMT attribute a_j, each record $g \in G$ is mapped by $\delta_j(g)$ to its corresponding tightest set of tags in $\text{dom}(a_j)$. If $\delta_j(g) = \{t_1, ..., t_n\}$, the record g is tagged *explicitly* by all the tags t_k for $k \in [1,n]$ and also *implicitly* by all their generalizations $\uparrow t_k$. Figure 3.2c illustrates this by reporting the flat representation of the collection of tagged records depicted in Figure 3.2b. It follows that a condition over a HMT attribute is defined as follows:

Definition 3.4.2 — Condition on a HMT attribute. (extends definition 2.2.12) Let G be a collection defined over the schema $\mathscr{A} = \{a_1, ..., a_m\}$

- If a_j a HMT attribute then **condition** r_j is a superset test of the form $a_j \supseteq \chi$ with $\chi \in \text{dom}(a_j)$.

Accordingly, a HMT condition can be depicted by a rooted sub-tree of T and a record supports such a condition if it contains at least all tags of the sub-tree. Moreover, it can be seen as a restricted itemset language (cf. Section 2.2.1). It follows that, the partial order between two HMT conditions r, r' denoted $r \sqsubseteq r'$ (r' *is a specialization* r) is valid if the sub-tree r covers the sub-tree r'. i.e. , $r \sqsubseteq r'$ means $\forall t \in r \ \exists t' \in r'$ s.t. $t' \in \downarrow t$.

Two ways are possible to take this attribute into account in the enumeration of descrip-



(a) Tree of tags - T

	tags
g^1	$\{1.20, 2.10\}$
g^2	$\{1, 3\}$
g^3	$\{1.10, 2.10, 3\}$
g^4	$\{2.10\}$
g^5	$\{1.20\}$

(b) Tagged records

	$*$	1	1.10	1.20	2	2.10	3
g^1	×	×		×	×	×	
g^2	×	×					×
g^3	×	×	×		×		×
g^4	×				×	×	
g^5	×	×		×			

(c) Flat representation

Figure 3.2: A collection of records labeled each by a set of tags and its flat representation.

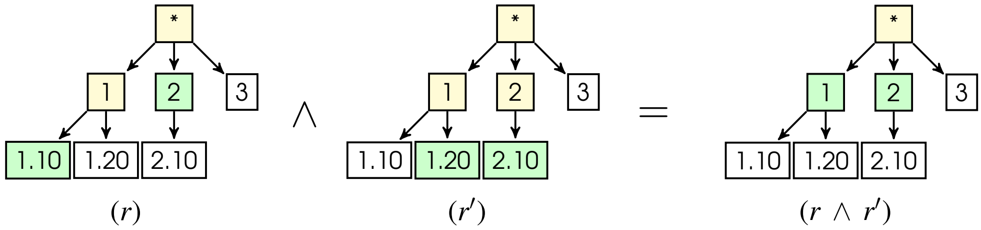


Figure 3.3: Illustration of the conjunction operator $\wedge$ between two HMT descriptions

tions from the complex search space aforementioned. One straightforward solution is to consider HMT attribute values as *itemsets* as depicted in the vector representation in Fig. 3.2c. However, such a solution ignores the taxonomy T implying the enumeration of *chain descriptions*. For instance, a chain description $\{1, 1.20.01\}$ is regarded as a different description than $\{1.20.01\}$. This stems from the fact that items are unrelated from the viewpoint of itemsets solution. As a consequence, a larger search space is explored while determining the same number of closed descriptions. To tackle this issue, we define a HMT description language.

Similarly to the aforementioned attributes, we define the conjunction operator $\wedge$ between two conditions which computes the *maximum common sub-tree* covering a set of conditions. Let $r = \{t_1, ..., t_n\}$ and $r' = \{t'_1, ..., t'_m\}$ be two HMT conditions, $r \wedge r' = \max(\cup_{t \in r} \uparrow t \cap \cup_{t' \in r'} \uparrow t')$ where $\max : 2^T \to 2^T$ maps a subset of tags $s \subseteq T$ to the leafs of the sub-tree induced by s: $\max(s) = \{t \in s | (\downarrow t \setminus \{t\}) \cap s = \emptyset\}$. Intuitively, $r \wedge r'$ is the set of the maximum explicit (green) and implicit tags (yellow) shared by the two descriptions. For instance, if $r = \{1.10, 2\}$ and $r' = \{1.10, 2.10\}$, we have $r \wedge r' = \{1, 2\}$ (cf. Fig. 3.3).

Moreover, we define an atomic refinement operation which enables calculating neighbors of a HMT condition r. A condition r' is said to be a neighbor of r if: either only one tag of r is refined in r' or a new tag is added in r' that shares a parent with a tag in r or with one of its ascendants. Formally:

$$\begin{cases} \exists! \ (t, u) \in r \times r' \ : \ t = p(u) \wedge \forall t' \in (r \setminus t) \ \exists u' \in r' \ : \ t' = u' \ \text{ if } & |r| = |r'| \\ \forall t \in r \ \exists u \in r' : t = u \wedge \exists! (t, u) \in r \times r' \text{ s.t. } \exists t' \in \uparrow t : p(u) = p(t') & |r| = |r'| + 1 \end{cases} \tag{3.5}$$

Finally, we define the lectic order between two conjunctions of tags $r = \{t_1, ..., t_n\}$ and its closure $r' = \{t'_1, ..., t'_n, ..., t'_m\}$ for the *canonicity test* to avoid the enumeration of already visited descriptions. Let r be generated after a refinement of the f^{th} tag, the lectic order is defined as: $r \lessdot_f r' \Leftrightarrow \forall i \in [1..f-1] : t_i = t'_i \wedge t_f \lessdot t'_f$. The linear order $\lessdot$ between tags can be provided by a depth first search order on T. These concepts being defined, the mapping function δ can be extended easily to handle HMT among other attributes. Note that HMT supports itemsets. This can be done simply by considering a flat tree T with all the items as leaves. Hence, HMT can be seen as a generalization of itemsets, where implications between items are known. Within this aim, we investigated a more generic generalization of itemsets with underlying implications in a recent work (Belfodil, Belfodil, and Kaytoue, 2019) which is out of the scope of this thesis.

3.4.3 Optimistic Estimates on Quality Measures

The enumeration of closed patterns enables a non-redundant traversal of the search space without pruning based on the quality measure. We present some pruning properties based on bounds on $\varphi_{consent}$ and $\varphi_{dissent}$.

Let u_1, u_2 be two descriptions from $\mathscr{D}_I$ that respectively cover the two groups $G_I^{u_1}$, $G_I^{u_2}$. We consider optimistic estimates (cf. Definition 2.4.1) only with regards to the description space $\mathscr{D}_E$. We assume that u_1 and u_2 are instantiated a priori. In the scope of this work, an optimistic estimate oe for a given quality measure φ is a function such that:

$$\forall\ c,d \in \mathscr{D}_E\ .\ c \sqsubseteq d \Rightarrow \varphi(d,u_1,u_2) \leq oe(c,u_1,u_2)$$

Tight optimistic estimates (cf. Definition 2.4.2) offer more pruning abilities than simple optimistic estimate. Without loss of generality, we assume that the input domains of oe and φ are defined over both the pattern space $\mathscr{P}$ and over $2^{G_E} \times 2^{G_I} \times 2^{G_I}$. This is possible, since the quality measure only depends on extents. In the scope of this work, a tight optimistic estimate oe is tight iff:

$$\forall c \in \mathscr{D}_E.\ \exists S \subseteq G_E^c\ :\ oe(G_E^c, G_I^{u_1}, G_I^{u_2}) = \varphi(S, G_I^{u_1}, G_I^{u_2})$$

.

3.4.3.1 Lower Bound and Upper Bound for the IAS Measure

The two quality measures $\varphi_{consent}$ and $\varphi_{dissent}$ rely on the IAS measure. Since u_1 and u_2 are considered to be instantiated for optimistic estimates, we can rewrite the IAS measure for a context $c \in \mathscr{D}_E$ and its extent G_E^c:

$$\mathrm{IAS}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \frac{\sum\limits_{e \in G_E^c} w_e \times \alpha(e)}{\sum\limits_{e \in G_E^c} w_e}\ \mathit{with} \begin{cases} \alpha(e) = & \mathrm{sim}(\theta(G_I^{u_1}, e), \theta(G_I^{u_2}, e)) \\ w_e = & w(e, G_I^{u_1}, G_I^{u_2}) \end{cases}.$$

We can now define a lower bound LB and an upper bound UB for the IAS measure based on the following operators that are defined for any context $c \in \mathscr{D}_E$ and for $n \in \mathbb{N}$:

- $\mathrm{m}(G_E^c, n) = \mathrm{Lowest}_{e \in G_E^c}(\{w_e \times \alpha(e) \mid e \in G_E^c\}, n)$ returns the set of the n distinct records e from G_E^c having the lowest values of $w_e \times \alpha(e)$.

- $\mathrm{M}(G_E^c, n) = \mathrm{Highest}_{e \in G_E^c}(\{w_e \times \alpha(e) \mid e \in G_E^c\}, n)$ returns the set of the n distinct records e from G_E^c having the highest values of $w_e \times \alpha(e)$.

- $\mathrm{mw}(G_E^c, n) = \mathrm{Lowest}_{e \in G_E^c}(\{w_e \mid e \in G_E^c\}, n)$ returns the set of the n distinct records e from G_E^c having the lowest values of w_e.

- $\mathrm{Mw}(G_E^c, n) = \mathrm{Highest}_{e \in G_E^c}(\{w_e \mid e \in G_E^c\}, n)$ returns the set of the n distinct records e from G_E^c having the highest values of w_e.

Proposition 3.4.1 — Lower bound LB for IAS. we define function LB as

$$\mathrm{LB}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \frac{\sum_{e \in \mathrm{m}(G_E^c, \sigma_E)} w_e \times \alpha(e)}{\sum_{e \in \mathrm{Mw}(G_E^c, \sigma_E)} w_e}$$

For any context c (corresponding to a subgroup G_E^c), LB provides a lower bound for IAS w.r.t. contexts with σ_E a minimum context support threshold:

$$\forall c, d \in \mathscr{D}_E.\ c \sqsubseteq d \Rightarrow \mathrm{LB}(G_E^c, G_I^{u_1}, G_I^{u_2}) \leq \mathrm{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2})$$

Before giving the proof of the proposition 3.4.1 we present the following lemma.

Lemma 3.4.2 Let $n \in \mathbb{N}^*$, $A = \{a_i\}_{1 \leq i \leq n}$ and $B = \{b_i\}_{1 \leq i \leq n}$ such that:

$$\begin{aligned} \forall i \in 1..n-1 &\ :\ 0 \leq a_i \leq a_{i+1} \\ \forall i \in 1..n-1 &\ :\ 0 < b_{i+1} \leq b_i \end{aligned}$$

we have:

$$\forall k \in 1..n \quad : \quad \frac{\sum_{i=1}^k a_i}{\sum_{i=1}^k b_i} \leq \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \leq \frac{\sum_{i=n-k+1}^n a_i}{\sum_{i=n-k+1}^n b_i}$$

Proof (lemma 3.4.2). Using the same notations of the lemma, we know that:

$$\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} - \frac{\sum_{i=1}^k a_i}{\sum_{i=1}^k b_i}$$

is of the same sign of:

$$\left(\sum_{i=1}^n a_i\right) \times \left(\sum_{i=1}^k b_i\right) - \left(\sum_{i=1}^k a_i\right) \times \left(\sum_{i=1}^n b_i\right)$$

This above quantity is equal to:

$$\left(\sum_{i=1}^k a_i + \sum_{i=k+1}^n a_i\right) \times \left(\sum_{i=1}^k b_i\right) - \left(\sum_{i=1}^k a_i\right) \times \left(\sum_{i=1}^k b_i + \sum_{i=k+1}^n b_i\right)$$

Which is equal to

$$\left(\sum_{i=k+1}^n a_i\right) \times \left(\sum_{i=1}^k b_i\right) - \left(\sum_{i=1}^k a_i\right) \times \left(\sum_{i=k+1}^n b_i\right)$$

Using the lemma hypotheses (orders between a_i's and b_i's), we have:

$$\begin{aligned} \sum_{i=k+1}^n a_i &\geq (n-k) \times a_k \\ \sum_{i=1}^k b_i &\geq k \times b_k \\ \sum_{i=1}^k a_i &\leq k \times a_k \\ \sum_{i=k+1}^n b_i &\leq (n-k) \times b_k \end{aligned}$$

Thus:

$$\left(\sum_{i=k+1}^{n} a_i\right) \times \left(\sum_{i=1}^{k} b_i\right) \geq (n-k) \times k \times a_k \times b_k$$
$$\left(\sum_{i=1}^{k} a_i\right) \times \left(\sum_{i=k+1}^{n} b_i\right) \leq (n-k) \times k \times a_k \times b_k$$

We conclude that

$$\left(\sum_{i=k+1}^{n} a_i\right) \times \left(\sum_{i=1}^{k} b_i\right) - \left(\sum_{i=1}^{k} a_i\right) \times \left(\sum_{i=k+1}^{n} b_i\right) \geq 0$$

Hence, we have:

$$\forall k \in 1..n \quad : \quad \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i} \leq \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Similarly the inequality $\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \frac{\sum_{i=n-k+1}^{n} a_i}{\sum_{i=n-k+1}^{n} b_i}$ can be easily proved following the same line of reasoning of the proof of the first part of the inequality.

■

Proof (Proposition 3.4.1). By a straightforward application of Lemma 3.4.2 we obtain for any d s.t. $|G_E^d| \geq \sigma_E$ the following inequality.

$$LB(G_E^d, G_I^{u_1}, G_I^{u_2}) \leq \text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}) \tag{3.6}$$

This stems from the fact that $LB(G_E^d, G_I^{u_1}, G_I^{u_2})$ takes the sum of the lowest σ_E quantities constituting the numerator of $\text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2})$ and divides them by the sum of the greatest σ_E quantities forming the denominator of $\text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2})$.

Moreover, we have that LB is monotonic w.r.t. $\sqsubseteq$ of $\mathcal{D}_E$. i.e.

$$c \sqsubseteq d \Rightarrow LB(G_E^c, G_I^{u_1}, G_I^{u_2}) \leq LB(G_E^d, G_I^{u_1}, G_I^{u_2}) \tag{3.7}$$

This results from $c \sqsubseteq d \Rightarrow G_E^d \subseteq G_E^c$. Hence, if we reorder values of G_E^c and G_E^d where $G_E^c = \{e_1^c, ..., e_{|G_E^c|}^c\}$ and $G_E^d = \{e_1^d, ..., e_{|G_E^d|}^d\}$ as such:

$$\begin{cases} w_{e_1^c}.\alpha(e_1^c) \leq w_{e_2^c}.\alpha(e_2^c) \leq ... \leq w_{e_{\sigma_E}^c}.\alpha(e_{\sigma_E}^c) \leq ... \leq w_{e_{|E_c|}^c}.\alpha(e_{|G_E^c|}^c) \\ w_{e_1^d}.\alpha(e_1^d) \leq w_{e_2^d}.\alpha(e_2^d) \leq ... \leq w_{e_{\sigma_E}^d}.\alpha(e_{\sigma_E}^d) \leq ... \leq w_{e_{|G_E^d|}^d}.\alpha(e_{|G_E^d|}^d) \end{cases}$$

Given that $G_E^d \subseteq G_E^c$, it is clear that: $\forall i \leq \sigma_E \mid w_{e_i^c}.\alpha(e_i^c) \leq w_{e_i^d}.\alpha(e_i^d)$. Having that $m(G_E^c, \sigma_E) = \{e_1^c, ..., e_{\sigma_E}^c\}$ and $m(G_E^d, \sigma_E) = \{e_1^d, ..., e_{\sigma_E}^d\}$, it follows that:

$$\sum_{e \in m(G_E^c, \sigma_E)} w_e \times \alpha(e) \leq \sum_{e \in m(G_E^d, \sigma_E)} w_e \times \alpha(e) \tag{3.8}$$

Similarly, if we reorder entities e in descending order w.r.t the weights w_e we have $\forall j \leq \sigma_E \mid w_{e^d_j} \leq w_{e^c_j}$. Resulting in:

$$\sum_{e \in Mw(G^c_E, \sigma_E)} w_e \geq \sum_{e \in Mw(G^d_E, \sigma_E)} w_e \tag{3.9}$$

Hence, from (3.8) and (3.9) we have $\mathrm{LB}(G^c_E, G^{u_1}_I, G^{u_2}_I) \leq \mathrm{LB}(G^d_E, G^{u_1}_I, G^{u_2}_I)$ and provided that $LB(G^d_E, G^{u_1}_I, G^{u_2}_I) \leq \mathrm{IAS}(G^d_E, G^{u_1}_I, G^{u_2}_I)$ from (3.6), we have: $\forall c, d \in \mathscr{D}_E.\ c \sqsubseteq d \Rightarrow \mathrm{LB}(G^c_E, G^{u_1}_I, G^{u_2}_I) \leq \mathrm{IAS}(G^d_E, G^{u_1}_I, G^{u_2}_I)$

■

Proposition 3.4.3 — Upper bound UB for IAS. we define function UB as

$$\mathrm{UB}(G^c_E, G^{u_1}_I, G^{u_2}_I) = \frac{\sum_{e \in \mathrm{M}(G^c_E, \sigma_E)} w_e \times \alpha(e)}{\sum_{e \in \mathrm{mw}(G^c_E, \sigma_E)} w_e}$$

For any context c (corresponding to a subgroup G^c_E), UB provides an upper bound for IAS w.r.t. contexts. i.e.

$$\forall c, d \in \mathscr{D}_E.\ c \sqsubseteq d \Rightarrow \mathrm{IAS}(G^d_E, G^{u_1}_I, G^{u_2}_I) \leq \mathrm{UB}(G^c_E, G^{u_1}_I, G^{u_2}_I)$$

Proof (proposition 3.4.3). This proof is similar to the proof of *Proposition 3.4.1*. For the sake of brevity, we give a *proof sketch*. By a direct application of Lemma 3.4.2, it is clear that for any d s.t. $|G^d_E| \geq \sigma_E$.

$$\mathrm{IAS}(G^d_E, G^{u_1}_I, G^{u_2}_I) \leq \mathrm{UB}(G^d_E, G^{u_1}_I, G^{u_2}_I) \tag{3.10}$$

We have that UB is anti-monotonic w.r.t. $\sqsubseteq$ of $\mathscr{D}_E$. i.e.

$$c \sqsubseteq d \Rightarrow \mathrm{UB}(G^c_E, G^{u_1}_I, G^{u_2}_I) \geq \mathrm{UB}(G^d_E, G^{u_1}_I, G^{u_2}_I) \tag{3.11}$$

This results from $c \sqsubseteq d \Rightarrow G^d_E \subseteq G^c_E$. Thus,

$$\sum_{e \in M(G^c_E, \sigma_E)} w_e \times \alpha(e) \geq \sum_{e \in M(G^d_E, \sigma_E)} w_e \times \alpha(e) \text{ and } \sum_{e \in mw(G^c_E, \sigma_E)} w_e \leq \sum_{e \in mw(G^d_E, \sigma_E)} w_e$$

Hence, given (3.10) and (3.11) it follows that:

$$\forall c, d \in \mathscr{D}_E.\ c \sqsubseteq d \Rightarrow \mathrm{IAS}(G^d_E, G^{u_1}_I, G^{u_2}_I) \leq \mathrm{UB}(G^c_E, G^{u_1}_I, G^{u_2}_I)$$

■

Now that both the lower bound and the upper bound of IAS are defined w.r.t. contexts, we define the optimistic estimates corresponding to φ_{consent} and φ_{dissent}.

3.4.3.2 Optimistic Estimates for Quality Measures

Proposition 3.4.4 — Optimistic estimate for φ_{consent} and φ_{dissent}. $\text{oe}_{\text{consent}}$ (resp. $\text{oe}_{\text{dissent}}$) is an **optimistic estimate** for φ_{consent} (resp. φ_{dissent}) with:

$$\text{oe}_{\text{consent}}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \max(\text{UB}(G_E^c, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}), 0)$$

$$\text{oe}_{\text{dissent}}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \max(\text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) - \text{LB}(G_E^c, G_I^{u_1}, G_I^{u_2}), 0)$$

Proof (proposition 3.4.4). given $c, d \in \mathscr{D}_E$ such that $c \sqsubseteq d$, using *proposition 3.4.1* we have:

$$\text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}) \leq \text{UB}(G_E^c, G_I^{u_1}, G_I^{u_2})$$
$$\text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) \leq \text{UB}(G_E^c, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2})$$

Since $\varphi_{\text{consent}}(G_E^d, G_I^{u_1}, G_I^{u_2}) = \max(\text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}), 0)$ thus
$\varphi_{\text{consent}}(G_E^d, G_I^{u_1}, G_I^{u_2}) \leq \text{oe}_{\text{consent}}(G_E^c, G_I^{u_1}, G_I^{u_2})$

Similarly we have:

$$\text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}) \geq \text{LB}(G_E^c, G_I^{u_1}, G_I^{u_2})$$
$$\text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}) \leq \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) - \text{LB}(G_E^c, G_I^{u_1}, G_I^{u_2})$$

Since $\varphi_{\text{dissent}}(G_E^d, G_I^{u_1}, G_I^{u_2}) = \max(\text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E^d, G_I^{u_1}, G_I^{u_2}), 0)$ we get:
$\varphi_{\text{dissent}}(G_E^d, G_I^{u_1}, G_I^{u_2}) \leq oe_{dissent}(G_E^c, G_I^{u_1}, G_I^{u_2})$ ■

The two defined optimistic estimates tight if the IAS measure is a simple average. i.e. all weights are equal to 1.

Proposition 3.4.5 If $\forall(\{e\}, G_I^{u_1}, G_I^{u_2}) \subseteq G_E \times G_I \times G_I \ : \ w(e, G_I^{u_1}, G_I^{u_2}) = 1$, $\text{oe}_{\text{consent}}$ (resp. $\text{oe}_{\text{dissent}}$) is a **tight optimistic estimate** for φ_{consent} (resp. φ_{dissent}).

Proof (proposition 3.4.5). Given that $\forall(e, G_I^{u_1}, G_I^{u_2}) \in E \times 2^I \times 2^I \ : \ w(e, G_I^{u_1}, G_I^{u_2}) = 1$, we have for any $c \in \mathscr{D}_E$ s.t. $|G_E^c| \geq \sigma_E$.

$$\text{IAS}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \frac{\sum_{e \in G_E^c} \alpha(e)}{|G_E^c|} \quad \textit{and} \quad \text{UB}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \frac{\sum_{e \in M(G_E^c, \sigma_E)} \alpha(e)}{\sigma_E}$$

It follows from the fact that $M(G_E^c, \sigma_E) \subseteq G_E^c$:

$$\begin{aligned}
\exists S \subseteq G_E^c \ : \text{UB}(G_E^c, G_I^{u_1}, G_I^{u_2}) &= \text{IAS}(S, G_I^{u_1}, G_I^{u_2}) \\
\text{UB}(G_E^c, G_I^{u_1}, G_I^{u_2}) - \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) &= \\
\text{IAS}(S, G_I^{u_1}, G_I^{u_2}) &- \text{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) \\
\text{oe}_{\text{consent}}(G_E^c, G_I^{u_1}, G_I^{u_2}) &= \varphi_{\text{consent}}(S, G_I^{u_1}, G_I^{u_2})
\end{aligned}$$

The subset S being for example the set $M(G_E^c, \sigma_E)$ itself. The same reasoning applies when considering $\text{oe}_{\text{dissent}}$. In this case we consider the lower bound LB. We have:

$$\text{LB}(G_E^c, G_I^{u_1}, G_I^{u_2}) = \frac{\sum_{e \in m(G_E^c, \sigma_E)} \alpha(e)}{\sigma_E}$$

Given that $m(G_E^c, \sigma_E) \subseteq E$, we have:

$$\begin{aligned}
\exists S \subseteq G_E^c \ : \mathrm{LB}(G_E^c, G_I^{u_1}, G_I^{u_2}) &= \mathrm{IAS}(S, G_I^{u_1}, G_I^{u_2}) \\
\mathrm{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) - \mathrm{LB}(G_E^c, G_I^{u_1}, G_I^{u_2}) &= \\
&\mathrm{IAS}(G_E, G_I^{u_1}, G_I^{u_2}) - \mathrm{IAS}(S, G_I^{u_1}, G_I^{u_2}) \\
\mathrm{oe}_{\mathrm{dissent}}(G_E^c, G_I^{u_1}, G_I^{u_2}) &= \varphi_{\mathrm{dissent}}(S, G_I^{u_1}, G_I^{u_2})
\end{aligned}$$

This proves that, if IAS is a simple mean, for any $c \in \mathscr{D}_E$ s.t. $|G_E^c| \geq \sigma_E$:

$$\exists S, S' \subseteq G_E^c \ : \ \begin{cases} \mathrm{oe}_{\mathrm{consent}}(G_E^c, G_I^{u_1}, G_I^{u_2}) & = \varphi_{\mathrm{consent}}(S, G_I^{u_1}, G_I^{u_2}) \\ {}_{dissent}(G_E^c, G_I^{u_1}, G_I^{u_2}) & = \varphi_{\mathrm{dissent}}(S', G_I^{u_1}, G_I^{u_2}) \end{cases}$$

Hence $\mathrm{oe}_{\mathrm{consent}}$ and $oe_{dissent}$ are tight optimistic estimates for respectively $\varphi_{\mathrm{consent}}$ and $\varphi_{\mathrm{dissent}}$ if the underlying IAS is a simple average. ■

3.4.4 Algorithm DEBuNk

DEBuNk is a Branch-and-Bound algorithm which returns the complete set of patterns as specified in the problem definition (Section 3.2). To this end, it takes benefit from the defined closure operator and optimistic estimates. DEBuNk follows the same line of reasoning of Algorithm B&B4SDEMM (cf. Algorithm 2). Relying on algorithm EnumCC (cf. Algorithm 1), DEBuNk starts by generating the couples of confronted groups of individuals that are large enough w.r.t. σ_I (lines 2-3). Then it computes the usual agreement observed between these two groups of individuals when considering all entities in G_E (line 4). Next, the context search space is explored to generate valid contexts c (line 5). Subsequently, the optimistic estimate oe is evaluated and the context sub search space is pruned if oe is lower than σ_φ (lines 7-8). Otherwise, the contextual inter-group agreement is computed and the quality measure is calculated (lines 9-10). If the pattern quality exceeds σ_φ then two scenarios are possible. Either the current pattern set P already contains a more general pattern, or it does not. In the former case, the pattern is discarded. In the latter, the new generated pattern is added to pattern set P while removing all previous generated patterns that are more specific than p w.r.t. extents (lines 11-14). Since the current pattern quality exceeds the threshold and all the remaining patterns in the current context sub search space are more specific than the current one, the sub search space is pruned (line 15). Eventually, if the quality measure is symmetric w.r.t. u_1 and u_2 (i.e. $\forall u_1, u_2 \in \mathscr{D}_I^2 \mid \varphi(c, u_1, u_2) = \varphi(c, u_2, u_1)$) there is no need to evaluate both qualities. As a consequence, it is possible to prune the sub search space of the couple descriptions (u_1, u_2) whenever $u_1 = u_2$ (lines 16-17).

DEBuNk and DSC algorithm (Belfodil et al., 2017a) differs on several levels. First, DEBuNk overcomes the limitations of lack of diversity of results provided by DSC which was designed to discover the top-k solutions. The present algorithm discards all patterns for which a generalization is already a solution. Second, DEBuNk handles a wider range of bounded quality measures (i.e. weighted mean IAS), in contrast to DSC algorithm which handles only a subset of these measures. Finally, DSC requires the prior definition of an aggregation level which makes it difficult to use and interpret. DEBuNk overcomes this issue by reducing the number of input parameters and integrating relevancy check between patterns. Hence, it requires less effort from the end-user both in terms of setting the parameters, and in terms of interpreting the quality of the resulting patterns.

Algorithm 3: DEBuNk($\mathscr{B}, \sigma_E, \sigma_I, \varphi, \sigma_\varphi$)

Inputs : $\mathscr{B} = \langle G_I, G_E, O, o\rangle$ a Behavioral dataset;
σ_E (resp. σ_I) minimum support threshold of a context (resp. group);
φ the quality measure; σ_φ quality threshold on the quality.
Output: P the set of exceptional inter-group agreement patterns.

$P \leftarrow \{\}$
foreach $(u_1, G_I^{u_1}, \mathtt{cont}_{u_1}) \in \text{EnumCC}(G_I, *, \sigma_I, 0, \text{True})$ **do**
    foreach $(u_2, G_I^{u_2}, \mathtt{cont}_{u_2}) \in \text{EnumCC}(G_I, *, \sigma_I, 0, \text{True})$ **do**
        $\text{IAS}_{\text{ref}} \leftarrow \text{IAS}(*, u_1, u_2)$
        foreach $(c, G_E^c, \mathtt{cont}_c) \in \text{EnumCC}(G_E, *, \sigma_E, 0, \text{True})$ **do**
            if $\text{oe}_\varphi(c, u_1, u_2) < \sigma_\varphi$ **then**
                $\mathtt{cont}_c \leftarrow \text{False}$; // Prune the sub-search space under c
            else
                $\text{IAS}_{\text{ref}} \leftarrow \text{IAS}(c, u_1, u_2)$
                $\text{quality} \leftarrow \varphi(c, u_1, u_2)$; // computed using IAS_{ref} and $\text{IAS}_{context}$
                if quality $\geq \sigma_\varphi$ **then**
                    $p_{\text{new}} \leftarrow (c, u_1, u_2)$
                    if $\nexists p_{\text{old}} \in P \mid \text{ext}(p_{\text{new}}) \subseteq \text{ext}(p_{\text{old}})$ **then**
                        $P \leftarrow (P \cup p_{\text{new}}) \setminus \{p_{\text{old}} \in P \mid \text{ext}(p_{\text{old}}) \subseteq \text{ext}(p_{\text{new}})\}$
                    $\mathtt{cont}_c \leftarrow \text{False}$; // Prune the sub search space
        if φ *is symmetric* ***and*** $u_1 = u_2$ **then**
            break ; // Prune the sub search space
return P

3.5 Sampling Inter-Group Agreement Patterns

The discovery of the complete set of interesting patterns as ensured by DEBuNk, has two disadvantages that limit the use of such methods in practice. It is time consuming to compute the complete set of solutions. Furthermore, this set can be absolutely huge and non-manageable for a human expert. To overcome this limitation, many approaches that can effectively sample the pattern space for interesting patterns have been proposed for a decade. These methods address some frequent or discriminant itemset mining tasks (Boley et al., 2011; Giacometti and Soulet, 2016; Li and Zaki, 2016; Moens and Goethals, 2013) offering some theoretical guarantees on the sampling quality or more generic ones (Al Hasan and Zaki, 2009; Boley, Gärtner, and Grosskreutz, 2010; Dzyuba, Leeuwen, and De Raedt, 2017). In (Dzyuba, Leeuwen, and De Raedt, 2017), the authors define the problem of sampling pattern sets and propose a method based on a SAT solver sampling solution. However, this approach only supports pattern languages that can be compactly represented by binary variables such as itemsets. It requires the discretization of numerical attributes. Authors in (Al Hasan and Zaki, 2009; Boley, Gärtner, and Grosskreutz, 2010) use a MCMC (Monte-Carlo Markov-Chain) based algorithm to achieve sampling with guarantees according to a desired probability distribution. Despite the generic nature and the interesting guarantees that MCMC algorithms provide, it requires a number of steps that grows exponentially in the input size to generate a single pattern (Boley, Gärtner, and Grosskreutz, 2010). This may prevent the user to obtain instant results. The problem we are interested in has several specificities. First, the search space involves attributes of different types (i.e., numerical, symbolical, HMT attributes) which prevents us to use sampling techniques based on itemset language. Second, the quality measure is not considered in the state-of-the-art methods that mainly support frequency and discriminative measures (Boley, Moens, and Gärtner, 2012; Boley et al., 2011). Finally, the method proposed in (Moens and Boley, 2014) for EMM is not suited to our problem since we have to simultaneously consider both description space and target space. To address this concerns, we devise Algorithm Quick-DEBuNk handles the specificity of the problem by yielding approximate solutions that improve over time. It combines an exploration step (Step 1) and an exploitation step while taking profit of the quality measures properties (Step 2). These two steps are summarized in Fig. 3.4.

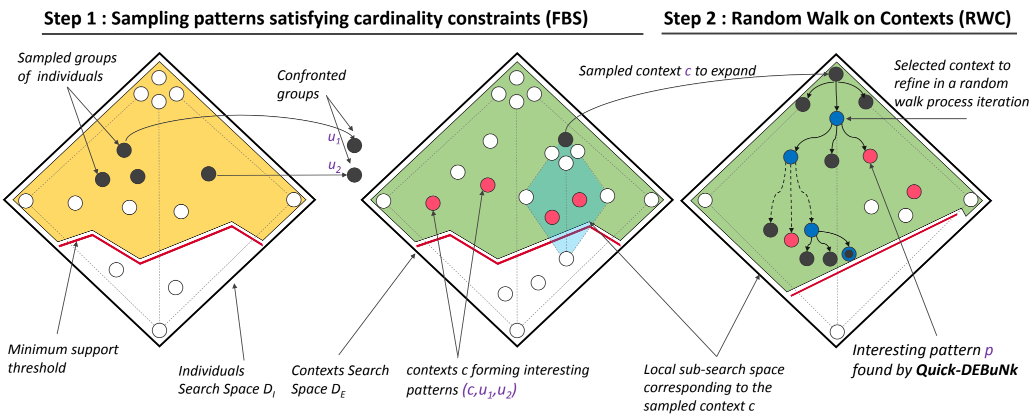


Figure 3.4: Quick-DEBuNk approach in a nutshell

Frequency-Based Sampling (Step 1). An inter-group agreement pattern $p \in \mathscr{P}$ is drawn with a probability proportional to the size of its extent (i.e. $|\text{ext}(p = (c, u_1, u_2))| = |G_E^c| \times |G_I^{u_1}| \times |G_I^{u_2}|$). The key insight is to provide more chance to patterns supported by larger groups and contexts which are less likely to be discarded by more general ones generated by future iterations. This technique is inspired by the direct frequency-based sampling algorithm proposed in (Boley et al., 2011) which considers only Boolean attributed datasets. Here, this method is extended to handle more complex data with HMT, categorical and numerical attributes.

Random Walk on Contexts (step 2). Starting from a context obtained in step 1, a random walk traverses the search tree corresponding to the contexts description space $\mathscr{D}_E$. We introduce some bias to fully take advantage of the devised quality measures and the optimistic estimates, this being done to reward high quality patterns by giving them more chance to be sampled by the algorithm.

3.5.1 Frequency-Based Sampling (Step 1)

To sample patterns of the form $p = (c, u_1, u_2)$, we aim to draw description c, respectively u_1 and u_2, from description space $\mathscr{D}_E$, respectively $\mathscr{D}_I$, with a probability proportional to their respective support size. To this end, we devise the algorithm FBS (Frequency-Based Sampling).

Algorithm 4: FBS(G)

Input: G a collection of records which may be G_E or G_I

Output: a description d from $\mathscr{D}$ with $\mathbb{P}(d) = \frac{|G^d|}{\sum_{d' \in \mathscr{D}} |G^{d'}|}$

```
draw g ∽ w_g from G ;                 // with w_g = |↓δ(g)|
draw d ∽ uniform(↓δ(g))
return d
```

In the following, for any $d \in \mathscr{D}$, $\downarrow d$ denotes the set of all descriptions subsuming d, i.e: $\downarrow d = \{d' \in \mathscr{D} \,:\, d' \sqsubseteq d\}$. Since the cartesian product[8] $\mathscr{D} = \mathscr{D}^1 \times \mathscr{D}^2 \times \ldots \times \mathscr{D}^m$, it follows that: $\downarrow d = \downarrow(r_1, r_2, \ldots, r_m) = \downarrow r_1 \times \downarrow r_2 \times \ldots \times \downarrow r_m$, where $\downarrow r_j$ is the set of conditions less specific than (implied by) r_j in the conditions space $\mathscr{D}^j$.

FBS generates a description d with a probability proportional to its frequency $\mathbb{P}(d) = \frac{|G^d|)}{\sum_{d' \in \mathscr{D}} |G^{d'}|}$ (formally defined in proposition 3.5.1). To this end, FBS performs two steps as depicted in Algorithm 4.

FBS starts by drawing a record g from G (line 1) with a probability proportional to the number of descriptions $d \in \mathscr{D}$ covering g (i.e: $|\downarrow \delta(g)|$). To enable this, each record $g \in G$ is weighted by $w_g = |\downarrow \delta(g)|$. For now, we use d^g to refer to $\delta(g)$. Knowing $d^g = (r_1^g, \ldots, r_m^g)$, the weight $w_g = |\downarrow d^g| = \prod_{j \in [1,m]} |\downarrow r_j^g|$ is the product of the numbers of restrictions subsuming each r_j^g. The size of $|\downarrow r_j^g|$ depends on the type of the related attribute a_j:

- *categorical attribute*: given that r_j^g corresponds to a value $v \in dom(a_j)$, we have $\downarrow r_j^g = \{*, v\}$ thus $|\downarrow r_j^g| = 2$.
- *numerical attribute*: given that r_j^g corresponds to an interval $[v,w]$ *with* $v,w \in dom(a_j)$, we have $\downarrow r_j^g$ is equal to the number of intervals having a left-bound $\underline{v} \leq v$ and a right-bound $\overline{w} \geq w$. More formally, $\downarrow r_j^g = \{[\underline{v}, \overline{w}] \mid \underline{v} \leq v \wedge \overline{w} \geq w\}$. Hence, the cardinal of this set is $|\downarrow r_j^g| = |\{\underline{v} \in dom(a_j) \,:\, \underline{v} \leq v\}| \times |\{\overline{w} \in dom(a_j) \,:\, \overline{w} \geq w\}|$.
- *HMT attribute*: given that r_j^g corresponds to a set of tags $\{t_1, t_2, \ldots t_l\} \in dom(a_j)$, with $t_k \in T$ and T a tree, the condition r_j^g can be conceptualized as a rooted subtree of T where the leaves are $\{t_1, t_2, \ldots t_l\}$. Thus, $\downarrow r_j^g$ represents the set of all possible rooted subtrees of r_j^g. The latter cardinality can be computed recursively by starting from the root $*$ using $\mathrm{nbs}(\mathrm{tree}, \mathrm{root}) = \prod_1^k (\mathrm{nbs}(\mathrm{tree}_i, \mathrm{neighbor}_i) + 1)$ where $\mathrm{neighbor}_i$ returns the child tags of a given root and tree_i the subtree rooted on $\mathrm{neighbor}_i$.

Given g the record returned from the first step and its corresponding description $d^g = \delta(g) = \langle r_1^g, \ldots, r_m^g \rangle$, FBS uniformly generates a description d from the set of descriptions covering g, that is $\downarrow d^g$. This can be done by uniformly drawing conditions r_j from $\downarrow r_j^g$, hence returning a description $d = \langle r_1, r_2, \ldots, r_m \rangle$. This comes from the fact that $\forall j \in [1,m] : \mathbb{P}(r_j) = \frac{1}{|\downarrow r_j^g|}$:

$$\mathbb{P}(d|g) = \prod_{j \in [1,m]} \mathbb{P}(r_j) = \frac{1}{\prod_{j \in [1,m]} |\downarrow r_j^g|} = \frac{1}{|\prod_{j \in [1,m]} \downarrow r_j^g|} = \frac{1}{|\downarrow d^g|} \,.$$

[8] Cartesian product of the m lattices related to attributes conditions spaces forms a lattice(Roman, 2008)

We now define the method used to uniformly draw a condition corresponding to an attribute a_j, according to its type:

- *categorical attribute*: given that $\downarrow r_j^g = \{*, v\}$ with $v \in dom(a_j)$, it is sufficient to uniformly draw an element r_j from $\{*, v\}$.
- *numerical attribute*: given that $\downarrow r_j^g = \{[\underline{v}, \overline{w}] \mid \underline{v} \leq v \wedge \overline{w} \geq w\}$, to generate an interval $[sv, sw]$ from $\downarrow r_j^g$ uniformly, one needs to uniformly draw a left-bound sv from the set $\{\underline{v} \in dom(a_j) \ : \ \underline{v} \leq v\}$ and a right-bound sw from the set $\{\overline{w} \in dom(a_j) \ : \ \overline{w} \geq w\}$.
- *HMT attribute*: given that $\downarrow r_j^g$ represents the set of rooted subtrees of r_j^g, we have to uniformly draw such rooted subtrees. A first way is to generate all the possible rooted subtrees and then uniformly draw an element from the resulting set. This does not scale. Hence we devised another method, relying on a stochastic process using the aforementioned function *nbs* (which counts the number of subtrees rooted on some given node). The algorithm takes the root $*$ as a starting tree. Next, the resulting subtree is augmented by a child c of $*$ with a chance equal to the number subtrees of $\downarrow r_j^g$ containing c. That is $\frac{\text{nbs}(r_j^g,*)-\text{nbs}(r_j^g-\{c\},*)}{\text{nbs}(r_j^g,*)}$. Recursively, the algorithm continues from a drawn candidate child c.

Proposition 3.5.1 A description $d \in \mathscr{D}$ has a probability of being generated by *FBS* equal to $\mathbb{P}(d) = \frac{G^d}{\sum_{d' \in \mathscr{D}} |G^{d'}|}$.

Before giving the proof of the proposition 3.5.1 we present the following lemma.

Lemma 3.5.2 The sums of the number of all descriptions covering each record in G is equal to the sum of the supports of all descriptions in $\mathscr{D}$. That is:

$$\sum_{g \in G} |\downarrow \delta(g)| = \sum_{d \in \mathscr{D}} |G^d|$$

Proof (lemma 3.5.2). For $g \in G$, we have $\downarrow \delta(g) = \{d \in \mathscr{D} \ : \ d \sqsubseteq \delta(g)\}$ and for $d \in \mathscr{D}$, we have $G^d = \{g \in G \mid d \sqsubseteq \delta(g)\}$. Let us define the indicator function on $\mathscr{D} \times G$:

$$\mathbb{1}_{\sqsubseteq}(d,g) = \begin{cases} 1 & \text{if } d \sqsubseteq \delta(g) \\ 0 & \text{else} \end{cases}$$

Hence, we have $|\downarrow \delta(g)| = \sum_{d \in \mathscr{D}} \mathbb{1}_{\sqsubseteq}(d,g)$ and $|G^d| = \sum_{g \in G} \mathbb{1}_{\sqsubseteq}(d,g)$ thus:

$$\sum_{g \in G} |\downarrow \delta(g)| = \sum_{g \in G} \sum_{d \in \mathscr{D}} \mathbb{1}_{\sqsubseteq}(d,g) = \sum_{d \in \mathscr{D}} \sum_{g \in G} \mathbb{1}_{\sqsubseteq}(d,g) = \sum_{d \in \mathscr{D}} |G^d|$$

■

Proof (proposition 3.5.1). We denote by **gs** the random record drawn in line 1 and by **ds** the random description drawn in line 2 of FBS.

$$\mathbb{P}(\mathbf{ds}=d) = \sum_{g\in G} \mathbb{P}((\mathbf{gs}=g)(\mathbf{ds}=d|g))$$
$$= \sum_{g\in G^d} \frac{1}{|{\downarrow}\delta(g)|} \times \underbrace{\frac{|{\downarrow}\delta(g)|}{\sum_{i\in G}|{\downarrow}\delta(i)|}}_{\text{weight } w_g \text{ normalized}} = \frac{|G^d|}{\sum_{g\in G}|{\downarrow}\delta(g)|}$$

It follows that from *Lemma 3.5.2* that $\mathbb{P}(\mathbf{ds}=d) = \dfrac{|G^d|}{\sum_{d'\in\mathscr{D}}|G^{d'}|}$

■

FBS algorithm makes it possible to generate valid patterns $p=(c,u_1,u_2)$ from the pattern space $\mathscr{P} = \mathscr{D}_E \times \mathscr{D}_I \times \mathscr{D}_I$. This is achieved in the first step of Quick-DEBuNk (lines 3-6 in *Algorithm 6*) by sampling two group descriptions u_1, u_2 from $\mathscr{D}_I$ and a context c from $\mathscr{D}_E$ followed by assessing if the three descriptions satisfy the cardinalities constraints $\mathscr{C}$ (min. support thresholds).

> **Proposition 3.5.3** Given the cardinality constraints $\mathscr{C}$, every valid pattern p is reachable by the first step of *Quick-DEBuNk*. i.e. $\forall p \in \mathscr{P} : p$ satistifies $\mathscr{C} \Rightarrow \mathbb{P}(p) > 0$

Proof (proposition 3.5.3). Given *Proposition 3.5.1*, it is clear that $\forall p \in \mathscr{P} : p=(c,u_1,u_2)$ *satisfies* $\mathscr{C} \Rightarrow \mathbb{P}(p) = \frac{|\text{ext}(p)|}{Z} > 0$. with $|\text{ext}(p)| = |G_E^c| \times |G_I^{u_1}| \times |G_I^{u_2}|$ and $Z = \sum_{p'\in\mathscr{P}} |\text{ext}(p')|$ a normalizing factor. ■

Step 1 of Quick-DEBuNk does not favor the sampling of high quality patterns as it does not involve an exploitation phase. The random walk process on contexts used in Step 2 enables a smarter traversal of the search space while taking into account the devised quality measures and optimistic estimates.

3.5.2 RWC - Random Walk on Contexts (Step 2)

RWC (Algorithm 5) enumerates contexts of the search space corresponding to $\mathscr{D}_E$ while considering closure and optimistic estimates. *RWC* takes as input two confronted groups of individuals described by u_1,u_2 for which it looks for relevant contexts (i.e., to form an inter-group agreement pattern) following a random walk process starting from a context c. Mainly, RWC has two steps that are recursively executed until a terminal node is reached. RWC starts by generating all neighbors d of the current context c (line 2). Next, RWC assesses whether the size of the corresponding support G_E^c and the optimistic estimates respectively exceed the support threshold σ_E and the quality threshold σ_φ (line 3). If appropriate, the closed description d is computed (line 4). The algorithm proceeds by evaluating the quality of pattern (line 5). If the quality exceeds the threshold σ_φ, the pattern is valid and is hence yielded (line 6). Otherwise, the pattern is added to NtE (*Neighbors to be Explored*) (line 8) as its related sub search space may contain interesting patterns (*i.e* $\text{oe}_\varphi(d,u_1,u_2) \geq \sigma_\varphi$). The second step of RWC consists in selecting a neighbor from NtE to be explored with a probability proportional to its quality (lines $10-12$). This process is recursively repeated until a terminal node is reached (i.e. $\text{NtE} = \emptyset$).

Algorithm 5: RWC($\mathscr{B}, c, u_1, u_2, \sigma_E, \varphi, \sigma_\varphi$)

Inputs : $\mathscr{B} = \langle G_I, G_E, O, o \rangle$ a Behavioral dataset;
c the current context;
(u_1, u_2) couple of confronted group descriptions of individuals;
σ_E threshold on support;
φ the quality measure;
σ_φ quality threshold.

Output: yield valid patterns (c, u_1, u_2)

NtE $\leftarrow \{\}$
foreach $d \in \eta(c)$ **do**
  if $|G_E^d| \geq \sigma_E$ *and* $\mathrm{oe}_\varphi(d, u_1, u_2) \geq \sigma_\varphi$ **then**
    `closure_d` $\leftarrow \delta(G_E^d)$
    if $\varphi(d, u_1, u_2) \geq \sigma_\varphi$ **then**
      yield `closure_d`
    else
      NtE $\leftarrow$ NtE $\cup \{d\}$
if NtE $\neq \emptyset$ **then**
  draw next $\backsim \varphi(\text{next}, u_1, u_2)$ ***from*** NtE
  foreach $c_{\text{next}} \in \mathrm{RWC}(\langle G_I, G_E, O, o \rangle, \text{next}, \sigma_E, \varphi, \sigma_\varphi, u_1, u_2)$ **do**
    yield c_{next}

3.5.3 Algorithm Quick-DEBuNk

Quick-DEBuNk (Algorithm 6) samples patterns from the full description space $\mathscr{D}_E \times \mathscr{D}_I \times \mathscr{D}_I$. It is based on FBS and RWC. It takes as input the same parameters as DEBuNk in addition to a *timebudget*. It starts by generating a couple of closed group descriptions of individuals u_1, u_2 that fulfill the support constraint (lines $3-5$) using FBS. Next, Quick-DEBuNk generates a context while only considering entities having a quality greater than the threshold σ_φ (line 6). The reason behind considering only $G_E^{\geq \sigma_\varphi}$ is clear: we have $\forall p \in \mathscr{P}$ *p satisfies* $\mathscr{C}$ and $\varphi(p) \geq \sigma_\varphi \Rightarrow \exists e \in G_E^c : \varphi(\{e\}, G_I^{u_1}, G_I^{u_2}) \geq \sigma_\varphi$ (since the quality measure is a weighted mean). If the context fulfills the cardinality constraint and its evaluated optimistic estimate is greater than the quality threshold (line 7), the algorithm then evaluates the quality of the sampled pattern (line 8). If this quality is greater than the threshold σ_φ, the pattern is appended to the resulting pattern set if and only if it is not more specific of an already found pattern w.r.t. extents (lines $9-11$). Otherwise, a random walk is launched starting from context c (line 13). This is done by relying on RWC. The algorithm continues by updating the resulting pattern set by each pattern yielded by RWC, as long as there is no more general pattern in the current pattern set P (lines $14-16$). Otherwise, RWC is interrupted (line 18). The process is repeated as long as the time budget allows.

Algorithm 6: Quick-DEBuNk($\mathscr{B}, \sigma_E, \sigma_I, \varphi, \sigma_\varphi$, *timebudget*)

Inputs : $\mathscr{B} = \langle G_I, G_E, O, o\rangle$ a Behavioral dataset;
σ_E (resp. σ_I) minimum support threshold of a context (resp. group);
φ the quality measure;
σ_φ threshold on the quality;
timebudget the maximum amount of time given to the algorithm.

Output: P the set of local relevant inter-group agreement patterns

$P \leftarrow \{\}$
while *executionTime* $<$ *timebudget* **do**
  $u_1 \leftarrow clo(\mathrm{FBS}(G_I))$
  $u_2 \leftarrow clo(\mathrm{FBS}(G_I))$
  if $|G_I^{u_1}| \geq \sigma_I \wedge |G_I^{u_2}| \geq \sigma_I$ **then**
    $c \leftarrow clo(\mathrm{FBS}(G_E^{\geq \sigma_\varphi}))$; // $G_E^{\geq \sigma_\varphi} = \{e \in G_E \mid \varphi(\{e\}, I_{u_1}, I_{u_2}) \geq \sigma_\varphi\}$
    if $|G_E^c| \geq \sigma_E \wedge \mathrm{oe}_\varphi(c, u_1, u_2) \geq \sigma_\varphi$ **then**
      if $\varphi(c, u_1, u_2) \geq \sigma_\varphi$ **then**
        $p_{\text{new}} \leftarrow (c, u_1, u_2)$
        if $\nexists p_{\text{old}} \in P \mid \mathrm{ext}(p_{\text{new}}) \subseteq \mathrm{ext}(p_{\text{old}})$ **then**
          $P \leftarrow (P \cup p_{\text{old}}) \setminus \{p_{\text{old}} \in P \mid \mathrm{ext}(p_{\text{old}}) \subseteq \mathrm{ext}(p_{\text{new}})\}$
      else
        foreach $d \in \mathrm{RWC}(\langle G_I, G_E, O, o\rangle, c, u_1, u_2, \sigma_E, \varphi, \sigma_\varphi)$ **do**
          $p_{\text{new}} \leftarrow (d, u_1, u_2)$
          if $\nexists p_{\text{old}} \in P \mid \mathrm{ext}(p_{\text{new}}) \subseteq \mathrm{ext}(p_{\text{old}})$ **then**
            $P \leftarrow (P \cup p_{\text{new}}) \setminus \{p_{\text{old}} \in P \mid \mathrm{ext}(p_{\text{old}}) \subseteq \mathrm{ext}(p_{\text{new}})\}$
          else
            break
          if *executionTime* $\geq$ *timebudget* **then**
            return P
return P

3.6 Empirical Study

In this section, we report on both quantitative and qualitative experiments over the implemented algorithms. For reproducibility purposes, source code (in Python) and data are made available in a companion page[9].

3.6.1 Aims and Datasets

The experiments aim to answer the following questions:

- Do the algorithms provide interpretable patterns?
- How effective is DEBuNk compared to classical SD/EMM algorithms and DSC?
- Are the closure operators and optimistic estimate based pruning efficient?
- How effective is HMT closed description enumeration?
- Does DEBuNk scale w.r.t. different parameters?
- How effective is Quick-DEBuNk at sampling patterns?

Most of the experiments were carried out on four real-world behavioral datasets whose main characteristics are given in Table 3.2. Each dataset involves entities and individuals described by an HMT (H) attribute together with categorical(C) and numerical(N) ones.

EPD8[10] features voting information of the eighth European Parliament about the 958 members who were elected in 2014 or after. The dataset records $2.7M$ tuples indicating the outcome (For, Against, Abstain) of a member voting during one of the 4161 sessions. Each session is described by its themes (H), a voting date (N) and the organizing committee (C). Individuals are described by a national party (C), a political group (C), an age group (C), a country(C) and additional information about countries (date of accession to the European Union (N) and currency (C)). To analyze inter-group agreement patterns in this dataset, we consider $\text{IAS}_{\text{voting}}$ which is defined by using θ_{majority} and $\text{sim}_{\text{voting}}$.

Movielens[11] is a movie review dataset (Harper and Konstan, 2016) consisting of $100K$ ratings (ranging from 1 to 5) expressed by 943 users on 1681 movies. A movie is characterized by its genres (H) and a release date (N), individuals are described with age group (C), gender (C) and occupation (C). To handle the numerical outcomes, we use the measure $\text{IAS}_{\text{rating}}$ which relies on θ_{wavg} and $\text{sim}_{\text{rating}}$.

Yelp[12] is a social network dataset featuring individuals who rate (scores ranging from 1 to 5) places (stores, restaurants, clinics) characterized by some categories (H) and a state (C). The dataset originally contains 1M users. We preprocessed the dataset to constitute 18 groups of individuals based on the size of their friends network (C), their seniority (C) in the platform and their account type (e.g., elites or not) (C). We also use $\text{IAS}_{\text{rating}}$ measure in this dataset.

[9] https://github.com/Adnene93/DEBuNk
[10] http://parltrack.euwiki.org/, last accessed on 17 November 2017
[11] https://grouplens.org/datasets/movielens/100k/
[12] https://www.yelp.com/dataset/challenge, last accessed on 25 April 2017

Openmedic[13] is a drug consumption monitoring dataset that has been recently made available by *Ameli*[14]. This dataset inventories the number of drug boxes (described by their Anatomical Therapeutic Chemical (ATC) Classification[15](H)) yearly administered to individuals (from 2014 to 2016). Individuals are described with demographic information such as age (C), gender (C) and region (C). We further discuss an adapted IAS measure.

Comparing the size and the complexity of these datasets is difficult because of the heterogeneity of the attributes. In particular, the hierarchies of the HMT attributes are very different, as well as the range of the numerical ones. To enable a fair comparison, we employ a conceptual scaling (Ganter and Wille, 1999). The attributes are “projected" on a set of items by transforming each one to a Boolean representation. Each possible value of a categorical attribute provides a single item (e.g., *gender* gives *male*, *female* and *unknown*). The items corresponding to an HMT attribute are all the nodes of the tag tree (T). Each numerical attribute is transformed to an itemset via *interordinal scaling* (Kaytoue et al., 2011). To a given set of values $[v_1, v_2, ... v_n]$, we associate $2n$ items $\{\leq v_1, \leq v_2, ... \leq v_n, \geq v_1, \geq v_2, ... \geq v_n\}$. Table 3.2 illustrates this step, while Table 3.3 shows the obtained comparable characteristics.

		Entities	Individuals	Outcomes
EPD8	Size (Nb. records)	4161	958	2.7M
	attribute types	$1H+1N+1C$	$1N+5C$	
	size after scaling	$347+26+40=413$	$16+285=301$	
	avg scaling per record	20.44	14	
Movielens	Size (Nb. records)	1681	943	100K
	attribute types	$1H+1N$	$3C$	
	size after scaling	$20+144=164$	$4+2+21=27$	
	avg scaling per record	75.72	3	
Yelp	Size (Nb. records)	127 000	18	750K
	attribute types	$1H+1C$	$3C$	
	size after scaling	$1175+29=1204$	$3+2+3=8$	
	avg scaling per record	5.77	3	
Openmedic	Size (Nb. records)	12 221	78	500K
	attribute types	$1H$	$3C$	
	size after scaling	14 094	$2+13+3=18$	
	avg scaling per record	7	3	

Table 3.2: Behavioral datasets characteristics before and after scaling.

[13] http://open-data-assurance-maladie.ameli.fr/, last accessed on 16 November 2017

[14] ***Ameli*** - France National Health Insurance and Social Security Organization

[15] The Anatomical Therapeutic Chemical classification system classifies therapeutic drugs according to the organ or system on which they act and their chemical, pharmaco- logical and therapeutic properties – https://www.whocc.no/atc/structure_and_principles/.

Dataset	Transactions	Items	AverageSize
EPD8	1 727 032 585	1 015	34.48
Movielens	16 807 109	218	79.37
Yelp	5 860 354	1 220	9.00
Openmedic	28 512 418	14 130	10.00

Table 3.3: Characteristics of the datasets considered as plain collections of itemsets records - the plain collections correspond to $G_E \times G_I \times G_I$ while considering only pairable individuals (i.e., the cartsian product contains a record (e, i_1, i_2) only if both individuals expressed an outcome on the entitiy e, that is $o(i_1, e)$ and $o(i_2, e)$ are given).

3.6.2 Qualitative Study

First, we focus on illustrating patterns discovered by DEBuNk. To this end, we report three real world case studies: (i) In collaborative rating platforms (Yelp, Movielens), we study the affinities between groups of users with regard to their expressed ratings. (ii) In a voting system (European Parliament Dataset), we show how the voting behavior of parlementarians can provide interesting insights about the cohesion and the polarization between groups of parliamentarians in different contexts. Such information can be valuable for journalists and political analysts. (iii) We give example patterns reporting substantial differences in medicine consumption behavior between groups. Such results can be leveraged by epidemiologists to study comparative prevalence of sicknesses among subpopulations.

3.6.2.1 Study of Collaborative Rating Data

Table 3.4 describes some patterns returned by DEBuNk on the Movielens dataset when looking for contexts that lead to a disagreement between groups of individuals labeled by their professional occupations. The first pattern describes that, while students and Health professionals agree 74% of the time, they tend to disagree for Horror and comedy-like movies released between 1986 and 1994 (e.g., *Evil Dead II*, *Braindead*). Figure 3.5 illustrates the usual and the contextual rating distribution of each groups. We observe from this rating distributions, that the students like the movies highlighted by the pattern, whereas the healthcare professionals dislike them.

	(c, u_1, u_2)	$\lvert G_E^c\rvert$	$\lvert G_I^{u_1}\rvert$	$\lvert G_I^{u_2}\rvert$	$o(i,e)$	φ_{dissent}
1	Student vs. Healthcare in ['11 Horror', '5 Comedy'] [1986, 1994]	6	196	16	106	$0.42 = 0.74 - 0.33$
2	Student vs. Healthcare in ['5 Comedy'] [1991, 1991]	5	196	16	40	$0.41 = 0.74 - 0.33$
3	Healthcare vs. Artist in ['5 Comedy', '8 Drama'] [1987, 1993]	5	16	28	28	$0.42 = 0.73 - 0.3$

Table 3.4: Top-3 w.r.t. number of expressed outcomes ($o(i,e)$ *column*) of disagreement patterns discovered on Movielens ($|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 1$, $\sigma_E = 5$, $\sigma_I = 10$ and $\sigma_\varphi = 0.4$).

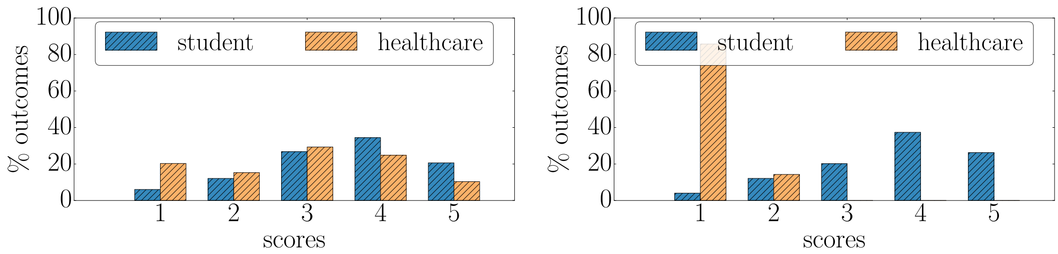


Figure 3.5: Pattern 1 Illustration – distribution of ratings of individuals constituting the group of students versus distribution of ratings of individuals constituting the group of health professionals. Left figure corresponds to the usual distribution observed over all movies. Right figure corresponds to the contextual distribution observed over the context of pattern 1 from Table 3.4.

In Table 3.5, we present some results provided by DEBuNk over Yelp dataset. The groups of individuals are labeled by the size of their friend network and their seniority in the Yelp platform. Notice that additional demographic data about users are missing. This prevents DEBuNk from obtaining concrete results similar to the ones obtained in Movielens. The resulting patterns highlight the places for which groups of individuals have divergent opinions. For example, pattern 2 states that Senior Yelp users (registered in Yelp before 2010) having a friend network of medium size (less than 100 friends) disagree with users registered in Yelp before 2015 having a large friend network (more than 100 friends) on Internal Medicines Clinics in Nevada (e.g., Las Vegas Urgent Care), contrary to the usual, where these two groups roughly share the same opinions about places (81% of the time).

	(c, u_1, u_2)	$\lvert G_E^c \rvert$	$\lvert G_I^{u_1} \rvert$	$\lvert G_I^{u_2} \rvert$	$o(i,e)$	φ_{dissent}
1	(Newcomer,*) vs. (Middler,*) in ['03 Automotive', '14.22 Electronics Repair', '22.06 Battery Stores', '22.21 Electronics'] *	10	6	6	43	$0.4 = 0.8 - 0.4$
2	(Senior, Medium) vs. (Middler, Large) in ['10.55.21 Internal Medicine'] Nevada	15	2	2	39	$0.43 = 0.81 - 0.38$
3	(Newcomer, Medium) vs. (Middler, Large) ['11.59.01 Apartments', '11.59.18 University Housing'] Arizona	14	2	2	30	$0.4 = 0.78 - 0.38$
4	(*, Small) vs. (Middler, Large),in ['10.55.50 Urologists'] *	10	6	2	30	$0.43 = 0.79 - 0.36$
5	(*, Large) vs. (Newcomer,*) in ['08 Financial Services', '22 Shopping'] AZ	12	6	6	30	$0.4 = 0.79 - 0.39$

Table 3.5: Top-5 w.r.t. number of expressed outcomes ($o(i,e)$ *column*) of disagreement patterns discovered on Yelp dataset ($|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 2$, $\sigma_E = 10$, $\sigma_I = 1$ and $\sigma_\varphi = 0.4$).

3.6.2.2 Analysis of the Voting Behavior in the European Parliament Dataset

The two past decades have witnessed an increasing emergence of Open Government Data[16] (OGD) promoting transparency and accountability in public institutions. Consequently, many researchers from different fields (e.g., information science, political and social sciences, data mining and machine learning) have studied such data (Charalabidis, Alexopoulos, and Loukis, 2016). For instance, Jakulin et al., 2009 uses hierarchical clustering and PCA to identify cohesion blocs and dissimilarity blocs of voters within the US Senate. Similar work was done on the Finnish (Pajala, Jakulin, and Buntine, 2004), the Italian (Amelio and Pizzuti, 2012) and the Swiss (Etter et al., 2014) parliaments to study the polarization and cohesion between parliamentarians. In the same spirit, Grosskreutz, Boley, and Krause-Traudes, 2010 investigates the voting behavior of citizens instead of politicians relying on subgroup discovery. The algorithms proposed in this chapter go further and supports the discovery of new insights in such data.

Table 3.6 reports patterns obtained by DEBuNk where the aim is to find contexts (subsets of voting sessions) that lead groups of parliamentarians (labeled by their countries and their corresponding date of accession to the European Union) to strong disagreement compared to the usually observed agreement. Note that we choose carefully $\sigma_E \geq 25$ to reach subgroups of the third level of the themes hierarchy which on average contain $\backsim$ 25 voting sessions. Such analysis can be valuable to political analysts and journalists as it enables to uncover subjects/thematics of votes on which countries have divergent opinions. For instance, the second pattern in Table 3.6 illustrated in Figure 3.6, states that the voting sessions about

	(c, u_1, u_2)	$\lvert G_E^c \rvert$	$\lvert G_I^{u_1} \rvert$	$\lvert G_I^{u_2} \rvert$	$o(i,e)$	φ_{dissent}
1	([1973, 1973] United Kingdom) vs. (*,*) ['4 Economic, social & territorial cohesion', '8.70 Budget of the Union']	47	88	958	30255	$0.54 = 0.68 - 0.14$
2	([1973, 1973] United Kingdom) vs. (*,*) ['4.15.05 Industrial restructuring, job losses, Globalization Adjustment Fund']	47	88	958	30250	$0.54 = 0.68 - 0.14$
3	([1958, 1958] Italy) vs. ([1981, 2013] *) ['3.40 Industrial policy', '6.20.02 Export /import control, trade defence']	79	99	433	29501	$0.51 = 0.87 - 0.35$
4	([1958, 1995] *) vs. ([1973, 2013] *) ['3.40.16 Raw materials']	44	709	547	28989	$0.55 = 0.91 - 0.36$
5	([1958, 1995] *) vs. ([1973, 2013] *) ['6.20 Common commercial policy' , '6.30 Development cooperation']	38	709	547	25268	$0.51 = 0.91 - 0.39$

Table 3.6: Top-5 w.r.t. number of expressed outcomes ($o(i,e)$ *column*) of inter-group agreement patterns discovered on EPD8 ($|\mathscr{A}_E| = 1$, $|\mathscr{A}_I| = 2$, $\sigma_E = 25$, $\sigma_I = 1$ and $\sigma_\varphi = 0.5$ using φ_{dissent}).

[16] http://www.oecd.org/gov/digital-government/open-government-data.htm

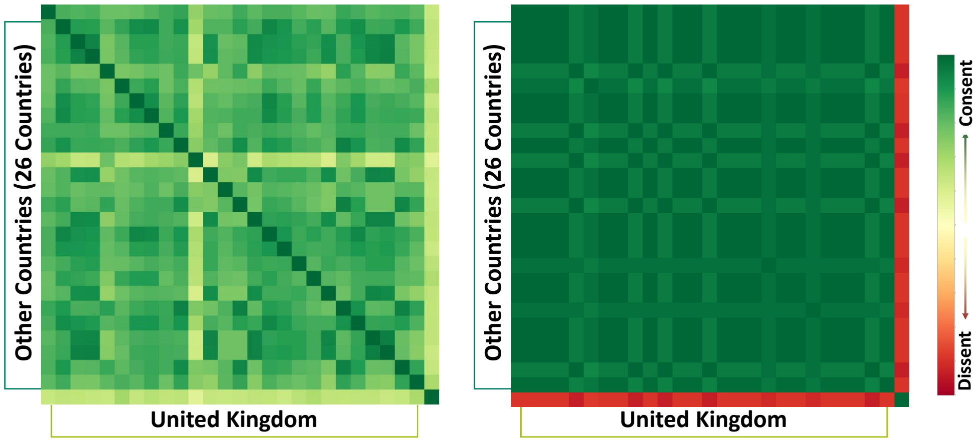


Figure 3.6: Illustration of pattern 2 reported in Table 3.6. The left matrix depicts the agreement observed in general between countries when considering all voting sessions. The right matrix corresponds to the inter country agreement for the context of pattern 2.

theme 4.15.05 (Industrial Restructuring, job losses, EGF, e.g., Mobilization of the European Globalization Adjustment Fund: redundancies in aircraft repair and installation services in Ireland) lead to strong disagreements between parliamentarians from the United Kingdom and their peers. In Figure 3.6, we provide a visualization of this pattern through a similarity matrix where each cell represents the similarity between two countries. This can be seen as a post-processing step where the end-user chooses to augment the pattern with more related information (similarities between other countries). Such visualization brings more context to the pattern. While the second pattern conveys that British parliamentarians are in strong disagreement with their peers, the visualization goes beyond by reporting that all other countries formed a coalition against the voting decision of British parlementarians. The Algorithms elaborated in this work also allow to discover patterns exhibiting consensual subjects, thanks to the quality measure φ_{consent}.

Algorithms elaborated in this work also enable the discovery of consensual subjects, thanks to the quality function φ_{consent}. In Table 3.7 , we report patterns where groups of

	(c, u_1, u_2)	$\lvert G_E^c \rvert$	$\lvert G_I^{u_1} \rvert$	$\lvert G_I^{u_2} \rvert$	$o(i,e)$	$\varphi_{consent}$
1	S&D vs. ECR in ['6.20.03 Bilateral economic and trade agreements and relations']	185	211	103	43162	$0.41 = 0.9 - 0.49$
2	PPE vs. GUE/NGL ['8.70.03.03 2013 discharge']	137	263	60	33664	$0.41 = 0.85 - 0.43$
3	ENF vs. * ['3', '8 State & evolution of the Union']	42	48	958	27191	$0.4 = 0.69 - 0.29$

Table 3.7: Top-3 w.r.t. number of expressed outcomes ($o(i,e)$ *column*) of relevant inter-group agreement patterns discovered over European Parliament Dataset considering by default the full dataset, $|\mathscr{A}_E| = 1$, $|\mathscr{A}_I| = 1$, $\sigma_E = 15$, $\sigma_I = 1$ and $\sigma_\varphi = 0.4$ using $\varphi_{consent}$.

parliamentarians agree more than what is observed in general. For example, pattern 1 of Table 3.7 shows that while *Socialists and Democrats* (S&D - *left-wing*) parlementarians are usually in disagreement ($\text{IAS}_{\text{voting}} = 0.41$) with European Conservatives and Reformists (ECR - *right-wing*), they tend to have convergent opinions ($\text{IAS}_{\text{voting}} = 0.9$) on ballots concerning theme 6.20.03 (bilateral agreement and relations with countries external to the union, e.g. Implementation of the Free Trade Agreement between the European Union and the Republic of Korea). In Figure 3.7, we illustrate the similarities between political groups for pattern 3 reported in Table 3.7. It is worth to note that, as part of a collaboration with political journalists, we provide an online tool[17], dubbed ANCORE (Lacombe et al., 2019), which makes it possible to analyze European parliament voting sessions.

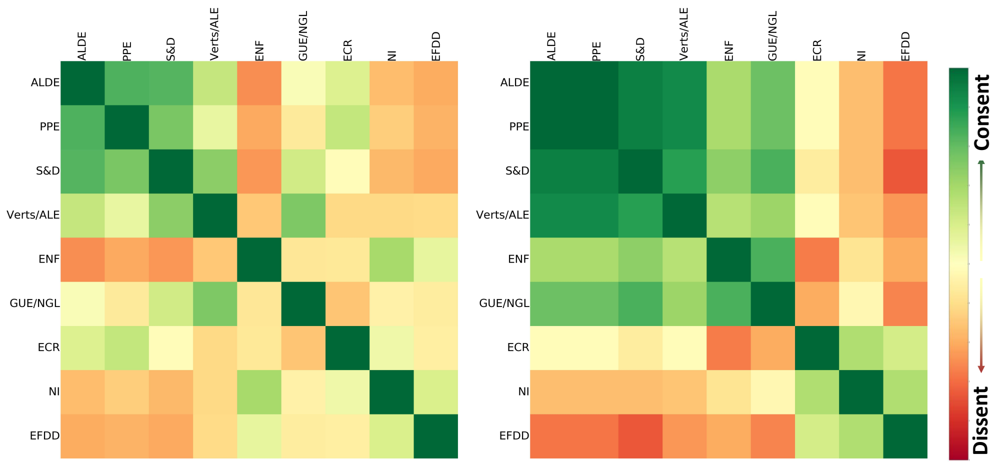


Figure 3.7: Illustration of pattern 3 reported in Table 3.7. The left matrix depicts the agreement observed in general between political groups when considering all ballots. The right matrix corresponds to the inter-group agreement between groups for the context pointed out by pattern 3. We observe that group ENF is in disagreement with ALDE, PPE and S&D who hold 63% of the seats in the 8^{th} European Parliament. The context of Pattern 3, which mainly covers EGF (European Globalisation Adjustment Fund) ballots, suggests an agreement between group ENF and the majority.

3.6.2.3 Illnesses Prevalence on the Basis of Medicine Consumption

Monitoring the disease prevalence is an important task. Many researchers dedicated their effort to analyze the prevalence of diseases considering different sources of data. Orueta et al., 2012 highlight the importance of considering outpatient data (e.g. medical prescriptions) in such epidemiology studies. With this in mind, one interesting analysis task to be conducted on *Openmedic* dataset is to look for subgroups of drugs where the ratio of intakes between two groups of individuals is substantially different than the one usually observed. For instance, we find that while *Females* take $1.32\times$ more drugs than *Males* in overall terms, this ratio increases up to $5\times$ when considering drugs prescribed for *Hyperthyroidism* (see Pattern 3 in Table 3.8). These results are similar to those reported in an epidemiology study by Wang and Crapo, 1997. Such a task can provide some insight regarding illness prevalence

[17] http://contentcheck.liris.cnrs.fr

for particular groups of individuals. In the behavioral dataset *Openmedic*, the outcomes expressed by individuals are depicted by numerical values reporting the count of medicine boxes. As we are interested in characterizing the agreement by the consumption ratio, we instantiate IAS as follows:

$$\mathrm{IAS}_{ratio}(c,u_1,u_2) = \frac{\sum_{e \in G_E^c} \theta_{avg}(G_I^{u_1},e)}{\sum_{e \in G_E^c} \theta_{avg}(G_I^{u_2},e)}$$

This ratio falls under the definition of IAS considered in *Definition 3.3.3* as it can be expressed as a weighted average:

$$\mathrm{IAS}_{ratio}(c,u_1,u_2) = \frac{\sum\limits_{e \in G_E^c} w(e,G_I^{u_1},G_I^{u_2}) \times sim_{ratio}(\theta_{avg}(G_I^{u_1},e),\theta_{avg}(G_I^{u_2},e))}{\sum_{e \in G_E^c} w(e,G_I^{u_1},G_I^{u_2})}$$

$$\text{with } w(e,G_I^{u_1},G_I^{u_2}) = \theta_{avg}(G_I^{u_2},e) \text{ and } sim_{ratio}(x,y) = \frac{x}{y}\,.$$

In order to provide interpretable patterns according to the aim of the study, we define an adapted quality measure φ_{ratio} as:

$$\varphi_{ratio}(p) = \frac{\mathrm{IAS}_{ratio}(p)}{\mathrm{IAS}_{ratio}(p^*)} \;\; \textit{with } p = (c,u_1,u_2) \in \mathscr{P} \textit{ and } p^* = (*,u_1,u_2)\,.$$

Drug boxes are labeled by tags in the ATC classification system. We aim at leveraging the medical consumption differences between groups of individuals to investigate the comparative prevalence[18] of illnesses between gender groups. Table 3.8 shows some patterns

	(c,u_1,u_2)	$\lvert G_E^c \rvert$	$\lvert G_I^{u_1} \rvert$	$\lvert G_I^{u_2} \rvert$	$o(i,e)$	φ_{ratio}
1	Men vs. Women in N07B - Drugs used in addictive disorders	138	39	39	4195	$4.59 = \frac{3.48}{0.76}$
2	Women vs. Men in A12A - Calcium	54	39	39	3174	$3.96 = \frac{5.21}{1.32}$
3	Women vs. Men in H03 - Thyroid Therapy	31	39	39	1981	$3.89 = \frac{5.13}{1.32}$
4	Men vs. Women in M04A - Antigout preparations	42	39	39	1940	$3.91 = \frac{2.97}{0.76}$

Table 3.8: Top-4 w.r.t. the number of expressed outcomes on Openmedic considering by default the full dataset, $|\mathscr{A}_E| = 1$, $|\mathscr{A}_I| = 1$, $\sigma_E = 10$, $\sigma_I = 1$ and $\sigma_\varphi = 3.5$ using φ_{ratio}.

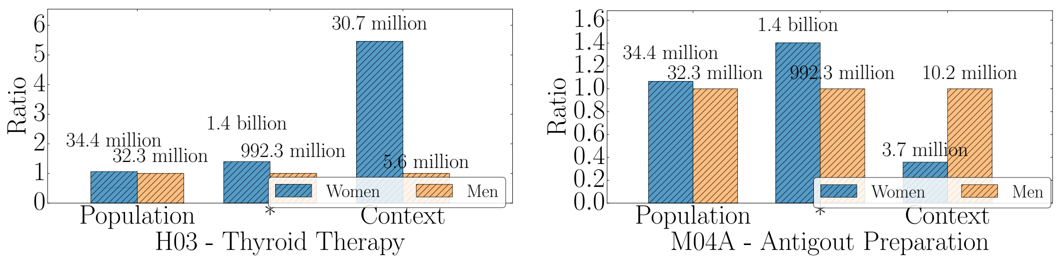


Figure 3.8: Drugs consumption behavior of gender groups in Patterns 3 (left) and 4 (right).

[18] http://www.med.uottawa.ca/sim/data/epidemiology_rates_e.htm

discovered by DEBuNk on Openmedic. Note that we carefully choose $\sigma_E \geq 10$ to reach subgroups of drugs of the fifth level of ATC tree which on average contains $\backsim 10$ drugs.

Pattern 4 states that, for drugs prescribed for *Gout*[19], men consume $3\times$ more drugs than women, whereas in overall terms, men consume $0.76\times$ less drugs than women. Similar results were reported by an epidemiology study of *Gout* in (Roddy and Doherty, 2010) giving an incidence of gout per 1,000 person-years of 1.4 in women and 4.0 in men. Patterns 3 and 4, depicted in Figure 3.8, report details on the differences between the two gender groups in terms of population size and number of drugs consumed both in overall and in the context highlighted by the pattern.

3.6.3 Quantitative study

In this section, we first start by comparing the devised algorithms against some standard SD/EMM techniques and against DSC (Belfodil et al., 2017a) in section 3.6.3.1. Next, we evaluate the efficiency of both the closure operator and the optimistic estimates proposed to improve the performance of DEBuNk in section 3.6.3.2. Moreover, in section 3.6.3.3, we investigate empirically the performance contribution of HMT descriptions enumerations compared to a standard itemsets enumeration when items are augmented with a taxonomy. Subsequently, we analyze in section 3.6.3.4, how DEBuNk scales with regards to different parameters. Finally, we compare the performance of Quick-DEBuNk in section 3.6.3.5. We wrap up by a discussion in section 3.6.4.

3.6.3.1 Comparison to classical SD/EMM techniques and to DSC

We have investigated the ability of classical SD/EMM techniques to tackle the problem of discovering exceptional (dis)agreement among groups of individuals. To this end, we have considered three appropriate SD/EMM adaptations[20] and tested them on synthetic datasets with ground truth. No existing quality measure (in classical SD) or model (in classical EMM) makes it possible to uncover exactly the inter-group agreement patterns, and these experiments obviously supported this observation (for more details, please refer to Appendix A). This is due to the fact that SD and EMM techniques are usually tailored to tackle a specific mining task. Therefore and for the interest of brevity, we report here only comparative experiments against our first attempt (Belfodil et al., 2017a) implemented by DSC.

DSC aims at discovering top-k patterns that elucidate exceptional (dis)agreement between groups of individuals. In addition, for a sufficiently large k, DSC solves the core problem tackled in this chapter limited to the two first conditions (i.e., validity and maximality). Note that, we disable the aggregation dimension parameter for DSC to obtain comparable pattern

[19] https://www.medicinenet.com/gout_gouty_arthritis/article.htm

[20] Since common SD techniques require flat representations of the underlying dataset augmented with a target attribute, we have proposed two adaptations: SD-Majority for discovering (dis)agreement with the majority and SD-Cartesian for discovering (dis)agreement between two groups on the cartesian product $G_E \times G_I \times G_I$. In both of the aforementioned adaptations, the target is equal to 1 if there is an agreement, 0 else. Experiments are performed using PySubgroup(Lemmerich and Becker, 2018) while utilizing the precision gain (Fürnkranz, Gamberger, and Lavrač, 2012) as a quality measure. Moreover, to take into account the usual agreement between groups, we adapt Exceptional Subgraph Mining(Kaytoue et al., 2017) to discover contextual (dis)agreement in subgraphs representing individuals group pairs.

sets. To compare between DEBuNk and DSC, we designed experiments to answer to the two following questions:

Q1. How concise is the patterns set provided by DEBuNk compared to the one provided by DSC?

Q2. How diversified is the patterns set, limited to k patterns, provided by DEBuNk compared to the one provided by DSC?

In order to answer (Q1), we evaluate the number of patterns returned by DEBuNk and DSC when looking for complete pattern set P (i.e., k sufficiently large for DSC). For this, we run both methods on EPD8 with various[21] quality thresholds σ_φ and descriptive attributes $\mathscr{A}_E$, $\mathscr{A}_I$. Figure 3.9 reports the results of these experiments. Results demonstrate that DEBuNk considerably reduces the desired pattern set while ensuring that each pattern returned by DSC is represented by a pattern returned by DEBuNk (according to the problem definition). On average, DSC returns 38 times more patterns than DEBuNk. Moreover, DEBuNk achieves better performance than DSC in terms of run time. This is explained by (i) the model simplification which reduces the complexity of computing the interestingness measure and (ii) the pruning property implemented by DEBuNk supported by condition (3) of the problem definition.

So far, we compared DEBuNk against DSC when looking for the complete pattern set. Experiments (Q1) demonstrated the fact that in such a setting DSC returns an overwhelmingly large results set. To tackle this problem, DSC implements a top-k algorithm to control the size of the returned pattern set. Of course, the main drawback of using a top-k algorithm is the lack of diversity even when redundancy is avoided by closure operators. This lack of diversity is induced by the fact that, most likely, the patterns observing the highest qualities are condensed in a small region of the dataset.

To fairly evaluate the diversity of patterns returned by both DSC and DEBuNk (Q2), we run both algorithms for several parameters[22] and compare the size of the datasets regions covered by both returned pattern sets. This quantity can be captured by the number of outcomes covered by a results set, that is $|o[P^k]| = |\{(i,e) \in G_I \times G_E \text{ s.t. } o(i,e) \text{ is expressed}\}|$ with P^k an arbitrary pattern set containing k patterns. For a fair comparison, we compare $|o[P^k_{\text{DSC}}]|$ (top-k patterns) against $|o[P^k_{\text{DEBuNk}}]|$. To obtain the latter quantity, we run DEBuNk

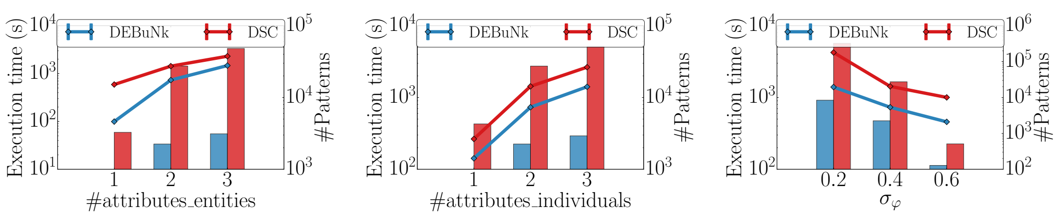


Figure 3.9: Comparison between DEBuNk and DSC for the task of discovering the complete set of the desired patterns on EPD8 dataset (default parameters are: $|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 2$, $\sigma_\varphi = 0.4$, $\sigma_E = 40$, $\sigma_I = 10$ and φ_{dissent}). Lines correspond to the execution time and bars correspond to the number of returned patterns.

[21] 27 runs for each method by varying $(|\mathscr{A}_E|, |\mathscr{A}_I|, \sigma_\varphi) \in [[1,2,3], [1,2,3], [0.2,0.4,0.6]]$
[22] 81 runs by varying $(k, |\mathscr{A}_E|, |\mathscr{A}_I|, \sigma_\varphi) \in [[10,50,100], [1,2,3], [1,2,3], [0.2,0.4,0.6]]$

so as to obtain the complete pattern set P_{DEBuNk}. Next, we draw 100 k-sized samples drawn uniformly from the obtained P_{DEBuNk} and then compute the average $|o[P^k_{\text{DEBuNk}}]|$. It is worth mentioning that comparison can be made also by taking the top-k patterns P_{DEBuNk} rather than an arbitrary k-sized sample. We decided to study the latter scenario, since the philosophy of DEBuNk is to retrieve the complete patterns set summarizing exceptional (dis)agreement in an underlying behavioral dataset.

Results are reported in Figure 3.10. Clearly, DEBuNk's k-sized pattern set covers larger (and different) parts of the dataset compared to DSC's top-k pattern set. We observe that DEBuNk surpasses DSC by one order of magnitude ($\times 12.5$ in average) when comparing the portions of the dataset covered by their respective k-sized pattern set. Simply put, when the pattern set related to DEBuNk covers 10% of the dataset, DSC patterns cover less than 1% of the underlying dataset records.

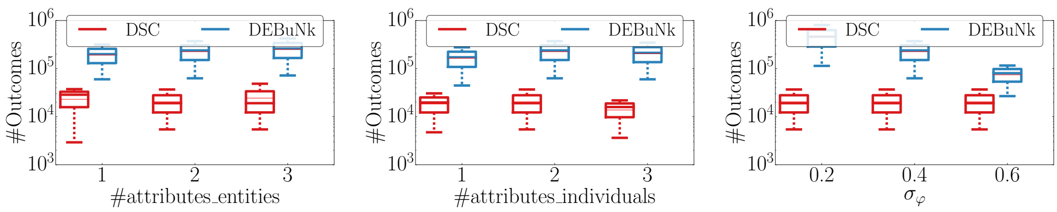


Figure 3.10: Comparison between DEBuNk and DSC (top-k) for the task of discovering k-sized pattern sets on EPD8 Dataset (default parameters: $|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 2$, $\sigma_\varphi = 0.4$, $\sigma_E = 40$, $\sigma_I = 10$ and φ_{dissent}). Box plots correspond to the size of $O[P^k]$ when varying k in $[10, 50, 100]$.

3.6.3.2 Efficiency of closure operators and optimistic estimates

To evaluate the efficiency of closure operators and optimistic estimates, we compare DEBuNk against two baseline algorithms. The first baseline, named *Baseline*, is obtained by disabling both closure operators and the pruning properties supported by the defined optimistic estimates. Thus, *Baseline* only pushes the anti-monotonic constraints (i.e., the set of cardinality constraints $\mathscr{C}$). The second baseline, *Baseline+Closed*, is proposed to study more precisely the efficiency of the optimistic estimates. Thus, it is obtained by disabling the optimistic estimate based pruning. In this experiments, we interrupt a method if its execution time exceeds 10 hours. Figures 3.11, 3.12 and 3.13, carried on respectively EPD8, Movielens and Yelp datasets, report the execution time and the number of candidate patterns processed by each of the three methods when varying the size of the dataset w.r.t. both the number of records and the size of the description space.

Experiments give evidence that the closure operator and the canonicity tests performed by EnumCC are effective as they drastically reduce the number of evaluated patterns. Additionally, DEBuNk is about one order of magnitude faster than the *Baseline+Closed* algorithm, thanks to the optimistic estimate-based pruning. This especially happens when the IAS measure is a simple average, which is the case of the IAS measure used for EPD8, Yelp and Movielens. This is explained by the fact that the corresponding optimistic estimate is tight.

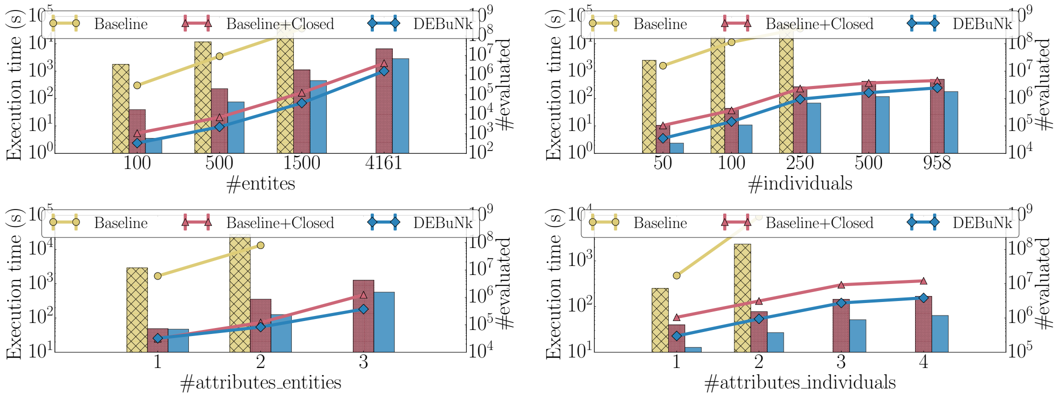


Figure 3.11: Effectiveness of DEBuNk considering EPD8 Dataset with $|G_E| = 2000$, $|G_I| = 500$, $|Outcomes| = 750k$, $|\mathscr{A}_E| = 3$, $|\mathscr{A}_I| = 4$, $\sigma_E = 40$, $\sigma_I = 10$, $\sigma_\varphi = 0.5$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of evaluated patterns.

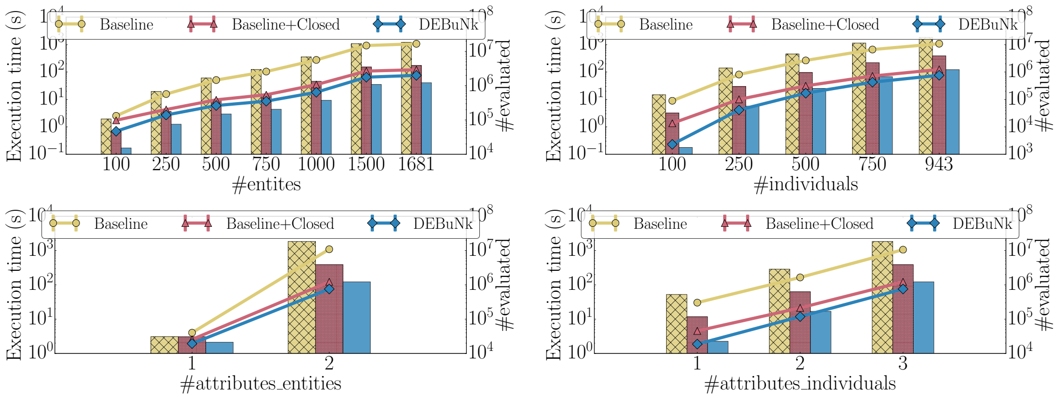


Figure 3.12: Effectiveness of DEBuNk considering Movielens Dataset with $|G_E| = 1681$, $|G_I| = 943$, $|Outcomes| = 100k$, $|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 3$, $\sigma_E = 8$, $\sigma_I = 50$, $\sigma_\varphi = 0.2$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of evaluated patterns.

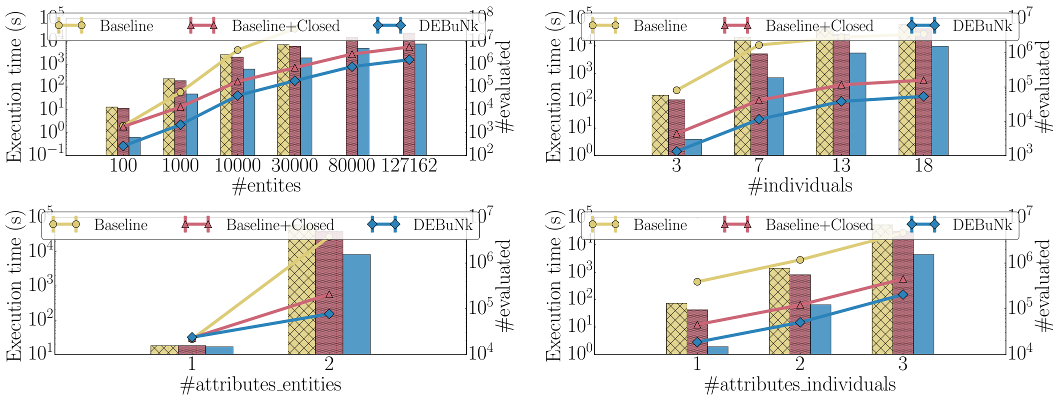


Figure 3.13: Effectiveness of DEBuNk considering Yelp Dataset with $|G_E| = 25000$, $|G_I| = 18$, $|Outcomes| = 146k$, $|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 3$, $\sigma_E = 5$, $\sigma_I = 1$, $\sigma_\varphi = 0.5$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of evaluated patterns.

3.6.3.3 Efficiency of HMT closed descriptions vs. closed itemsets enumeration

In order to evaluate the performance of the closed descriptions enumeration in the presence of a taxonomy linking the tags (items), we study the behavior of DEBuNk (i.e. execution time and the number of explored patterns) both with and without leveraging the hierarchy between items. The latter can be done by scaling the HMT values (as illustrated in Fig. 3.2) using a vector representation for each tagged record. Experiments are carried out on EPD8 and Yelp datasets whose entities are characterized by a hierarchy of 347 tags and 1175 tags respectively. To vary the number of items/tags constituting the hierarchy, we remove tags from the tree in a bottom-up fashion until the desired number of tags/items is reached, followed by replacing the HMT values of each entity by the set of ascendants tags remaining in the obtained tree.

Experiments reported in Figure 3.14 demonstrate that taking into account the hierarchy of tags significantly improves the performance of DEBuNk (5× faster). This results from the fact that, in contrast to itemsets enumeration, HMT descriptions enumeration exploits the structure of the hierarchy and therefore avoids considering chain descriptions (e.g., $\{1, 1.10.40\}$). Note that the bars depict the number of patterns that are visited by EnumCC used in DEBuNk to generate the closed patterns. Obviously, the HMT and Itemset closed description enumeration return the same number of closed patterns. We choose to represent the number of visited patterns rather than the number of closed patterns to illustrate the differences between the HMT and Itemset enumeration in terms of the size of the explored search space.

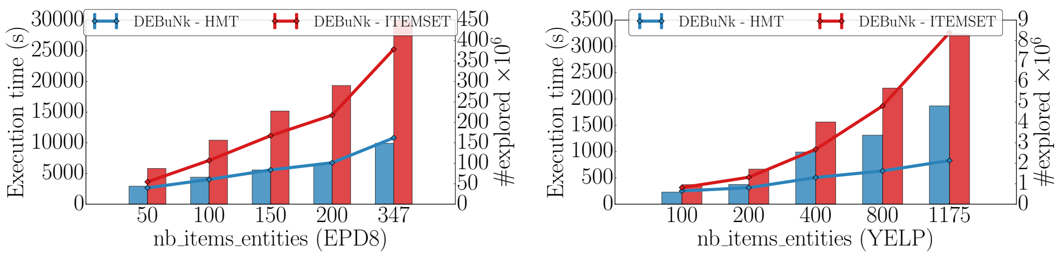


Figure 3.14: Efficiency of HMT against itemsets closed descriptions enumeration according to the number of items/tags constituting the hierarchy for the two datasets EPD8 (left) and Yelp (right). For both datasets we only consider the HMT attribute for entities $|\mathscr{A}_E| = 1$. The used parameters for EPD8 are: $|\mathscr{A}_I| = 6$, $\sigma_E = 1$, $\sigma_I = 10$, $\sigma_\varphi = 0.5$ and φ_{dissent}. The used parameters for Yelp are: $|\mathscr{A}_I| = 3$, $\sigma_E = 5$, $\sigma_I = 1$, $\sigma_\varphi = 0.5$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of visited patterns.

3.6.3.4 Performance study of DEBuNk

We now focus on the study of DEBuNk according to the size of the description spaces ($\mathscr{D}_E$, $\mathscr{D}_I$), the support thresholds, the quality threshold and the quality measures. To study the behavior of DEBuNk according to the size of the description spaces, we choose to vary the number of items resulting from projecting the attributes values of each record (entity/individual) on their corresponding vector representation. To this end, we select values from each attribute according to the size of its corresponding domain so as to obtain the required number of items. We follow the same approach as in the experiments reported in Figure 3.14 to select the required number of tags for an HMT attribute. Numerical attributes domains are discretized according to the required number of items. Subsets of values of categorical attributes are regrouped under single categories in order to obtain the desired number of values.

Figures 3.15, 3.16 and 3.17 report the behavior of DEBuNk when carried on EPD8, Movielens and Yelp. Clearly, the number of evaluated candidates and the execution time increase with regards to the size of description spaces $\mathscr{D}_I$ and $\mathscr{D}_E$ and also the size of the datasets (i.e. $|G_I|$ and $|G_E|$). These experiments confirm that pushing monotonic constraints (i.e. supports threshold σ_E, σ_I) drastically improves the efficiency of DEBuNk. Finally, a higher threshold on the quality σ_φ leads to an important reduction of the number of visited patterns and therefore to a better execution time. This demonstrates the effectiveness of the pruning properties enabled by the use of optimistic estimates. We also notice that φ_{consent} performs slightly better than φ_{dissent}. This effect arises mainly from the fact that, in the EU Parliament dataset, the overall observed agreement between groups of individuals is rather consensual.

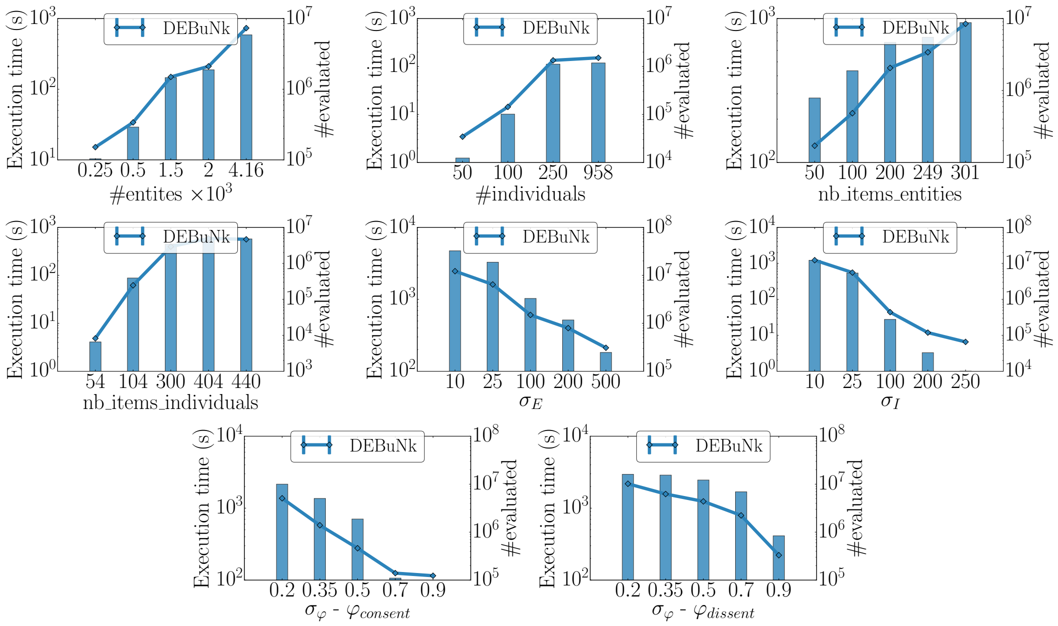


Figure 3.15: Effectiveness of DEBuNk on EPD8 according to the sizes of G_E, G_I, $\mathscr{D}_E$, $\mathscr{D}_I$, the supports and quality measures thresholds. Considering by default $|G_E| = 4161$, $|G_I| = 958$, $|Outcomes| = 2.7M$, $|\mathscr{A}_E| = 3$, $|\mathscr{A}_I| = 6$. $\sigma_E = 40$, $\sigma_I = 10$, $\sigma_\varphi = 0.5$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of evaluated patterns.

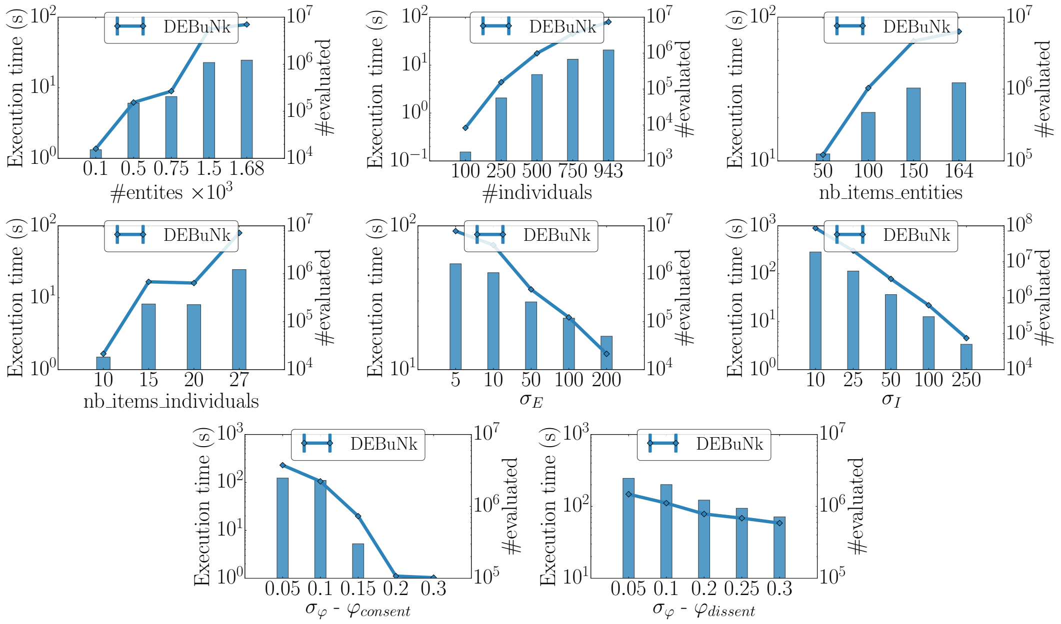


Figure 3.16: Effectiveness of DEBuNk on Movielens according to the sizes of G_E, G_I, $\mathscr{D}_E$, $\mathscr{D}_I$, the supports and quality measures thresholds. Considering by default $|G_E| = 1681$, $|G_I| = 943$, $|Outcomes| = 100k$, $|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 3$. $\sigma_E = 8$, $\sigma_I = 50$, $\sigma_\varphi = 0.2$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of evaluated patterns.

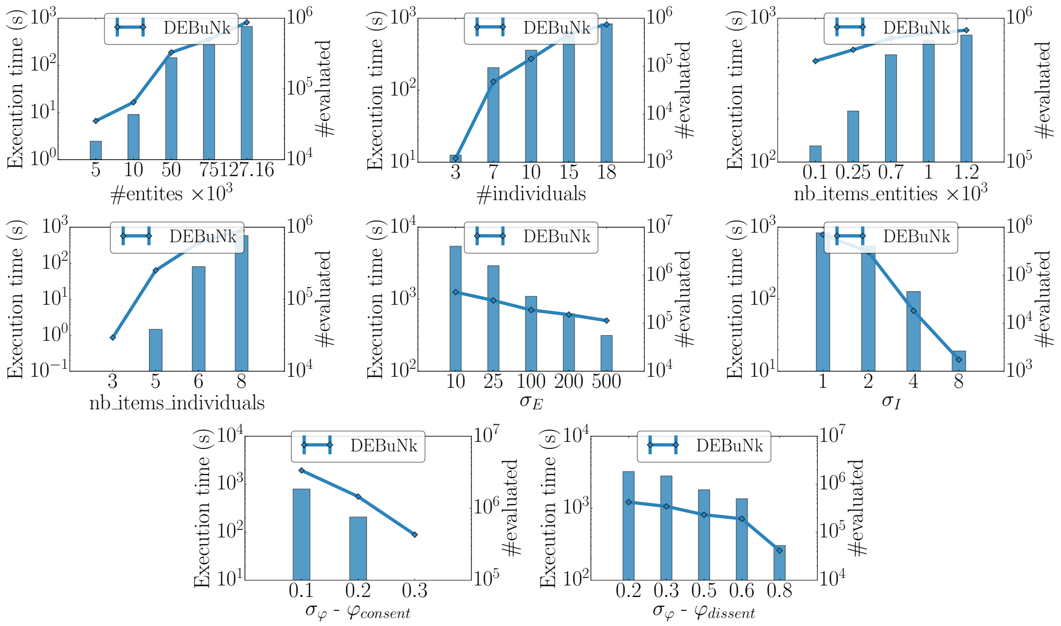


Figure 3.17: Effectiveness of DEBuNk on Yelp according to the sizes of G_E, G_I, $\mathscr{D}_E$, $\mathscr{D}_I$, the supports and quality measures thresholds. Considering by default $|G_E| = 127k$, $|G_I| = 18$, $|Outcomes| = 750k$, $|\mathscr{A}_E| = 2$, $|\mathscr{A}_I| = 3$. $\sigma_E = 50$, $\sigma_I = 1$, $\sigma_\varphi = 0.5$ and φ_{dissent}. Lines correspond to the execution time and bars correspond to the number of evaluated patterns.

3.6.3.5 Quick-DEBuNk vs. DEBuNk

To evaluate the efficiency of Quick-DEBuNk, we compare it against the exhaustive search algorithm DEBuNk over different time budgets. To objectively measure how well Quick-DEBuNk results approximates DEBuNk results, let us first define a similarity measure $\mathrm{sim}_{\mathscr{P}}$ between two patterns $p = (c, u_1, u_2)$ and $p' = (c', u'_1, u'_2)$ from $\mathscr{P}$. It captures to what extent two patterns covers the same context and groups and relies on a Jacquard Index (J in what follows):

$$\mathrm{sim}_{\mathscr{P}}(p,p') = \sqrt{J(G_E^c, G_E^{c'}) \times \frac{1}{2} \cdot \left(J(G_I^{u_1}, G_I^{u'_1}) + J(G_I^{u_2}, G_I^{u'_2})\right)} \text{ with } J(G,G') = \frac{|G \cap G'|}{|G \cup G'|} \,.$$

Note that, the quantity $(J(G_I^{u_1}, G_I^{u'_1}) + J(G_I^{u_2}, G_I^{u'_2}))$ is replaced by the following measure if the quality measure φ is symmetric:

$$\max(J(G_I^{u_1}, G_I^{u'_1}) + J(G_I^{u_2}, G_I^{u'_2}), J(G_I^{u_1}, G_I^{u'_2}) + J(G_I^{u_2}, G_I^{u'_1}))\,.$$

For comparing two pattern sets P, P' returned by respectively DEBuNk and Quick-DEBuNk, we use an F_1 score defined as follows.

$$F_1(P,P') = 2 \cdot \frac{\mathrm{precision}(P,P') \cdot \mathrm{recall}(P,P')}{\mathrm{precision}(P,P') + \mathrm{recall}(P,P')}\,, \tag{3.12}$$

$$\text{with} \begin{cases} \mathrm{precision}(P,P') & = \dfrac{\sum_{p \in P} \max(\{\mathrm{sim}_{\mathscr{P}}(p,p') \mid p' \in P'\})}{|P|}\,, \\ \mathrm{recall}(P,P') & = \dfrac{\sum_{p' \in P'} \max(\{\mathrm{sim}_{\mathscr{P}}(p',p) \mid p \in P\})}{|P'|}\,. \end{cases}$$

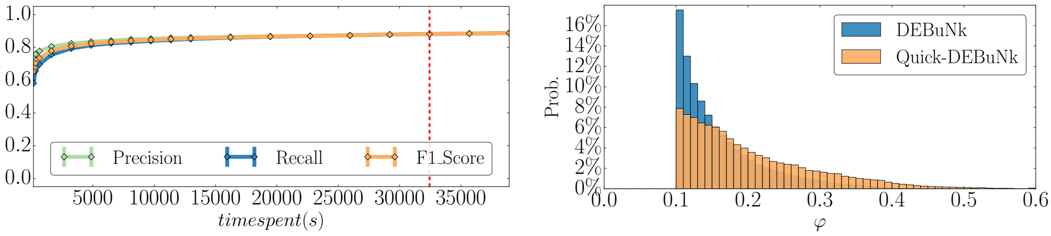
A similar measure to recall has been proposed by Bosc et al., 2018 to evaluate the ability of their algorithm to retrieve ground-truth patterns. We extend this measure with the precision to evaluate not only that all the patterns returned by DEBuNk have been retrieved by Quick-DEBuNk (i.e. recall=1) but also the conciseness of the returned set (i.e. precision=1 if and only if all returned patterns by Quick-DEBuNk are actually present in the ground-truth results set, namely the returned patterns by DEBuNk).

Figures 3.18a, 3.19a and 3.20a report the comparative study between DEBuNk and Quick-DEBuNk carried out on respectively EPD8, Movielens and Yelp. We notice that in all situations, Quick-DEBuNk is able to promptly returning high quality patterns. Interestingly, some differences can be observed between datasets. Quick-DEBuNk is less efficient on Yelp dataset. We argue that this is due to the fact that the corresponding context search space is much larger than the three other behavioral datasets (see Table 3.2) which might impede random walk step RWC for finding high quality patterns.

We also investigate the empirical distribution from which the patterns are sampled when using Quick-DEBuNk. This requires the true distribution of the qualities of valid patterns in the corresponding datasets. To this end, we run DEBuNk by disabling the generality condition (see Problem definition). This makes it possible to identify all interesting inter-group agreement patterns in the dataset. In these experiments, we choose an arbitrary threshold set to $\sigma_\varphi = 0.1$. Similarly, we run Quick-DEBuNk so as to obtain a sufficiently large pattern set, and calculate the sampling distribution from the retrieved patterns' qualities.

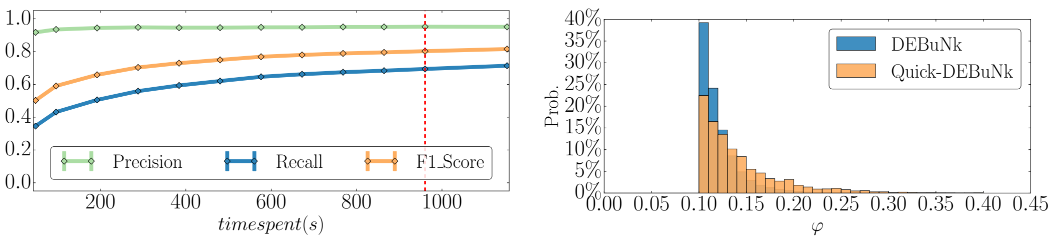
We observe from the empirical distributions depicted in Figures 3.18b, 3.19b and 3.20b that Quick-DEBuNk rewards high quality patterns by giving them a better chance to be sampled.

Finally, to evaluate the importance of the RWC (Random Walk on Contexts) step in Quick-DEBuNk, we perform the same experiments with the same time budgets with the RWC step disabled. This configuration, Quick-DEBuNk without RWC returned only 3 472, 389 and 120 valid patterns compared to 408 610, 64 198 and 75 398 valid patterns when carried out on, respectively, EPD8, Movielens and Yelp. In average, Quick-DEBuNk without RWC retrieved 20× fewer valid patterns than the original Quick-DEBuNk. This clearly



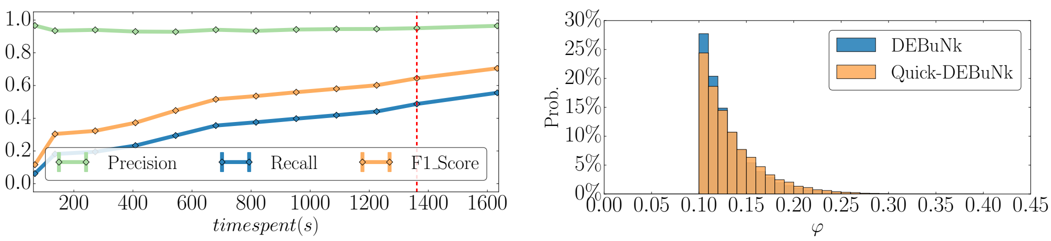
(a) $F_1(P_{\text{Quick-DEBuNk}}, P_{\text{DEBuNk}})$ (b) Empirical Distribution

Figure 3.18: Efficiency of Quick-DEBuNk compared to DEBuNk on EPD8. Parameters used are $\sigma_E = 40$, $\sigma_I = 10$, $\sigma_\varphi = 0.5$ and φ_{dissent}. The red line corresponds to the required time by DEBuNk to perform an exhaustive search.



(a) $F_1(P_{\text{Quick-DEBuNk}}, P_{\text{DEBuNk}})$ (b) Empirical Distribution

Figure 3.19: Efficiency of Quick-DEBuNk compared to DEBuNk on Movielens. Parameters used are $\sigma_E = 5$, $\sigma_I = 10$, $\sigma_\varphi = 0.25$ and φ_{dissent}. The red line corresponds to the required time by DEBuNk to perform an exhaustive search.



(a) $F_1(P_{\text{Quick-DEBuNk}}, P_{\text{DEBuNk}})$ (b) Empirical Distribution

Figure 3.20: Efficiency of Quick-DEBuNk compared to DEBuNk over Yelp. Parameters used are $\sigma_E = 15$, $\sigma_I = 1$, $\sigma_\varphi = 0.1$ and φ_{dissent}. The red line corresponds to the required time by DEBuNk to perform an exhaustive search.

indicates that RWC improves the performance of Quick-DEBuNk. This stems from the fact that when the first step (FBS step) generates a pattern, most of the time the pattern is not of a sufficient quality. RWC tackles this issue by locally searching for interesting patterns, starting from the generated pattern.

3.6.4 Discussion

DEBuNk scales well w.r.t. the size of the search space corresponding to the entities collection thanks to the defined optimistic estimates which enable to prune unpromising parts of the search space. However, DEBuNk does not scale according to the size of the description spaces related to the individuals. This limits its application when behavioral datasets have a large number of individuals described with many attributes. This is due to the need of taking into account the usual inter-group agreement in the interestingness measures. As a consequence, it is notoriously difficult to define an optimistic estimate which not only works on the entities related search space, but also on the one corresponding to the confronted couples of groups of individuals. This should be the scope of future research, starting with definition of bounds on the usual agreement quantity. Algorithm Quick-DEBuNk partially addresses this scalability issue by sampling the couples of groups directly from the patterns space rather than starting from the search tree root. Interestingly, the experiments demonstrated that Quick-DEBuNk makes it possible to retrieve most of the interesting patterns in a relatively small amount of time (i.e. compared to what returns the exhaustive search algorithm DEBuNk and the ground truth in artificial data). This is particularly observed for EPD8 dataset involving the largest description space $\mathscr{D}_I \times \mathscr{D}_I$, hence empirically demonstrating its interest. Nevertheless, Quick-DEBuNk does not have theoretical guarantees on the distribution of the sampled patterns (we only proved that all valid patterns are reachable and are generated proportionally to their size). This shortcoming is due to two reasons. On the one hand, the three-set format of the patterns makes them challenging to be sampled proportionally to their interestingness measure since the value is computed only when the context is known (no information is available before the instantiation of the two groups). On the other hand, quality measures that are expressed as average functions are complex to apprehend under direct pattern sampling framework. Dealing with this two issues is required to obtain theoretical guarantees.

To avoid misleading interpretations, it is important to be aware of the data sparsity. Remind that the proposed approaches enable to discard some patterns that involve too small subset of entities on which the two confronted groups haven't expressed enough outcomes. Moreover, the strength of the claim related to the pattern should be assessed according not only to the data sparsity but also to the representativeness of the two subpopulation of interest (e.g., the claims drawn from the EU parliament votes are usually consistent even though the data are fairly sparse).

3.7 Summary

In this chapter, we have defined the problem of discovering exceptional (dis)agreement in behavioral data and tailored an approach rooted in SD/EMM with a novel pattern domain and associated quality measures for the discovery of exceptional inter-group agreement patterns (cf. figure 3.21). We have defined DEBuNk, a branch-and-bound algorithm which takes benefit from closure operators, properties of the underlying description space (as for HMT attributes) and (tight) optimistic estimates to efficiently enumerate the patterns. Alternatively, we devised Quick-DEBuNk that samples the space of patterns instead of returning the complete set of inter-group agreement patterns. We have investigated several quality measures to assess inter-group agreement. The extensive experimental study demonstrates the efficiency of our algorithms as well as their ability to provide new insights in three case-studies: (i) the investigation of contexts that impact the inter-group agreement between parliamentarians, (ii) the characterization of affinities and contrasted opinions between reviewers in rating platforms and (iii) the study of prevalence of certain sicknesses that can be pointed out by high discrepancies between the medicine consumption rates of two subpopulations.

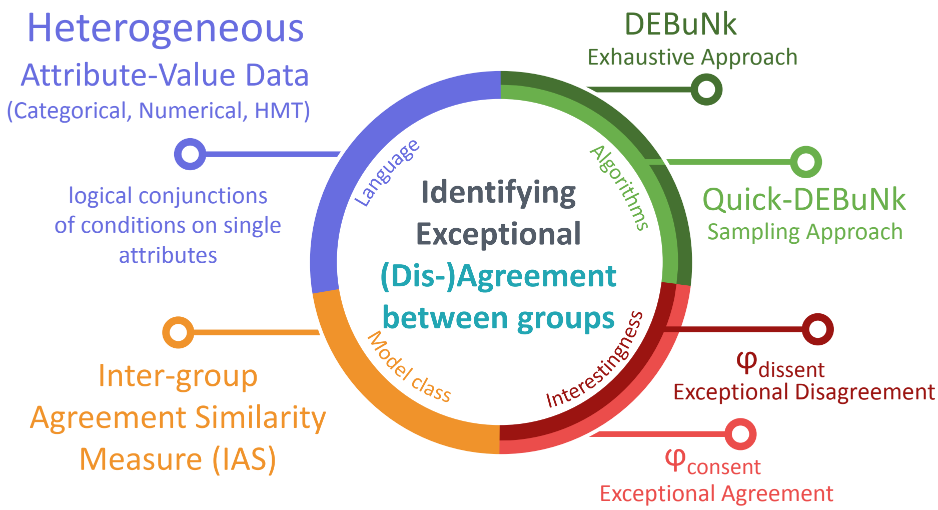


Figure 3.21: Exceptional Model Mining for Identifying exceptional (dis-)agreement between groups.

We believe that this work opens new directions for future research. First, while our method is able to analyze behavioral datasets with large collections of entities (e.g., Yelp), tackling large collections of individuals still remains challenging to ensure the scalability of both DEBuNk and Quick-DEBuNk. Indeed, the search space related to individuals does not have, according to our problem definition, properties that can be leveraged to prune unpromising parts of this search space. Another interesting future direction is to take into account the temporal dimension into the analysis of behavioral data. This can offer the opportunity to investigate how the relationship (e.g., inter-group agreement) between groups of individuals evolves through time.

This generic framework allows to discover *exceptional inter-group agreement* in several kind of behavioral datasets. By following the same reasoning as for this chapter, one can pay particular attention to the analysis of *intra-group agreement* within a group of individuals. It may support the discovery of contexts that divide a political group. This requires the definition and the integration of suited similarity measures. For instance, the cohesion of a political group can be assessed by the "agreement index" (Hix, Noury, and Roland, 2005), which is an application-specific measure to the study the European parliament. More generic measures could also be investigated to tackle a broader range of behavioral data. This is the scope of the next chapter.

4

Identifying exceptional (dis)agreement within groups

In this chapter, we devise a method which enables the discovery of exceptional (dis)agreement patterns within groups by searching for exceptional intra-group agreement patterns. We strive to find contexts (i.e., subgroups of entities) under which exceptional (dis-)agreement occurs within a group of individuals, in any type of data featuring individuals (e.g., parliamentarians, customers) performing observable actions (e.g., votes, ratings) on entities (e.g., legislative procedures, movies). To this end, we introduce the problem of discovering statistically significant exceptional contextual intra-group agreement patterns. To handle the sparsity inherent to voting and rating data, we use Krippendorff's Alpha measure for assessing the agreement among individuals. We devise a branch-and-bound algorithm, named DEvIANT, to discover such patterns. DEvIANT exploits both closure operators and tight optimistic estimates. We derive analytic approximations for the confidence intervals (CIs) associated with patterns for a computationally efficient significance assessment. We prove that these approximate CIs are nested along specialization of patterns. This allows to incorporate pruning properties in DEvIANT to quickly discard non-significant patterns. Empirical study on several datasets demonstrates the efficiency and the usefulness of DEvIANT.

4.1 Introduction

The previous chapter discussed how SD/EMM framework can be used to formalize and discover exceptional (dis)agreement **between** groups. While this technique is generic enough to handle several kind of behavioral datasets, it does not allow to discover exceptional (dis)agreement **within** groups. This chapter aims to extend the capabilities of SD/EMM to efficiently discover exceptional intra-group agreement patterns.

Consider a behavioral dataset (cf. Definition 1.1.1) describing voting behavior in the European Parliament (EP). Such a dataset records the votes of each member (MEP) in voting sessions held in the parliament, as well as the information on the parliamentarians (e.g., gender, national party, European party alliance) and the sessions (e.g., topic, date). This dataset offers opportunities to study the agreement or disagreement of coherent subgroups, especially to highlight unexpected behavior. It is to be expected that on the majority of voting sessions, MEPs will vote along the lines of their European party alliance. However, when matters are of interest to a specific nation within Europe, alignments may change and agreements can be formed or dissolved. For instance, when a legislative procedure on fishing rights is put before the MEPs, the island nation of the UK can be expected to agree on a specific course of action regardless of their party alliance, fostering an exceptional agreement where strong polarization exists otherwise.

We aim to discover such exceptional (dis-)agreements. This is not limited to just EP or voting data: members of the US congress also votes on bills, while Amazon-like customers post ratings or reviews of products. A challenge when considering such voting or rating data is to effectively handle the absence of outcomes (sparsity), which is inherently high. For instance, in the European parliament data, MEPs vote on average on only ¾ of all sessions. These outcomes are not missing at random: special workgroups are often formed of MEPs tasked with studying a specific topic, and members of these workgroups are more likely to vote on their topic of expertise. Hence, present values are likely associated with more pressing votes, which means that missing values need to be treated carefully. This problem becomes much worse when looking at Amazon or Yelp rating data: the vast majority of customers will not have rated the vast majority of products/places.

In this chapter, we introduce the problem of discovering significantly exceptional contextual intra-group agreement patterns in behavioral data, rooted in the Subgroup Discovey (SD) (Wrobel, 1997)/ Exceptional Model Mining (EMM) (Duivesteijn, Feelders, and Knobbe, 2016) framework. To tackle the data sparsity issue, we measure the agreement within groups with *Krippendorff's alpha*, a measure developed in the context of content analysis (Krippendorff, 2004) which handles missing outcomes elegantly. We develop a branch-and-bound algorithm to find subgroups featuring statistically significantly exceptional (dis-)agreement within groups. This algorithm enables discarding non-significant subgroups by pruning unpromising branches of the search space.

Figure 4.1 gives an overview of the approach we devise to discover exceptional (dis-)agreement within groups. At a high level of description, seven steps are necessary to discover significantly exceptional contextual intra-group agreement patterns. First a group of individuals g is selected by intent (1) followed by the computation of overall intra-group agreement using Krippendorff's Alpha and the confidence region of such measurement for statistical soundness (2). In order to find significantly exceptional (dis-)agreement between

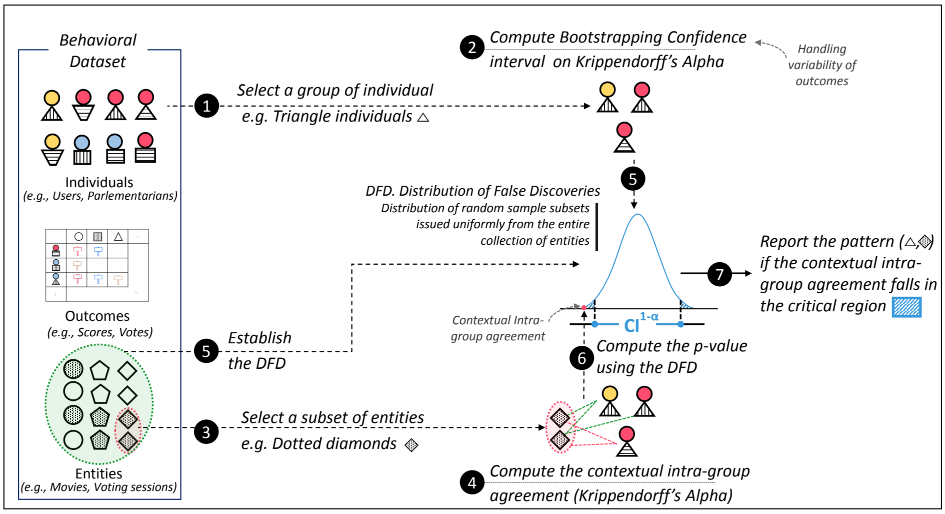


Figure 4.1: Overview of the task of discovering statistically significant exceptional (dis)agreement within groups

members of the selected group, all characterizable subsets of entities (e.g. voting sessions) are (conceptually) enumerated (3). Each characterizable subset corresponds to a context reflecting the shared properties between the entities of the subset (e.g. voting sessions of judicial matters). In each enumerated context, the intra-group agreement is evaluated (4). To gauge the excpetionality of such a contextual intra-group agreement, we compute its *p-value*: the probability that for a random subset of entities, we observe an agreement at least as extreme as the one observed for the context. Thus we avoid reporting subgroups observing a low/high intra-agreement due to chance only. To achieve this, we estimate the empirical distribution of the intra-agreement of random subsets (DFD: Distribution of False Discoveries, cf. (Duivesteijn and Knobbe, 2011)) (5) and establish, for a chosen critical value α, a confidence interval $CI^{1-\alpha}$ over the corresponding distribution under the null hypothesis (6). If the subgroup intra-agreement is outside $CI^{1-\alpha}$, the context is statistically significant (*p-value* $\leq \alpha$) and should be reported (7); otherwise the subgroup is a spurious finding.

Contributions. The main contributions of this chapter are threefold:

Problem formulation. We define the novel problem of discovering significantly exceptional contextual (dis)agreement within groups when considering a particular subset of outcomes compared to the whole set of outcomes.

Algorithms. We derive an analytical approximation of the confidence intervals associated with subgroups. This allows a computationally efficient assessment of the statistical significance of the findings. Moreover, we define tight optimistic estimates for the intra-group agreement measure (Krippendorff's Alpha) and prove that analytical approximate confidence intervals are nested. These two notions are leveraged in the devised branch-and-bound algorithm, named DEvIANT, to safe-prune unpromising branches of the search space.

Evaluation. We report a thorough empirical study to demonstrate the efficiency of the proposed algorithm as well as the interest of the found patterns over four real-world behavioral datasets (Two voting datasets and two collaborative rating datasets).

The following content is based on our article on DEvIANT (Belfodil et al., 2019a).

Roadmap. The rest of this chapter is organized as follows. The problem formulation is given in Section 4.2. We present the *intra-group agreement* measure in Section 4.3. Next, we show how such a measure is used to gauge the exceptionality of a found context for some selected group in Section 4.4 while discussing the safe-pruning properties. Then, we give particular attention to the variability of outcomes among rater in Section 4.5. Subsequently, in Section 4.6, we present the branch and bound algorithm called DEvIANT which implements the discovery of exceptional (dis)agreement within groups. Eventually, empirical evaluation is conducted in Section 4.7 to study the qualitative and quantitative performance of DEvIANT. We wrap up this chapter in Section 4.8 with some concluding thought.

Note: Notations used in this chapter are listed in Appendix D and Appendix E.

4.2 Setup and Problem Formalization

Here, we first define the fundamental concepts that we use throughout the paper in Section 4.2.1, followed by the formal problem statement in Section 4.2.2. Some definitions were given previously, although, we recall some of them for the convenience of the reader.

4.2.1 Preliminaries

A Behavioral dataset (cf. Definition 1.1.1) $\mathscr{B} = \langle G_I, G_E, O, o \rangle$ consists of a set of individuals G_I (*e.g., social network users, parliamentarians*) who give outcomes $o : G_I \times G_E \to O$ (*e.g., ratings, votes*) on entities G_E (*e.g., movies, ballots*). An example dataset is given in Table 4.1 (a modified version of Table 1.1 for the sake of toy examples).

ide	themes	date
e_1	1.20 Citizen's rights	20/04/16
e_2	5.05 Economic growth	16/05/16
e_3	1.20 Citizen's rights; 7.30 Judicial Coop	04/06/16
e_4	7 Security and Justice	11/06/16
e_5	7.30 Judicial Coop	03/07/16
e_6	7.30 Judicial Coop	29/07/16

(a) Entities (Voting sessions)

idi	country	group	age
i_1	France	S&D	26
i_2	France	PPE	30
i_3	Germany	S&D	40
i_4	Germany	ALDE	45

(b) Individuals (Parliamentarians)

idi	ide	outcome
i_1	e_2	Against
i_1	e_5	For
i_1	e_6	Against
i_2	e_1	For
i_2	e_3	Against
i_2	e_4	For
i_2	e_5	For
i_3	e_1	For
i_3	e_2	Against
i_3	e_3	For
i_3	e_5	Against
i_4	e_1	For
i_4	e_4	For
i_4	e_6	Against

(c) Outcomes

Table 4.1: Example of behavioral dataset - European Parliament Voting dataset

Similarly as in Chapter 3, from now on we refer to both sets G_E and G_I by the generic term *collection of records* denoted G if no confusion can arise. Elements from G are augmented with descriptive attributes $\mathscr{A} = (a_1, \ldots, a_m)$. Attributes $a \in \mathscr{A}$ can be boolean, numerical or categorical, potentially organized in a taxonomy (cf. Section 3.4.2). The domains of possible values of each attribute a_j ,denoted $\mathrm{dom}(a_j)$, define altogether a description space $\mathscr{D}$ (cf. Section 2.2.1) which is the set of all possible descriptions that one can use to characterize subgroups of records $\in G$. A description is a conjunction of conditions of the form $d = \langle r_1, \ldots, r_m \rangle$ where r_j depends on the type of the attribute a_j (cf. Definition 2.2.2 and Definition 2.2.12). Descriptions are partially ordered with a specialization operator denoted $\sqsubseteq$ (cf. Definition 2.2.4). $(G, (\mathscr{D}, \sqsubseteq), \delta)$ forms a pattern structure (cf. Definition 2.2.7) with δ a mapping function $\delta : G \to \mathscr{D}$ which maps each record $g \in G$ to the tightest (maximum) description $\delta(g)$ in $\mathscr{D}$ with regard to $\sqsubseteq$. A description d in $\mathscr{D}$ characterizes a subgroup of records $G^d = \{g \in G \text{ s.t. } d \sqsubseteq \delta(g)\}$.

In this chapter, we are interested in finding patterns where each one highlights a *context* in which an exceptional (dis-)agreement is observed between some *group* members. Hence, the sought patterns are defined as follows:

Definition 4.2.1 — Intra-Group Agreement Pattern. An intra-group agreement pattern is defined by intent by $p = (u, c)$ where $u \in \mathscr{D}_I$ is a *group description* and $c \in \mathscr{D}_E$ is a *context*.

The collection of all intra-group agreement pattern is denoted $\mathscr{P} = \mathscr{D}_I \times \mathscr{D}_E$ and is called the pattern space. An intra-group agreement pattern $p = (u, c)$ is defined by extent by $ext(p) = (G_I^u, G_E^c)$ with G_I^u the set of individuals supporting the group description u and G_E^c the set of entities satisfying the conditions of context c.

Each pattern depicts a group whose members express outcomes on the entities identified by the pattern's context. These outcomes are the input of an intra-group agreement measure which is required to evaluate to what extent members of a group are in (dis)agreement over the context's entities. Below, we only give the generic definition of an intra-group agreement measure in the scope of this thesis, delaying its proper instantiation to section 4.3.

Definition 4.2.2 — Intra-group Agreement Measure. An intra-group agreement measure $A : \mathscr{P} \to \mathbb{R}$ assigns to each pattern $p = (u, c) \in \mathscr{P}$ a real number $A^u(c) \in \mathbb{R}$.

This quantity captures the agreement observed among members of the group g in the context c and is computed exclusively using the outcomes $o(i, e)$ expressed by individuals $i \in G_I$ on the entities $e \in G_E$.

4.2.2 Formal Problem Definition

We are interested in finding patterns of the form $(u, c) \in \mathscr{P}$ (with $\mathscr{P} = \mathscr{D}_I \times \mathscr{D}_E$), highlighting an exceptional intra-agreement between members of a group of individuals u over a context c. We formalize this problem using the well-established framework of SD/EMM (Duivesteijn, Feelders, and Knobbe, 2016), while giving particular attention to the statistical significance and soundness of the discovered patterns (Hämäläinen and Webb, 2019).

Statistical assessment of patterns has received attention for a decade (Hämäläinen and Webb, 2019; Webb, 2007), especially for association rules (Hämäläinen, 2010b; Minato et al., 2014). Some work focused on statistical significance of results in SD/EMM during enumeration (Duivesteijn and Knobbe, 2011; Lemmerich et al., 2016) or a posteriori (Duivesteijn et al., 2010) for statistical validation of the found subgroups. This work goes in the same line of the first collection of methods, where statistical significance of patterns is assessed during enumeration.

Given a group of individuals $u \in \mathscr{D}_I$, we strive to find contexts $c \in \mathscr{D}_E$ where the observed intra-agreement, denoted $A^u(G_E^c)$, *significantly* differs from the expected intra-agreement occurring due to chance alone. In the spirit of (Duivesteijn and Knobbe, 2011; Lemmerich et al., 2016; Webb, 2007), we evaluate pattern interestingness by statistical significance of the contextual intra-agreement: we estimate the probability to observe the intra-agreement $A^u(G_E^c)$ or a more extreme value, which corresponds to the *p-value* for some null hypothesis H_0. The pattern is said to be *significant* if the estimated probability is low enough (i.e., under some critical value α). The relevant null hypothesis H_0 is: the observed intra-agreement is generated by the distribution of intra-agreements observed on a bag of i.i.d. random subsets drawn from the entire collection of entities (DFD: Distributions of False Discoveries, cf. (Duivesteijn and Knobbe, 2011)).

The choice of evaluating the interestingness intra-group agreement pattern by statistical significance is motivated by: (i) the desire to not specify to the algorithm an arbitrary threshold on the distance from the overall observed intra-group agreement, since fixing the critical value α is more intuitive, (ii) the recommendations of Krippendorff (Hayes and Krippendorff, 2007) to provide a confidence interval on the alpha metric rather than a point-value.

Problem 4.2.1 (*Discovering Exceptional Contextual Intra-group Agreement Patterns*). Given a behavioral dataset $\mathscr{B} = \langle G_I, G_E, O, o \rangle$, a minimum group support threshold σ_I, a minimum context support threshold σ_E, a significance critical value $\alpha \in]0,1]$, and the null hypothesis H_0 (the observed intra-agreement is generated by the DFD); find the pattern set $P \subseteq \mathscr{P}$ such that:

$$P = \{(u,c) \in \mathscr{D}_I \times \mathscr{D}_E \ : \ |G_I^u| \geq \sigma_I \text{ and } |G_E^c| \geq \sigma_E \text{ and } p\text{-}value^u(c) \leq \alpha\}$$

where $p\text{-}value^u(c)$ is the probability (under H_0) of obtaining an intra-agreement A at least as extreme as $A^u(G_E^c)$, the one observed over the current context.

4.3 Intra-Group Agreement Measure

Measuring the agreement between two things is historically well-studied. The most famous version is Pearson's correlation coefficient (Pearson et al., 1901), a measure of the degree of linear relationship between two variables. If one is not necessarily interested in linear, but rather monotone relationships, one can consider rank correlation instead, for instance Spearman's ρ (Spearman, 1904) or Kendall's τ (Kendall, 1938). All these measures primarily target two variables that are continuous; an equivalent of Pearson's correlation coefficient for two variables that are categorical is Association (Goodman, 1970). All these measures of agreement focus on two targets, and cannot handle missing values well. As pointed out by Krippendorff (Krippendorff, 1980, page 145), using association and correlation measures to assess agreement leads to particularly misleading conclusions. When all data falls along a line $Y = aX + b$, correlation is perfect, but agreement requires that $Y = X$ which is not what correlation coefficients measure.

Cohen's κ (Cohen, 1960) is a seminal measure of agreement between two raters who classify items into a fixed number of mutually exclusive categories. The Fleiss κ (Fleiss, 1971) extends this notion to multiple raters. We will see the fundamental definition of Krippendorff's α in Section 4.3. A modified definition (Krippendorff et al., 2016) is able to assess the reliabilities of diverse properties of unitized continua, making alpha available for time series of texts, videos, or sounds.

It has been shown (Krippendorff, 1980, page 138) that when the number of (dis-)agreeing observers is exactly two, various variants of Krippendorff's alpha are strongly related to various famous other measures. If the observations have unordered categories (nominal attribute), then alpha is aymptotically equal to Scott's Pi (Scott, 1955) (which, in turn, differs from Cohen's Kappa only by the way the probability of agreement by chance is computed). If the observations are ordinal, alpha is identical to Spearman's ρ without ties in ranks. If the observations are interval data, alpha is identical to Pearson et al.'s intraclass-correlation coefficient (Pearson et al., 1901). For more than two entities, Krippendorff's alpha formalizes a method suggested by Spiegelman et al. (Spiegelman, Terwilliger, and Fearing, 1953). More on Krippendorff's alpha, similar measures, their relations and design principles can be found in (Hayes and Krippendorff, 2007).

The simplest example to illustrate how Krippendorff's alpha (hereforth denoted A) measures agreement, concerns two observes who each mark each of then documents as relevant or not to a specific topic. Hence, the outcome is binary, there are two observers and ten documents to mark, and each observer assigns a binary mark to each document. One can simply count the percentage of documents on which both observers agree, but this is not such a meaningful number: the contingency table marginals matter too, to assess whether a certain agreement is significant or not. For instance, if the observers agree on all documents, this is much more significant if the total fraction of ones in cells is 50% than if it were 80%. Instead, one would want to assess how the agreement compares to chance.

Summarizing the actual marks in an observer-outcome contingency table is easy to do. Given the relative proportions of zeroes and ones in the dataset, we can construct the hypothetical contingency table of maximum agreement as well, with all off-diagonal entries equal to zero. Instead, we can also determine the contingency table of chance agreement, by generating observer-outcome contingency table cells through the corresponding process of simulating drawing balls from urns.

Krippendorff's alpha now uses these three contingency tables[1] to quantify the degree of observed agreement. The core idea is that a proper measure would be: on a scale from the contingency table of chance agreement to the contingency table of maximum agreement, how far along that line do we find the contingency table of observed agreement? More formally:

$$\text{observed co-occurrences} = A(\text{maximum agreement}) + (1 - A)(\text{chance agreement})$$

Hence, when $A = 1$, the agreement is as large as it can possibly be (given the class prior), and when $A = 0$, the agreement is indistinguishable to agreement by chance. We can also

[1] to make the contingency table math work out, one must balance the disagreement (off-diagonal) cells, but this does not alter the outcome

have $A < 0$, where disagreement is larger than expected by chance and which corresponds to systematic disagreement. Simple algebra gives us the direct formula:

$$A = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}} = 1 - \frac{D_{\text{obs}}}{D_{\text{exp}}} \tag{4.1}$$

In summary, Krippendorff's alpha (A) measures the agreement among raters. This measure has several properties that make it attractive in our setting, namely: (i) it is applicable to any number of observers; (ii) it handles various domains of outcomes (ordinal, numerical, categorical, time series); (iii) it handles missing values; (iv) it corrects for the agreement expected by chance.

Given a behavioral dataset $\mathscr{B}$, we want to measure Krippendorff's alpha for a given context $c \in \mathscr{D}_E$ characterizing a subset of entities $G_E^c \subseteq G_E$, which indicates to what extent the individuals who comprise some selected group are in agreement $g \in \mathscr{D}_I$. From Equation (4.1), we have: $A(S) = 1 - \frac{D_{\text{obs}}(S)}{D_{\text{exp}}}$ for any $S \subseteq G_E$. Note that the measure only considers entities having at least two outcomes; we assume the entities not fulfilling this requirement to be removed upfront by a preprocessing phase. We capture observed disagreement by:

$$D_{\text{obs}}(S) = \frac{1}{\sum_{e \in S} m_e} \sum_{o_1 o_2 \in O^2} \Delta_{o_1 o_2} \cdot \sum_{e \in S} \frac{m_e^{o_1} \cdot m_e^{o_2}}{m_e - 1} \tag{4.2}$$

Where m_e is the number of expressed outcomes for the entity e and $m_e^{o_1}$ (resp. $m_e^{o_2}$) represents the number of outcomes equal to o_1 (resp. o_2) expressed for the entity e. $\Delta_{o_1 o_2}$ is a distance measure between outcomes, which can be defined according to the domain of the outcomes (e.g., $\Delta_{o_1 o_2}$ can correspond to the Iverson bracket indicator function $[o_1 \neq o_2]$ for categorical outcomes or distance between ordinal values for ratings. Choices for the distance measure are discussed in (Krippendorff, 2004)). In the following, we define two distance measures that are used in this chapter.

1. **Distance between categorical outcomes:** this distance measure is appropriate when the underlying behavioral data consider categorical outcomes such as in voting datasets. Given O a set of possible categorical votes and o_1, o_2 two outcomes in O, the expression of such a distance is given as following:

$$\Delta_{o_1 o_2} = \begin{cases} 0 & \text{iff} \quad o_1 = o_2 \\ 1 & \text{iff} \quad o_1 \neq o_2 \end{cases} \tag{4.3}$$

2. **Distance between ordinal outcomes:** this distance measure is appropriate when the underlying behavioral data consider ordinal outcomes such as in rating datasets. Given O a set of possible totally ordered ratings and o_1, o_2 two outcomes in O^2, the expression of such a distance is given as following:

$$\Delta_{o_1 o_2} = \left(\sum_{z=o_1}^{z=o_2} m^z - \frac{m^{o_1} + m^{o_2}}{2} \right)^2 \tag{4.4}$$

We define below D_{exp} that represents the disagreement expected by chance in Krippendorff's alpha:

$$D_{\text{exp}} = \frac{1}{m \cdot (m-1)} \sum_{o_1, o_2 \in O^2} \Delta_{o_1 o_2} \cdot m^{o_1} \cdot m^{o_2} \tag{4.5}$$

Where m is the number of all expressed outcomes, m^{o_1} (resp. m^{o_2}) is the number of expressed outcomes equal to o_1 (resp. o_2) observed in the entire behavioral dataset. This corresponds to the disagreement by chance observed on the overall marginal distribution of outcomes.

Example:
Table 4.2 summarizes the behavioral data from Table 4.1. The disagreement expected by chance equals (given: $m^F = 8$, $m^A = 6$): $D_{\text{exp}} = 48/91$. To evaluate intra-agreement among the four individuals in the global context (considering all entities), first we need to compute the observed disagreement $D_{\text{obs}}(G_E)$. This equals the weighted average of the two last lines by considering the quantities m_e as the weights: $D_{\text{obs}}(G_E) = \frac{4}{14}$.

Hence, for the global context, $A(G_E) = 0.46$. Now, consider the context $c = \langle \textit{themes} \supseteq \{\text{7.30 Judicial Coop.}\}\rangle$, having as support: $G_E^c = \{e_3, e_5, e_6\}$. The observed disagreement is obtained by computing the weighted average, only considering the entities belonging to the context: $D_{\text{obs}}(G_E^c) = \frac{4}{7}$. Hence, the contextual intra-agreement is: $A(G_E^c) = -0.08$.

Comparing $A(G_E^c)$ and $A(G_E)$ leads to the following statement: *"while parliamentarians are slightly in agreement in overall terms, matters of judicial cooperation create systematic disagreement among them"*.

	[F]or		**[A]**gainst			
	e_1	e_2	e_3	e_4	e_5	e_6
i_1		A			F	A
i_2	F		A	F	F	
i_3	F	A	F		A	
i_4	F			F		A
m_e	3	2	2	2	3	2
$D_{\text{obs}}(e)$	0	0	1	0	$\frac{2}{3}$	0

Table 4.2: Summarized Behavioral Data; $D_{\text{obs}}(e) = \sum_{o_1,o_2 \in O^2} \Delta_{o_1 o_2} \dfrac{m_e^{o_1} \cdot m_e^{o_2}}{m_e \cdot (m_e - 1)}$

4.4 Exceptional Contexts: Evaluation and Pruning

To avoid overloading notation and for the sake of simplicity, from now on we omit the exponent g if no confusion can arise, while keeping in mind a selected group of individuals $u \in \mathscr{D}_I$ related to a subset $G_I^u \subseteq G_I$.

4.4.1 Gauging Exceptionality of a Subgroup

To evaluate the extent to which our findings are exceptional, we follow the significant pattern mining paradigm. That is, we consider each context c as a hypothesis test which returns a *p-value*. The *p-value* is the probability of obtaining an intra-agreement at least as extreme as the one observed over the current context $A(G_E^c)$, assuming the truth of the null hypothesis H_0. The pattern is accepted if H_0 is rejected. This happens if the *p-value* is under a critical significance value α which amounts to test if the observed intra-agreement $A(G_E^c)$ is outside the confidence interval $\text{CI}^{1-\alpha}$ established using the distribution assumed under H_0.

H_0 corresponds to the baseline finding: the observed contextual intra-agreement is generated by the distribution of random subsets equally likely to occur, a.k.a. *Distribution of False Discoveries* (DFD, cf. (Duivesteijn and Knobbe, 2011)). We evaluate the *p-value* of the observed A against the distribution of random subsets of a cardinality equal to the size of the observed subgroup G_E^c. The subsets are issued by uniform sampling without replacement from the entire entities collection. The rationale behind using sampling without replacement is that the observed subgroup does not contain multiple instances of the same entity. Moreover, drawing samples only from the collection of subsets of size equal to $|G_E^c|$ allows to drive more judicious conclusions: the variability of the statistic A is impacted by the size of the considered subgroups, since smaller subgroups are more likely to observe low/high values of A. The same reasoning was followed in (Lemmerich et al., 2016).

We define $\theta_k : F_k \to \mathbb{R}$ as the random variable corresponding to the observed intra-agreement A of k-sized subsets $S \in G_E$. I.e., for any $k \in [1,n]$ with $n = |G_E|$, we have $\theta_k(S) = A(S)$ and $F_k = \{S \in G_E \ s.t. \ |S| = k\}$. F_k is then the set of possible subsets which are equally likely to occur under the null hypothesis H_0. That is, $\mathbb{P}(S \in F_k) = \binom{n}{k}^{-1}$. We denote by $CI_k^{1-\alpha}$ the $(1-\alpha)$ confidence interval related to the probability distribution of θ_k under the null hypothesis H_0. To easily manipulate θ_k, we reformulate A using Equations (4.1)-(4.5):

$$A(S) = \frac{\sum_{e \in S} v_e}{\sum_{e \in S} w_e} \mid w_e = m_e \text{ and } v_e = m_e - \frac{1}{D_{\text{exp}}} \sum_{o_1, o_2 \in O^2} \Delta_{o_1 o_2} \cdot \frac{m_e^{o_1} \cdot m_e^{o_2}}{(m_e - 1)} \tag{4.6}$$

Under the null hypothesis H_0 and the assumption that the underlying distribution of intra-agreements is a Normal distribution[2] $\mathcal{N}(\mu_k, \sigma_k^2)$, one can define $CI_k^{1-\alpha}$ by computing $\mu_k = E[\theta_k]$ and $\sigma_k^2 = \text{Var}[\theta_k]$. Doing so requires either empirically calculating estimators of such moments by drawing a large number r of uniformly generated samples from F_k, or analytically deriving the formula of $E[\theta_k]$ and $\text{Var}[\theta_k]$. In the former case, the confidence interval $CI_k^{1-\alpha}$ endpoints are given by Geisser, 1993, p.9: $\mu_k \pm t_{1-\frac{\alpha}{2}, r-1} \sigma_k \sqrt{1 + (1/r)}$, with μ_k and σ_k empirically estimated on the r samples, and $t_{1-\frac{\alpha}{2}, r-1}$ the $(1-\frac{\alpha}{2})$ percentile of Student's t-distribution with $r-1$ degrees of freedom. In the latter case, (μ_k and σ_k are known/derived analytically), the $(1-\alpha)$ confidence interval can be computed in its most basic form, that is $CI_k^{1-\alpha} = [\mu_k - z_{(1-\frac{\alpha}{2})}\sigma_k, \mu_k + z_{(1-\frac{\alpha}{2})}\sigma_k]$ with $z_{(1-\frac{\alpha}{2})}$ the $(1-\frac{\alpha}{2})$ percentile of $\mathcal{N}(0,1)$.

However, due to the problem setting, empirically establishing the confidence interval is computationally expensive, since it must be calculated for each enumerated context. Even for relatively small behavioral datasets, this quickly becomes intractable. Alternatively, analytically deriving a computationally efficient form of $E[\theta_k]$ is notoriously difficult, given that $E[\theta_k] = \binom{n}{k}^{-1} \sum_{S \in F_k} \frac{\sum_{e \in S} v_e}{\sum_{e \in S} w_e}$ and $\text{Var}[\theta_k] = \binom{n}{k}^{-1} \sum_{S \in F_k} \left(\frac{\sum_{e \in S} v_e}{\sum_{e \in S} w_e} - E[\theta_k] \right)^2$.

Since θ_k can be seen as a weighted arithmetic mean, one can model the random variable θ_k as the ratio $\frac{V_k}{W_k}$, where V_k and W_k are two random variables $V_k : F_k \to \mathbb{R}$ and $W_k : F_k \to \mathbb{R}$

[2] In the same line of reasoning of (Bie, 2011a; Lijffijt et al., 2018), one can assume that the underlying distribution can be derived from what prior beliefs the end-user may have on such distribution. If only the observed expectation μ and variance σ^2 are given as constraints which must hold for the underlying distribution, the maximum entropy distribution (*taking into account no other prior information than the given constraints*) is known to be the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ (Cover and Thomas, 2012, p.413).

with $V_k(S) = \frac{1}{k}\sum_{e\in S} v_e$ and $W_k(S) = \frac{1}{k}\sum_{e\in S} w_e$. An elegant way to deal with a ratio of two random variables is to approximate its moments using the *Taylor series* following the line of reasoning of (Duris et al., 2018) and (Kendall, Stuart, and Ord, 1994, p.351), since no easy analytic expression of $E[\theta_k]$ and $\mathrm{Var}[\theta_k]$ can be derived.

Proposition 4.4.1 — An Approximate Confidence Interval $\widehat{CI}_k^{1-\alpha}$ for θ_k. Given $k \in [1,n]$ and $\alpha \in]0,1]$ (significance critical value), $\widehat{CI}_k^{1-\alpha}$ is given by:

$$\widehat{CI}_k^{1-\alpha} = \left[\widehat{E}[\theta_k] - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}[\theta_k]}, \widehat{E}[\theta_k] + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}[\theta_k]}\right] \tag{4.7}$$

with $\widehat{E}[\theta_k]$ a Taylor approximation for the expectation $E[\theta_k]$ expanded around (μ_{V_k}, μ_{W_k}), and $\widehat{\mathrm{Var}}[\theta_k]$ a Taylor approximation for $\mathrm{Var}[\theta_k]$ given by:

$$\widehat{E}[\theta_k] = \left(\frac{n}{k} - 1\right)\frac{\mu_v}{\mu_w}\beta_w + \frac{\mu_v}{\mu_w} \qquad \widehat{\mathrm{Var}}[\theta_k] = \left(\frac{n}{k} - 1\right)\frac{\mu_v^2}{\mu_w^2}(\beta_v + \beta_w) \tag{4.8}$$

with:

$$\mu_v = \frac{1}{n}\sum_{e\in G_E} v_e \qquad \mu_w = \frac{1}{n}\sum_{e\in G_E} w_e \qquad n = |G_E|$$

$$\mu_{v^2} = \frac{1}{n}\sum_{e\in G_E} v_e^2 \qquad \mu_{w^2} = \frac{1}{n}\sum_{e\in G_E} w_e^2 \qquad \mu_{vw} = \frac{1}{n}\sum_{e\in G_E} v_e w_e$$

and:

$$\beta_v = \frac{1}{n-1}\left(\frac{\mu_{v^2}}{\mu_v^2} - \frac{\mu_{vw}}{\mu_v\mu_w}\right) \qquad \beta_w = \frac{1}{n-1}\left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v\mu_w}\right)$$

Note that the complexity of the computation of the approximate confidence interval $\widehat{CI}_k^{1-\alpha}$ is $\mathscr{O}(n)$, with n the size of entities collection G_E.

Proof (proposition 4.4.1). For any $f(x,y)$, the bivariate second order Taylor expansion about any $\lambda = (\lambda_x;\lambda_y)$ is:[3]

$$\begin{aligned} f(x,y) &= f(\lambda) + f'_x(\lambda)(x-\lambda_x) + f'_y(\lambda)(y-\lambda_y) \\ &+ \frac{1}{2}\left(f''_{xx}(\lambda)(x-\lambda_x)^2 + 2f''_{xy}(\lambda)(x-\lambda_x)(y-\lambda_y) + f''_{yy}(\lambda)(y-\lambda_y)^2\right) + \varepsilon \end{aligned} \tag{4.9}$$

where ε is a remainder of smaller order than the term of the equation.
An approximation of the expectation $E[f(x,y)]$ expanded around $\lambda = (\lambda_x;\lambda_y)$ is:

$$E[f(x,y)] \approx f(\lambda) + \frac{1}{2}\left[f''_{xx}(\lambda)\mathrm{Var}[X] + 2f''_{xy}(\lambda)\mathrm{Cov}[X,Y] + f''_{yy}(\lambda)\mathrm{Var}[Y]\right]$$

Given that $f(x,y) = \frac{x}{y}$ and using the fact that $E[X-\mu_x] = 0$ (which is valid for both V and W), we have: $\mathrm{Var}[X] = E[(X-\mu_x)^2]$ and $\mathrm{Cov}[X,Y] = (X-\mu_x)(Y-\mu_y)$. We can derive an approximation of $E[\theta_k] = E[\frac{V_k}{W_k}]$ around (μ_{V_k}, μ_{W_k}):

$$E[\theta_k] = E[\frac{V_k}{W_k}] = E[f(V_k,W_k)] \approx \frac{\mu_{V_k}}{\mu_{W_k}} - \frac{\mathrm{Cov}[V_k,W_k]}{\mu_{W_k}^2} + \frac{\mathrm{Var}[W_k]\mu_{V_k}}{\mu_{V_k}^3} \tag{4.10}$$

[3] a concise lecture note follows the same reasoning and explains the derivations; see http://www.stat.cmu.edu/~hseltman/files/ratio.pdf

The formulas of $E[V_k]$ (resp. $E[W_k]$) and $\mathrm{Var}[V_k]$ (resp. $V[W_k]$) can be derived analytically. We denote by μ_v (resp. μ_w) the arithmetic mean of the values (resp. weights) corresponding to each entity $e \in G_E$, i.e.: $\mu_v = \frac{1}{n}\sum_{e\in G_E} v_e$ and $\mu_w = \frac{1}{n}\sum_{e\in G_E} w_e$ with $n = |G_E|$.

$$E[V_k] = \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \frac{1}{k}\sum_{e\in S} v_e = \frac{1}{n}\sum_{e\in G_E} v_e = \mu_v \tag{4.11}$$

$$\begin{aligned}\mathrm{Var}[V_k] &= \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - E[V_k]\right)^2 = \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - \mu_v\right)^2 \\ &= \frac{1}{k}\left(\frac{n}{n-1}\left(\mu_{v^2} - \mu_v^2\right)\right) - \frac{1}{n-1}\left(\mu_{v^2} - \mu_v^2\right) \text{ with } \mu_{v^2} = \frac{1}{n}\sum_{e\in G_E} v_e^2\end{aligned} \tag{4.12}$$

The same reasoning applies to compute the expected value and the variance related to W_k:

$$E[W_k] = \frac{1}{n}\sum_{e\in G_E} w_e = \mu_w \tag{4.13}$$

$$\begin{aligned}\mathrm{Var}[W_k] &= \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} w_e - E[W_k]\right)^2 \\ &= \frac{1}{k}\left(\frac{n}{n-1}\left(\mu_{w^2} - \mu_w^2\right)\right) - \frac{1}{n-1}\left(\mu_{w^2} - \mu_w^2\right) \text{ with } \mu_{w^2} = \frac{1}{n}\sum_{e\in G_E} w_e^2\end{aligned} \tag{4.14}$$

We now derive the formula for $\mathrm{Cov}(V_k, W_k)$. The same line of reasoning for the computation of the variance of V_k and W_k applies. We obtain:

$$\begin{aligned}\mathrm{Cov}[V_k, W_k] &= \frac{1}{\binom{n}{k}} \sum_{S\in F_k} \left(\frac{1}{k}\sum_{e\in S} v_e - E[V_k]\right)\left(\frac{1}{k}\sum_{e\in S} w_e - E[W_k]\right) \\ &= \frac{1}{k}\left(\frac{n}{n-1}\left(\mu_{vw} - \mu_v\mu_w\right)\right) - \frac{1}{n-1}\left(\mu_{vw} - \mu_v\mu_w\right) \\ &\text{with } \mu_{vw} = \frac{1}{n}\sum_{e\in G_E} w_e v_e\end{aligned} \tag{4.15}$$

Using Equations *(4.11)*, *(4.12)*, *(4.13)*, *(4.14)*, *(4.15)*, we derive the approximation of $E[\theta_k]$ after simplifications of *(4.10)*:

$$E[\theta_k] \approx \widehat{E}[\theta_k] = \left(\frac{n}{k} - 1\right)\frac{\mu_v}{\mu_w}\beta_w + \frac{\mu_v}{\mu_w} \text{ with } \beta_w = \frac{1}{n-1}\left(\frac{\mu_{w^2}}{\mu_w^2} - \frac{\mu_{vw}}{\mu_v\mu_w}\right) \tag{4.16}$$

The same reasoning applies to approximate $\mathrm{Var}[\theta_k]$ using Taylor expansions. We will confine ourselves to a first-order Taylor expansion around (μ_v, μ_w) to make the analytic derivation of the approximation of $\mathrm{Var}[\theta_k]$ feasible. The same observation has been made by (Duris et al., 2018; Kempen and Vliet, 2000) and (Kendall, Stuart, and Ord, 1994, p. 351) to approximate the variance for a ratio random variable. We obtain:

$$\mathrm{Var}[\theta_k] = \mathrm{Var}[f(V_k, W_k)] \approx \frac{\mathrm{Var}[V_k]}{\mu_{W_k}^2} - 2\frac{\mu_{V_k}\mathrm{Cov}[V_k, W_k]}{\mu_{W_k}^3} + \frac{\mu_{V_k}^2\mathrm{Var}[W_k]}{\mu_{W_k}^4} \tag{4.17}$$

After simplifications and by using the same line of reasoning when deriving the expected value approximation reported in Equation (4.16), we obtain:

$$\begin{aligned}\mathrm{Var}[\theta_k] \approx \widehat{\mathrm{Var}}[\theta_k] = \left(\frac{n}{k}-1\right)\frac{\mu_v^2}{\mu_w^2}(\beta_v+\beta_w)\\ \text{with } \beta_w = \frac{1}{n-1}\left(\frac{\mu_{w^2}}{\mu_w^2}-\frac{\mu_{vw}}{\mu_v\mu_w}\right) \text{ and } \beta_v = \frac{1}{n-1}\left(\frac{\mu_{v^2}}{\mu_v^2}-\frac{\mu_{vw}}{\mu_v\mu_w}\right)\end{aligned} \tag{4.18}$$

We denote by $\widehat{CI}_k^{1-\alpha}$ the approximate confidence interval calculated using the approximations from Equations (4.16) and (4.18) of the expected value $\widehat{E}[\theta_k]$ and the variance $\widehat{\mathrm{Var}}[\theta_k]$, respectively. This results in:

$$\widehat{CI}_k^{1-\alpha} = \left[\widehat{E}[\theta_k] - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}[\theta_k]}, \widehat{E}[\theta_k] + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}[\theta_k]}\right]$$

It is worth mentioning that the complexity of the computation of this approximate confidence interval is linear to the size n. ■

4.4.2 Pruning the Search Space

Before introducing formally the main components required for pruning areas of the search space, we give an overview of how these components are leveraged for safe-pruning unpromising branches (cf. Figure 4.2).

Suppose that we are interested in subgroups of entities (context) whose sizes are greater than a support threshold σ. Recall that the exceptionality of a given subgroup of size $X \geq \sigma$, by its *p-value*: the probability that for a random subset of entities, we observe

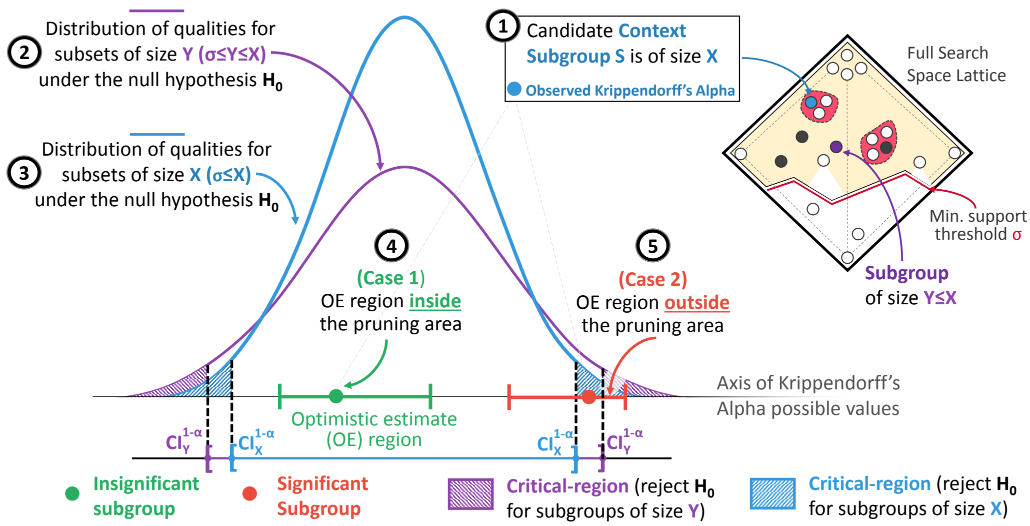


Figure 4.2: Main DEvIANT properties for safe sub-search space pruning. A subgroup is reported as significant if its related Krippendorff's Alpha falls in the critical region of the corresponding empirical distribution of random subsets (DFD). When traversing the search space downward (decreasing support size), the approximate confidence intervals are nested. If the optimistic estimates region falls into the confidence interval computed on the related DFD, the sub-search space can be safely pruned.

an intra-agreement at least as extreme as the one observed for the subgroup. To achieve this, we estimate the empirical distribution of the intra-agreement of random subsets (DFD: Distribution of False Discoveries) for a chosen critical value α, a confidence interval $CI_X^{1-\alpha}$ over the corresponding distribution under the null hypothesis. If the subgroup intra-agreement is outside $CI_X^{1-\alpha}$, the subgroup is statistically significant (*p-value* $\leq \alpha$); otherwise the subgroup is a spurious finding. In section 4.4.2.1, we compute a tight optimistic estimate (OE) (Grosskreutz, Rüping, and Wrobel, 2008) to define a lower and upper bounds of Krippendorff's Alpha for any specialization of a subgroup having its size greater than σ. In section 4.4.2.2, we prove that the analytic approximate confidence intervals are nested: $\sigma \leq Y \leq X \Rightarrow \widehat{CI}_X^{1-\alpha} \subseteq \widehat{CI}_Y^{1-\alpha}$ (i.e., when the support size grows, the confidence interval shrinks). Combining these properties, if the OE region falls into the corresponding CI, we can safely prune large parts of the search space that do not contain significant subgroups. The latter point is discussed in the concluding section 4.4.2.3.

4.4.2.1 Optimistic Estimate on Krippendorff's Alpha (A)

To quickly prune unpromising areas of the search space, we define a tight optimistic estimate (Grosskreutz, Rüping, and Wrobel, 2008) on Krippendorff's alpha. Eppstein and Hirschberg, 1997 propose a smart *linear algorithm* `Random-SMWA`[4] to find subsets with maximum weighted average. Recall that A can be seen as a weighted average (cf. Equation (4.6)).

In a nutshell, `Random-SMWA` seeks to remove k values to find a subset of S having $|S|-k$ values with maximum weighted average. The authors model the problem as such: given $|S|$ values decreasing linearly with time, find the time at which the $|S|-k$ maximum values add to zero. In the scope of this work, given a user-defined support threshold σ_E on the minimum allowed size of context extents, k is fixed to $|S|-\sigma_E$. The obtained subset corresponds to the smallest allowed subset having support $\geq \sigma_E$ maximizing the weighted average quantity A. The `Random-SMWA` algorithm can be tweaked[5] to retrieve the smallest subset of size $\geq \sigma_E$ having analogously the minimum possible weighted average quantity A. We refer to the algorithm returning the maximum (resp. minimum) possible weighted average by $\texttt{RandomSMWA}^{\texttt{max}}$ (resp. $\texttt{RandomSMWA}^{\texttt{min}}$).

Proposition 4.4.2 — Upper and Lower Bounds for A. Given $S \subseteq G_E$, minimum context support threshold σ_E, and the following functions:

$$UB(S) = A\left(\texttt{RandomSMWA}^{\texttt{max}}(S,\sigma_E)\right) \qquad LB(S) = A\left(\texttt{RandomSMWA}^{\texttt{min}}(S,\sigma_E)\right)$$

we know that LB (resp. UB) is a lower (resp. upper) bound for A, i.e.:

$$\forall c,d \in \mathscr{D}_E \;:\; c \sqsubseteq d \,\wedge\, |G_E^c| \geq |G_E^d| \geq \sigma_E \Rightarrow LB(G_E^c) \leq A(G_E^d) \leq UB(G_E^c)$$

Before giving the proof of Proposition 4.4.2, we present the following lemma:

[4] `Random-SMWA`: Randomized algorithm - Subset with Maximum Weighted Average.

[5] Finding the subset having the minimum weighted average is a dual problem to finding the subset having the maximum weighted average. To solve the former problem using `Random-SMWA`, we modify the values of v_i to $-v_i$ and keep the same weights w_i.

Lemma 4.4.3 Let $n \in \mathbb{N}^*$, $A = \{a_i\}_{1 \leq j \leq n}$ and $B = \{b_i\}_{1 \leq j \leq n}$ such that: $\forall i \in 1..n : b_j \geq 0$

$$\text{Given } M = \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j} \text{ and } M(i) = \frac{\sum_{j=1;j\neq i}^{n} a_j}{\sum_{j=1;j\neq i}^{n} b_j}$$

$$\text{we have: } \exists i \in 1..n \text{ s.t. } M(i) \ \geq \ M$$

Informally, their exists always an element i that can be removed to increase the function M (weighted arithmetic mean).

Proof (lemma 4.4.3). This can be proved by reductio ad absurdum. Assume that $\forall i \in 1..n$: $M(i) < M$. Given that:

$$\begin{aligned}
\forall i \in 1..n \quad &: M &&= \left(1 - \frac{b_i}{\sum_1^n b_j}\right) M(i) + \frac{a_i}{\sum_1^n b_j} \\
\forall i \in 1..n \quad &: M(i) - M &&= \frac{1}{\sum_1^n b_j}(b_i M(i) - a_i)
\end{aligned}$$

Provided that, $\forall i \in 1..n$ we have: $M(i) - M < 0$. It follows that for any $i \in 1..n$, we have:

$$\begin{aligned}
b_i M(i) - a_i &< 0 && \Rightarrow \\
M(i) &< \frac{a_i}{b_i} && \Rightarrow \\
\frac{\sum_{j=1}^{n} a_j - a_i}{\sum_{j=1}^{n} b_j - b_i} &< \frac{a_i}{b_i} && \Rightarrow \\
\frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} b_j} &< \frac{a_i}{b_i} && \Rightarrow \\
M &< \frac{a_i}{b_i} && \Rightarrow \\
b_i M &< a_i && \Rightarrow \\
\sum_1^n b_i M &< \sum_1^n a_i && \Rightarrow \\
M &< \frac{\sum_1^n a_i}{\sum_1^n b_i} && \Rightarrow \\
M &< M && \text{which is absurd.}
\end{aligned}$$

■

Hence, given the results of lemma 4.4.3, we know that we can always find an element e to remove from S so as to increase the weighted average quantity A. It follows that, the subset $S_{\max} \subseteq S$ having its support $\geq \sigma_E$ maximizing the weighted average quantity belongs to the minimal frontier, i.e. $|S_{\max}| = \sigma_E$. Such subset is returned by $\texttt{RandomSMWA}^{\texttt{max}}$ as proved by Eppstein and Hirschberg, 1997.

Proof (proposition 4.4.2). To simplify the text, we will omit σ_E as a parameter in the proof and keep in mind that we consider the minimum support threshold σ_E. Given that $c \sqsubseteq d$, with c, d two descriptions from $\mathscr{D}$, we have $G_E^d \subseteq G_E^c$. The proposition stems from the fact that:

1. $A(G_E^c) \leq UB(G_E^c)$, since `RandomSMWA`$^{\mathtt{max}}$ computes the subset $S^c_{\max}$ having the maximum weighted average A as proven by Eppstein and Hirschberg, 1997.

2. UB is monotonic w.r.t. the partial order $\subseteq$ between sets. That is:

$$\forall S, S' \subseteq G_E \ : \ S' \subseteq S \Rightarrow UB(S') \leq UB(S)$$

This can be proved by reductio ad absurdum. We denote by $S'_{\max} \subseteq S'$ (resp. $S_{\max} \subseteq S$) the optimal subset of S' (resp. S) having its size $\geq \sigma_E$ and the maximum possible weighted average A. Suppose that $\exists S, S' \subseteq G_E \ : \ S' \subseteq S \wedge UB(S') > UB(S)$ $(A(S'_{\max}) > A(S_{\max}))$. Since $S' \subseteq S$, this means that there is another subset in S, namely $S'_{\max}$, that observes a greater weighted average A than the actual optimal subset $S_{\max}$, which is absurd.

From properties 1. and 2. we have: $A(G_E^d) \leq UB(G_E^d) \leq UB(G_E^c)$. The same reasoning holds to prove that LB is a lower bound.

■

Using these results, we define the optimistic estimate for A as an interval bounded by the minimum and the maximum A measure that one can observe from the subsets of a given subset $S \subseteq G_E$, that is: $OE(S, \sigma_E) = [LB(S), UB(S)]$.

4.4.2.2 Nested Confidence Intervals for Krippendorff's Alpha (A)

The desired property between two confidence intervals of the same significance level α related to respectively k_1, k_2 with $k_1 \leq k_2$ is that $CI_{k_1}^{1-\alpha}$ encompasses $CI_{k_2}^{1-\alpha}$. Colloquially speaking, larger samples lead to "narrower" confidence intervals. This property is intuitively plausible, since the dispersion of the observed intra-agreement for smaller samples is likely to be higher than the dispersion for larger samples. Having such a property allows to prune the search subspace related to a context c when traversing the search space downward if $OE(G_E^c, \sigma_E) \subseteq CI^{1-\alpha}_{|G_E^c|}$.

Proving $CI_{k_2}^{1-\alpha} \subseteq CI_{k_1}^{1-\alpha}$ for $k_1 \leq k_2$ for the exact confidence interval is nontrivial, since it requires to analytically derive $E[\theta_k]$ and $\mathrm{Var}[\theta_k]$ for any $1 \leq k \leq n$. Note that the expected value $E[\theta_k]$ varies when k varies. We study such a property for the approximate confidence interval $\widehat{CI}_k^{1-\alpha}$.

Proposition 4.4.4 — Minimum Cardinality Constraint for Nested Approximate Confidence Intervals. Given a context support threshold σ_E and α.

$$\text{If } \sigma_E \geq C^\alpha = \frac{4n\beta_w^2}{z^2_{1-\frac{\alpha}{2}}(\beta_v + \beta_w) + 4\beta_w^2},$$

$$\text{then } \forall k_1, k_2 \in \mathbb{N} : \sigma_E \leq k_1 \leq k_2 \Rightarrow \widehat{CI}_{k_2}^{1-\alpha} \subseteq \widehat{CI}_{k_1}^{1-\alpha}$$

Proof (proposition 4.4.4). In order to prove the desired property for the approximate confidence intervals, we first must determine if the variance decreases when k increases.

$$k_1, k_2 \in \mathbb{N} : \text{ if } k_1 \leq k_2 \Rightarrow \widehat{\mathrm{Var}}[\theta_{k_1}] \geq \widehat{\mathrm{Var}}[\theta_{k_2}] \tag{4.19}$$

From Equation (4.18), $\widehat{\mathrm{Var}}[\theta_k] = \left(\frac{n}{k} - 1\right) \frac{\mu_v^2}{\mu_w^2} (\beta_v + \beta_w)$. Given that $\frac{n}{k} - 1$ is a decreasing function w.r.t. k, proving Equation (4.19) requires that $\beta_v + \beta_w$ is a positive quantity. This stems from the fact that the original formula of the approximate variance given in Equation (4.17) is positive. This can be proved by a direct application of the Covariance inequality (Mukhopadhyay, 2000, p. 149), which itself is an application of the Cauchy-Schwarz inequality (Steele, 2004). Since $\beta_v + \beta_w$ is of the same sign of Equation (4.18), we have $\beta_v + \beta_w \geq 0$. For the sake of a self-contained proof. We give the proof of this assertion below:

From Equations (4.17) and (4.18), we have: $\beta_v + \beta_w$ is of the same sign of:

$$\frac{\mathrm{Var}[V_k]}{\mu_{V_k}^2} - 2\frac{\mathrm{Cov}[V_k, W_k]}{\mu_{V_k}\mu_{W_k}} + \frac{\mathrm{Var}[W_k]}{\mu_{W_k}^2} \tag{4.20}$$

From the Covariance inequality, we have $\mathrm{Cov}[V_k, W_k] \leq \sigma[V_k]\sigma[W_k]$ with $\sigma^2[V_k] = \mathrm{Var}[V_k]$ and $\sigma^2[W_k] = \mathrm{Var}[W_k]$, hence Equation (4.20) is greater than:

$$\begin{aligned}
&\frac{\sigma^2[V_k]}{\mu_{V_k}^2} - 2\frac{\sigma[V_k]\sigma[W_k]}{\mu_{V_k}\mu_{W_k}} + \frac{\sigma^2[W_k]}{\mu_{W_k}^2} \\
&= \frac{\sigma[V_k]}{\mu_{V_k}}\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right) - \frac{\sigma[W_k]}{\mu_{W_k}}\left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right) \\
&= \left(\frac{\sigma[V_k]}{\mu_{V_k}} - \frac{\sigma[W_k]}{\mu_{W_k}}\right)^2 \\
&\geq 0
\end{aligned}$$

Hence $\beta_v + \beta_w \geq 0$, which confirms that the variance is decreasing under increasing size k, as stated in Equation (4.19).

Recall that, by approximation, we want to ensure that for $\sigma_E \leq k_1 \leq k_2$ with σ_E a threshold on the context support, we have $\widehat{CI}_{k_2}^{1-\alpha} \subseteq \widehat{CI}_{k_1}^{1-\alpha}$. Hence, we need to find the minimum σ_E above which such property is valid. This amounts to finding a lower bound for σ_E such that:

$$z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}[\theta_{k_1}]} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}[\theta_{k_2}]} \geq \left|\widehat{E}[\theta_{k_1}] - \widehat{E}[\theta_{k_2}]\right| \tag{4.21}$$

Using the definitions of $\widehat{\mathrm{Var}}[\theta_k]$ and $\widehat{E}[\theta_k]$ from Equations (4.16) and (4.18), the Equation (4.21) can be rewritten to:

$$\left(\sqrt{\frac{n}{k_1} - 1} + \sqrt{\frac{n}{k_2} - 1}\right) \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\beta_v + \beta_w}{\beta_w^2}}$$

Since $\sigma_E \leq k_1 \leq k_2$, we require that:

$$2\sqrt{\frac{n}{\sigma_E} - 1} \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\beta_v + \beta_w}{\beta_w^2}}$$

After simplifications, we obtain that σ_E must satisfy the following constraint:

$$\sigma_E \geq C^\alpha = \frac{4n\beta_w^2}{z_{1-\frac{\alpha}{2}}^2(\beta_v + \beta_w) + 4\beta_w^2} \tag{4.22}$$

■

4.4.2.3 Pruning branches using Optimistic estimates regions and nested CIs for A

Combining Propositions 4.4.1, 4.4.2 and 4.4.4, we formalize the pruning region property which answers: *when to prune the sub-search space under a context c?*

Corollary 4.4.5 — Pruning Regions. Given a behavioral dataset $\mathscr{B}$, a context support threshold $\sigma_E \geq C^\alpha$, and a significance critical value $\alpha \in]0,1]$. For any $c,d \in \mathscr{D}_E$ such that $c \sqsubseteq d$ with $|G_E^c| \geq |G_E^d| \geq \sigma_E$, we have:

$$OE(G_E^c, \sigma_E) \subseteq \widehat{CI}^{1-\alpha}_{|G_E^c|} \Rightarrow A(G_E^d) \in \widehat{CI}^{1-\alpha}_{|G_E^d|} \Rightarrow \text{p-value}(d) > \alpha$$

Proof (corollary 4.4.5). The proof is straightforward. From Proposition 2, we have that for any $c,d \in \mathscr{D}_E$ s.t. $c \sqsubseteq d$, if $G^c \geq G^d \geq \sigma_E$ then:

$$A\left(G_E^d\right) \in OE\left(G_E^c, \sigma_E\right) \tag{4.23}$$

From Proposition 3, if $\sigma_E \geq C^\alpha$ we have:

$$CI^{1-\alpha}_{|G_E^c|} \subseteq \widehat{CI}^{1-\alpha}_{|G_E^d|} \tag{4.24}$$

From Equations (4.23) and (4.24) and the fact that $OE(G_E^c, \sigma_E) \subseteq \widehat{CI}^{1-\alpha}_{|G_E^c|}$, it follows that $A\left(G_E^d\right) \in OE\left(G_E^c, \sigma_E\right) \subseteq \widehat{CI}^{1-\alpha}_{|G_E^c|} \subseteq \widehat{CI}^{1-\alpha}_{|G_E^d|}$, hence *p-value(d)* $> \alpha$.

■

4.5 On Handling Variability of Outcomes Among Raters

In Section 4.4, we defined the confidence interval $CI^{1-\alpha}$ established over the DFD. By taking into consideration the variability induced by the selection of a subset of entities, such a confidence interval enables to avoid reporting subgroups indicating an intra-agreement likely (w.r.t. the critical value α) to be observed by a random subset of entities. For more statistically sound results, one should not only take into account the variability induced by the selection of subsets of entities, but also the variability induced by the outcomes of the selected group of individuals. This is well summarized by Hayes and Krippendorff, 2007: "The obtained value of A is subject to random sampling variability—specifically variability attributable to the selection of units (i.e., entities) in the reliability data (i.e., behavioral data) and the variability of their judgments". To address these two questions, they recommend to employ a standard Efron & Tibshirani *bootstrapping approach* (Efron and Tibshirani, 1994) to empirically generate the sampling distribution of A and produce an empirical confidence interval $\text{CI}^{1-\alpha}_{\text{bootstrap}}$.

Recall that we consider here a behavioral dataset $\mathscr{B}$ reduced to the outcomes of a selected group of individuals u. Following the bootstrapping scheme proposed by Krippendorff (Hayes and Krippendorff, 2007; Krippendorff, 2004), the empirical confidence interval is computed by repeatedly performing the following steps: (1) resample n entities from G_E with replacement; (2) for each sampled entity, draw uniformly $m_e \cdot (m_e - 1)$ pairs of outcomes according to the distribution of the observed pairs of outcomes; (3) compute the observed disagreement and calculate Krippendorff's alpha on the resulting resample. This process, repeated b times, leads to a vector of bootstrap estimates (sorted in ascending order) $\hat{B} =$

$[\hat{A}_1, \ldots, \hat{A}_b]$. Given the empirical distribution $\hat{B}$, the empirical confidence interval $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}}$ is defined by the percentiles of $\hat{B}$, i.e., $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}} = [\hat{A}_{\lfloor \frac{\alpha}{2} \cdot b \rfloor}, \hat{A}_{\lceil (1-\frac{\alpha}{2}) \cdot b \rceil}]$. We denote by $\mathrm{MCI}^{1-\alpha}$ (Merged CI) the confidence interval that takes into consideration both $CI^{1-\alpha} = [\mathrm{le}_1, \mathrm{re}_1]$ and $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}} = [\mathrm{le}_2, \mathrm{re}_2]$. We have $\mathrm{MCI}^{1-\alpha} = [\min(\mathrm{le}_1, \mathrm{le}_2), \max(\mathrm{re}_1, \mathrm{re}_2)]$.

Bootstrapping is a computationally expensive operation. To speed up such a step, we employ the BLB (Bag of Little Bootstraps) procedure (Kleiner et al., 2012). BLB is a simple technique to implement. In a nutshell, the techniques consists of two major steps: (1) Repeatedly subsample $n' < n$ without replacement from the original dataset of size n (G_E in our setting). (2) For each subsample, perform a standard Efron & Tibshirani bootstrapping approach and compute an estimate of the statistic of interest (the confidence interval endpoints). Finally, the obtained estimates of each subsample are aggregated to output the final estimate of the statistic of interest. Three hyper-parameters are required to be fixed upfront to run BLB: the number of subsamples s, the size of each subsample n' and the number of Monte-Carlo iterations in each bootstrap r'. Kleiner et al., 2012 recommend to have $s \simeq 5$, $n' \simeq n^{0.7}$ and $r' \simeq 50$ to achieve low-relative error compared to standard Bootstrapping . In this work, we have fixed these hyper-parameters as follows: $s = 5$, $n' = n^{0.7}$ and $r' = 80$.

4.6 A Branch-and-bound Solution: Algorithm DEvIANT

We start by recalling how candidate subgroups of individuals (groups) and candidate subgroups of entities (contexts) are enumerated in section 4.6.1. Subsequently, in section 4.6.2, we present algorithm DEvIANT tailored for the discovery of statistically significant exceptional (dis)agreement among groups.

4.6.1 Enumerating Candidate Subgroups

In order to detect exceptional contextual intra-group agreement patterns, we need to enumerate candidates $p = (u,c) \in (\mathscr{D}_I, \mathscr{D}_E)$. Several enumeration algorithms exist in the literature, ranging from heuristic (e.g., beam search (Leeuwen and Knobbe, 2012)) to exhaustive techniques (e.g., GP-growth (Lemmerich, Becker, and Atzmueller, 2012)). In this paper, we exhaustively enumerate all candidate subgroups while leveraging closure operators (Ganter and Kuznetsov, 2001) (since A computation only depends on the extent of a pattern). This makes it possible to avoid redundancy and to substantially reduce the number of visited patterns. With this aim in mind, and since $(G_E, (\mathscr{D}_E, \sqsubseteq), \delta^E)$ and $(G_I, (\mathscr{D}_I, \sqsubseteq), \delta^I)$ are two pattern structures (cf. Definition 2.2.7), we apply EnumCC (Enumerate Closed Candidates) (cf. Algorithm 1) to enumerate subgroups u (resp. c) in $\mathscr{D}_I$ (resp. $\mathscr{D}_E$). Recalls that EnumCC follows the line of the CloseByOne algorithm (Kuznetsov, 1999). Given a collection G of records (which can be either G_E or G_I), EnumCC traverses the search space $(\mathscr{D}, \sqsubseteq)$ (which can be either $\mathscr{D}_E$ or $\mathscr{D}_I$) depth-first and enumerates only once all closed descriptions fulfilling the minimum support constraint σ.

4.6.2 Algorithm DEvIANT

DEvIANT implements an efficient branch-and-bound algorithm to **D**iscover statistically significant **E**xceptional **I**ntra-group **A**greement pa**T**terns while leveraging closure, tight optimistic estimates and pruning properties. It follows the same principles as B&B4SDEMM (cf. Algorithm 2). DEvIANT (Algorithm 7) starts by selecting a group u of individuals. Next, the corresponding behavioral dataset $\mathscr{B}^u$ is established by reducing the original dataset $\mathscr{B}$ to elements concerning solely the individuals comprising G_I^u and entities having at least two outcomes. Subsequently, the bootstrap confidence interval $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}}$ is calculated.

Before searching for exceptional contexts, the minimum context support threshold σ_E is adjusted to $C^\alpha(u)$ (cf. Proposition 4.4.4) if it is lower than $C^\alpha(u)$. While in practice $C^\alpha(u) \ll \sigma_E$, we keep this correction for algorithm soundness. Next, contexts are enumerated by EnumCC. For each candidate context c, the optimistic estimate interval $OE(G_E^c)$ is computed (cf. Proposition 4.4.2). According to Corollary 4.4.5, if $OE(G_E^c, \sigma_E^u) \subseteq \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$, the search subspace under c can be pruned. Otherwise, $A^u(G_E^c)$ is computed and evaluated against $\mathrm{MCI}^{1-\alpha}_{|G_E^c|}$. If $A^u(G_E^c) \notin \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$, then (u,c) is significant and kept in the result set P. To reduce the number of reported patterns, we keep only the most general patterns while ensuring that each significant pattern in $\mathscr{P}$ is represented by a pattern in P. This formally translates to: $\forall p' = (u',c') \in \mathscr{P} \setminus P : p\text{-}value^{u'}(c') \leq \alpha \Rightarrow \exists p = (u,c) \in P$ s.t. $\mathrm{ext}(q) \subseteq \mathrm{ext}(p)$, with $\mathrm{ext}(q = (u',c')) \subseteq \mathrm{ext}(p = (u,c))$ defined by $G_I^{u'} \subseteq G_I^u$ and $G_E^{c'} \subseteq G_E^c$. This is based on the following postulate: the end-user is more interested by exceptional (dis-)agreement within larger groups and/or for larger contexts rather than local exceptional (dis-)agreement. Moreover, the end-user can always refine their analysis to obtain more fine-grained results by re-launching the algorithm starting from a specific context or group.

Algorithm 7: DEvIANT($\mathscr{B}, \sigma_E, \sigma_I, \alpha$)

Inputs : $\mathscr{B} = \langle G_I, G_E, O, o \rangle$ a Behavioral dataset;
σ_E minimum support threshold of a context;
σ_I of minimum support threshold a group;
α critical significance value.

Output: Set of exceptional intra-group agreement patterns P.

$P \leftarrow \{\}$
foreach $(u, G_I^u, cont_u) \in EnumCC(G_I, *, \sigma_I, 0, True)$ **do**
  $G_E(u) = \{e \in G_E \text{ s.t. } m_e(u) \geq 2\}$; // $m_e(u)$: number of individuals of group u who expressed an outcome on e
  $\mathscr{B}^u = \langle G_E(g), G_I^u, O, o \rangle$
  $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}} = [\hat{A}_{\lfloor \frac{\alpha}{2} \cdot b \rfloor}, \hat{A}_{\lceil (1-\frac{\alpha}{2}) \cdot b \rceil}]$; // With $\hat{B} = [\hat{A}_1^u, \ldots, \hat{A}_b^u]$ computed on respectively b resamples of $\mathscr{B}^u$
  $\sigma_E^u = \max(C^\alpha(u), \sigma_E)$
  foreach $(c, G_E^c, cont_c) \in EnumCC(G_E(u), *, \sigma_E^u, 0, True)$ **do**
    $\mathrm{MCI}^{1-\alpha}_{|G_E^c|} = \mathrm{merge}\left(\widehat{CI}^{1-\alpha}_{|G_E^c|}, \mathrm{CI}^{1-\alpha}_{\text{bootstrap}}\right)$
    if $\mathrm{OE}(G_E^c, \sigma_E^u) \subseteq \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$ **then**
      $\mathrm{cont}_c \leftarrow \mathrm{False}$; // Prune the unpromising search space under c
    else if $A^u(G_E^c) \notin \mathrm{MCI}^{1-\alpha}_{|G_E^c|}$ **then**
      $p_{\text{new}} \leftarrow (u,c)$
      if $\nexists p_{\text{old}} \in P$ *s.t.* $\mathrm{ext}(p_{\text{new}}) \subseteq \mathrm{ext}(p_{\text{old}})$ **then**
        $P \leftarrow (P \cup p_{\text{new}}) \setminus \{p_{\text{old}} \in P \mid \mathrm{ext}(p_{\text{old}}) \subseteq \mathrm{ext}(p_{\text{new}})\}$
      $\mathrm{cont}_c \leftarrow \mathrm{False}$; // Prune the sub search space (generality)
return P

4.7 Empirical Study

In this section, we report on both quantitative and qualitative experiments over the implemented algorithms. For reproducibility purposes, source code (in Python) and data are made available in a companion page[6].

4.7.1 Aims and Datasets

The experiments aim to answer the following questions:

- Does DEvIANT provide interpretable patterns?
- How well does the Taylor-approximated CI approach the empirical CI?
- How efficient is the Taylor-approximated CI and the pruning properties?
- Does DEvIANT scale w.r.t. different parameters?

Most of the experiments were carried out on four real-world behavioral datasets whose main characteristics are given in Table 4.3. Each dataset involves entities and individuals described by an HMT (H) attribute together with categorical(C) and numerical(N) ones.

	$\|G_E\|$	$\mathscr{A}_E$ (Items-Scaling)	$\|G_I\|$	$\mathscr{A}_I$ (Items-Scaling)	Outcomes	Sparsity	$C^{0.05}$
EPD8	4704	$1H+1N+1C$ (437)	848	$3C$ (82)	$3.1M$ (C)	78.6%	$\simeq 10^{-6}$
CHUS	17350	$1H+2N$ (307)	1373	$2C$ (261)	$3M$ (C)	31.2%	$\simeq 10^{-4}$
Movielens	1681	$1H+1N$ (161)	943	$3C$ (27)	$100K$ (O)	06.3%	$\simeq 0.065$
Yelp	$127K$	$1H+1C$ (851)	$1M$	$3C$ (6)	$4.15M$ (O)	0.003%	$\simeq 1.14$

Table 4.3: Main characteristics of the behavioral datasets. $C^{0.05}$ represents the minimum context support threshold over which we have nested approximate CI property.

EPD8[7] features voting information of the eighth European Parliament about the 848 members who were elected in 2014 or after. The dataset records $3.1M$ tuples indicating the outcome (For, Against, Abstain) of a member voting during one of the 4704 sessions. Each session is described by its themes (H), a voting date (N) and the organizing committee (C). Individuals are described by a national party (C), a political group (C), an age group (C), a country(C) and additional information about countries (date of accession to the European Union (N) and currency (C)).

CHUS[8] features voting information of the United States House of Representatives about the 1373 members who were elected in between 1991 and 2015. The dataset records $3M$ tuples indicating the outcome (Yea, Nay) of a member voting during one of the 17350 sessions. Each session is described by its topic[9] (H), the session (N) and the year (N). Individuals are described by a political party (C) and a state (C).

Movielens[10] is a movie review dataset (Harper and Konstan, 2016) consisting of $100K$ ratings (ranging from 1 to 5) expressed by 943 users on 1681 movies. A movie is

[6] https://github.com/Adnene93/Deviant
[7] http://parltrack.euwiki.org/, last accessed on 04 October 2018
[8] https://voteview.com/data, last accessed on 09 January 2019
[9] https://www.comparativeagendas.net/
[10] https://grouplens.org/datasets/movielens/100k/

characterized by its genres (H) and a release date (N), while individuals are described with demographic information such as age group (C), gender (C) and occupation (C).

Yelp[11] is a social network dataset featuring individuals who rate (scores ranging from 1 to 5) places (stores, restaurants, clinics) characterized by some categories (H) and a state (C). The dataset originally contains 1M users. We preprocessed the dataset to constitute 18 groups of individuals based on the size of their friends network (C), their seniority (C) in the platform and their account type (e.g., elites or not) (C).

4.7.2 Qualitative Study

In this section, we focus on illustrating some patterns discovered by DEvIANT when carried out on the four behavioral datasets. First, we show how DEvIANT can provide interesting insights when analyzing voting datasets (EU Parliament Dataset and U.S. House of Representatives).

Table 4.4 reports exceptional contexts observed among House Republicans during the 115^{th} Congress. Pattern p_1, illustrated in Figure 4.3, highlights a collection of voting sessions addressing Government and Administrative issues where a clear polarization is observed between two clusters of Republicans. A roll call vote in this context featuring significant disagreement between Republicans is "**House Vote 417**"[12] which was closely watched by the media[13].

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	Republicans	20.11 Government Branch Relations, Admin. Issues, and Constitutional Reforms	0.83	0.32	$<.001$	Conflict
p_2	Republicans	5 Labor	0.83	0.63	$<.01$	Conflict
p_3	Republicans	20.05 Nominations and Appointments	0.83	0.92	$<.001$	Consensus

Table 4.4: Exceptional consensual/conflictual subjects among Republicans Party representatives in the 115^{th} congress of the US House of Representatives. $\alpha = 0.01$

DEvIANT can detect interesting highlights on exceptionally conflictual or consensual topics between parliamenterians. For instance, Table 4.5 reports 10 patterns suggesting such peculiarities between country representatives in the Eighth European Parliament. A valuable pattern that emerges when conducting such a study in the EU parliament voting dataset is Pattern 5. The latter draws attention on an exceptional conflict between Slovakia's Parliamentarians on EU Fundamental rights matters. An interesting news article[14] covers some aspects of an ongoing discussion in the European Parliament about the human right situation in Slovakia. Similarly, one can analyze the cohesion of political groups using DEvIANT, a sample set of patterns is depicted in Table 4.6. It is worth mentioning that recently Krippendorff's Alpha as an intra-group agreement measure was also used to analyze cohesion within political groups (Cherepnalkoski et al., 2016).

[11] https://www.yelp.com/dataset/challenge, last accessed on 25 April 2017
[12] https://projects.propublica.org/represent/votes/115/house/1/417
[13] Washington Post:https://wapo.st/2W32I9c; Reuters:https://reut.rs/2TF0dgV
[14] https://www.dw.com/en/slovakia-has-the-eu-looked-the-other-way-for-too-long/a-43015470

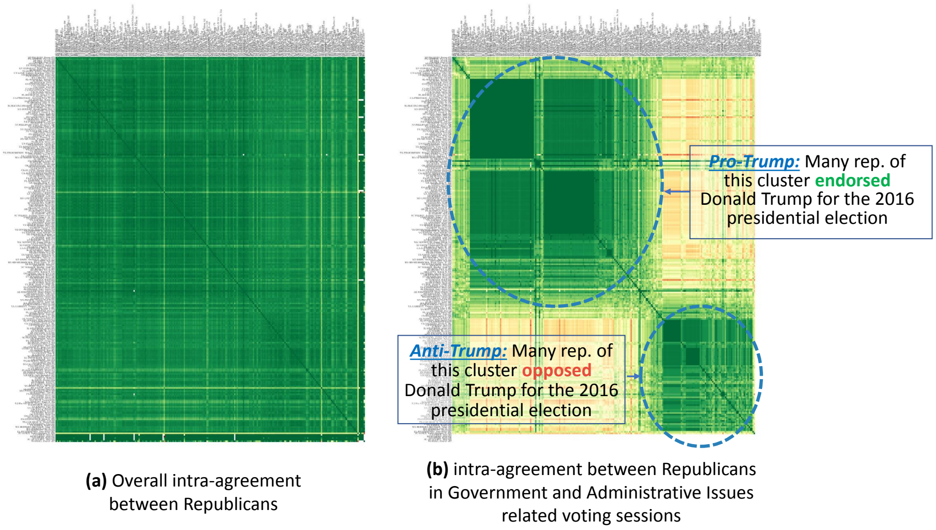


Figure 4.3: Illustrating Pattern 1 from Table 4.4 with a similarity matrix between Republicans. Each cell represents the ratio of voting sessions in which Republicans agreed. Green cells report strong agreement; red cells highlight strong disagreement.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	Sweden	4 Economic, social and territorial cohesion 6.30 Development cooperation	0.3	0.84	<0.0001	Consensus
p_2	Finland	4 Economic, social and territorial cohesion 6.30 Development cooperation	0.36	0.87	<0.0001	Consensus
p_3	Finland	8.20.04 Pre-accession and partnership	0.36	0.75	<0.01	Consensus
p_4	Sweden	8.20 Enlargement of the Union	0.3	0.66	<0.0001	Consensus
p_5	Slovakia	1.10 Fundamental rights in the EU, Charter	0.48	0.13	<0.0001	Conflict
p_6	Malta	4.60.06 Consumers economic and legal interests	0.63	0.97	<0.0001	Consensus
p_7	Malta	2.10 Free movement of goods	0.63	0.34	<0.0001	Conflict
p_8	Latvia	4.60.06 Consumers economic and legal interests	0.42	0.69	<0.0001	Consensus
p_9	Luxembourg	1.20 Citizen's rights, 8 State and evolution of the Union	0.51	0.23	<0.01	Conflict
p_{10}	*	2 Internal market, single market 6 External relations of the Union	0.27	0.54	<0.001	Consensus

Table 4.5: Top-10 exceptional consensual/conflictual subjects among countries' parlementarians in the 8$^{\text{th}}$ EU parliament. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	S&D	8.10 Revision of the Treaties and intergovernmental conferences	0.81	0.44	< 0.001	Conflict
p_2	*	2 Internal market, single market 6 External relations of the Union	0.27	0.54	< 0.001	Consensus
p_3	S&D	8.30 Treaties in general	0.81	0.55	< 0.001	Conflict
p_4	*	2 Internal market, single market, 4.15 Employment policy, act. combat unemployment	0.27	0.53	< 0.001	Consensus
p_5	ALDE	1.20.09 Protection of privacy and data protection 8 State and evolution of the Union	0.73	0.48	< 0.001	Conflict

Table 4.6: Top-5 exceptional consensual/conflictual subjects among European Political Groups in the 8$^{\text{th}}$ EU parliament. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

DEvIANT also enables the discovery of exceptional intra-group (dis)agreement patterns in collaborative rating data. As an example, table 4.7 reports patterns returned by DEvIANT on the Movielens dataset. Pattern p_2 reports that "Middle-aged Men" observe an intra-group agreement significantly higher than overall, for movies labeled with both adventure and musical genres (e.g., The Wizard of Oz (1939)). A similar exceptional (dis)agreement analysis was conducted on Yelp dataset whose results are depicted in Table 4.8.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	Old	1.Action & 2.Adventure & 6.Crime Movies	-0.06	-0.29	< 0.01	Conflict
p_2	Middle-aged Men	2.Adventure & 12.Musical Movies	0.05	0.21	< 0.01	Consensus
p_3	Old	4.Children & 12.Musical Movies	-0.06	-0.21	< 0.01	Conflict

Table 4.7: Top-3 exceptionally consensual/conflictual genres between Movielens raters, α=0.01. Patterns are ranked by absolute difference between $A^g(c)$ and $A^g(*)$.

id	group (g)	context (c)	$A^g(*)$	$A^g(c)$	*p-value*	IA
p_1	*	03 Automotive	0.14	-0.16	<0.0001	Conflict
p_2	*	10 Health & Medical	0.14	-0.14	<0.0001	Conflict
p_3	*	08 Financial Services	0.14	-0.11	<0.0001	Conflict
p_4	newcomer	09.38.07 Health Markets, 09.47 Juice Bars & Smoothies	0.14	-0.07	<0.01	Conflict
p_5	*	El Dorado Hills, California	0.14	0.35	<0.0001	Consensus
p_6	*	14 Local Services	0.14	-0.06	<0.0001	Conflict
p_7	*	04 Beauty & Spas	0.14	-0.06	<0.0001	Conflict
p_8	*	15 Mass Media	0.14	-0.05	<0.01	Conflict
p_9	*	11 Home Services'	0.14	-0.05	<0.0001	Conflict
p_{10}	*	Midlothian, Edinburgh	0.14	0.31	<0.0001	Consensus

Table 4.8: Top-10 exceptional consensual/conflictual places/categories/states among Yelp users. $\alpha = 0.01$. Patterns are ranked by the absolute difference between $A^g(c)$ and $A^g(*)$.

4.7.3 Quantitative Study

Before studying the performance of DEvIANT, we give an overview of the empirical distributions of Krippendorff's Alpha for 1000 draws from F_k equally likely to occur. Recall that F_k represents the subsets of the entire collection of entities of size k over which we define the random variable $\theta_k : F_k \to \mathbb{R}$. Thus, the distributions presented here illustrate the values observed on 1000 trials of θ_k. To illustrate the fact that the confidence intervals

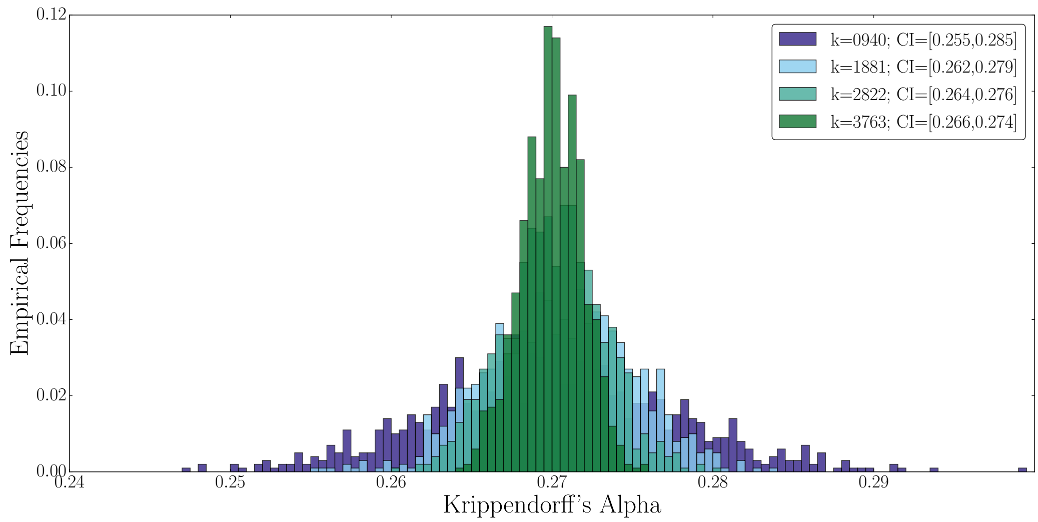


Figure 4.4: Empirical distribution of the observed values of 1000 trials of θ_k for four valuations of k (DFD), experiments were carried on EPD8. We observe that the distributions are encapsulated when k decreases. Also, the dispersion of A increases and the corresponding empirical confidence interval grows in size.

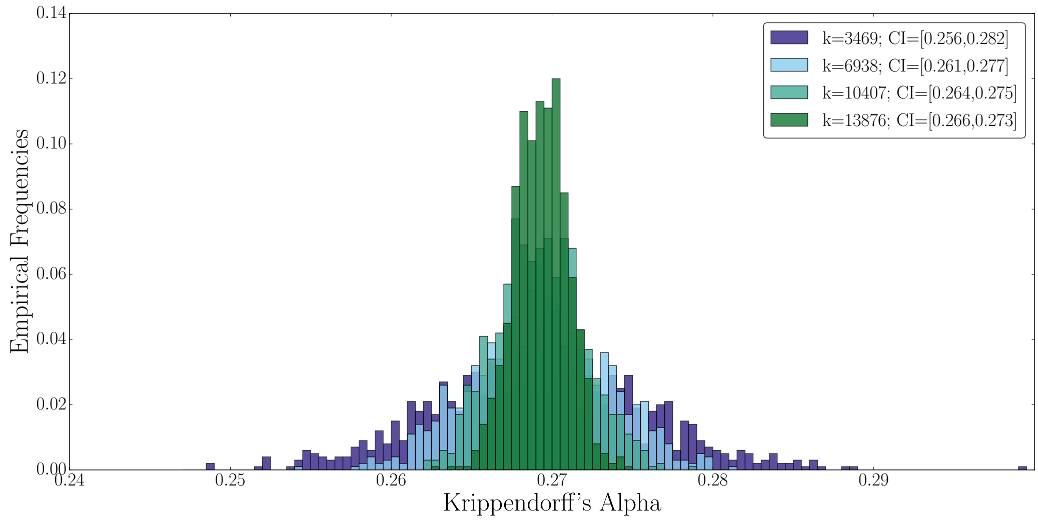


Figure 4.5: Empirical distribution of the observed values of 1000 trials of θ_k for four valuations of k (DFD), experiments were carried on CHUS. We observe that the distributions are encapsulated when k decreases. Also, the dispersion of A increases and the corresponding empirical confidence interval grows in size.

associated with θ_k (considering its distribution under the null hypothesis) are nested (when k grows, the confidence interval shrinks), we perform the experiments for various valuations of k. Figures 4.4, 4.5, 4.6 and 4.7 depict the results of such experiments carried on the four underlying behavioral datasets. We observe that the distributions are bell-shaped and resemble the normal distribution. Moreover, normality test (Shapiro-Wilk-Test (Shapiro and Wilk, 1965)) was not rejected at the $\alpha = 0.05$ for these distributions. It is important to note that empirical confidence intervals are also nested w.r.t. increasing k.

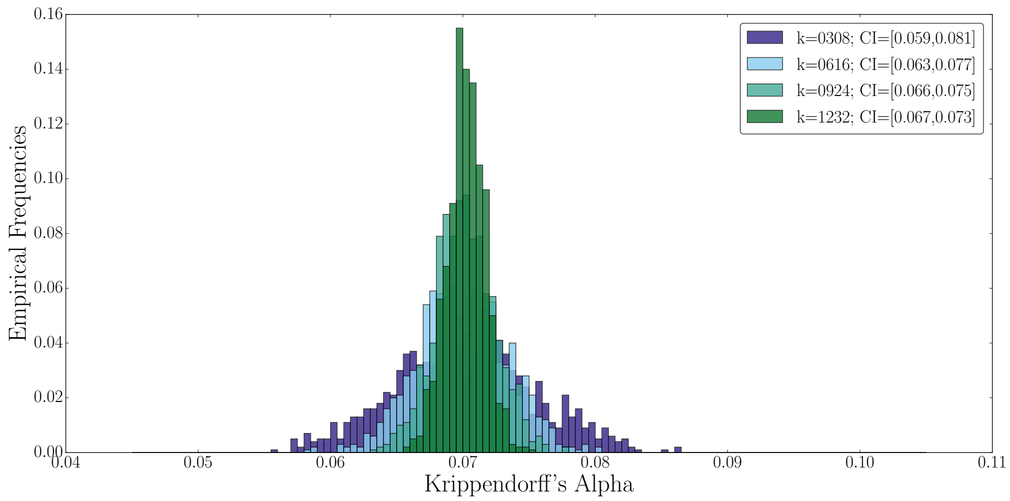


Figure 4.6: Empirical distribution of the observed values of 1000 trials of θ_k for four valuations of k (DFD), experiments were carried on Movielens. We observe that the distributions are encapsulated when k decreases. Also, the dispersion of A increases and the corresponding empirical confidence interval grows in size.

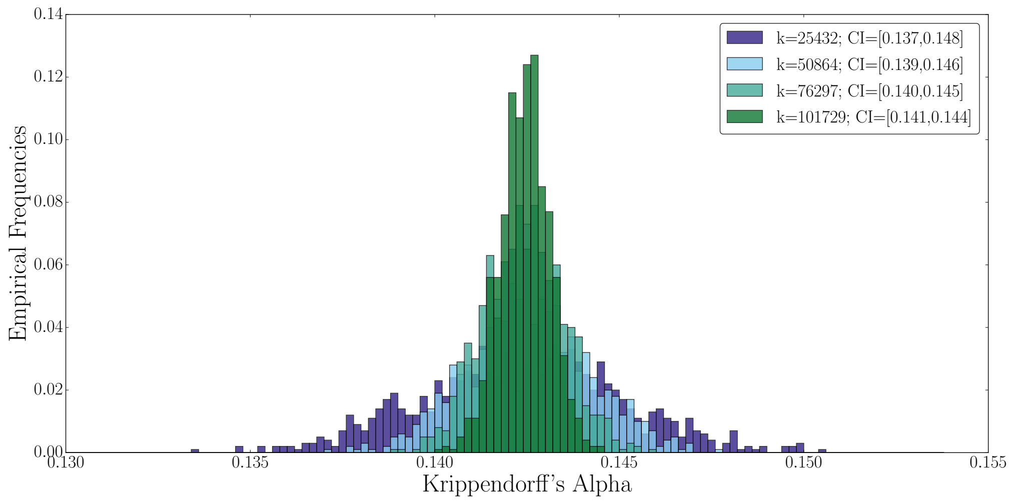


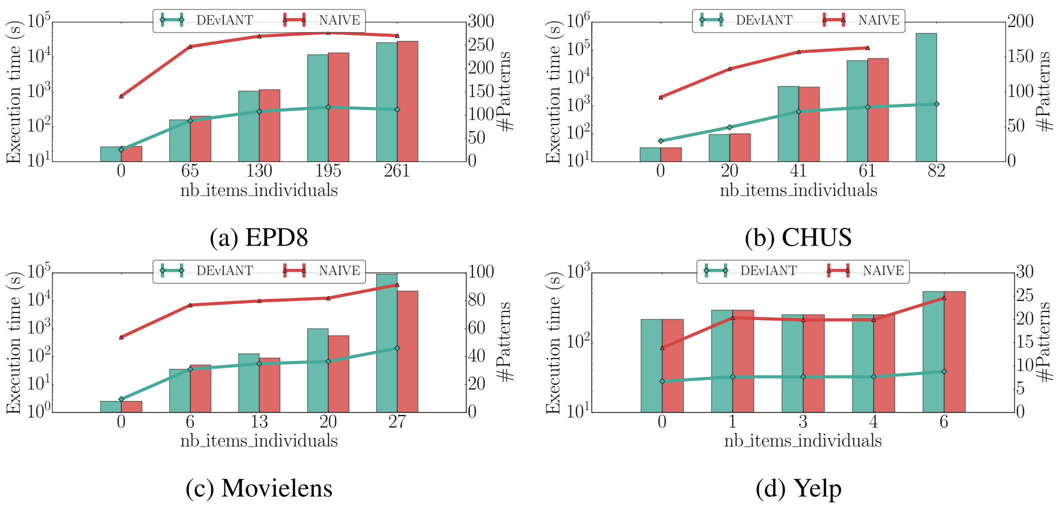
Figure 4.7: Empirical distribution of the observed values of 1000 trials of θ_k for four valuations of k (DFD), experiments were carried on Yelp. We observe that the distributions are encapsulated when k decreases. Also, the dispersion of A increases and the corresponding empirical confidence interval grows in size.

Now, we evaluate to what extent the empirically computed confidence interval approximates the confidence interval computed by Taylor approximations. We run 1000 experiments for subset sizes k uniformly randomly distributed in $[1, n = |G_E|]$. For each k, we compute the corresponding Taylor approximation $\widehat{CI}_k^{1-\alpha} = [a^T, b^T]$ and empirical confidence interval $\mathrm{ECI}_k^{1-\alpha} = [a^E, b^E]$. The latter is calculated over 10^4 samples of size k from G_E, on which we compute the observed A which are then used to estimate the moments of the empirical distribution required for establishing $\mathrm{ECI}_k^{1-\alpha}$. Once both CIs are computed, we measure their distance by Jaccard index, i.e., $\mathrm{dist}(\mathrm{ECI}_k^{1-\alpha}, \widehat{\mathrm{CI}}_k^{1-\alpha}) = 1 - \frac{(\min(b^E, b^T) - \max(a^E, a^T))}{(\max(b^E, b^T) - \min(a^E, a^T))}$. Table 4.9 reports the average μ_{err} and the standard deviation σ_{err} of the observed distances (coverage error) over the 1000 experiments. Note that the difference between the analytic Taylor approximation and the empirical approximation is negligible (μ_{err} is less than 10^-2). Therefore, the CIs approximated by the two methods are so close, that it does not matter which method is used. Hence, the choice is guided by the computational efficiency.

$\mathscr{B}$	μ_{err}	σ_{err}
CHUS	0.007	0.004
EPD8	0.007	0.004
Movielens	0.0075	0.0045
Yelp	0.007	0.004

Table 4.9: Coverage error between empirical CIs and Taylor CIs.

To evaluate the pruning properties' efficiency (**(i)** Taylor-approximated CI, **(ii)** optimistic estimates and **(iii)** nested approximated CIs), we compare DEvIANT with a `Naive` approach where the three aforementioned properties are disabled. For a fair comparison, `Naive` pushes monotonic constraints (minimum support threshold) and employs closure operators while



(a) EPD8 (b) CHUS

(c) Movielens (d) Yelp

Figure 4.8: Comparison between DEvIANT and `Naive` when varying the size of the description space $\mathscr{D}_I$. Lines correspond to the execution time and bars correspond to the number of output patterns. Parameters: $\sigma_E = \sigma_I = 1\%$ and $\alpha = 0.05$.

empirically estimating the CI by successive random trials from F_k. In both algorithms we disable the bootstrap $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}}$ computation, since its overhead is equal for both algorithms. We vary the description space size related to groups of individuals $\mathscr{D}_I$ while considering the full entity description space. Figure 4.8 displays the results: DEvIANT outperforms `Naive` in terms of runtime by nearly two orders of magnitude while outputting the same number of the desired patterns.

Figures 4.9, 4.10, 4.11, and 4.12, report respectively the performance of DEvIANT in terms of runtime and number of output patterns when carried on EPD7, CHUS, Movielens and Yelp datasets. When varying the description space size, DEvIANT requires more time to finish. Note that the size of individuals description space $\mathscr{D}_I$ substantially affects the runtime of DEvIANT. This is mainly because larger $\mathscr{D}_I$ leads to more candidate groups of individuals g which require DEvIANT to: (i) generate $\mathrm{CI}^{1-\alpha}_{\text{bootstrap}}$ and (ii) mine for exceptional contexts c concerning the candidate group g. Also, when α decreases, the execution time required for DEvIANT to finish increases while returning more patterns. This may seem counter-intuitive, since fewer patterns are significant when alpha decreases. It is a consequence of DEvIANT considering only the most general patterns. Hence, when α decreases, DEvIANT goes deeper in the context search space, implying thus much more candidate patterns to be tested and thus a larger result set. Finally, we observe that the bootstrap confidence interval computation induces an overhead by a factor of about 1.5x to 3x, such overhead is mainly impacted by the number of evaluated groups of individuals which is determined by the size of the individuals description space $\mathscr{D}$.

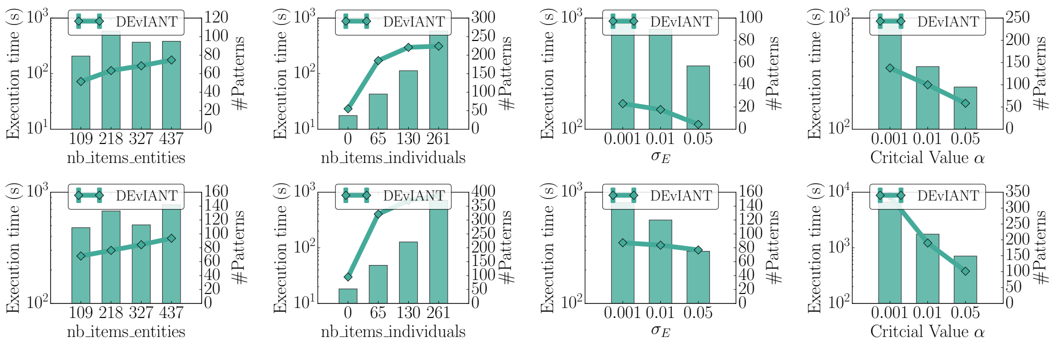


Figure 4.9: Effectiveness of DEvIANT on EPD8 when varying sizes of both description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

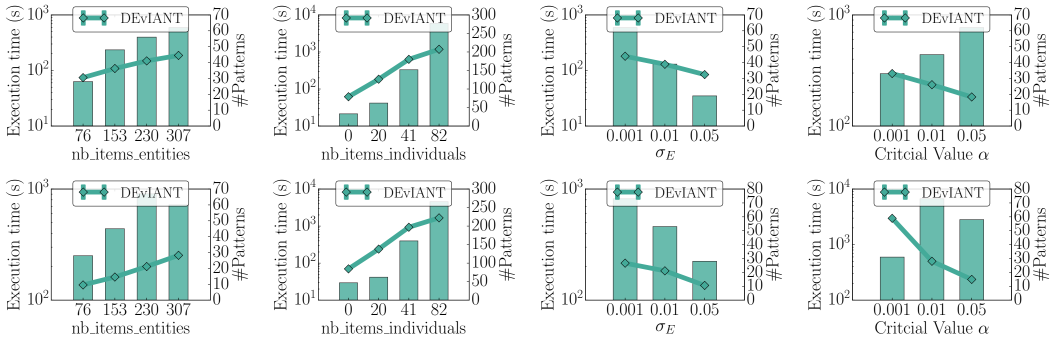


Figure 4.10: Effectiveness of DEvIANT on CHUS when varying sizes of both description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

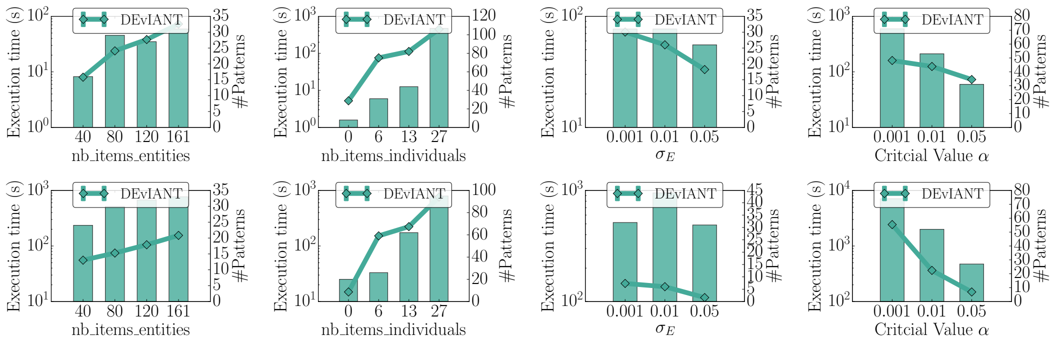


Figure 4.11: Effectiveness of DEvIANT on Movielens when varying sizes of both description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

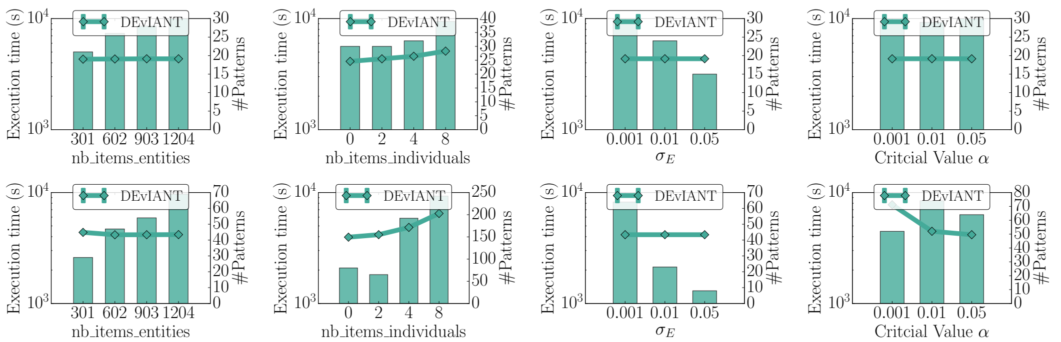


Figure 4.12: Effectiveness of DEvIANT on Yelp when varying sizes of both description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, minimum context support threshold σ_E and the critical value α. Default parameters: full description spaces $\mathscr{D}_E$ and $\mathscr{D}_I$, $\sigma_E = 0.1\%$, $\sigma_I = 1\%$ and $\alpha = 0.05$. Bootstrapping Confidence intervals for handling variability of outcomes is disabled in the figures on the top row, and enabled in the figures on the bottom row.

4.8 Summary

In this chapter, we have defined the problem of discovering exceptional (dis)agreement inside groups in behavioral data and tailored an approach rooted in SD/EMM with a novel pattern domain and associated interestingness measure for the discovery of exceptional intra-group agreement patterns (cf. Figure 4.13). To efficiently search for such patterns, we devise DEvIANT, a branch-and-bound algorithm leveraging closure operators, approximate confidence intervals, tight optimistic estimates on Krippendorff's Alpha measure, and the property of nested CIs. Experiments demonstrate DEvIANT's performance on behavioral datasets in domains ranging from political analysis to rating data analysis.

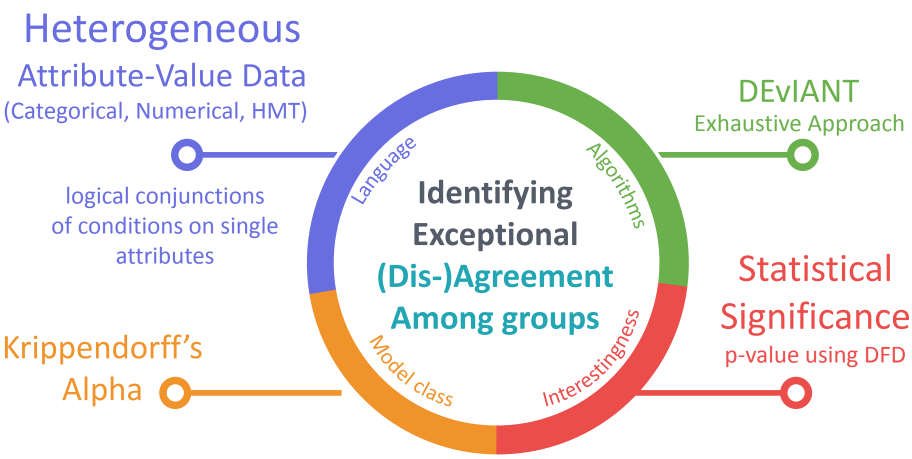


Figure 4.13: Exceptional Model Mining for Identifying exceptional (dis-)agreement among groups (Summary)

In future work, we plan to (i) tackle the multiple comparison problem (MCP[15]) (Hämäläinen and Webb, 2019), (ii) investigate intra-group agreement which is exceptional w.r.t. all individuals *over the same context*, and (iii) integrate the option to choose which kind of exceptional consensus the end-user wants: is the exceptional consensus caused by common preference or dislike for the context-related entities? All this is to be done within a comprehensive framework and tool[16] for behavioral data analysis alongside exceptional inter-group agreement pattern discovery. Such a tool, dubbed ANCORE, is presented and developed in the following chapter.

[15] MCP is a non-trivial task in our setting, and solving it requires an extension of the significant pattern mining paradigm as a whole: its scope is bigger than this work. We provide a brief discussion in Appendix B.

[16] A prototype is available online in http://contentcheck.liris.cnrs.fr

5

Behavioral Data Analysis for Computational Journalism

In this chapter, we motivate the usage of the two proposed approaches, namely DEBuNk and DEvIANT, in the context of computational journalism, where the analysis is conducted on voting data perceived as behavioral data. We introduce ANCORE, a web platform tailored for the discovery of exceptional (dis)agreement within and among groups in voting data. The objective of this tool is to facilitate both fact checking and lead finding tasks. We present several scenarii illustrating its use in data-driven fact checking/lead finding. The web platform is available online on `https://contentcheck.liris.cnrs.fr`.

5.1 Introduction

Hamilton and Turner, 2009 define computational journalism as: "the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism". In the last few years, much efforts have been done by journalists and computer scientists in the development of computational journalism tools and algorithms to assist journalists in the process of investigating the data and fact-check claims. Fact-checking is the act of asserting the correctness of factual claims. Fact-checking has become increasingly common in political journalism which aroused much interest amongst researchers in the computational journalism community. The survey (Cazalens et al., 2018) provides an extensive overview of the recent research in the area. Figure 5.1 depicts, in a brief manner, the different stages of an end-to-end Computational Fact-Checking system.

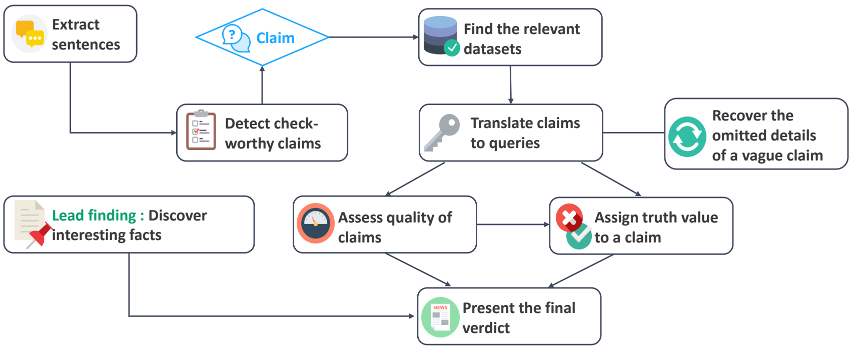


Figure 5.1: Overview of Computational Fact-checking major steps

While the quest of building a fully automated fact-checking framework remains utopian, several works in the state of art tackle different parts of the fact checking process. The different stages of an end-to-end Computational Fact-Checking system depicted in figure 5.1, can be summarized into three major steps. The first step focuses on extracting check-worthy claims from scripted texts (Ennals, Trushkowsky, and Agosta, 2010; Hassan, Li, and Tremayne, 2015; Hassan et al., 2017b). The second step takes as input a claim and searches for relevant datasets (one or more) by relying, for instance, on some underlying knowledge base (Bonaque et al., 2016). The third and last step exploits relevant data to provide perspectives and insights that can be leveraged in the claim quality assessment task (Ciampaglia et al., 2015; Wu, 2015; Wu et al., 2014; Wu et al., 2017). The results can be consolidated to output the final verdict (Ennals, Trushkowsky, and Agosta, 2010; Hassan et al., 2017a). Some projects emerged recently to combine all these components in order to provide an end-to-end fact-checking tool: ClaimBuster (Hassan et al., 2017a), DeFacto (Lehmann et al., 2012) or ClaimChecker (Nguyen et al., 2018), to name a few.

The work presented in this chapter falls within the scope of the third step. Our main objective is to provide a data mining tool that helps putting into perspective some investigated claim in voting data by unraveling insights about exceptional (dis)agreements. This can serve, for instance, to disentangle what is false from what is true by bringing more context to a studied claim which pertains to one category of fake news reported in the typology of figure 5.2. Moreover, our endeavor is to provide a tool which allows also to query voting datasets for interesting facts without having a particular claim in mind to investigate. This falls within the task of computational lead-finding (Wu, 2015), whose main goal is to find interesting information nuggets from raw data that lead to further investigation and/or news stories around them.

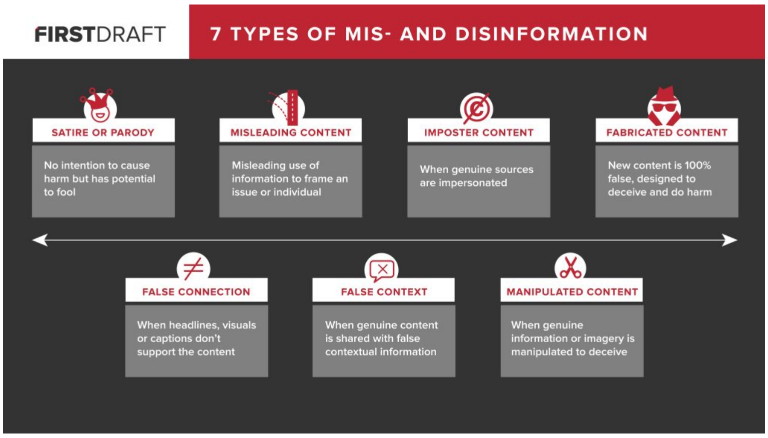


Figure 5.2: Typologie of fake news (source[1])

The web platform, dubbed ANCORE, presented in this chapter can be specifically tailored for putting into perspective *behavioral comparison claims* (BCC). The latter category of claims encompasses any claim stating a comparison of behavior between groups, individuals, countries, populations, etc. Such claims can be investigated by leveraging the contents of some underlying behavioral datasets. For example, the *Vote Correlation Claim* (Wu et al., 2014): *"Jim Marshall, a Democratic incumbent from Georgia voted the same as Republican leaders 65 percent of the time"* can be seen as a BCC since it states a comparison between the voting behavior of two individuals. Such a claim can be investigated by using the U.S. congress roll call votes data[2]. Since DEBuNk (Chapter 3) aims to discover exceptional inter-group agreement patterns, it can be used to look for contexts (time periods, specific themes or topics) to shed more light on the claim, by providing contextual counter-arguments or elements reinforcing the claim from the data. Moreover, an analyst can goes beyond by using DEvIANT (Chapter 4) to analyze intra-group agreement patterns among republicans or democrats to study, among others, the cohesion within such political groups.

This chapter gives a brief overview of how the algorithms developed in this thesis can serve in data-driven fact checking or lead finding. Recall that, the task of fact-checking aims to evaluate to what extent some objective claim is valid (Vlachos and Riedel, 2014). Lead-finding, in turn, aims to uncover interesting facts from some given collection of data.

Contributions. The contributions of this chapter are:

Tools. We introduce ANCORE, a platform which enables to integrates the two approaches presented in the previous chapters (i.e. DEBuNk and DEvIANT) into an easy-to-use and interactive tool for exceptional intra-group and inter-group analysis in voting data.

Use Cases. We demonstrate the usefulness of ANCORE for computational journalism through multiple real-world use cases in the context of fact-checking and lead-finding.

[1] https://firstdraftnews.org/fake-news-complicated/
[2] https://voteview.com/data

The following content extends our article on ANCORE (Lacombe et al., 2019).

Roadmap. The remainder of this chapter is organized as follows. Section 5.2 describes platform ANCORE and develops its building components. Section 5.3 demonstrates two exemplary applications of ANCORE by developing multiple scenarii of fact-checking and lead finding using voting data. We wrap up by summarizing the chapter conclusions in Section 5.4.

5.2 Platform ANCORE

In order to provide a system facilitating the investigation of exceptional behaviors in voting data, we design ANCORE whose overview is depicted in Figure 5.3.

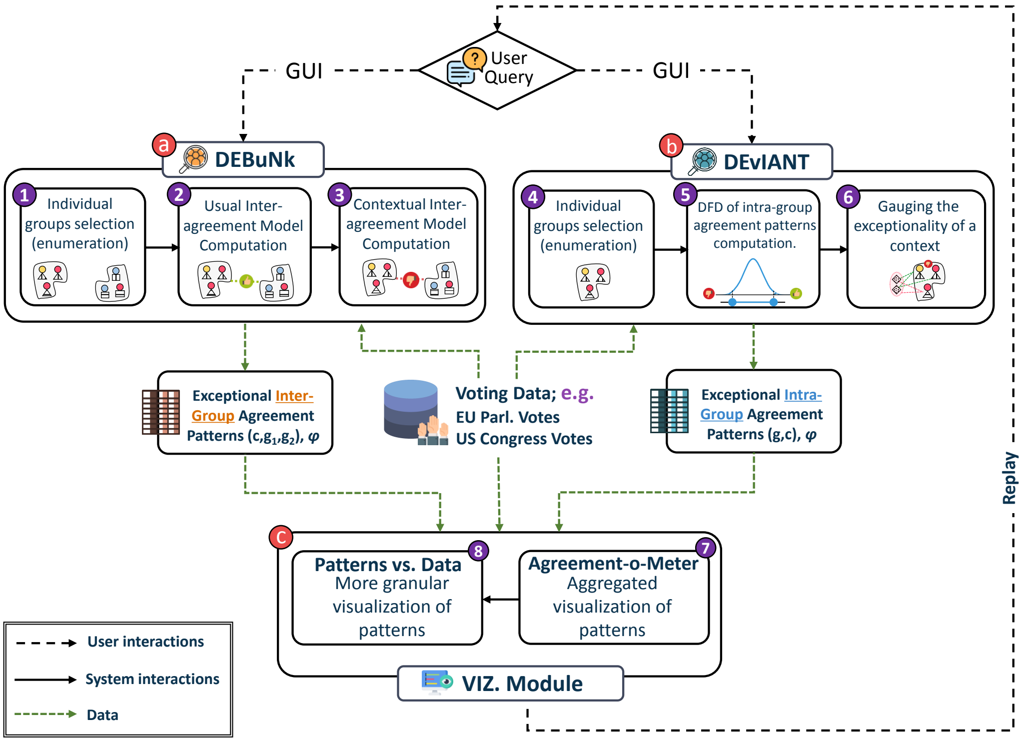


Figure 5.3: Global overview of Platform ANCORE

The platform not only provides an easy-tool to query voting datasets[3] for the discovery of exceptional inter-group and intra-group agreement patterns, but also displays the results in an intuitive fashion. All this being done to help the user to understand and interpret the found patterns. ANCORE relies on two main modules, namely DEBuNk and DEvIANT to mine for the exceptional patterns which are queried via a dedicated GUI:

Module DEBuNk (Chapter 3) addresses the problem of discovering exceptional **inter-group** agreement patterns. Such patterns exhibit an unexpected contextual agreement between two groups of individuals compared to their overall agreement. DEBuNk considers a *voting dataset*, perceived as a behavioral dataset. Voting sessions and voting members are characterized by descriptive attributes (e.g., numerical, categorical). The

[3]Currently, the platform ANCORE maintains: (1) the Eighth and the Seventh European Parliament roll call votes and (2) the US House of Representatives Votes ranging from the 102th(1991) to the 115th(2017) congresses.

patterns are of the form (c, g_1, g_2) with c a context and g_1, g_2 two groups. DEBuNk enumerates conceptually all the patterns and outputs the most interesting ones. A pattern interestingness is measured by a quality measure which enables to rank the pattern in the result set from the most to the least interesting one according to some given query. It evaluates the deviation between *(i)* the overall agreement between the two groups g_1, g_2 observed when considering all the voting sessions and *(ii)* the contextual agreement between the same two groups over the voting sessions supporting context c. To facilitate the interpretation of the patterns, the contextual (resp. overall) agreement is measured by the percentage of the context corresponding (resp. all) voting sessions on which the two majorities of the two confronted groups agree.

In ANCORE, the input of DEBuNk is specified by an end-user's query through the configuration GUI (see Figure 5.4), where she can select: (1) Which voting dataset she is interested in (EU Parliament or U.S. House of Representatives); (2) Which groups of voting members she wants to confront in her investigation (e.g. France v.s. Germany); (3) Which contexts she is interested in (e.g. Time period ranging from 2012 to 2016); (4) Which dimensions of study she wants to use to characterize the

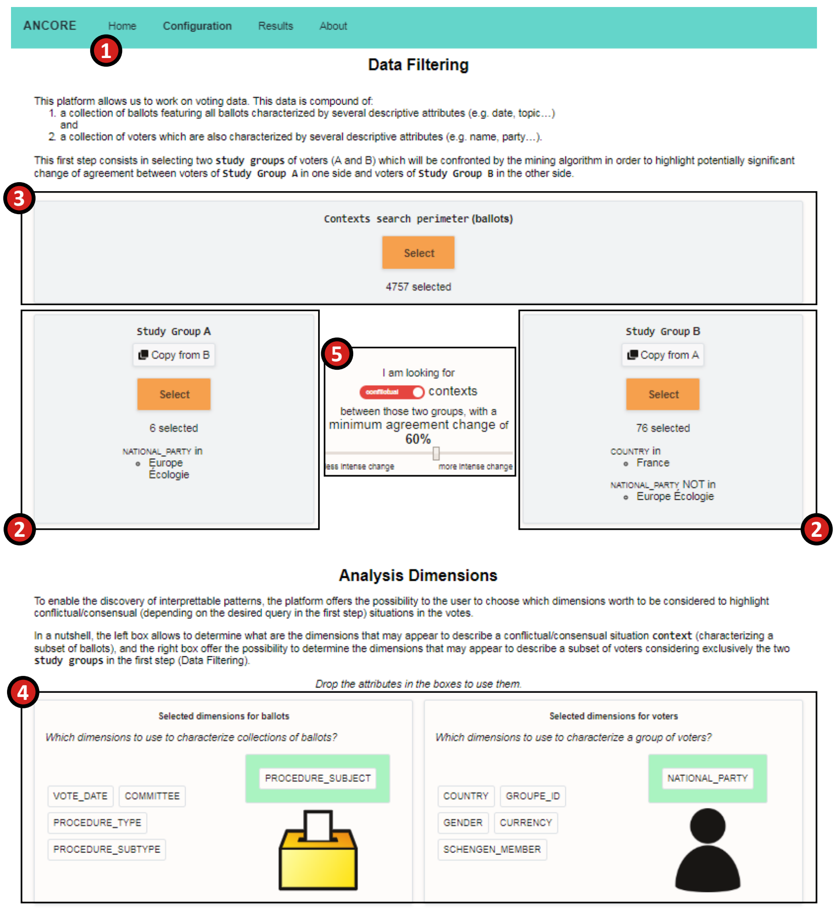


Figure 5.4: GUI for querying DEBuNk in ANCORE - in case of the EU parliament is selected as an underlying voting dataset.

desired exceptional inter-group agreement patterns (e.g. contexts described by the addressed topics). Eventually, (5) she determines which type of contexts she is looking for (e.g. conflictual or consensual) while specifying the intensity of changes (i.e. the minimum quality threshold) required to consider a pattern as exceptional.

Module DEvIANT (Chapter 4) addresses the problem of discovering exceptional **intra-group** agreement patterns. Such patterns highlights a statistically significant contextual intra-group agreement pattern. DEvIANT takes as input, a voting dataset seen as a behavioral dataset where sessions and members are characterized by descriptive attributes. The patterns are of the form (g, c) with g a group of voters and c a context regrouping a subgroup of voting sessions. The intra-group agreement is measures by Krippendorff's Alpha. A pattern interestingness is measured by its p-value: the probability to observe for a random subset of voting sessions an intra-group agreement between members of g as extreme as the one observed for the subset of voting sessions characterized by the context c.

In ANCORE, the input of DEvIANT is specified by an end-user's query through the configuration GUI (see Figure 5.5), where she can select: (1) Which voting dataset she

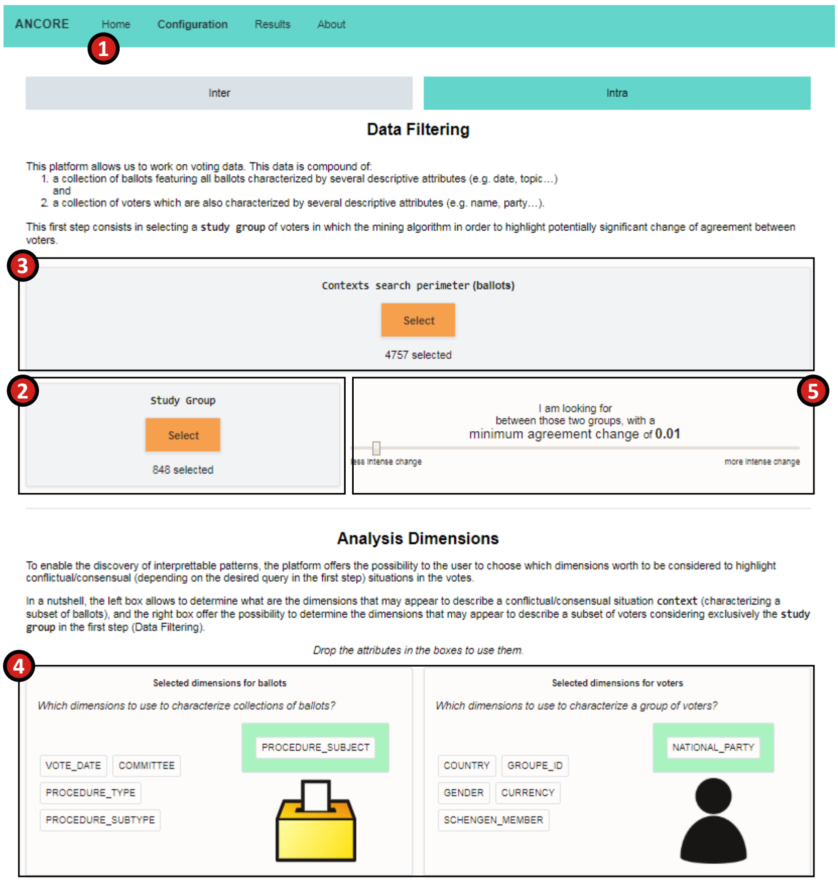


Figure 5.5: GUI for querying DEvIANT in ANCORE - in case of the EU parliament is selected as an underlying voting dataset.

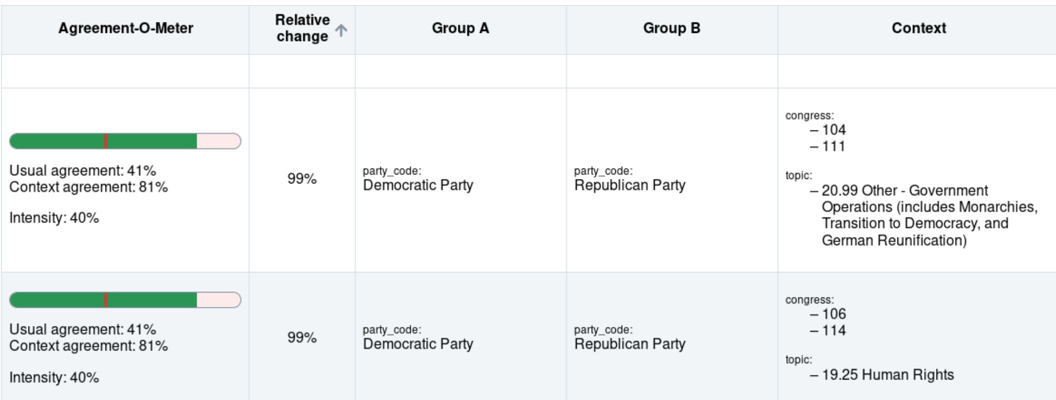
is interested in (EU Parliament or U.S. House of Representatives); (2) Which group of voting members she wants to study in her investigation (e.g. S&D); (3) Which contexts she is interested in (e.g. Judicial matters); (4) Which dimensions of study she wants to use to characterize the desired exceptional intra-group agreement patterns. Eventually, (5) she determines the intensity of changes required to consider a pattern as exceptional by fixing the critical value α, under which a returned pattern is considered as statistically significant.

The exceptional patterns once computed by one of the two modules, are processed by the visualization module (VIZ. Module, cf. Fig 5.3). A visual rendering of the retrieved patterns should enable to understand and interpret the patterns. To this end the visualization module presents the results with different levels of granularity. Indeed the visual rendering depends on which kind of patterns are given to the module. First, an aggregated view (see Fig. 5.6), enables to summarize the set of patterns, by consolidating the following details in a table:

Agreement-o-meter. It depicts, with a gauge, the overall inter-group/intra-group agreement level and the contextual inter-group/intra-group agreement level.

Pattern' descriptions. In case of inter-group agreement patterns are analyzed, the descriptions characterizing the confronted voting members groups g_1, g_2 and the exceptional context c are given. Similarly, for the visualization of intra-group agreement patterns, the descriptions corresponding to the voting group g and the context c are displayed.

Textual description. A natural language text is optionally given as a supplementary material for each returned exceptional inter-agreement pattern by adopting a "data to text generation approach" (Portet et al., 2009; Vizzini, Labbé, and Portet, 2017). Sentence templates take the form of a tree that convey the syntactic structure of the sentence. These templates are completed according to the returned patterns and a SimpleNLG (Gatt and Reiter, 2009) surface realization engine is used to generate the final phrases by applying grammatical rules: number and gender concordance of verbs and adjectives.



Agreement-O-Meter	Relative change	Group A	Group B	Context
Usual agreement: 41% Context agreement: 81% Intensity: 40%	99%	party_code: Democratic Party	party_code: Republican Party	congress: – 104 – 111 topic: – 20.99 Other - Government Operations (includes Monarchies, Transition to Democracy, and German Reunification)
Usual agreement: 41% Context agreement: 81% Intensity: 40%	99%	party_code: Democratic Party	party_code: Republican Party	congress: – 106 – 114 topic: – 19.25 Human Rights

Figure 5.6: Aggregated view summarizing the list of retrieved exceptional **inter-group agreement patterns**. They correspond to exceptional **consensual contexts** between **Democrats** and **Republicans** in the U.S. House, for **the time period 1991-2017**.

Second, for a better understanding and interpretation of each pattern, the user is offered a more fined-grained visualization where she can navigate the data used to compute the intensity of changes between the contextual inter-group/intra-group agreement and the overall one. In this detailed view, the set of voting sessions supporting the context is ranked from the most consensual to the most conflictual one (see **(1)** in Figures. 5.7 and 5.8). Each voting session, represented by a colored square, can be selected by the user to provide additional information. For instance, the voting decision made by each voting member is reported (see **(3)** in Figures. 5.7 and 5.8). For the EU Parliament, the link to the official procedure file concerning the voting session is given (see **(2)** in Figures. 5.7 and 5.8), so as to help the user navigate through all the context surrounding a reported pattern.

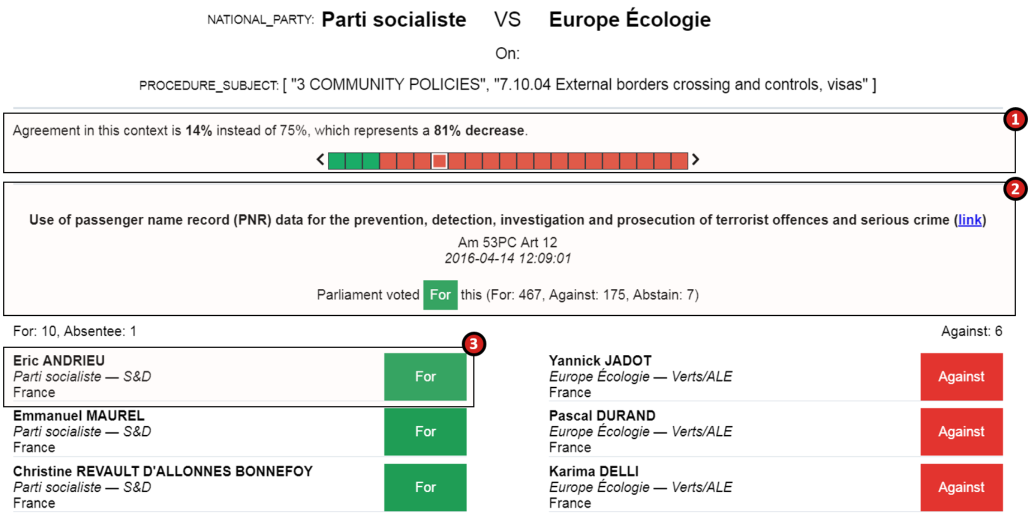


Figure 5.7: Detailed view of an **inter-group agreement pattern**, reporting the context, all the voting sessions and the vote of every voting member.

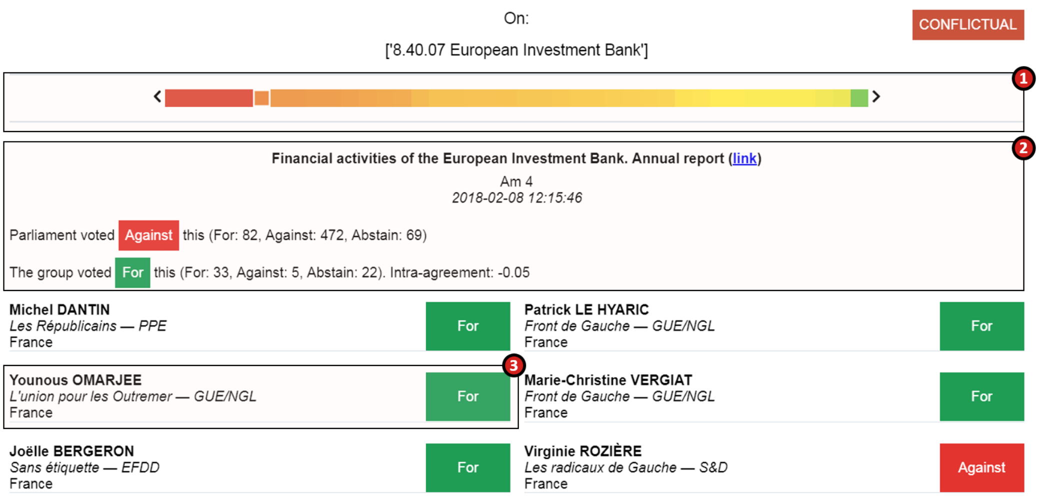


Figure 5.8: Detailed view of an **intra-group agreement pattern**, reporting the context, all the voting sessions and the vote of every voting member.

5.3 Use cases: Computational Fact Checking/Lead Finding

We envision platform ANCORE as a computational journalism tool which enables political analysts and data journalists to investigate exceptional behavior in voting data. Insights provided by ANCORE can be used to put into perspective and to assess the quality of some given claim in the context of a fact-checking process. For instance, the claim: "In the European parliament, French deputies vote always following the voting recommendation given by their respective national parties" can be studied using ANCORE. We develop this point in Section 5.3.1. Furthermore, highlights brought up by ANCORE can raise further investigations in the context of a lead-finding process. For example, by answering to the question: "What are the most conflictual subjects between countries in the European parliament?". We present two lead-finding use-case scenarii in Section 5.3.2.

5.3.1 Fact Checking using ANCORE

First, in Section 5.3.1.1, we start by giving some examples of claims that can be evaluated using (dis)agreement patterns that can be returned by ANCORE. Next, in Section 5.3.1.2, we particularly focus on studying claims reported in a real news article to demonstrate how ANCORE can assists an analyst in a real-world case scenario.

5.3.1.1 From Behavioral Comparison Claims to (Dis)Agreement Patterns

Earlier in this chapter, we presented briefly the category of claims dubbed **Behavior Comparison claims** (BCC) which covers any claim reporting a comparison of behavior between individuals or groups. We give below a set of claims that can be straightforwardly seen as BCCs.

- **Claim 1:** *Parliamentarian X votes always the same as parliamentarian Y.*
- **Claim 2:** *German and French S&D representatives share the same political line in most of the subjects treated in the European Parliament.*
- **Claim 3:** *The majority of the french far-right party Front National (FN) deputies vote always the same as their political leader Marine Le Pen (MLP).*

For instance, in order to evaluate **Claim 3**, we can first compute the overall agreement between MLP and the majority. If we observe a low percentage of agreement then we can conclude that the claim is not valid. As a second step, DEBuNk algorithm can be used to look for contexts in which a weakening of agreement between MLP and her peers in FN, thereby providing a set of patterns that can be presented as contextual counter-arguments. Figure 5.9 illustrates the exceptional inter-group agreement patterns found between FN' parlementarians and their leader MLP when using DEBuNk via ANCORE. Overall, we observe that MLP and the majority of FN are in strong agreement (i.e. MLP agrees with the majority in 98% of the voting sessions of the Eighth EU parliament). Still, in the three inter-group agreement patterns featured in Figure 5.9, we observe that MLP do not express the same voting outcome as her FN peers. For example, for the 18 sessions concerning both themes *"4 - Economic, social and territorial cohesion"* and *"7.40 - Judicial Coop"*, we note that MLP disagrees with the majority of FN in 8 sessions out of 18. Although, when investigating the voting decisions of FN members in these sessions, we observe that MLP abstained while her peers voted "for" the legislative procedures concerned.

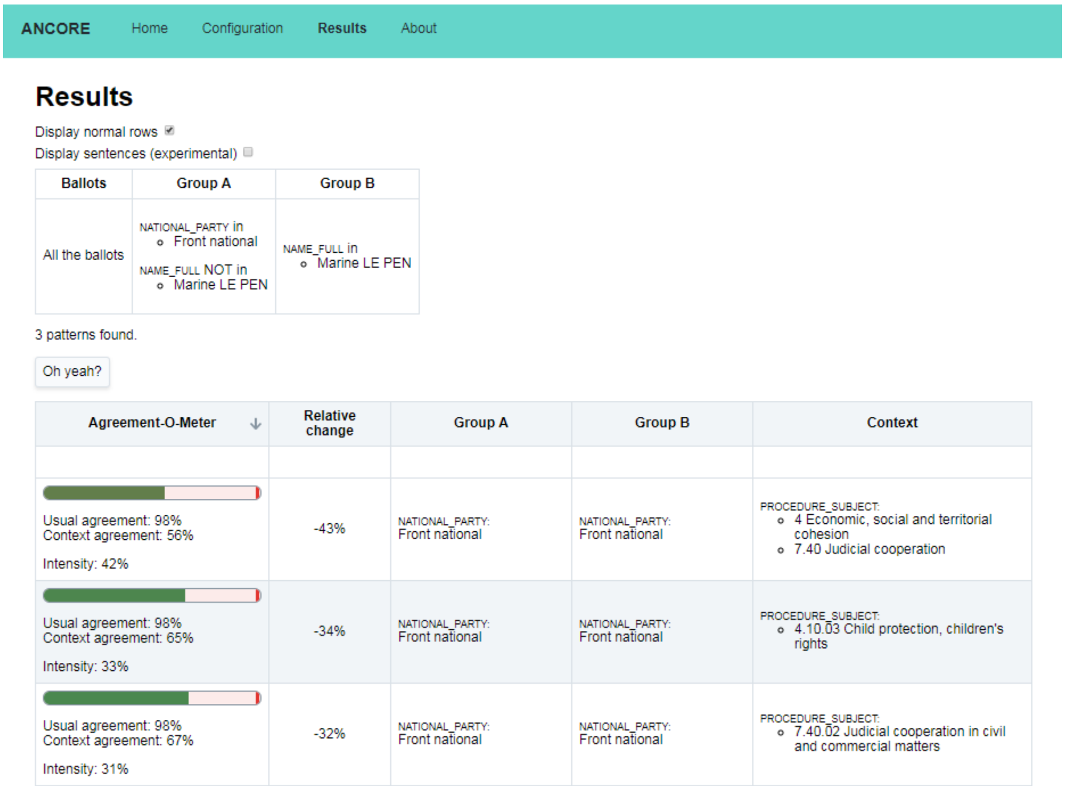


Figure 5.9: Patterns illustrating **conflictual contexts** between **MLP** and the majority of the **Front National Party** (Stripped from MLP) in the **eighth European parliament**. The **minimum threshold** for inter-group agreement measure change is fixed to 0.5 (50%).

With these information, the end-user (e.g. journalist) is well informed on the voting sessions where MLP has a different voting outcome than the majority of her national party. Hence, providing a sharper vision on the contexts surrounding the investigated claim.

Several other claims can be studied using ANCORE even if they are not explicitly expressed as comparisons as it was the case in the three former claims.

- **Claim 4:** *In the European parliament, French parliamentarians vote always following the voting recommendation given by their respective national parties.*
- **Claim 5:** *There is no national position when it comes to votes of EU Political Groups.*
- **Claim 6:** *Migration policy is one of the most controversial topics between countries in the European Parliament.*

For example, **Claim 5** can be examined using ANCORE in various ways. For instance, the claim can be investigated across each of the eight political groups composing the EU parliament. This can be done either by using DEBuNk by confronting countries' representatives in each political groups and then look for conflictual inter-group agreement patterns to provide contextual counter-arguments; or by using DEvIANT to look for exceptional intra-group agreement patterns among each political group and then investigate the voting behavior of each country representatives in the discovered pattern.

In Figure 5.10, we illustrate an example pattern uncovered by DEBuNk when considering parliamentarians of the S&D political group. While, in the overall case, we observe an agreement between countries majorities (i.e. countries majorities in S&D votes the same in more than 80% of the cases). Though, several conflictual patterns between countries emerges (13 patterns) as illustrated in Figure 5.10. Such patterns besides other patterns returned by DEvIANT, can constitute relevant materials to provide deeper insights on the situations between parliamentarians in each group and their cohesiveness in particular contexts.

Here, we purposely choose to analyze a particular claim (**Claim 5**) to demonstrate how complex is the task of fact checking even when the relevant data are available. Moreover, no peremptory verdict can be given on the claim. Although, depending on the resulting (dis)agreement patterns given by ANCORE, one can assess the quality of the claim, by providing a better understanding of the situation as a whole of the agreement between parlementarians within their respective political groups while considering the countries dimension. Still, efforts need to be invested by the analyst in terms of (i) formulating the proper queries, (2) consolidating the results and (3) combining the results with materials from other sources in a such complex fact-checking scenario.

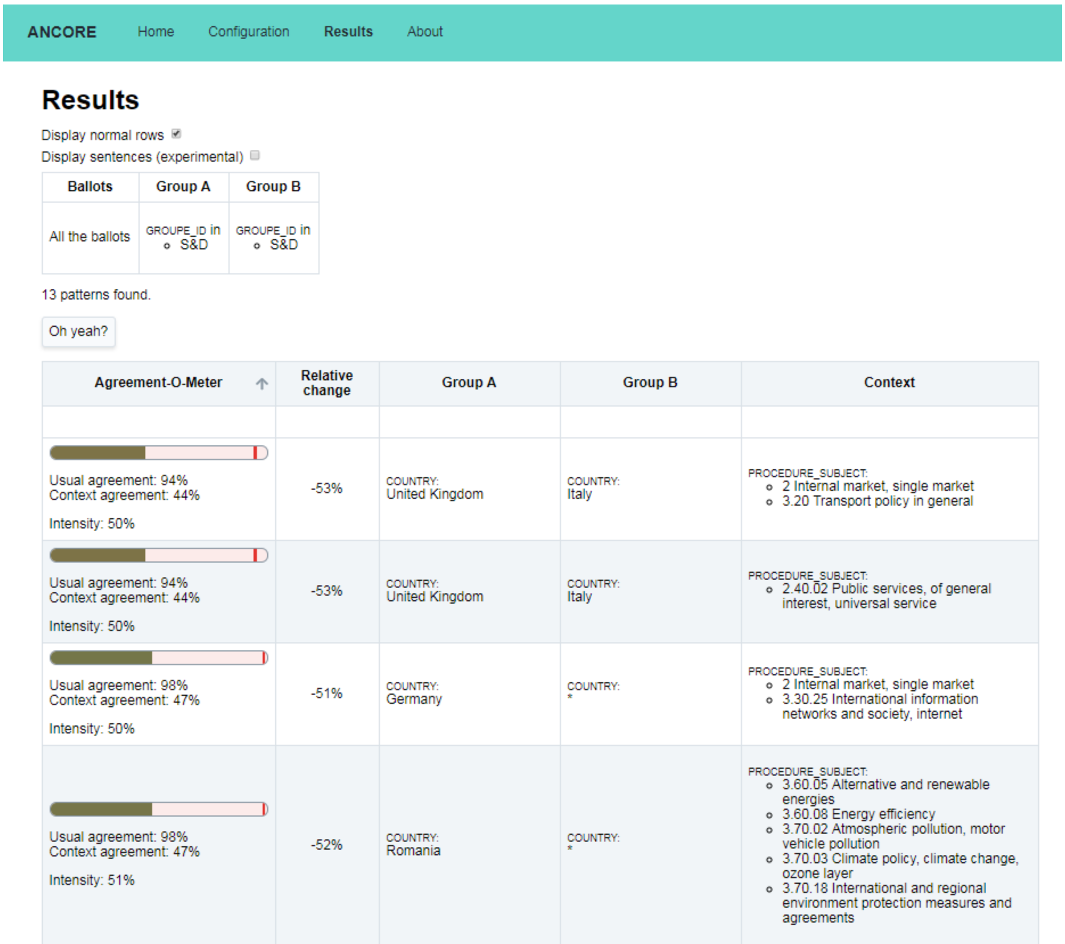


Figure 5.10: Patterns illustrating **conflictual contexts** between **countries** in the **Socialists & Democrats (S&D) political group**. The **minimum threshold** for inter-group agreement measure change is fixed to 0.6 (60%).

5.3.1.2 From a Real-World News Article to (Dis)Agreement Patterns

In this section, we choose to demonstrate the platform capabilities by analyzing the news article "Groups in the European Parliament, sometimes surprising alliances"[4]. It refers particularly to the EPP (European People's Party, the majority group of the 8th legislature of the EU parliament), and argues that the desire of some political group to bring together as many parties as possible leads sometimes to "surprising alliances". One specific party is brought to the fore in the article, the Fidesz party (Hungary) which belongs to EPP. This raises several questions that one can study using ANCORE:

- Is the Fidesz in conflict with the rest of the EPP?
- Does the Fidesz have any conflicts with specific EPP parties?
- Are there any other conflicts within the EPP?

Fidesz against the rest of EPP

We first confront the Fidesz MEPs with the rest of the EPP members, by looking for conflictual contexts. By analyzing the results provided by ANCORE, the first insight that emerges, is that the Fidesz MEPs are in agreement with their EPP peers in 94% of the cases. The most conflictual subjects highlighted by the system were agricultural measures and the administrative processes of the EU.

Fidesz against other EPP parties

We now focus on contexts that oppose the Fidesz to other EPP parties. The most intense change of intra-group agreement is observed between the Fidesz and the Partido Popular (Spain). The returned pattern shows that, while the Fidesz and the Partido Popular are in strong agreement (91%), the following contexts lead to strong disagreement (cf. Figure 5.11):

- *2 Internal market* and *4.10 Social policy, charter and protocol*.
- *4.10.07 The elderly*.

To investigate in more depth the relationship between the Fidesz and other national parties, we look for consensual contexts. When analyzing the results provided by ANCORE, we observe that the EPP has an overall consistent political line. Moreover, the results highlight two national parties: the Partido da Terra (Portugal), and the Centre Démocrate Humaniste (Belgium, mentioned in the article), both represented by one single MEP and respectively having a usual agreement of 75% and 76% with the Fidesz.

[4] goo.gl/43MM3k, article published on the RTBF website on 23 Oct. 2015

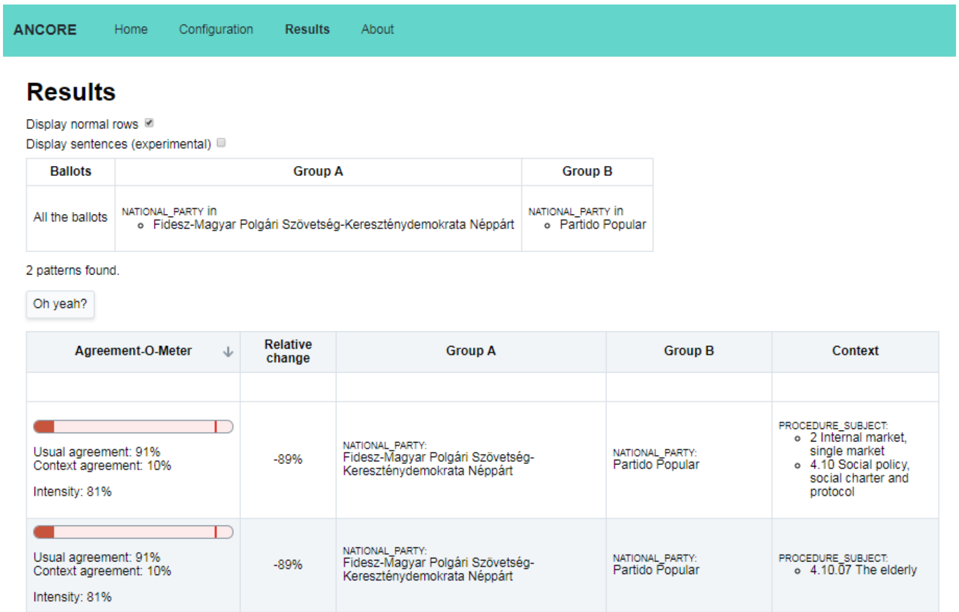


Figure 5.11: Patterns highlighting **conflictual contexts** between **Fidesz** and **Partido Popular**. The **minimum threshold** for inter-group agreement measure change is fixed to 0.8 (80%).

Conflicts within the EPP

We are now interested in the conflicts within the EPP as a whole, without emphasizing on the Fidesz. When investigating the results, two patterns arise which oppose the Partido Popular with the rest of the group. The patterns highlight the same contexts observed when analyzing conflictual contexts with the Fidesz. This demonstrates that the conflict was rather on the side of Partido Popular, since the Fidesz was in agreement with the majority decision. An example pattern is visualized in detail in Figure 5.12. Another important conflict within EPP

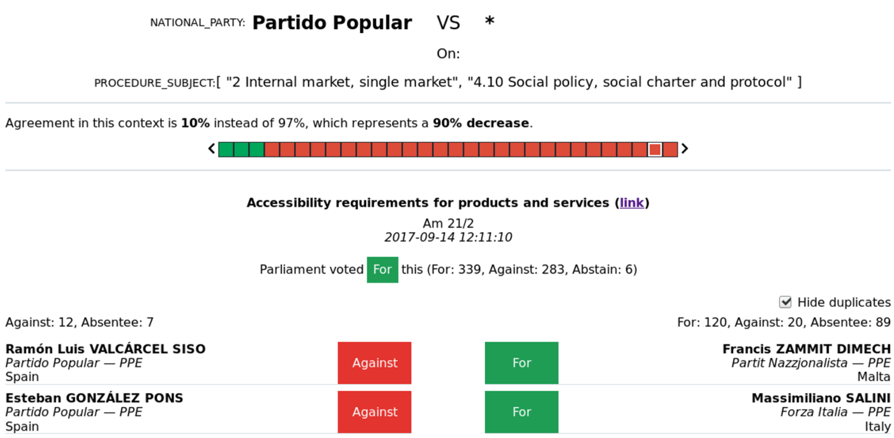


Figure 5.12: Detailed view of an exceptional **intra-group agreement** pattern, showing the context (defined by the procedure subject), all the voting sessions and the vote of every voting member. It corresponds to an exceptionally **conflictual context** in the European Parliament between the **Spanish National Party "Partido Popular"** and the **EPP Group**.

is revealed by DEBuNk and concerns the Forza Italia party with the rest over relations with Russia. Overall, these two parties are in agreement with the rest of EPP in 97% of the voting sessions. Furthermore, when investigating the conflict among EPP representative using DEvIANT we obtain 19 significantly conflictual contexts in EPP as illustrated in Figure 5.13.

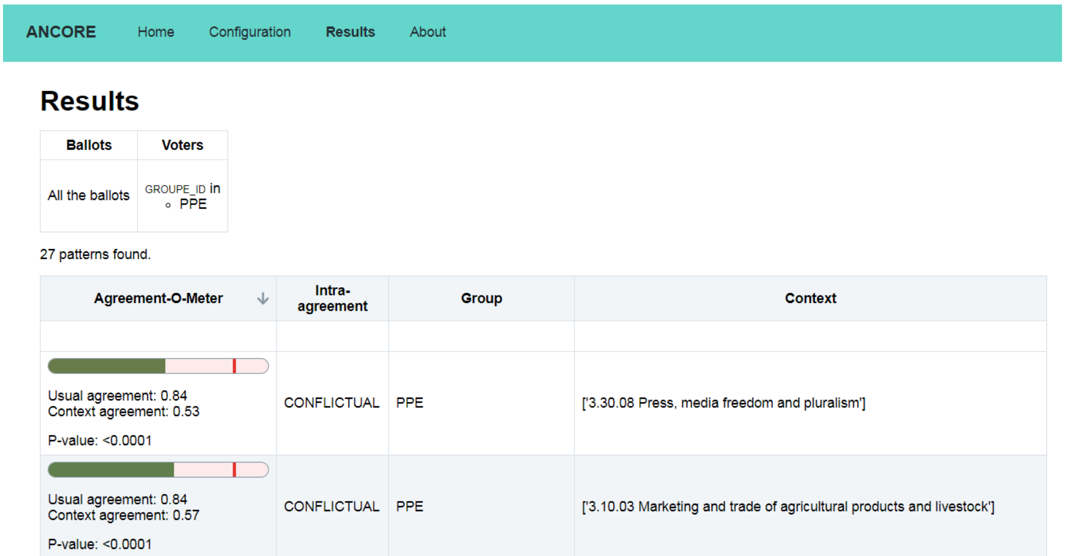
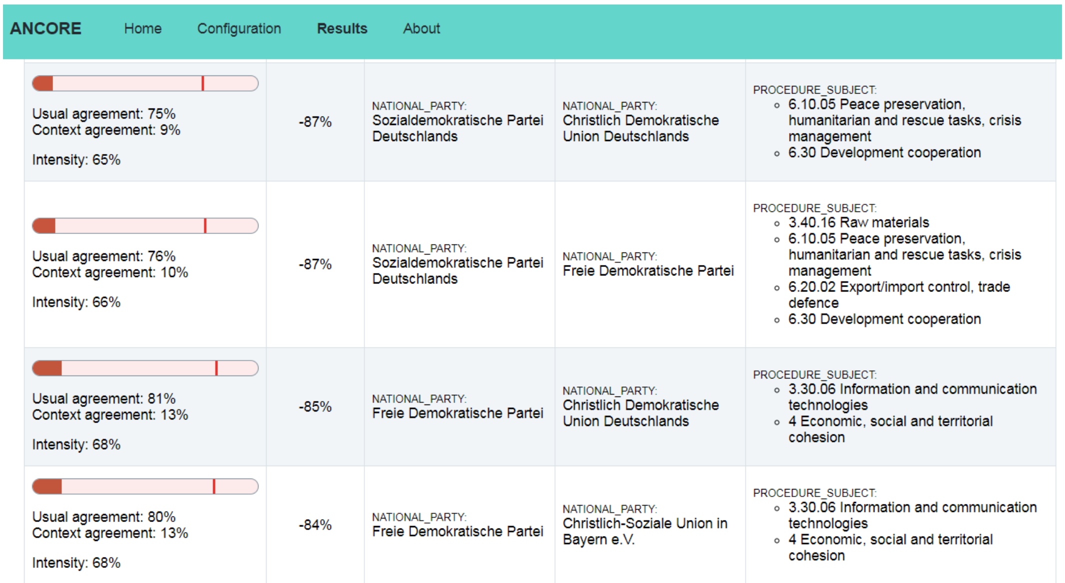


Figure 5.13: Two **conflictual contexts** among the 19 conflictual contexts in the **EPP group** during the eighth European Parliament voting dataset. The **critical value alpha** is fixed to 0.05.

5.3.2 LEAD FINDING USING ANCORE

Lead finding, as defined by (Wu, 2015; Wu et al., 2017) in the context of computational journalism, is ”the task of finding interesting information nuggets from raw data that lead to further investigation and/or news stories around them". In the scope of this chapter, and more generally, in the scope of this thesis, we define the lead-finding as: ”the task of discovering exceptional (dis)agreement between or among groups from behavioral data“.

Practically, patterns exhibited in the qualitative experimental sections 3.6.2 and 4.7.2 provide some good examples where raw data corresponding to roll call votes are transformed to interpretable and actionable insights. In a more general scope, The philosophy behind our Subgroup Discovery/Exceptional Model Mining approaches is rather close to the philosophy behind computational lead-finding, since our end-user persona (e.g. data-journalist), in this thesis, is interested in finding exceptional areas in some underlying behavioral data without knowing upfront what these patterns look like. While computational lead finding covers various types of interesting pieces of information that one can extract from a dataset (examples are given in (Wu, 2015; Wu et al., 2014; Wu et al., 2017)), in this section, we are interested in uncovering exceptional (dis)agreement patterns in voting data. For instance, the end-user can use ANCORE to look for high-conflict/high-consensual topics between or within national parties, political groups or countries when considering European Parliament voting dataset. In Figure 5.14, we give some example patterns returned by ANCORE when looking for high-controversial contexts between German national parties. For instance, we observe that while FDP (Free Democratic Party) and CDU (Christian Democratic Union of Germany) agree most of the time ($\backsim$ 81%), they express diverging opinions on procedures voted under the theme "3.30.06 Information and Communication Technologies" covering most importantly the dossier: *open internet access*[5]. These pieces of information alongside other materials may help a journalist investigating the failure of the so-called *"Jamaica" coalition* after the 2017' German federal election by studying the relationship between its constituting parties.



Usual agreement: 75% Context agreement: 9% Intensity: 65%	-87%	NATIONAL_PARTY: Sozialdemokratische Partei Deutschlands	NATIONAL_PARTY: Christlich Demokratische Union Deutschlands	PROCEDURE_SUBJECT: ◦ 6.10.05 Peace preservation, humanitarian and rescue tasks, crisis management ◦ 6.30 Development cooperation
Usual agreement: 76% Context agreement: 10% Intensity: 66%	-87%	NATIONAL_PARTY: Sozialdemokratische Partei Deutschlands	NATIONAL_PARTY: Freie Demokratische Partei	PROCEDURE_SUBJECT: ◦ 3.40.16 Raw materials ◦ 6.10.05 Peace preservation, humanitarian and rescue tasks, crisis management ◦ 6.20.02 Export/import control, trade defence ◦ 6.30 Development cooperation
Usual agreement: 81% Context agreement: 13% Intensity: 68%	-85%	NATIONAL_PARTY: Freie Demokratische Partei	NATIONAL_PARTY: Christlich Demokratische Union Deutschlands	PROCEDURE_SUBJECT: ◦ 3.30.06 Information and communication technologies ◦ 4 Economic, social and territorial cohesion
Usual agreement: 80% Context agreement: 13% Intensity: 68%	-84%	NATIONAL_PARTY: Freie Demokratische Partei	NATIONAL_PARTY: Christlich-Soziale Union in Bayern e.V.	PROCEDURE_SUBJECT: ◦ 3.30.06 Information and communication technologies ◦ 4 Economic, social and territorial cohesion

Figure 5.14: Exceptional **conflictual contexts** between **German national parties** in the eighth European parliament voting dataset. The **minimum threshold** for inter-group agreement measure change is fixed to 0.8 (80%).

Similarly, as in Section 5.3.1, we give an example of using ANCORE in the context of a real-world computational lead-finding case scenario. Let us consider, the recent news article "European Elections 2019 : How did the 82 French MEPs voted since 2014 ?"[6] published on Le Monde website on 10 Mai 2019. Exceptional (dis)agreement patterns both within and between French national parties representatives in the EU, can provide valuable information for the analysis of French MEPs votes. This enables the analysis to go beyond by outlining highlights on the inter-group and intra-group voting behavior of french MEPs in the European parliament. For example, when using DEvIANT via ANCORE to mine for exceptionally conflictual or consensual topics among french national parties, several exceptional intra-group agreement patterns are brought to the fore (cf. Figure 5.15), some of which are relevant to the investigation conducted on French MEPs voting behavior in the news article. For instance, voting sessions related to judicial cooperation in criminal matters led to a conflict between members of the French left-wing party "Front de Gauche". Additionally, DEBuNk was able to retrieve exceptional conflict between French national

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32015R2120

[6] https://www.lemonde.fr/les-decodeurs/article/2019/05/10/europeennes-2019-comment-ont-vote-les-deputes-europeens-francais-depuis-2014_5460395_4355770.html

parties (cf. Figure 5.16). As an example, we observe that while French Social-Democratic MEPs and Green Party MEPs are in agreement in the overall terms, matters of external borders crossing, controls and visas create a strong disagreement between these two parties. A controversial legislative procedure in this context was “**2011/0023(COD)**”[7] which was raised in the considered news article.

Usual agreement: 0.92 Context agreement: 0.51 P-value: <0.0001	CONFLICTUAL	Mouvement Démocrate	['3.45.01 Company law', '3.45.04 Company taxation']
Usual agreement: 0.72 Context agreement: 0.33 P-value: <0.05	CONFLICTUAL	Front de Gauche	['2.80 Cooperation between administrations', '7 Area of freedom, security and justice']
Usual agreement: 0.72 Context agreement: 0.37 P-value: <0.01	CONFLICTUAL	Front de Gauche	['3.10.06 Crop products in general, floriculture']
Usual agreement: 0.72 Context agreement: 0.41 P-value: <0.001	CONFLICTUAL	Front de Gauche	['7.40.04 Judicial cooperation in criminal matters']
Usual agreement: 0.92 Context agreement: 0.63 P-value: <0.0001	CONFLICTUAL	Mouvement Démocrate	['4.15.15 Health and safety at work, occupational medicine']

Figure 5.15: Exceptional **conflictual contexts** within **french national parties** during the eighth European Parliament voting dataset. The **critical value alpha** is fixed to 0.05.

Usual agreement: 86% Context agreement: 29% Intensity: 57%	-67%	NATIONAL_PARTY: Parti socialiste	NATIONAL_PARTY: *	PROCEDURE_SUBJECT: ◦ 6.10 Common foreign and security policy (CFSP) ◦ 6.30 Development cooperation
Usual agreement: 67% Context agreement: 10% Intensity: 57%	-85%	NATIONAL_PARTY: Europe Écologie	NATIONAL_PARTY: *	PROCEDURE_SUBJECT: ◦ 6.10.02 Common security and defence policy; WEU, NATO ◦ 8 State and evolution of the Union
Usual agreement: 75% Context agreement: 18% Intensity: 57%	-76%	NATIONAL_PARTY: Parti socialiste	NATIONAL_PARTY: Europe Écologie	PROCEDURE_SUBJECT: ◦ 1 European citizenship ◦ 7.10.04 External borders crossing and controls, visas
Usual agreement: 85% Context agreement: 29% Intensity: 56%	-66%	NATIONAL_PARTY: Mouvement Démocrate	NATIONAL_PARTY: Les Républicains	PROCEDURE_SUBJECT: ◦ 4.10.04 Gender equality ◦ 4.20.02 Medical research
Usual agreement: 86% Context agreement: 30% Intensity: 56%	-66%	NATIONAL_PARTY: Parti socialiste	NATIONAL_PARTY: *	PROCEDURE_SUBJECT: ◦ 3.40.16 Raw materials

Figure 5.16: Exceptional **conflictual contexts** between **french national parties** in the eighth European parliament voting dataset. The **minimum threshold** for inter-group agreement measure change is fixed to 0.6 (60%).

[7]Use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime, available on https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2011/0023(COD)&l=en.

5.4 Summary

In this chapter, we devised platform ANCORE, which enables discovering patterns exhibiting exceptional (dis)agreement within and among groups of voting members by leveraging DEBuNk or DEvIANT depending on the aim of the study. The proposed platform not only computes and returns the list of exceptional patterns, but also, for better interpretability, it enriches the provided results by offering a visualization tool covering the reported patterns from various perspectives. To demonstrates the capabilities offered by ANCORE, we showed how this analytic tool can serve as a basis to quickly put claims into perspective in the context of a fact-checking process, or to uncover insights from voting data for lead-finding.

Note that, besides lead-finding examples outlined from the EU voting dataset presented here, the study of medicines consumption discrepancies between french sub-population presented in Table 3.8 (leveraging Openmedic data) in Section 3.6.2.3 (Chapter 3) is an interesting illustration of computational lead finding in the context of epidemiology studies and health monitoring applications. For instance, a news article, entitled "Medicines and refunds: Openmedic dataset in six points"[8] covering Openmedic dataset was published on Le Monde website on 28 November 2017. The investigation conducted in this article can be enriched by highlighting, for example, substantial differences in medicine consumptions between subpopulations of interest (e.g., age-, gender- or region-specific) by using DEBuNk.

We believe that this work sets the ground for many interesting improvements. First, the visualization module can be enhanced for better usability and user-friendliness of ANCORE tool. Also the visualization tool can be improved by allowing richer graphical representations. For example, *Nominate* (Hix, Noury, and Roland, 2006; Poole and Rosenthal, 1985; Voeten, 2009) can be implemented to depict and compare the positions (*Left/Right, Conservative/Liberals, Pro-integration/Anti-Integration*) of voting members in both the overall terms and the discovered contexts. Second, additional unsupervised learning methods can be investigated to improve the interpretability of the found patterns. For instance, clustering can summarize in a compact way the agreement between parlementarians both in overall terms and in the contexts uncovered by DEBuNk or DEvIANT. In this perspective, several dissimilarity metrics can be used ranging from a simple Iverson bracket $[o_1 \neq o_2]$ (with o_1, o_2 are two voting outcomes) to Rajski's Distance (Jakulin et al., 2009) to cluster parliamentarians and study, among others, how contexts impacts the cohesion and the polarization in/between political groups, countries, national parties, etc. Third, ANCORE can benefit from interactive pattern mining paradigm (Dzyuba, 2017; Dzyuba et al., 2014; Leeuwen, 2014) to actively involve the user in the exploratory data mining process, therefore providing more interesting results.

[8] https://www.lemonde.fr/les-decodeurs/article/2017/11/28/medicaments-et-remboursements-la-base-de-donnees-open-medic-en-6-points_5221378_4355770.html

6

Conclusion

This thesis brings several contributions to the following task:

■ *Discovering and characterizing* ***Exceptional (Dis-)Agreement between*** *and* ***within sub-populations*** *in* ***Behavioral data***. ■

In this vein, we proposed two approaches for the efficient and optimal discovery of such insights from behavioral data and consolidated them into a web-platform for the analysis of exceptional voting behaviors in voting data. Section 6.1 summarizes the contributions of this thesis, and Section 6.2 discusses opportunities for future works.

6.1 Summary

In order to tackle the aforementioned problem, we have proposed two novel and complementary approaches to mine for exceptional (dis)agreement between and within groups in behavioral data. All this has been done by relying on the frameworks of Subgroup Discovery (SD) and Exceptional Model Mining (EMM) that have been discussed in Chapter 2.

We now review the research questions outlined in the introduction of this thesis and highlight how each chapter of this thesis contributes to answering the addressed questions.

> **Research Question. 1** How to characterize, discover, summarize and present **exceptional (dis) agreement <u>between</u> groups** (sub-populations) in **behavioral data** ?

In this thesis, the research question has been brought down to four questions:

- *How to* ***characterize*** *exceptional (dis)agreement between groups in behavioral data ?*
- (Chapter 2 and Chapter 3) The characterization exceptional (dis) agreement between groups requires the definition of the syntax and the semantic of patterns conveying such kind of insights.

For the syntax of the desired patterns, we choose to structure them as triplets: (c, u_1, u_2) with c a context describing a subset of entities and (u_1,u_2) two confronted groups of individuals. For easier interpretation, Groups and Contexts are conjunctive selection predicates over the corresponding attributes. This induces a description language which was introduced and established in Chapter 2.

Once the syntax of the patterns had been formalized, we relied on SD/EMM framework to objectively define what exceptional (dis)agreement means in the scope of this thesis, a.k.a the semantics of the patterns. Recall that, under the umbrella of EMM framework, two aspects need to be appropriately instantiated in order to convey the meaning of a pattern: the model class and the interestingness measure. In Chapter 3, the chosen model class was IAS (Inter-group Agreement Similarity measure) which captures to what extent two groups of individuals (u_1, u_2) are in agreement with regards to entities characterized by some context c.

To assess the exceptionality of a pattern, we defined several interestingness measures (φ_{consent}, φ_{dissent} and φ_{ratio}) which evaluate the deviation between the contextual inter-group agreement measure and the overall one observed over the entire entities collection. For instance φ_{consent} gives better score to patterns where there is more consensus between the two confronted (u_1, u_2) groups in the context c compared to the consensus observed in the overall terms. Conversely, φ_{dissent} is associated to the discovery of conflictual situations rather than consensual ones.

? *How to **discover** exceptional (dis)agreement between groups in behavioral data ?*

? (Chapter 3) The previous answer formalized the description language for exceptional inter-group (dis)agreement patterns and the interestingness measure used to evaluate the exceptionality of such patterns. In order to discover these patterns, we devised two algorithmic solutions DEBuNk and Quick-DEBuNk.

DEBuNk (cf. Algorithm 3) is an exhaustive search algorithms which uses EnumCC to generate candidate subgroups. It uses several optimization techniques in order to efficiently return the most interesting patterns as defined in Problem 3.2.1 (Chapter 3). First, closure operators are used to avoid redundancy in the discovery process, this is possible since the interestingness measure is extent-based and the underlying description language induces a pattern structure. Moreover, DEBuNk relies on (tight) optimistic estimates for the proposed interestingness measures to prune as soon as possible unpromising branches of the search space.

Quick-DEBuNk (cf. Algorithm 6), in turn, offers an alternative and tractable solution to the problem 3.2.1 of discovering exceptional (dis)agreement between groups. The end-user is given the possibility to specify a time-budget to the algorithm within which the algorithm is required to stop and return the currently found patterns. Quick-DEBuNk is a stochastic algorithm which combines exploitation and exploration techniques in order to (quickly) find the desired patterns. For exploration, the algorithm relies on direct sampling paradigm via FBS (cf. Algorithm 4) where the patterns (c, u_1, u_2) are drawn randomly and with a chance proportional to the product of the support size of each description. For exploitation, Quick-DEBuNk uses RWC (cf. Algorithm 5) to search for exceptional (dis)agreement between groups starting from the pattern (c,u_1,u_2) returned by FBS. In order to give more chance to high-quality patterns, RWC chooses to expand neighbor search nodes (from a given search node, i.e. context) with a probability proportional to their quality. Moreover it relies on closure operator and optimistic estimates to avoid generating uninteresting patterns.

3 *How to **summarize** exceptional (dis)agreement between groups in behavioral data ?*

3 (Chapter 3) As a first step, the summarization of exceptional (dis)agreement between groups have been defined through a set of constraints that need to be satisfied in the returned list of patterns. Redundancy is avoided by using closure operators. Furthermore, only the most general patterns are returned. This is motivated by the following postulate: the end-user is more interested by (dis)agreement observed between larger groups in larger context rather than local (dis)agreements.

4 *How to **present** exceptional (dis)agreement between groups in behavioral data ?*

4 (Chapter 5) In Chapter 3, the proposed algorithms returned the list of exceptional inter-group agreement patterns in the form of a raw table (csv file). This requires effort from the end-user to interpret the results and find proper explanation of why a pattern from the final result set has been declared exceptional. Within this aim, we proposed ANCORE, a web platform for discovering exceptional (dis)agreement in voting data. ANCORE has been presented and detailed in Chapter 5. ANCORE provides an easy-to-use tool to search for exceptional inter-group agreement patterns and to interpret them. It enables the end-user to have an in-depth understanding of why an intra-group agreement pattern is considered as exceptional, by bringing to the fore the data used to evaluate the interestingness of pattern. For instance, for every exceptional pattern (c,u_1,u_2), the visualization tool of ANCORE prints out every outcome expressed by the individuals comprising the two confronted groups in every entity covered by the context in question.

Research Question. 2 How to characterize, discover, summarize and present **exceptional (dis) agreement <u>within</u> groups** (sub-populations) in **behavioral data** ?

Similarly as **Research Question 1**, this question was brought down to four question:

5 *How to **characterize** exceptional (dis)agreement within groups in behavioral data ?*

5 (Chapter 2 and Chapter 4) Exceptional (dis)agreement within groups is captured by patterns of the form (u,c) where c is a context and u a group. Contexts and groups are formalized as conjunctive selection predicates (syntax) as detailed in Chapter 2.

The semantic of exceptional intra-group agreement patterns (c,u) is instantiated via EMM framework by the definition of an appropriate model class and its associated interestingness measure. In order to evaluate to what extent the members comprising a group are in agreement when considering the entities related to the context c, we used Krippendorff's Alpha measure. The latter measure is adapted to our setting as it handles the sparsity usually encountered in behavioral data and various possible domains of outcomes.

To evaluate the exceptionality of an intra-group (dis)agreement pattern, we used statistical significance (p-value) of the contextual intra-group agreement measure (Krippendorff's alpha). In short, the proposed interestingness measure is the probability of observing an intra-group agreement for a random subset of the collection of entities which is at least as extreme as the one observed for the context c. If such a probability is under a critical value α (usually 0.05) the pattern is declared exceptional, otherwise, it is considered as a spurious finding.

6 *How to **discover** exceptional (dis)agreement within groups in behavioral data ?*

6 (Chapter 4) the former point addressed the syntax and semantics of pattern conveying exceptional (dis)agreement within agreement. In order to discover the desired patterns, we devised DEvIANT (cf. Algorithm 7) to solve the problem of finding exceptional (dis)agreement within groups in behavioral data as defined in Problem 4.2.1. DEvIANT is a branch and bound algorithm which relies on EnumCC (cf. Algorithm 1) to generate closed candidate subgroups without redundancy. For further optimization, DEvIANT uses tight optimistic estimates on Krippendorff Alpha to establish the interval within which the contextual intra-group agreement varies when considering the search space under some context c. Along this interval, an interesting property between confidence intervals is leveraged which states, in brief, that confidence intervals grow in size and are encompassed when going downward in the search tree. These concepts, when combined, ensure a safe-pruning strategy to avoid generating and evaluating unpromising candidate subgroups.

7 *How to **summarize** exceptional (dis)agreement within groups in behavioral data ?*

7 (Chapter 4) As for Question 3, two concepts are used to provide a concise list of exceptional intra-group (dis)agreement patterns. First, redundancy is avoided via closure operators. Second, only the most general patterns are kept in the final result set, that is, if an exceptional (dis)agreement is observed in the pattern (c,u), no specialization of this pattern is included in the results set.

8 *How to **present** exceptional (dis)agreement within groups in behavioral data ?*

8 (Chapter 5) In Chapter 4, only a raw list in csv format is returned at the end of execution of DEvIANT. This requires further processing by the end-user to explore exceptional (dis)agreement within groups in behavioral data. To facilitate the reading and the interpretation of the patterns, we use ANCORE. its visualization tool enables to explore in details the outcomes used to assess the exceptionality of a pattern (u,c).

In order to demonstrate the usefulness of exceptional inter-group (dis)agreement patterns and exceptional intra-group (dis)agreement patterns, we conducted several qualitative experiments in this thesis. Chapter 3 illustrates the search for exceptional (dis)agreement between groups within three different types of behavioral data: political analysis using European parliament voting data, rating data analysis using yelp rating data and movielens rating data, healthcare surveillance using Openmedic dataset. Chapter 4 depicts examples of exceptional (dis)agreement within groups in two different types of behavioral data: political analysis using European parliament voting data and United States Congress votes in the House of representatives; rating data analysis using yelp rating data and movielens rating data. Finally, Chapter 5 focus on computational journalism, highlighting how exceptional (dis)agreement between and within groups can empower and help journalists in fact-checking claims or finding interesting facts from data in a lead-finding process.

6.2 Outlook

The contributions of this thesis set the ground for many improvements and instigate new research venues that we foresee could to lead to interesting results. In the following, we review some of the promising perspectives of this work.

6.2.1 Enriching the Visualization tool of ANCORE

We proposed in ANCORE a visualization tool which provides an in-depth consultation of every outcomes expressed by the individuals in each entity covered by the context in question. This allows to study the impact of each context' entity on the contextual inter/intra-group agreement. An interesting improvement that we started to investigate recently is the integration of new graphical representations of exceptional intra/inter-group agreement patterns. For instance, for both kinds of patterns, we can depict at a high-granularity level the agreement between the individuals (if applicable) in a heatmap. For inter-group agreement patterns (c, u_1, u_2), one can confront the individuals of group u_1 against the individuals of group u_2 and draws two associated similarity matrices: the contextual similarity matrix and the overall one. The similarity measure can rely on the inter-group agreement similarity (IAS). Similarly, two heatmaps can be associated to each exceptional intra-group agreement pattern (u, c) by confronting the individuals of the group u by an adapted similarity measure.

Other graphical representations can leverage the above similarity matrices. For instance, Multi-Dimensional Scaling (Cox and Cox, 2000) Techniques (MDS) can be used to represent in two-dimensional space the individuals of the considered groups. This can help to identify quickly the disagreeing /agreeing parties when comparing the contextual representation against the overall one. An application of MDS is Nominate (Poole and Rosenthal, 1985) which is widely used to analyze the legislative roll-call voting behavior of parliamentarians in the United States Congresses. The interesting feature of Nominate is the fact that the projection dimensions convey more meaning than a standard MDS technique and can be used to describe political ideology of parlementarians (Hix, 2001; Hix, Noury, and Roland, 2006; Poole and Rosenthal, 2000). For example, in the U.S. congress and in the European parliament, the first dimension usually represents the Left/right positions which is the most used dimension to describe the voting behavior of parliamentarians.

Other unsupervised learning approaches can be used to provide additional insights on the voting behavior of individuals. For instance, Hierarchical Clustering (Murtagh and Contreras, 2012) and K-Nearest Neighbors (Fukunage and Narendra, 1975) can be leveraged to identify clusters of individuals in both the context and the entire collection of entities. This offers the possibility to compare and study how alignment may change between groups and how agreement can be formed or dissolved from a context to another.

6.2.2 Discovering Exceptional Contextual Clusters in Behavioral Data

In the same spirit of the final point discussed above, we can leverage clustering algorithms to provide contextual insights of the behavior of individuals and bring to the fore exceptional ones. This can fill the gap between local and global behavior models by providing a deep understanding of peculiar comportment of the whole population of interest.

For this aim, we can define via the EMM framework a new model class which uses a clustering algorithm for characterizing the (dis)agreement between individuals. The input considered for the clustering algorithm (e.g. Hierarchical Clustering (Murtagh and Contreras, 2012), K-Nearest Neighbors (Fukunage and Narendra, 1975), Community Detection (Blondel et al., 2008)) consists in a similarity/distance matrix. The latter leverages a defined similarity/distance between the outcomes expressed by individuals of the population of interest of interest. Exceptional Contextual Clusters patterns can be formalized similarly to the ones returned by DEvIANT as such, (u,c) which reads: "there is an exceptional clustering of individuals of group u in the context c when compared to the clustering of the same group in overall terms".

Once the clustering algorithm (model class) is appropriately defined, we need to define the associated interestingness measure which instantiates the meaning of "exceptional" in this setting. One can compare two clusterings by using variation of information (Meilă, 2007) which measures the amount of information lost and gained in changing from the overall clustering to the contextual clustering.

In this thesis, we were mainly interested in providing exhaustive search algorithms which rely on efficient pruning properties to avoid enumerating unpromising areas of the search space. For the problem of discovering exceptional contextual clusters, we need to investigate the properties of the interestingness measure to define proper optimistic estimates. Moreover, determining an incremental computation of the clustering from a context to a sub-context can be essential to the functioning of an algorithm which solves this problem since clustering algorithms are computationally expensive.

6.2.3 Discovering Change and Trends of Intra/Inter-Group Agreement

In the study of exceptional inter-group and intra-group agreement patterns (Chapter 3 and Chapter 4), time (if present) was simply considered as a numerical attribute. Hence time attribute was perceived as a static variable. While such considerations enabled the discovery of interesting and exceptional local patterns, it do not offer the possibility to uncover how time affects the behavior of groups in behavioral data. For this aim, a dynamic representation of time is required.

In this perspective work, both inter-group and intra-group agreement works can be extended by a dynamic representation of time to enable a longitudinal study of the interactions between individuals of the population of interest. For intra-group agreement patterns (u,c), one can transform time into a sequence of bins and evaluate the intra-group agreement measure (e.g. Krippendorff Alpha - cf. Chapter 4) for each bin both in the context c and in overall terms. Having this two sequences of measurements, one can use an EMM regression model class (Duivesteijn, Feelders, and Knobbe, 2012; Duivesteijn, Feelders, and Knobbe, 2016) induced on the context sequence and the overall one; and compare between the two regression models by using one of the proposed interestingness measures for this EMM instance (e.g. Cook's distance (Duivesteijn, Feelders, and Knobbe, 2012), Significance of Slope Difference (Leman, Feelders, and Knobbe, 2008) - see Section 2.3 of Chapter 2).

Similarly as for dynamic intra-group agreement patterns, one can consider EMM regression model to study the inter-group agreement patterns (c, u_1, u_2). For this objective, first a transformation of time to a sequence of time bins is required. This is followed by the evaluation of inter-group agreement measures (e.g. IAS) in each time bins both in the context c and in overall terms. Once this measuring is achieved, we can apply the regression model over both sequences and evaluate how exceptional the deviation is between the contextual regression model and the one evaluated in the overall terms.

Additional interesting measures in this setting can be investigated to enable the discovery of how time and context impact inter-group agreements or intra-group agreements. For instance, one can compare locally or globally the two built sequences of measurements as explained beforehand and then compare the two curves representing the sequences using a Freshet Distance (Alt and Godau, 1995; Eiter and Mannila, 1994). Moreover, to enable an exhaustive search algorithm when considering Freshet Distance, we need to investigate optimistic estimates to prune, as soon as possible, uninteresting areas of the search space induced by the context description language.

6.2.4 Anytime Exceptional Behaviors Mining

In this thesis, we have been mainly interested in providing exhaustive search algorithms that ensures the discovery of all the desired patterns for some given SD/EMM task. However, even when several optimization techniques are used to improve the efficiency of the exhaustive search algorithms, they become unfeasible when the search space grows in size (e.g. above descriptive attributes for entities collection and individuals collection). To alleviate this problem, we proposed in Chapter 3 Quick-DEBuNk which is a stochastic algorithm that makes tractable the discovery of exceptional inter-group agreement patterns. A similar approach, dubbed Quick-DEvIANT, can be devised to heuristically approximate the complete solution of the problem of discovering exceptional intra-group agreement patterns. This can offer an alternative tractable version of DEvIANT.

However, although these heuristic solutions offer a good trade-off between efficiency and effectiveness, they do not provide guarantees upon interruption on how far they are from the exact solution. In this spirit, Anytime Algorithms (Zilberstein, 1996) with guarantees can be used to provide error bounds of the best pattern found compared to the best pattern existing in the underlying search space. Towards this objective, we started investigating such paradigm in numerical data (all descriptive attributes are numerical) for standard Subgroup Discovey (e.g. using WRAcc (Lavrac, Flach, and Zupan, 1999) as an interesting measure). We proposed Refine&Mine (Belfodil, Belfodil, and Kaytoue, 2018), an anytime algorithm with four key properties: It yields progressively interval patterns whose quality improves over time; (ii) It can be interrupted anytime and always gives a guarantee bounding the error on the top pattern quality and (iii) It always bounds a distance to the exhaustive exploration; (iv) It converges to an exhaustive search algorithm if enough time is given, hence ensuring completeness. These are compelling properties that need to be investigated to see how they can be extended to our algorithms (Namely DEBuNk and DEvIANT) for providing anytime solutions to the problem of discovering exceptional (dis)agreement in behavioral data.

A

Study of DEBuNk and Quick-DEBuNk on synthetic data

In this appendix, we qualitatively compare DEBuNk and Quick-DEBuNk with standard state-of-the-art methods using artificially generated behavioral data. Additionally, we study their ability of finding the sought exceptional (dis)agreement patterns when confronted to noisy data.

Some questions we aim to answer require data for which the ground truth is known. Since it is notoriously difficult to obtain such data, we designed an artificial behavior data generator. The generator works as follows. It first generates `nb_hidden_patterns` inter-group agreement patterns. Each pattern is represented by two group descriptions (u_1, u_2) and a context (c) where u_1, u_2 and c are defined over random categorical descriptions and are of random size. For each pattern, the extent is generated (i.e., `context_support_size` entities for the context and the two groups involving `group_support_size` individuals). The extents are generated as follows: first, a random description **ds** is uniformly drawn from $\mathscr{D}_E$ (resp. $\mathscr{D}_I$). Next, `support_size` records are generated, which have a description equal to or subsumed by **ds**. This process is repeated for each component of the pattern so as to built `nb_hidden_patterns` inter-group agreement patterns. Note that the pattern generation process avoids overlapping between groups and contexts between differents patterns. These patterns describe conflictual situations, i.e., the individuals of one group in the pattern context express a voting outcome which is different from the other group' voting outcome. Conversely, the two groups are in agreement in the usual case, i.e., their votes over the entities outside the pattern context are similar. To achieve this, for each planted pattern (c, u_1, u_2) and for each entity $e \in G_E$, a random outcome **os** is drawn from the pool of possible outcomes (in here we consider $O = \{\text{Yes}, \text{No}\}$). Subsequently, each member comprising $G_I^{u_1}$ votes **os** for the entity e. Accordingly, individuals from $G_I^{u_2}$ cast a different outcome, if e is described by the context c. Otherwise, they cast the same outcome **os**. Once these patterns are generated, the rest of the dataset is generated by adding entities and individuals randomly while preserving the exceptionality of the patterns (i.e., the patterns must remain the most general exceptional patterns) till the desired size of the dataset is reached (i.e. $|G_E| =$ nb_entities and $|G_I| =$ nb_individuals). As described, the hidden patterns are pure. A last step enables to add noise within the patterns. For each pattern, the expressed outcome of individuals are randomly replaced with a noise_rate probability. Similarly, noise is added outside the patterns. Eventually, to get as close as possible to real-world behavioral dataset, we add sparsity in the data. To perform this task, each outcome of each pair $(i,e) \in G_I \times G_E$ have a probability of data_sparisty to be removed from the generated artificial behavioral dataset. The parameters used are summarized in Table A.1.

Parameter	Description	Default value
$\|G_E\|$ (**nb_entities**)	Number of entities	2000
$\|G_I\|$ (**nb_individuals**)	Number of individuals	500
$\|O\|$	Number of possible categorical outcomes	2
$\|\mathscr{A}_E\|$	Number of categorical attributes for entities	2
$\|dom(a_j)\|$ with $a_j \in \mathscr{A}_E$	Domain size of a categorical attribute $a_j \in \mathscr{A}_E$	4
$\|\mathscr{A}_I\|$	Number of categorical attributes for individuals	2
$\|dom(a_j)\|$ with $a_j \in \mathscr{A}_I$	Domain size of a categorical attribute $a_j \in \mathscr{A}_I$	4
nb_hidden_patterns	Number of planted conflictual patterns	3
context_support_size	Support size of a hidden pattern context	5
group_support_size	Support size of a hidden pattern group	5
noise_rate	Noise rate in/out the ground truth patterns	0
data_sparsity	Probability of an individual not to cast an outcome	0.33

Table A.1: Default Parameters Used for Generating Artificial Behavioral Data

A.1 Comparison to SD/EMM methods

We aim to study how the SD/EMM methods are able to discover relevant inter-group agreement patterns. SD algorithms available in public implementations (e.g., Vikamine(Atzmueller and Lemmerich, 2012), Cortana (Meeng and Knobbe, 2011), PySubgroup (Lemmerich and Becker, 2018)) only consider one flat table with a target attribute. However, behavioral datasets involve three relations (Entities, Individuals, Outcomes) which are all processed by DEBuNk and its sampling alternative Quick-DEBuNk to discover the interesting inter-group agreement patterns. To handle the problem we defined with a classical SD algorithm, we need to preprocess the data. We discuss and compare several problem adaptations.

SD-Majority: SD to discover contextual disagreements with the majority. The most direct way to apply SD on behavioral data is to consider the discovery of *groups* of individuals who express disagreement with the majority vote. This enables to discover patterns (c, g_1) where c is a context describing a set of entities and g_1 is a description of a group of individuals. To this end, we preprocess the behavioral data to obtain a Flat Behavioral Dataset (FBD) with a single table and a singe target class SAME_AS_MAJORITY as following: (1) we combine the entities and individuals tables using a join operation with the outcomes collection. (2) We compute the majority vote by aggregating the votes expressed on each entity. (3) We use this information to extend each record in the newly generated FBD with the attribute SAME_AS_MAJORITY which is equal to $+$, indicating that the individual voted in agreement with the majority in the considered entity. Otherwise SAME_AS_MAJORITY is equal to $-$. Example of FBD after such preprocessing is given in Table A.2. Having this FBD augmented with the target class SAME_AS_MAJORITY offers the possibility to run common SD techniques to identify subgroups with a high prevalence of disagreement with the majority (Target label = $'-'$). The most adapted interestingness measure in this case is the precision gain (Fürnkranz, Gamberger, and Lavrač, 2012), i.e. $Precision(subgroup) - \alpha^-$, which is high when there is a high disagreement in a subgroup compared to the disagreement observed in the full dataset. Note that this model does not fit perfectly our problem setting. It enables only the discovery of bi-set patterns (c, g_1) rather than the desired three-set patterns (c, g_1, u_2). Nevertheless, highlighting this type of pattern may help to partially identify interesting inter-group agreement patterns in a behavioral dataset. Furthermore, this adaptation does not takes into account the usual behavior of the group against the majority. This might clearly lead to the discovery of obvious patterns highlighting the individuals that are known to be a systematic opposition.

Entities			***Individuals***			***Outcomes***	
ide	theme	date	idi	country	group	outcome	SAME_AS_MAJORITY
e_1	1.20 Citizen's rights	20/04/16	i_1	France	S&D	For	$+$
...	...	...	...	...	...	...	...

Table A.2: Example of input data format for SD-Majority after transforming the behavioral dataset given in Table 3.1.

SD-Cartesian: SD to discover contextual disagreement between two groups. We propose a second modeling to enable the discovery of three-set patterns (c, u_1, u_2) with SD techniques. To this end, the behavioral dataset is transformed into a flat table equivalent to the Cartesian product $G_E \times G_I \times G_I$. This flat table is then augmented with a target class attribute SAME_VOTE which captures the (dis-)agreement between each couple of individuals on each entity for which both expressed an outcome. SAME_VOTE is thus equal to $+$ if both individuals expressed the same outcome for the entity, $-$ otherwise. This modeling – illustrated in Table A.3 – makes it possible to discover patterns (c, u_1, u_2) which identify two groups of individuals and a context regrouping a set of entities over which the individuals in the first group disagrees with the ones composing the second group. This can be done using the precision gain as the interestingness measure. Even if the syntax of the patterns is similar to ours, the usual agreement between the two selected groups is not take into account. Hence, the semantics conveyed by these patterns is different from ours. Another major drawback of such modeling is the size of the table resulting from the Cartesian product. For instance, a small behavioral dataset with 200 entities and 100 individuals can contain up to 2×10^6 records which clearly make this setting not adapted and not scalable for real-world behavioral data.

Entities		*Individuals*		*Individuals*		*Outcomes*		
ide	theme	idi_1	country_1	idi_2	country_2	$outcome_1$	$outcome_2$	SAME_VOTE
e_5	7.30	i_1	France	i_2	France	*For*	*For*	+
e_5	7.30	i_1	France	i_3	France	*For*	*Against*	−
...	...	...	...	...	...	...	...	

Table A.3: Example of input data format for SD-Cartesian after transforming the behavioral dataset given in Table 3.1 to a Cartesian product $G_E \times G_I \times G_I$.

Exceptional Contextual Subgraph Mining to discover contextual disagreement between two groups. Applying SD in the two aforementioned modelings does not allow to take into account the usual inter-group agreement in the model. A way to overcome this issue is to model the behavioral dataset as an attributed graph and looking for exceptional contextual subgraphs (Kaytoue et al., 2017). The so-called COSMIC algorithm is rooted in SD/EMM and aims at discovering contextual subgraphs whose edges have weights larger than expected. To this end, we transform the behavioral dataset to the Cartesian product $G_E \times G_I \times G_I$ extended with SAME_VOTE attribute like in *SD-Cartesian* formalization. This table is then used to build a bipartite graph where each side represents the collection of individuals G_I and an edge is instantiated between two vertices (individuals) for each entity on which the two individuals expressed conflicting outcomes. The set of transactions from $G_E \times G_I \times G_I$ where two individuals are disagree are associated to the edge between the two corresponding vertices (see Fig. A.1). Once this transactions set obtained, COSMIC algorithm can be used to obtain exceptional contextual subgraphs. Note that, in this problem setting, an exceptional contextual subgraph corresponds to two groups of individuals which exhibit a higher disagreement rate in the considered context compared to the disagreement expected in a similar sized subgraph. Several interestingness measures have been proposed in the COSMIC framework (Bendimerad et al., 2017b; Kaytoue et al., 2017). For the aim of this study, the

ide	themes	date
e_1	1.20	20/04/16
e_2	2.10	16/05/16
e_3	1.20; 7.30	04/06/16
...	...	...

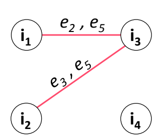
(a) Entities

idi	country	group	age
i_1	France	S&D	26
i_2	France	PPE	30
...	...	...	

(b) Individuals

id_edge	ide	idi_1	idi_2
t_1	e_2	i_1	i_3
t_2	e_5	i_1	i_3
t_3	e_3	i_2	i_3
t_4	e_5	i_2	i_3

(c) Transactions set (edges)



(d) Augmented Graph

Figure A.1: Example of input data format for Cosmic after transforming the behavioral dataset given in Table 3.1 to an augmented graph and its corresponding transactions set according to the observed discords.

lift measure is the most adapted: $\varphi(S) = \frac{\mathbb{P}(S|C)}{\mathbb{P}(S)}$ with S is the connected contextual subgraph induced by the selection performed by the description C. Note that: $\mathbb{P}(S|C)$ is the probability that a random drawn edge from all the edges in the full graph supporting the selection C falls in the induced contextual subgraph, $\mathbb{P}(S)$ is the relative weights in terms of the number of edges of the full subgraph S (the subgraph with the most general context). Note that a post-processing is necessary to transform exceptional contextual subgraphs into inter-group agreement patterns (c, u_1, u_2). Applying contextual subgraph mining given this modeling has some limitations: (1) the expected disagreement between two groups is computed from all the individuals instead of the individuals of the two groups. This can lead to the discovery of obvious patterns. (2) it considers as an input a transaction dataset computed from the Cartesian product $G_E \times G_I \times G_I$ which limits its usage, even for relatively small behavioral dataset.

We aim to compare how state-of-the-art methods perform in this three modelings and compare them to DEBuNk and Quick-DEBuNk. To this end, we generated 81 artificial dataset with 3 hidden patterns by varying several parameters (see Fig. A.2). Note that the behavioral datasets are relatively small to be sure to obtain results for each modeling, especially ones that requires to build a Cartesian product. For SD-Majority and SD-Cartesian modelings, we used PySubgroup(Lemmerich and Becker, 2018) to discover subgroups for the following reasons: the implementation is available online[1] as well as the easiness of its use. We ran the exhaustive search algorithm BSD (Lemmerich, Rohlfs, and Atzmueller, 2010) which is tailored to find relevant subgroups (Garriga, Kralj, and Lavrač, 2008), this choice is also motivated by the fact that the selected interestingness measure is the Precision gain. For the attributed graph modeling, we used an implementation of COSMIC algorithm provided by the authors (Kaytoue et al., 2017).

To evaluate the ability of the different approaches of uncovering planted patterns, we first define a similarity measure $sim_{\mathscr{P}}$ between two patterns $p = (c, u_1, u_2)$ and $p' = (c', u'_1, u'_2)$ from $\mathscr{P}$. It captures to what extent two patterns provide similar insights about changes of inter-group agreement.

$$\mathrm{sim}_{\mathscr{P}}(p,p') = \sqrt{J(G_E^c, G_E^{c'}) \times \frac{1}{2} \cdot \left(J(G_I^{u_1}, G_I^{u'_1}) + J(G_I^{u_2}, G_I^{u'_2})\right)} \text{ with } J(G,G') = \frac{|G \cap G'|}{|G \cup G'|} .$$

Note that, the quantity $(J(G_I^{u_1}, G_I^{u'_1}) + J(G_I^{u_2}, G_I^{u'_2}))$ is replaced by the following measure if the quality measure φ is symmetric:

$$\max(J(G_I^{u_1}, G_I^{u'_1}) + J(G_I^{u_2}, G_I^{u'_2}), J(G_I^{u_1}, G_I^{u'_2}) + J(G_I^{u_2}, G_I^{u'_1})) .$$

For comparing two pattern sets P, P' returned by respectively DEBuNk and Quick-DEBuNk, we use an F_1 score defined as follows.

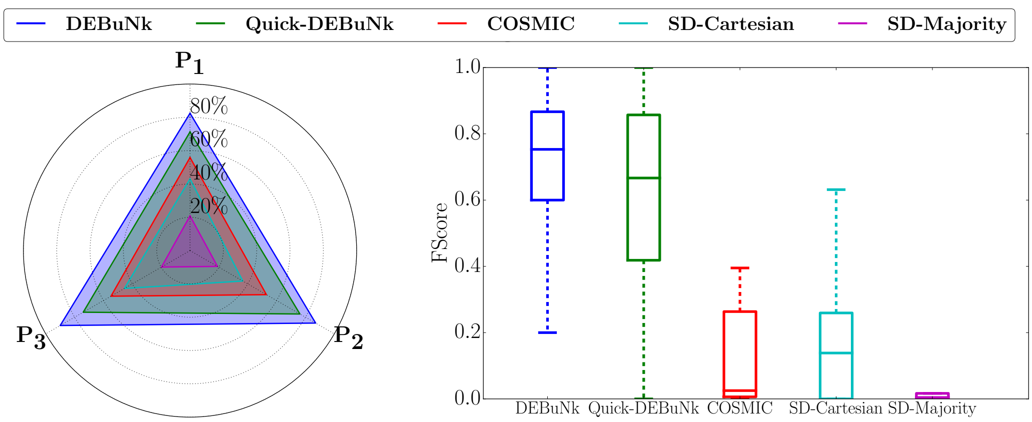
$$F_1(P,P') = 2 \cdot \frac{\mathrm{precision}(P,P') \cdot \mathrm{recall}(P,P')}{\mathrm{precision}(P,P') + \mathrm{recall}(P,P')} , \tag{A.1}$$

[1] https://bitbucket.org/florian_lemmerich/pysubgroup

$$\text{with} \begin{cases} \text{precision}(P,P') & = \dfrac{\sum_{p \in P} \max(\{\text{sim}_{\mathscr{P}}(p,p') \mid p' \in P'\})}{|P|}, \\ \text{recall}(P,P') & = \dfrac{\sum_{p' \in P'} \max(\{\text{sim}_{\mathscr{P}}(p',p) \mid p \in P\})}{|P'|}. \end{cases}$$

A similar measure to the recall has been proposed by Bosc et al., 2018 to evaluate the ability of their algorithm to retrieve the ground-truth patterns. We extend this measure with the precision to evaluate not only that all the hidden patterns have been discovered by an algorithm (i.e. recall=1.) but also the conciseness of the returned set (i.e. precision=1 if and only if all returned patterns are actually present in the behavioral dataset).

We report in Figure A.2a the comparative experiments between DEBuNk, Quick-DEBuNk, SD-Cartesian, SD-Majority and COSMIC in terms of their ability to retrieve each planted pattern in synthetic behavioral datasets. We report for each method the average similarity (over the 81 artificial data) between one of the three hidden patterns and its nearest representative in the result set. As expected, DEBuNk and Quick-DEBuNk outperforms other approaches. Moreover, the order between the approaches/modelings is sound. Majority-SD has the worst results due to the fact that this method, in the best case scenario, is only able to identify two of the three restrictions of a inter-group agreement pattern which impact on its performance. COSMIC performs slightly better than its alternative SD technique over the Cartesian product $G_E \times G_I \times G_I$ thanks to a more sophisticated model to capture the usual behavior.



(a) Average similarity between the planted patterns and their representatives returned by each method.

(b) Boxplots of F-score comparing the top-10 discovered patterns set by each method on each generated artificial data and the corresponding ground truth.

Figure A.2: Comparative qualitative performance study between DEBuNk ($\sigma_E = 3$, $\sigma_I = 3$, $\sigma_\varphi = 0.5$ and the quality measure φ_{dissent}), Quick-DEBuNk (same parameters as DEBuNk with *timebudget* $= 5$ *seconds*), SD-Majority (`resultSetSize`$= 50$, i.e. Top-50), SD-Cartesian (`resultSetSize`$= 25$, i.e. Top-25) and Cosmic (Default parameters) performed over 81 artificial behavioral data with 3 hidden patterns by varying the number of individuals in $[100, 125, 150]$, the number of entities in $[100, 150, 200]$, the sparsity factor in $[0., 0.25, 0.5]$ and the noise in $[0., 0.2, 0.4]$.

Figure A.2b summarizes the results obtained after running the five approaches. For a fair comparison (i.e., the problem of setting the good thresholds), we report the average F-Score of the only top-10 results for each approach. We observe that DEBuNk and Quick-DEBuNk achieves to return high-quality results compared to the other approaches. Interestingly, COSMIC adaptation is of less quality than SD-Cartesian adaptation when analyzing both their conciseness and exactitude in terms of hidden pattern identification. Finally SD-Majority performs the worst due to its fundamental difference with the other approaches when comparing the provided patterns format.

A.2 Robustness to noise and ability to discover hidden patterns

We now study the ability of DEBuNk and Quick-DEBuNk to discover hidden patterns for larger behavioral datasets as well as their robustness to noise. To this end, we carried out DEBuNk and Quick-DEBuNk over several artificial datasets varying the noise rate from 0 to 0.8. The results illustrated in Figure A.3 demonstrates that the exhaustive search approach DEBuNk is able to discover almost exclusively all the hidden patterns ($F1\_Score > 0.8$) even if the noise rate is rather high (≤ 0.6). Indeed when the noise rate is substantially high, *DEBuNk* does not retrieve the noisy hidden patterns. This clearly results from the evidence that several planted patterns disappear in the underlying artificially generated data after adding too much noise. This is an advantage for DEBuNk since the quality threshold is able to remove nonsensical patterns from the final set. In contrast, from these experiments, we observe that Quick-DEBuNk less robust to noise than DEBuNk. The performance of Quick-DEBuNk in terms of finding hidden patterns decreases faster with regard to the noise rate compared to DEBuNk. This is mainly due to the random walk procedure (RWC) which considers other sub search space than the one actually containing a hidden context as the noise reduces the quality of its subsuming parents. Still, it is worth mentioning that Quick-DEBuNk is able to retrieve partially planted patterns even when the noise is rather high. Interestingly, the sampling approach achieves a comparable precision to the exhaustive approach, this demonstrates that most of returned patterns are valid.

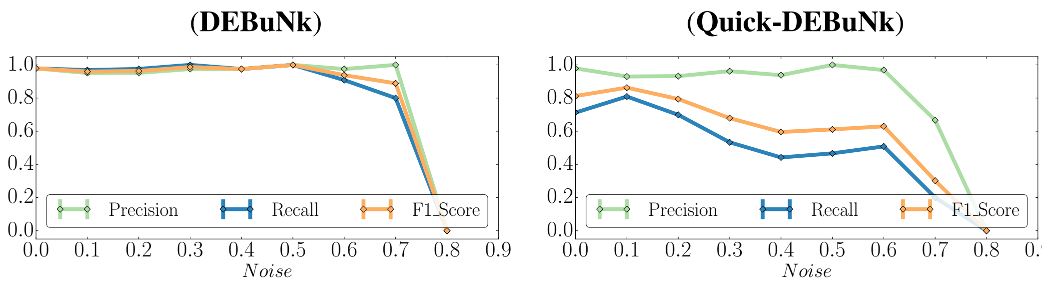


Figure A.3: Efficiency of DEBuNk ($\sigma_E = 7$, $\sigma_I = 7$, $\sigma_\varphi = 0.5$ and $\varphi_{dissent}$) and Quick-DEBuNk ($\sigma_E = 7$, $\sigma_I = 7$, $\sigma_\varphi = 0.5$, $timebudget =$ 3 mn and $\varphi_{dissent}$) performed over 21 behavioral artificial data generated with the following default parameters ($|G_E| = 2000$, $|G_I| = 500$, $|\mathscr{A}_E|$ = $|\mathscr{A}_I|$ =3, ***size_dom_entities_attributes*** = ***size_dom_individuals_attributes*** =4, ***nb_hidden_patterns*** =5, ***context_support_size*** =10, ***group_support_size*** =10).

B

Multiple Comparisons Problem

In what follows, each pattern $H_i = (u_i, c_i)$ is seen as a hypothesis test which returns a p-value p_i. Recall that, in this thesis (Chapter 4), the list of hypotheses to test corresponds to the full search space $L = \{(u,c) \in \mathscr{D}_I \times \mathscr{D}_E : |G_I^u| \geq \sigma_I \text{ and } |G_E^c| \geq \sigma_E\}$ where u (resp. c) is a closed description (i.e. the maximum description w.r.t. $\sqsubseteq$) in the equivalence class $[u]$ (resp. $[c]$) of descriptions having their extent equal to G_I^u (resp. G_E^c), i.e. $[u] = \{u' \in \mathscr{D}_I \text{ s.t. } G_I^{u'} = G_I^u\}$ (resp. $[c] = \{c' \in \mathscr{D}_E \text{ s.t. } G_E^{c'} = G_E^c\}$). Having this in mind, in what follows, the content of L is shortly denoted by $L = \{H_1, \ldots, H_\omega\}$ and comprises ω hypotheses. Hypotheses in L are ordered by their p-values $\{p_1, \ldots, p_\omega\}$ where $p_i = \textit{p-value}^{u_i}(c_i)$.

The Multiple Comparisons Problem (MCP) (Holm, 1979) is a critical issue in significant pattern mining (Hämäläinen and Webb, 2019). In a nutshell, given the critical value α which roughly corresponds to the probability of type 1 error (rejecting a true null hypothesis which is equivalent to accepting a spurious pattern), it is to be expected that $\omega \cdot \alpha$ hypotheses will erroneously pass the test, i.e., $\omega \cdot \alpha$ hypotheses suffer a type 1 error. The classic approach to deal with the MCP is to control the *family wise error rate* (FWER), which is the probability of accepting at least one false discovery. Other approaches control the *false discovery rate* (FDR), which corresponds to the expected proportion of false discoveries. We give an overview of relevant existing approaches that deal with the MCP and point out why using them in our setting is a non-trivial task. For a survey on methods dealing with the MCP, we refer the interested reader to (Hämäläinen and Webb, 2019).

The most famous method to control FWER at $\leq \gamma$ (typically 0.05) is Bonferroni adjustments (Dunn, 1961). The critical α used to test the significance of a pattern is adjusted to $\frac{\gamma}{\omega}$ so as to have FWER at $\leq \gamma$ with ω the number of all patterns to test in L. The problem with this approach is that when ω is huge[1], Bonferroni adjusts α to a value very close to 0. This leads to a high number of false negatives as most interesting pattern will be considered

[1]Which is the case in the general setting of pattern mining even if we consider only closed patterns satisfying the support size threshold constraint.

spurious (high Type 2 error rate). Clearly, ω is unknown and needs, in the most trivial way, to be bounded by a quantity ω_0 which is **larger** than ω. Usually, ω_0 corresponds to the maximum size of the search space: it is equal to $2^{\#\text{items}}$ in the case of an itemset dataset. Webb, 2007 gives a bound on the size of the search space when dealing with the MCP in attribute-value datasets when the description length is bounded. Using this reasoning without bounding the description length and considering the specification of each attribute (numerical, categorical, ...), in the smallest of our datasets (Movielens; see Table 3) we have $\omega_0 = 72\,349\,200$. This requires α to be equal to 6.92×10^{-10} for the FWER to be at ≤ 0.05. All the other datasets require α to be $\leq 10^{-76}$ when bounding ω with the size of the search space. Clearly, such settings for α prohibit the discovery of any meaningful information from the datasets, which cannot possibly be the desired effect of attempts at solving the MCP.

Several techniques exist in the literature to relax the requirements on α while ensuring a FWER at $\leq \gamma$ in order to increase the statistical power:

1. Terada et al. (Terada, duVerle, and Tsuda, 2016; Terada et al., 2013) propose the LAMP technique, relying on Tarone's Exclusion Principle (TEP) (Tarone, 1990). This principle stipulates that in the list of m hypotheses in L to be tested, one must ignore *untestable patterns* for multiple comparisons. A pattern H_i is said to be *untestable* if the **lower bound of its p-value**, denoted p_i^*, is under the adjusted $\alpha = \frac{\gamma}{m}$. Terada et al., 2013 proposed this lower bound p_i^* for the particular task of finding significant rules[2] (Webb, 2006) where significance is commonly assessed using a Fisher exact test (Hämäläinen, 2010a; Hämäläinen, 2010b), since a 2×2 contingency table is available. The lower bound p_i^* computation depends on this contingency table. Clearly, there is no trivial mapping of our problem to the problem of finding significant rules. Hence, adapting the LAMP algorithm to have an efficient branch and bound technique, incorporating both the proposed bounds in this work (the DEvIANT algorithm) and LAMP reasoning, is clearly a daunting task that requires an in-depth investigation and a new devoted approach which is beyond the scope of this work.

2. Similarly, most of the existing work measuring the interestingness of patterns with statistical significance while efficiently handling the MCP, deals with the significant rule discovery setting (Komiyama et al., 2017; Llinares-López et al., 2015; Pellegrina and Vandin, 2018; Terada, Tsuda, and Sese, 2013). Some of these methods (Llinares-López et al., 2015; Pellegrina and Vandin, 2018; Terada, Tsuda, and Sese, 2013) rely on the Westfall-Young permutation testing method (Westfall and Young, 1993) to increase statistical power. Still, no straightforward application of these techniques in our setting is possible: these methods perform random permutations on the class label, and no class label is given in the problem addressed in our work.

3. Other state-of-the-art techniques follow a multi-stage procedure (Hämäläinen and Webb, 2019) to tackle the MCP. A first step constrains L to a subset of patterns (e.g., testable under TEP). A subsequent post-processing phase controls the FWER (Webb, 2007) or FDR (Komiyama et al., 2017; Webb, 2007). For example, Webb, 2007 proposes to divide the data into Exploratory and Holdout data. Hypotheses are sought

[2]Each record in the underlying dataset is associated with a binary target label and the objective is to find rules that have significant association with one of the two labels.

by analyzing solely the exploratory data. Eventually, a constrained number of patterns are found which are validated against the holdout data. In our setting, one needs to investigate how to divide the data into these two parts, since we have two dimensions: context space and group space. In this configuration, a question of crucial importance must be answered: do we need to consider each group independently and divide the entities dataset (defining the context space) into exploratory vs holdout data for each group? Or do we need to jointly consider both these dimensions? This clearly requires a thorough investigation to avoid proposing a naive solution.

4. Layered critical values (Bay and Pazzani, 2001; Webb, 2008) propose to consider a varying adjustment factor for each level of the search space as long as the sum of all critical values is not above γ. This requires:

   - estimating the size of each level (which could be done by following the reasoning of Webb, 2008);
   - identifying what is a level of the search space: do we consider levels jointly between group and context search space?

   Choosing joint consideration in the latter bullet point implies ignoring (most of the time) the level-1 groups in the search space: the level will grow in size after considering all the contexts corresponding to the group characterizing the whole collection of individuals. Otherwise, the question raised in the former bullet point needs to be answered to provide an appropriate algorithm. Furthermore, combining the layered critical values along with DEvIANT is not straightforward as it requires re-investigation of the proposed pruning properties.

As we can see, several fundamental questions remain to be answered before one could incorporate a solution to the MCP in the task of finding significant exceptional contextual intra-group agreement patterns. We argue that the scope of this problem is bigger than the work introduced in Chapter 4; it is a non-trivial task that deserves proper attention in the wider context of the significant pattern mining paradigm. We plan to investigate this in future work.

C

Symbol Table (Chapter 1 and 2)

Symbol	Definition
G_E	A finite collection of records depicting entities
G_I	A finite collection of records depicting individuals
O	The domain of possible outcomes
o	A function $o : G_E \times G_I \to O$ returning the outcome $o(i,e)$ of an individual i over an entity e
$\mathscr{B}$	$= \langle G_I, G_E, O, o \rangle$; A behavioral dataset (cf. Definition 1.1.1)
$\mathscr{A}$	$\mathscr{A}_E$ (resp. $\mathscr{A}_I$): Descriptive attributes of entities (resp. individuals)
$\mathscr{D}$	$\mathscr{D}_E$ (resp. $\mathscr{D}_I$): The description domain of contexts (resp. groups)
u	$\in \mathscr{D}_I$; A description (cf. Definitions 2.2.2 and 2.2.12) of a group (cf. Definition 1.1.2)
c	$\in \mathscr{D}_E$; a description defining a context (cf. Definition 1.1.3)
G_E^c	A subgroup of entities corresponding to the extent (cf. Definition 2.2.3) a context $c \in \mathscr{D}_E$
G_I^u	A subgroup of individuals corresponding to the extent of a group description $g \in \mathscr{D}_I$
	Now, we omit the indices I or E in the notations and we consider that we have a collection of records G, its schema of attributes $\mathscr{A}$ and the related description space $\mathscr{D}$
δ	a mapping function $\delta : G \to \mathscr{D}$ which maps each record g to its maximum corresponding description $\delta(g) \in \mathscr{D}$ w.r.t. $\sqsubseteq$. The definition is extended to return the maximum description shared between records in some subset in G
$\sqsubseteq$	read "less restrictive than" is a partial order (cf. Definition 2.2.4) between descriptions in some description space $\mathscr{D}$
G^d	$= \text{ext}(d)$ is the extent (subgroup; cf. Definition 2.2.3) of a description $d \in \mathscr{D}$ in G, i.e. $G^d = \{g \in G \text{ s.t. } d \sqsubseteq \delta(g)\}$.
$\langle G, (\mathscr{D}, \sqsubseteq), \delta \rangle$	a pattern structure (cf. Definition 2.2.7)
$\text{clo}(d)$	$= \delta(G^d)$ a closure operator in $\mathscr{D}$.
$\eta(d)$	a refinement operator (cf. Definition 2.2.5) which return the neighbors $\eta(d) \subseteq \mathscr{D}$ of a description $d \in \mathscr{D}$ w.r.t. $\sqsubseteq$; i.e. $\eta(d) = \{d' \in \mathscr{D} \text{ s.t. } d \sqsubset d' \wedge \nexists e \in \mathscr{D} : d \sqsubset e \sqsubset d'\}$
φ	$\varphi : \mathscr{D} \to \mathbb{R}$ is the interestingness (quality) measure (cf. Definition 2.2.6). The quality measure is extent-based, hence we can define φ as such: $\varphi : 2^G \to \mathbb{R}$ with: $\forall d \in \mathscr{D}. \varphi(d) = \varphi(G^d)$
oe	$\text{oe} : \mathscr{D} \to \mathbb{R}$ is the optimistic estimate (cf. Definition 2.4.1) associated to the quality measure φ.

Table C.1: Symbol table related to Chapter 1 and Chapter 2

D

Symbol Table (Chapter 3)

Symbol	Definition
$\mathscr{P}$	$= \mathscr{D}_E \times \mathscr{D}_I \times \mathscr{D}_I$ and denotes the pattern space
p	$= (c, u_1, u_2) \in \mathscr{P}$ is an inter-group agreement pattern where
	c is a context and (u_1, u_2) two group of individuals
p^*	$= (*, u_1, u_2) \in \mathscr{P}$ is the referential inter-group agreement pattern
	related to some pattern $p = (c, u_1, u_2)$
P	$\subseteq \mathscr{P}$ denotes a pattern set returned by DEBuNk or Quick-DEBuNk
θ	An outcome aggregation measure
sim	a similarity function between two aggregated outcomes
IAS	Inter-group Agreement Similarity Measure
φ	An interestingness measure

Table D.1: Symbol Table related to Chapter 3

E

Symbol Table (Chapter 4)

Symbol	Definition
$\mathscr{P}$	$= \mathscr{D}_I \times \mathscr{D}_E$ and denotes the pattern space
p	$= (u, c) \in \mathscr{P}$; is an intra-group agreement pattern where u is a group and c a context
P	$\subseteq \mathscr{P}$ denotes the returned pattern set by DEvIANT
$\mathscr{B}^g$	The reduced behavioral dataset for individuals comprising G_I^g
A	Intra-group agreement measure - Krippendorff's Alpha
$A^u(G_E^c)$	Intra-group agreement of a group u over a context c
$p\text{-}value^u(c)$	p-value of an observed $A^u(G_E^c)$ of a group g over a context c considering the DFD
	We omit the exponent g in the notations and we assume that
	we have a group of individuals g in mind (we use $\mathscr{B}^g$)
D_{exp}	Expected disagreement (via marginal distribution) between individuals
D_{obs}	Observed disagreement between individuals
n	Number of entities in G_E, i.e., $\lvert G_E \rvert$
m	Number of all expressed outcomes
m^{o_1}	Number of expressed outcomes equal to o_1
m_e	Number of expressed outcomes for entity e (also denoted w_e)
$m_e^{o_1}$	Number of expressed outcomes equal to o_1 for entity e
$\delta_{o_1 o_2}$	Distance between two outcomes in O
DFD	Distribution of False discoveries
F_k	$F_k = \{S \subseteq G_E \text{ s.t. } \lvert S \rvert = k\}$
θ_k	Random variable $\theta_k : F_k \to \mathbb{R}$ with $S \mapsto A(S)$. Also $\theta_k = \frac{V_k}{W_k}$
v_e	Intra-group agreement (Krippendorff's Alpha) for one entity,
V_k	Random variable $V_k : F_k \to \mathbb{R}$ with $S \mapsto \frac{1}{k} \sum_{e \in S} v_e$
W_k	Random variable $W_k : F_k \to \mathbb{R}$ with $S \mapsto \frac{1}{k} \sum_{e \in S} w_e$
α	Critical value
$CI_k^{1-\alpha}$	The $1-\alpha$ confidence interval associated with the DFD of θ_k.
$\widehat{CI}_k^{1-\alpha}$	The $1-\alpha$ Taylor-approximated confidence interval of $CI_k^{1-\alpha}$.
$\widehat{CI}_{\text{bootstrap}}^{1-\alpha}$	The bootstrap confidence interval.
$LB(S, \sigma_E)$	Lower bound of A for any specialization of a subgroup having its size greater than σ_E
$UB(S, \sigma_E)$	Upper bound of A for any specialization of a subgroup having its size greater than σ_E
$OE(S, \sigma_E)$	= $[LB(S, \sigma_E), UB(S, \sigma_E)]$. Optimistic estimate region of A

Table E.1: Symbol Table related to Chapter 4

References

Abudawood, Tarek and Peter A. Flach (2009). "Evaluation Measures for Multi-class Subgroup Discovery". In: *ECML/PKDD (1)*. Volume 5781. Lecture Notes in Computer Science. Springer, pages 35–50 (cited on page 49).

Agrawal, Rakesh, Tomasz Imielinski, and Arun N. Swami (1993). "Mining Association Rules between Sets of Items in Large Databases". In: *SIGMOD Conference*. ACM Press, pages 207–216 (cited on pages 21, 29, 31).

Agrawal, Rakesh and Ramakrishnan Srikant (1994). "Fast Algorithms for Mining Association Rules in Large Databases". In: *VLDB*. Morgan Kaufmann, pages 487–499 (cited on page 35).

Agrawal, Rakesh, Ramakrishnan Srikant, et al. (1995). "Mining sequential patterns". In: *icde*. Volume 95, pages 3–14 (cited on page 21).

Al Hasan, Mohammad and Mohammed J Zaki (2009). "Output space sampling for graph patterns". In: *Proceedings of the VLDB Endowment* 2.1, pages 730–741 (cited on pages 34, 36, 71).

Alejandro, Jennifer (2010). "Journalism in the age of social media". In: *Reuters Institute Fellowship Paper*, page 5 (cited on page 1).

Alt, Helmut and Michael Godau (1995). "Computing the Fréchet distance between two polygonal curves". In: *International Journal of Computational Geometry & Applications* 5.01n02, pages 75–91 (cited on page 153).

Amelio, Alessia and Clara Pizzuti (2012). "Analyzing voting behavior in Italian Parliament: Group cohesion and evolution". In: *ASONAM*. IEEE, pages 140–146 (cited on pages 8, 82).

Amer-Yahia, Sihem, Sofia Kleisarchaki, Naresh Kumar Kolloju, Laks V. S. Lakshmanan, and Ruben H. Zamar (2017). "Exploring Rated Datasets with Rating Maps". In: *WWW*. ACM, pages 1411–1419 (cited on pages 10, 62).

Arora, Sanjeev and Boaz Barak (2009). *Computational complexity: a modern approach*. Cambridge University Press (cited on page 46).

Atzmueller, Martin (2015). "Subgroup discovery". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.1, pages 35–49 (cited on pages 10, 18, 33).

— (2017). "Descriptive Community Detection". In: *Formal Concept Analysis of Social Networks*. Springer, pages 41–58 (cited on page 8).

Atzmueller, Martin, Stephan Doerfel, and Folke Mitzlaff (2016). "Description-oriented community detection using exhaustive subgroup discovery". In: *Inf. Sci.* 329, pages 965–984 (cited on page 8).

Atzmueller, Martin and Florian Lemmerich (2012). "VIKAMINE–open-source subgroup discovery, pattern mining, and analytics". In: *Joint European Conference on Machine

Learning and Knowledge Discovery in Databases. Springer, pages 842–845 (cited on pages 35, 156).

Atzmüller, Martin and Florian Lemmerich (2009). "Fast Subgroup Discovery for Continuous Target Concepts". In: *ISMIS*, pages 35–44 (cited on pages 35, 45).

Atzmüller, Martin and Frank Puppe (2006). "SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 6–17 (cited on pages 18, 22, 34, 35, 45).

Bay, Stephen D and Michael J Pazzani (2001). "Detecting group differences: Mining contrast sets". In: *Data mining and knowledge discovery* 5.3, pages 213–246 (cited on pages 18, 165).

Becker, Martin, Florian Lemmerich, Philipp Singer, Markus Strohmaier, and Andreas Hotho (2017). "MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data". In: *Data Min. Knowl. Discov.* 31.5, pages 1359–1390 (cited on page 11).

Behrens, John T (1997). "Principles and procedures of exploratory data analysis." In: *Psychological Methods* 2.2, page 131 (cited on page 7).

Belfodil, Adnene, Sylvie Cazalens, Philippe Lamarre, and Marc Plantevit (2017a). "Flash Points: Discovering Exceptional Pairwise Behaviors in Vote or Rating Data". In: *ECML/PKDD (2)*. Volume 10535. Lecture Notes in Computer Science. Springer, pages 442–458 (cited on pages 14, 46, 55, 56, 59, 70, 86).

Belfodil, Adnene, Wouter Duivesteijn, Marc Plantevit, Sylvie Cazalens, and Philippe Lamarre (2019a). "DEvIANT: Discovering significant exceptional (dis-)agreement within groups". In: *ECML/PKDD* (cited on pages 14, 102).

Belfodil, Adnene, Aimene Belfodil, Anes Bendimerad, Philippe Lamarre, Celine Robardet, Mehdi Kaytoue, and Marc Plantevit (2019b). "FSSD – A Fast and Efficient Algorithm for Subgroup Set Discovery". In: *DSAA*. IEEE (cited on page 45).

Belfodil, Adnene, Sylvie Cazalens, Philippe Lamarre, and Marc Plantevit (Feb. 2019c). *Identifying exceptional (dis)agreement between groups (Technical Report)*. Technical report. LIRIS UMR CNRS 5205. URL: https://contentcheck.liris.cnrs.fr/ (cited on pages 14, 55, 56).

Belfodil, Aimene, Adnene Belfodil, and Mehdi Kaytoue (2018). "Anytime Subgroup Discovery in Numerical Domains with Guarantees". In: *ECML/PKDD (2)*. Volume 11052. Lecture Notes in Computer Science. Springer, pages 500–516 (cited on pages 34, 37, 45, 153).

— (2019). "Mining Formal Concepts using Implications between Items". In: *International Conference on Formal Concept Analysis (ICFCA 2019)* (cited on page 64).

Belfodil, Aimene, Sergei O. Kuznetsov, and Mehdi Kaytoue (2018). "Pattern Setups and Their Completions". In: *CLA*. Volume 2123. CEUR Workshop Proceedings. CEUR-WS.org, pages 243–253 (cited on page 25).

— (2019). "On Pattern Setups and Pattern Multistructures". In: *CoRR* abs/1906.02963 (cited on page 25).

Belfodil, Aimene, Sergei O. Kuznetsov, Céline Robardet, and Mehdi Kaytoue (2017b). "Mining Convex Polygon Patterns with Formal Concept Analysis". In: *IJCAI*. ijcai.org, pages 1425–1432 (cited on page 21).

Bendimerad, Ahmed Anes, Marc Plantevit, and Céline Robardet (2016). "Unsupervised Exceptional Attributed Sub-Graph Mining in Urban Data". In: *ICDM*, pages 21–30 (cited on pages 18, 43, 51).

— (2018). "Mining exceptional closed patterns in attributed graphs". In: *Knowl. Inf. Syst.* 56.1, pages 1–25 (cited on page 43).

Bendimerad, Ahmed Anes, Rémy Cazabet, Marc Plantevit, and Céline Robardet (2017a). "Contextual Subgraph Discovery with Mobility Models". In: *COMPLEX NETWORKS*, pages 477–489 (cited on page 11).

Bendimerad, Anes (2019). "Mining Useful Patterns in Attributed Graphs". PhD thesis. University of Lyon, France (cited on page 42).

Bendimerad, Anes, Rémy Cazabet, Marc Plantevit, and Céline Robardet (2017b). "Contextual Subgraph Discovery With Mobility Models". In: *International Workshop on Complex Networks and their Applications*. Springer, pages 477–489 (cited on pages 4, 18, 51, 158).

Bie, Tijl De (2011a). "An information theoretic framework for data mining". In: *KDD*. ACM, pages 564–572 (cited on pages 23, 34, 108).

— (2011b). "Maximum entropy models and subjective interestingness: an application to tiles in binary databases". In: *Data Min. Knowl. Discov.* 23.3, pages 407–446 (cited on page 34).

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008 (cited on pages 8, 152).

Boley, Mario, Thomas Gärtner, and Henrik Grosskreutz (2010). "Formal concept sampling for counting and threshold-free local pattern mining". In: *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, pages 177–188 (cited on pages 34, 36, 71, 72).

Boley, Mario, Sandy Moens, and Thomas Gärtner (2012). "Linear space direct pattern sampling using coupling from the past". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 69–77 (cited on pages 34, 37, 72).

Boley, Mario, Tamás Horváth, Axel Poigné, and Stefan Wrobel (2010). "Listing closed sets of strongly accessible set systems with applications to data mining". In: *Theoretical Computer Science* 411.3, pages 691–700 (cited on pages 30, 36, 40, 46).

Boley, Mario, Claudio Lucchese, Daniel Paurat, and Thomas Gärtner (2011). "Direct local pattern sampling by efficient two-step random procedures". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 582–590 (cited on pages 34, 37, 45, 71, 72).

Boley, Mario, Bryan R. Goldsmith, Luca M. Ghiringhelli, and Jilles Vreeken (2017). "Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery". In: *Data Min. Knowl. Discov.* 31.5, pages 1391–1418 (cited on page 34).

Bonaque, Raphaël, T. D. Cao, Bogdan Cautis, François Goasdoué, J. Letelier, Ioana Manolescu, O. Mendoza, S. Ribeiro, Xavier Tannier, and Michaël Thomazo (2016). "Mixed-instance querying: a lightweight integration architecture for data journalism". In: *PVLDB* 9.13, pages 1513–1516 (cited on page 130).

Bonchi, Francesco, Fosca Giannotti, Alessio Mazzanti, and Dino Pedreschi (2003). "ExAMiner: Optimized level-wise frequent pattern mining with monotone constraints". In: *Third IEEE International Conference on Data Mining*. IEEE, pages 11–18 (cited on page 35).

Bonchi, Francesco, Fosca Giannotti, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Roberto Trasarti (2009). "A constraint-based querying system for exploratory pattern discovery". In: *Information Systems* 34.1, pages 3–27 (cited on page 35).

Bosc, Guillaume, Jérôme Golebiowski, Moustafa Bensafi, Céline Robardet, Marc Plantevit, Jean-François Boulicaut, and Mehdi Kaytoue (2016). "Local subgroup discovery for eliciting and understanding new structure-odor relationships". In: *International Conference on Discovery Science*. Springer, pages 19–34 (cited on page 51).

Bosc, Guillaume, Jean-François Boulicaut, Chedy Raïssi, and Mehdi Kaytoue (2018). "Anytime discovery of a diverse set of patterns with Monte Carlo tree search". In: *Data Mining and Knowledge Discovery* 32.3, pages 604–650 (cited on pages 34, 37, 45, 93, 160).

Boulicaut, Jean-François and Baptiste Jeudy (2009). "Constraint-based data mining". In: *Data mining and knowledge discovery handbook*. Springer, pages 339–354 (cited on page 35).

Breiman, Leo, J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth (cited on page 9).

Cao, Nan, Yu-Ru Lin, Liangyue Li, and Hanghang Tong (2015). "g-Miner: Interactive Visual Group Mining on Multivariate Graphs". In: *CHI*. ACM, pages 279–288 (cited on page 9).

Carmona, Cristóbal J., M. J. del Jesus, and Francisco Herrera (2018). "A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy". In: *Knowl.-Based Syst.* 139, pages 89–100 (cited on page 18).

Carmona, Cristóbal J., Pedro González, María José del Jesús, and Francisco Herrera (2014). "Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms". In: *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 4.2, pages 87–103 (cited on pages 34, 36).

Caswell, David and Konstantin Dörr (2018). "Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories". In: *Journalism practice* 12.4, pages 477–496 (cited on page 2).

Cazalens, Sylvie, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier (2018). "A Content Management Perspective on Fact-Checking". In: *" Journalism, Misinformation and Fact Checking" alternate paper track of" The Web Conference"* (cited on pages 2, 130).

Charalabidis, Yannis, Charalampos Alexopoulos, and Euripidis Loukis (2016). "A taxonomy of open government data research areas and topics". In: *Journal of Organizational Computing and Electronic Commerce* 26.1-2, pages 41–63 (cited on pages 2, 82).

Cherepnalkoski, Darko, Andreas Karpf, Igor Mozetič, and Miha Grčar (2016). "Cohesion and coalition formation in the European Parliament: roll-call votes and Twitter activities". In: *PloS one* 11.11, e0166586 (cited on page 120).

Ciampaglia, Giovanni Luca, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini (2015). "Computational fact checking from knowledge networks". In: *PloS one* 10.6, e0128193 (cited on page 130).

Clinton, Joshua, Simon Jackman, and Douglas Rivers (2004). “The statistical analysis of roll call data”. In: *American Political Science Review* 98.2, pages 355–370 (cited on page 7).

Coddington, Mark (2015). “Clarifying journalism’s quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting”. In: *Digital journalism* 3.3, pages 331–348 (cited on page 2).

Cohen, J. (1960). “A Coefficient of Agreement for Nominal Scales”. In: *Education and Psychological Measurement* 20, pages 37–46 (cited on page 105).

Cohen, Sarah, Chengkai Li, Jun Yang, and Cong Yu (2011). “Computational Journalism: A Call to Arms to Database Researchers”. In: *CIDR*. www.cidrdb.org, pages 148–151 (cited on pages 2, 3, 7).

Cover, Thomas and Joy Thomas (2012). *Elements of information theory*. John Wiley & Sons (cited on page 108).

Cox, Trevor F and Michael AA Cox (2000). *Multidimensional scaling*. Chapman and hall/CRC (cited on pages 9, 151).

Csisz, I et al. (1967). “Information-type measures of difference of probability distributions and indirect observations”. In: *Studia Sci. Math. Hungar.* 2, pages 299–318 (cited on page 62).

Das, Mahashweta, Sihem Amer-Yahia, Gautam Das, and Cong Yu (2011). “Mri: Meaningful interpretations of collaborative ratings”. In: *PVLDB* 4.11, pages 1063–1074 (cited on pages 4, 8, 10).

Davey, Brian A and Hilary A Priestley (2002). *Introduction to lattices and order*. Cambridge university press (cited on page 25).

De Nooy, Wouter, Andrej Mrvar, and Vladimir Batagelj (2018). *Exploratory social network analysis with Pajek: Revised and expanded edition for updated software*. Volume 46. Cambridge University Press (cited on page 7).

Demsar, Janez, Tomaz Curk, Ales Erjavec, Crtomir Gorup, Tomaz Hocevar, Mitar Milutinovic, Martin Mozina, Matija Polajnar, Marko Toplak, Anze Staric, Miha Stajdohar, Lan Umek, Lan Zagar, Jure Zbontar, Marinka Zitnik, and Blaz Zupan (2013). “Orange: data mining toolbox in python”. In: *Journal of Machine Learning Research* 14.1, pages 2349–2353 (cited on page 35).

Dimitriadou, Kyriaki, Olga Papaemmanouil, and Yanlei Diao (2014). “Explore-by-example: an automatic query steering framework for interactive data exploration”. In: *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. ACM, pages 517–528 (cited on page 9).

Diop, Lamine, Cheikh Talibouya Diop, Arnaud Giacometti, Dominique Li, and Arnaud Soulet (2018). “Sequential Pattern Sampling with Norm Constraints”. In: *ICDM*. IEEE Computer Society, pages 89–98 (cited on pages 34, 37).

Dong, Guozhu and Jinyan Li (1999). “Efficient mining of emerging patterns: Discovering trends and differences”. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 43–52 (cited on page 18).

Downar, Lennart and Wouter Duivesteijn (2015). “Exceptionally Monotone Models - The Rank Correlation Model Class for Exceptional Model Mining”. In: *ICDM*. IEEE Computer Society, pages 111–120 (cited on page 41).

Downar, Lennart and Wouter Duivesteijn (2017). “Exceptionally monotone models—the rank correlation model class for exceptional model mining”. In: *Knowledge and Information Systems* 51.2, pages 369–394 (cited on page 41).

Du, Xin, Wouter Duivesteijn, and Mykola Pechenizkiy (2018). “ELBA: Exceptional Learning Behavior Analysis”. In: *EDM*. International Educational Data Mining Society (IEDMS) (cited on page 10).

Duivesteijn, Wouter, Ad Feelders, and Arno J. Knobbe (2012). “Different slopes for different folks: mining for exceptional regression models with cook’s distance”. In: *KDD*. ACM, pages 868–876 (cited on pages 42, 152, 153).

Duivesteijn, Wouter, Ad J Feelders, and Arno Knobbe (2016). “Exceptional model mining”. In: *Data Mining and Knowledge Discovery* 30.1, pages 47–98 (cited on pages 3, 7, 11, 18, 23, 38, 39, 41, 42, 45, 100, 103, 152).

Duivesteijn, Wouter and Arno J. Knobbe (2011). “Exploiting False Discoveries - Statistical Validation of Patterns and Quality Measures in Subgroup Discovery”. In: *ICDM*. IEEE Computer Society, pages 151–160 (cited on pages 52, 101, 104, 108).

Duivesteijn, Wouter, Arno J. Knobbe, Ad Feelders, and Matthijs van Leeuwen (2010). “Subgroup Discovery Meets Bayesian Networks – An Exceptional Model Mining Approach”. In: *ICDM*. IEEE Computer Society, pages 158–167 (cited on pages 18, 42, 43, 104).

Dunn, Olive Jean (1961). “Multiple comparisons among means”. In: *Journal of the American statistical association* 56.293, pages 52–64 (cited on page 163).

Duris, Frantisek, Juraj Gazdarica, Iveta Gazdaricova, Lucia Strieskova, Jaroslav Budis, Jan Turna, and Tomas Szemes (2018). “Mean and variance of ratios of proportions from categories of a multinomial distribution”. In: *Journal of Statistical Distributions and Applications* 5 (cited on pages 109, 110).

Dzyuba, Vladimir (2017). “Mine, Interact, Learn, Repeat: Interactive Pattern-based Data Exploration ; Zoek, Interacteer, Leer, Herhaal: interactieve data-exploratie met patronen”. PhD thesis. Katholieke Universiteit Leuven, Belgium (cited on pages 9, 145).

Dzyuba, Vladimir, Matthijs van Leeuwen, and Luc De Raedt (2017). “Flexible constrained sampling with guarantees for pattern mining”. In: *Data Mining and Knowledge Discovery* 31.5, pages 1266–1293 (cited on pages 34, 37, 71).

Dzyuba, Vladimir, Matthijs van Leeuwen, Siegfried Nijssen, and Luc De Raedt (2014). “Interactive Learning of Pattern Rankings”. In: *International Journal on Artificial Intelligence Tools* 23.6 (cited on page 145).

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press (cited on page 116).

Eiter, Thomas and Heikki Mannila (1994). *Computing discrete Fréchet distance*. Technical report. Citeseer (cited on page 153).

Ennals, Rob, Beth Trushkowsky, and John Mark Agosta (2010). “Highlighting disputed claims on the web”. In: *Proceedings of the 19th international conference on World wide web*. ACM, pages 341–350 (cited on page 130).

Eppstein, David and Daniel S Hirschberg (1997). “Choosing subsets with maximum weighted average”. In: *J. Algorithms* 24.1, pages 177–193 (cited on pages 112–114).

Erevelles, Sunil, Nobuyuki Fukawa, and Linda Swayne (2016). “Big Data consumer analytics and the transformation of marketing”. In: *Journal of Business Research* 69.2, pages 897–904 (cited on page 7).

Etter, Vincent, Julien Herzen, Matthias Grossglauser, and Patrick Thiran (2014). “Mining democracy”. In: *COSN*. ACM, pages 1–12 (cited on pages 3, 82).

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). “From data mining to knowledge discovery in databases”. In: *AI magazine* 17.3, pages 37–37 (cited on pages 2, 13).

Fleiss, J. L. (1971). “Measuring nominal scale agreement among many raters”. In: *Psychological Bulletin* 76.5, pages 378–382 (cited on page 105).

Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift (2012). “The promise of computational journalism”. In: *Journalism Practice* 6.2, pages 157–171 (cited on page 2).

Fortunato, Santo (2010). “Community detection in graphs”. In: *Physics reports* 486.3-5, pages 75–174 (cited on page 8).

Freeman, Linton C (1977). “A set of measures of centrality based on betweenness”. In: *Sociometry*, pages 35–41 (cited on page 8).

Fukunage, K and Patrenahalli M. Narendra (1975). “A branch and bound algorithm for computing k-nearest neighbors”. In: *IEEE transactions on computers* 7, pages 750–753 (cited on pages 151, 152).

Fürnkranz, Johannes and Peter A Flach (2005). “Roc ‘n’rule learning—towards a better understanding of covering algorithms”. In: *Machine Learning* 58.1, pages 39–77 (cited on pages 32, 33, 45, 49).

Fürnkranz, Johannes, Dragan Gamberger, and Nada Lavrač (2012). *Foundations of rule learning*. Springer Science & Business Media (cited on pages 86, 157).

Galbrun, Esther and Pauli Miettinen (2017). *Redescription Mining*. Springer Briefs in Computer Science. Springer (cited on page 9).

— (2018). “Mining Redescriptions with Siren”. In: *TKDD* 12.1, 6:1–6:30 (cited on page 9).

Ganter, Bernhard and Sergei O. Kuznetsov (2001). “Pattern Structures and Their Projections”. In: *ICCS*. Volume 2120. Lecture Notes in Computer Science. Springer, pages 129–142 (cited on pages 25, 27, 44, 45, 59, 117).

Ganter, Bernhard and Rudolf Wille (1999). *Formal concept analysis - mathematical foundations*. Springer. ISBN: 978-3-540-62771-5 (cited on pages 25, 27, 45, 47, 79).

Ganter, Bernhard, Sergei Obiedkov, Sebastian Rudolph, and Gerd Stumme (2016). *Conceptual exploration*. Springer (cited on pages 27, 30, 46).

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis (2018). “Quantifying Controversy on Social Media”. In: *ACM Trans. Social Computing* 1.1, 3:1–3:27 (cited on page 8).

Garriga, Gemma C, Petra Kralj, and Nada Lavrač (2008). “Closed sets for labeled data”. In: *Journal of Machine Learning Research* 9.Apr, pages 559–580 (cited on pages 36, 159).

Gatt, Albert and Ehud Reiter (2009). “SimpleNLG: A Realisation Engine for Practical Applications”. In: *The 12th European Workshop on Natural Language Generation*. ENLG (cited on page 135).

Geisser, Seymour (1993). *Predictive Inference*. Volume 55. CRC Press (cited on page 108).

Geng, Liqiang and Howard J. Hamilton (2006). “Interestingness measures for data mining: A survey”. In: *ACM Comput. Surv.* 38.3, page 9 (cited on pages 31, 33, 45).

Giacometti, Arnaud and Arnaud Soulet (2016). “Frequent pattern outlier detection without exhaustive mining”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pages 196–207 (cited on page 71).

— (2018). “Dense Neighborhood Pattern Sampling in Numerical Data”. In: *SDM*. SIAM, pages 756–764 (cited on pages 34, 37).

Goodman, L. A. (1970). “The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications”. In: *Journal of the American Statistical Association* 65, pages 226–256 (cited on page 104).

Grosskreutz, Henrik (2008). “Cascaded subgroups discovery with an application to regression”. In: *Proc. ECML/PKDD*. Volume 5211. Citeseer (cited on pages 33, 35).

Grosskreutz, Henrik, Mario Boley, and Maike Krause-Traudes (2010). “Subgroup discovery for election analysis: a case study in descriptive data mining”. In: *International Conference on Discovery Science*. Springer, pages 57–71 (cited on pages 3, 8, 10, 82).

Grosskreutz, Henrik, Bastian Lang, and Daniel Trabold (2013). “A Relevance Criterion for Sequential Patterns”. In: *ECML/PKDD (1)*. Volume 8188. Lecture Notes in Computer Science. Springer, pages 369–384 (cited on pages 18, 21).

Grosskreutz, Henrik and Stefan Rüping (2009). “On subgroup discovery in numerical domains”. In: *Data Min. Knowl. Discov.* 19.2, pages 210–226 (cited on pages 18, 21, 34).

Grosskreutz, Henrik, Stefan Rüping, and Stefan Wrobel (2008). “Tight Optimistic Estimates for Fast Subgroup Discovery”. In: *ECML/PKDD (1)*. Volume 5211. Lecture Notes in Computer Science. Springer, pages 440–456 (cited on pages 35, 48, 49, 112).

Hämäläinen, Wilhelmiina (2010a). “Efficient Discovery of the Top-K Optimal Dependency Rules with Fisher’s Exact Test of Significance”. In: *ICDM*. IEEE Computer Society, pages 196–205 (cited on page 164).

— (2010b). “StatApriori: an efficient algorithm for searching statistically significant association rules”. In: *Knowl. Inf. Syst.* 23.3, pages 373–399 (cited on pages 34, 52, 103, 164).

— (2012). “Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures”. In: *Knowl. Inf. Syst.* 32.2, pages 383–414 (cited on page 34).

Hämäläinen, Wilhelmiina and Geoffrey I. Webb (2019). “A tutorial on statistically sound pattern discovery”. In: *Data Min. Knowl. Discov.* 33.2, pages 325–377 (cited on pages 34, 52, 103, 128, 163, 164).

Hamilton, James T and Fred Turner (2009). “Accountability through algorithm: Developing the field of computational journalism”. In: *Report from the Center for Advanced Study in the Behavioral Sciences, Summer Workshop*, pages 27–41 (cited on pages 2, 130).

Hammal, Mohamed Ali, Hélène Mathian, Luc Merchez, Marc Plantevit, and Céline Robardet (2019). “Rank correlated subgroup discovery”. In: *Journal of Intelligent Information Systems*, pages 1–24 (cited on page 41).

Han, Jiawei, Jian Pei, and Yiwen Yin (2000). “Mining frequent patterns without candidate generation”. In: *ACM sigmod record*. Volume 29. 2. ACM, pages 1–12 (cited on pages 35, 41).

Harper, F. Maxwell and Joseph A. Konstan (2016). “The MovieLens Datasets: History and Context”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5.4, 19:1–19:19 (cited on pages 78, 119).

Hassan, Naeemul, Chengkai Li, and Mark Tremayne (2015). “Detecting check-worthy factual claims in presidential debates”. In: ACM, pages 1835–1838 (cited on page 130).

Hassan, Naeemul, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. (2017a). “ClaimBuster: The First-ever End-to-end Fact-checking System”. In: *Proceedings of the VLDB Endowment* 10.7 (cited on page 130).

Hassan, Naeemul, Fatma Arslan, Chengkai Li, and Mark Tremayne (2017b). “Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 1803–1812 (cited on page 130).

Hayes, Andrew F and Klaus Krippendorff (2007). “Answering the call for a standard reliability measure for coding data”. In: *Communication methods and measures* 1.1, pages 77–89 (cited on pages 104, 105, 116).

Hébert, Céline and Bruno Crémilleux (2007). “A Unified View of Objective Interestingness Measures”. In: *MLDM*. Volume 4571. Lecture Notes in Computer Science. Springer, pages 533–547 (cited on page 31).

Heer, Jeffrey and Danah Boyd (2005). “Vizster: Visualizing Online Social Networks”. In: *INFOVIS*. IEEE Computer Society, pages 32–39 (cited on page 9).

Herman, Ivan, Guy Melançon, and M Scott Marshall (2000). “Graph visualization and navigation in information visualization: A survey”. In: *IEEE Transactions on visualization and computer graphics* 6.1, pages 24–43 (cited on page 9).

Herrera, Francisco, Cristóbal José Carmona, Pedro González, and María José Del Jesus (2011). “An overview on subgroup discovery: foundations and applications”. In: *Knowledge and information systems* 29.3, pages 495–525 (cited on pages 3, 10, 18).

Hix, Simon (2001). “Legislative behaviour and party competition in the European Parliament: An application of nominate to the EU”. In: *JCMS: Journal of Common Market Studies* 39.4, pages 663–688 (cited on page 151).

Hix, Simon, Abdul Noury, and Gérard Roland (2005). “Power to the parties: cohesion and competition in the European Parliament, 1979–2001”. In: *British Journal of Political Science* 35.2, pages 209–234 (cited on page 97).

Hix, Simon, Abdul Noury, and Gerard Roland (2006). “Dimensions of politics in the European Parliament”. In: *American Journal of Political Science* 50 (cited on pages 145, 151).

Hix, Simon, Abdul G Noury, and Gérard Roland (2007). *Democratic politics in the European Parliament*. Cambridge University Press (cited on page 2).

Holm, Sture (1979). “A simple sequentially rejective multiple test procedure”. In: *Scandinavian journal of statistics*, pages 65–70 (cited on page 163).

Hu, Qiong and Tomasz Imielinski (2017). "Alpine: Progressive itemset mining with definite guarantees". In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, pages 63–71 (cited on page 37).

Huang, Enhui, Liping Peng, Luciano Di Palma, Ahmed Abdelkafi, Anna Liu, and Yanlei Diao (2018). "Optimization for Active Learning-based Interactive Database Exploration". In: *PVLDB* 12.1, pages 71–84 (cited on page 9).

Ireton, Cherilyn and Julie Posetti (2018). *Journalism, fake news & disinformation: handbook for journalism education and training*. UNESCO Publishing (cited on page 1).

Jakulin, Aleks, Wray Buntine, Timothy M La Pira, and Holly Brasher (2009). "Analyzing the us senate in 2003: Similarities, clusters, and blocs". In: *Political Analysis* 17.3, pages 291–310 (cited on pages 8, 9, 82, 145).

Janssen, Frederik and Johannes Fürnkranz (2006). "On Trading Off Consistency and Coverage in Inductive Rule Learning". In: *LWA*. Volume 1/2006. Hildesheimer Informatik-Berichte. University of Hildesheim, Institute of Computer Science, pages 306–313 (cited on page 31).

Jesús, María José del, Pedro González, Francisco Herrera, and Mikel Mesonero (2007). "Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A Case Study in Marketing". In: *IEEE Trans. Fuzzy Systems* 15.4, pages 578–592 (cited on page 34).

Jeudy, Baptiste and Jean-François Boulicaut (2002). "Optimization of association rule mining queries". In: *Intelligent Data Analysis* 6.4, pages 341–357 (cited on page 35).

Johnson, David S., Christos H. Papadimitriou, and Mihalis Yannakakis (1988). "On Generating All Maximal Independent Sets". In: *Inf. Process. Lett.* 27.3, pages 119–123 (cited on page 46).

Johnson, Don and Sinan Sinanovic (2001). "Symmetrizing the kullback-leibler distance". In: *IEEE Transactions on Information Theory* (cited on page 62).

Karypis, George and Vipin Kumar (1995). *METIS – Unstructured Graph Partitioning and Sparse Matrix Ordering System*. Technical report (cited on page 8).

Kavšek, Branko and Nada Lavrač (2006). "APRIORI-SD: Adapting association rule learning to subgroup discovery". In: *Applied Artificial Intelligence* 20.7, pages 543–583 (cited on pages 34, 35).

Kaytoue, Mehdi, Sergei O. Kuznetsov, and Amedeo Napoli (2011). "Revisiting Numerical Pattern Mining with Formal Concept Analysis". In: *IJCAI*. IJCAI/AAAI, pages 1342–1347 (cited on pages 21, 29).

Kaytoue, Mehdi, Sergei O Kuznetsov, Amedeo Napoli, and Sébastien Duplessis (2011). "Mining gene expression data with pattern structures in formal concept analysis". In: *Information Sciences* 181.10, pages 1989–2001 (cited on pages 21, 29, 79).

Kaytoue, Mehdi, Marc Plantevit, Albrecht Zimmermann, Anes Bendimerad, and Céline Robardet (2017). "Exceptional contextual subgraph mining". In: *Machine Learning*, pages 1–41 (cited on pages 11, 18, 21, 43, 51, 86, 158, 159).

Kempen, G. M. P. Kempenvan and L. J. van Vliet (2000). "Mean and variance of ratio estimators used in fluorescence ratio imaging". In: *Cytometry: The Journal of the International Society for Analytical Cytology* 39.4, pages 300–305 (cited on page 110).

Kendall, M. G. (1938). "A New Measure of Rank Correlation". In: *Biometrika* 30.1, pages 81–93 (cited on page 104).

Kendall, Maurice George (1948). “Rank correlation methods.” In: (cited on page 41).

Kendall, M.G., A. Stuart, and J.K. Ord (1994). “Kendall’s advanced theory of statistics. v. 1: Distribution theory”. In: (cited on pages 109, 110).

Kirchgessner, Martin, Vincent Leroy, Sihem Amer-Yahia, and Shashwat Mishra (2016). “Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns”. In: *DSAA*. IEEE, pages 547–556 (cited on page 31).

Kleiner, Ariel, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan (2012). “The Big Data Bootstrap”. In: *ICML*. icml.cc / Omnipress (cited on page 117).

Kloesgen, W (2000). “Subgroup Mining”. In: *Computational Intelligence in Data Mining*. Springer, pages 39–49 (cited on pages 7, 21, 44).

Klösgen, Willi (1996). “Explora: A Multipattern and Multistrategy Discovery Assistant”. In: *Advances in Knowledge Discovery and Data Mining*, pages 249–271 (cited on pages 2, 10, 13, 18, 33, 34, 44).

Klösgen, Willi (2002). “Data mining tasks and methods: subgroup discovery: deviation analysis”. In: *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc., pages 354–361 (cited on page 33).

Klösgen, Willi and Michael May (2002). “Spatial Subgroup Mining Integrated in an Object-Relational Spatial Database”. In: *PKDD*. Volume 2431. Lecture Notes in Computer Science. Springer, pages 275–286 (cited on page 34).

Kohavi, Ron (1995). “The Power of Decision Tables”. In: *ECML*. Volume 912. Lecture Notes in Computer Science. Springer, pages 174–189 (cited on page 41).

— (2001). “Mining e-commerce data: the good, the bad, and the ugly”. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 8–13 (cited on page 7).

Komiyama, Junpei, Masakazu Ishihata, Hiroki Arimura, Takashi Nishibayashi, and Shin-ichi Minato (2017). “Statistical emerging pattern mining with multiple testing correction”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 897–906 (cited on page 164).

Kovach, Bill and Tom Rosenstiel (2014). “The Elements of Journalism”. In: (cited on page 1).

Krak, Thomas E. and Ad Feelders (2015). “Exceptional Model Mining with Tree-Constrained Gradient Ascent”. In: *SDM*. SIAM, pages 487–495 (cited on page 40).

Kralj-Novak, Petra, Nada Lavrac, and Geoffrey I. Webb (2009). “Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining”. In: *Journal of Machine Learning Research* 10, pages 377–403 (cited on pages 31, 38, 39, 45, 51).

Kralj Novak, Petra, Nada Lavrač, and Geoffrey I. Webb (2009). “Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining”. In: *J. Mach. Learn. Res.* 10 (cited on pages 18, 21, 32).

Krippendorff, Klaus (1980). *Context Analysis: An Introduction to Its Methodology*. Sage (cited on pages 104, 105).

— (2004). “Content Analysis, An introduction to its methodology”. In: (cited on pages 100, 106, 116).

Krippendorff, Klaus, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher (2016). “On the Reliability of Unitizing Textual Continua: Further Developments”. In: *Qual. Quant.* 50, pages 2347–2364 (cited on page 105).

Kucharski, Adam (2016). “Post-truth: Study epidemiology of fake news”. In: *Nature* 540.7634, page 525 (cited on page 1).

Kuznetsov, Sergei O. (1993). “A Fast Algorithm for Computing All Intersections of Objects in a Finite Semi-lattice”. In: *Nauchno-Tekhnicheskaya Informatsiya* ser. 2.1, pages 17–20 (cited on page 46).

— (1999). “Learning of Simple Conceptual Graphs from Positive and Negative Examples”. In: *PKDD*. Volume 1704. Lecture Notes in Computer Science. Springer, pages 384–391 (cited on pages 46, 117).

Kuznetsov, Sergei O and Sergei A Obiedkov (2002). “Comparing performance of algorithms for generating concept lattices”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 14.2-3, pages 189–216 (cited on pages 27, 30, 40, 46).

Lacombe, Charles de, Antoine Morel, Adnene Belfodil, François Portet, Cyril Labbé, Sylvie Cazalens, Marc Plantevit, and Philippe Lamarre (2019). “Analyse de comportements relatifs exceptionnels expliquée par des textes : les votes du parlement européen”. In: *Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019*, pages 437–440 (cited on pages 15, 84, 132).

Land, AH and AG Doig (1960). “An Automatic Method of Solving Discrete Programming Problems”. In: *Econometrica* 28.3, pages 497–520 (cited on page 35).

Lavrač, Nada, Peter Flach, and Blaz Zupan (1999). “Rule evaluation measures: A unifying view”. In: *International Conference on Inductive Logic Programming*. Springer, pages 174–185 (cited on page 10).

Lavrac, Nada, Peter A. Flach, and Blaz Zupan (1999). “Rule Evaluation Measures: A Unifying View”. In: *Inductive Logic Programming, 9th International Workshop, ILP-99, Bled, Slovenia, June 24-27, 1999, Proceedings*. Volume 1634. Lecture Notes in Computer Science. Springer, pages 174–185 (cited on pages 31, 32, 43, 45, 153).

Lavrač, Nada, Branko Kavšek, Peter Flach, and Ljupčo Todorovski (2004). “Subgroup discovery with CN2-SD”. In: *Journal of Machine Learning Research* 5.Feb, pages 153–188 (cited on page 18).

Lavrac, Nada, Branko Kavsek, Peter A. Flach, and Ljupco Todorovski (2004). “Subgroup Discovery with CN2-SD”. In: *Journal of Machine Learning Research* 5, pages 153–188 (cited on pages 34, 36, 45).

Le Cam, Lucien and Grace Lo Yang (2012). *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media (cited on page 42).

Leeuwen, Matthijs van (2014). “Interactive Data Exploration Using Pattern Mining”. In: *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Volume 8401. Lecture Notes in Computer Science. Springer, pages 169–182 (cited on page 145).

Leeuwen, Matthijs van and Arno J. Knobbe (2011). “Non-redundant Subgroup Discovery in Large and Complex Data”. In: *ECML/PKDD (3)*. Volume 6913. Lecture Notes in Computer Science. Springer, pages 459–474 (cited on pages 34, 36, 45).

— (2012). “Diverse subgroup set discovery”. In: *Data Min. Knowl. Discov.* 25.2, pages 208–242 (cited on pages 18, 22, 34, 36, 44, 45, 117).

Lehmann, Jens, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo (2012). “Defacto-deep fact validation”. In: *International semantic web conference*. Springer, pages 312–327 (cited on page 130).

Leman, Dennis, Ad Feelders, and Arno Knobbe (2008). “Exceptional model mining”. In: *ECMLPKDD*. Springer, pages 1–16 (cited on pages 13, 18, 38, 41, 42, 44, 153).

Lemmerich, Florian, Martin Atzmueller, and Frank Puppe (2016). “Fast exhaustive subgroup discovery with numerical target concepts”. In: *Data Min. Knowl. Discov.* 30.3, pages 711–762 (cited on pages 18, 33–35, 45, 49).

Lemmerich, Florian and Martin Becker (2018). “pysubgroup: Easy-to-Use Subgroup Discovery in Python”. In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part III*. Edited by Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley. Volume 11053. Lecture Notes in Computer Science. Springer, pages 658–662 (cited on pages 35, 86, 156, 159).

Lemmerich, Florian, Martin Becker, and Martin Atzmueller (2012). “Generic Pattern Trees for Exhaustive Exceptional Model Mining”. In: *ECML/PKDD (2)*. Volume 7524. Lecture Notes in Computer Science. Springer, pages 277–292 (cited on pages 40, 117).

Lemmerich, Florian, Marianus Ifl, and Frank Puppe (2011). “Identifying influence factors on students success by subgroup discovery”. In: *Educational Data Mining* (cited on page 8).

Lemmerich, Florian, Mathias Rohlfs, and Martin Atzmueller (2010). “Fast Discovery of Relevant Subgroup Patterns.” In: *FLAIRS Conference* (cited on pages 34, 159).

Lemmerich, Florian, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier (2016). “Mining Subgroups with Exceptional Transition Behavior”. In: *KDD*. ACM, pages 965–974 (cited on pages 4, 7, 11, 43, 52, 104, 108).

Lenca, Philippe, Patrick Meyer, Benoît Vaillant, and Stéphane Lallich (2008). “On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid”. In: *European Journal of Operational Research* 184.2, pages 610–626 (cited on page 31).

Li, Geng and Mohammed J Zaki (2016). “Sampling frequent and minimal boolean patterns: theory and application in classification”. In: *Data Mining and Knowledge Discovery* 30.1, pages 181–225 (cited on pages 34, 71).

Li, Jiuyong, Ada Wai-Chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Graham J. Williams, Ross Sparks, and Chris Kelman (2005). “Mining risk patterns in medical data”. In: *KDD*. ACM, pages 770–775 (cited on page 8).

Lijffijt, Jefrey, Bo Kang, Wouter Duivesteijn, Kai Puolamäki, Emilia Oikarinen, and Tijl De Bie (2018). “Subjectively Interesting Subgroup Discovery on Real-Valued Targets”. In: *ICDE*. IEEE Computer Society, pages 1352–1355 (cited on pages 23, 34, 108).

Little, John DC, Katta G Murty, Dura W Sweeney, and Caroline Karel (1963). “An algorithm for the traveling salesman problem”. In: *Operations research* 11.6, pages 972–989 (cited on page 35).

Liu, Bing, Wynne Hsu, and Yiming Ma (1998). “Integrating Classification and Association Rule Mining”. In: *KDD*, pages 80–86 (cited on page 18).

Llinares-López, Felipe, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten Borgwardt (2015). “Fast and memory-efficient significant pattern mining via permutation testing”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pages 725–734 (cited on page 164).

Lowerre, Bruce T (1976). “The HARPY speech recognition system”. In: *PhD thesis, Carnegie Mellon University* (cited on page 36).

Lucas, Tarcísio, Renato Vimieiro, and Teresa Bernarda Ludermir (2018). “SSDP+: A Diverse and More Informative Subgroup Discovery Approach for High Dimensional Data”. In: *CEC*. IEEE, pages 1–8 (cited on pages 34, 36).

Lumpe, Lars and Stefan E. Schmidt (2015). “Pattern Structures and Their Morphisms”. In: *CLA*. Volume 1466. CEUR Workshop Proceedings. CEUR-WS.org, pages 171–179 (cited on page 25).

Luna, José María, José Raúl Romero, Cristóbal Romero, and Sebastián Ventura (2013). “Discovering Subgroups by Means of Genetic Programming”. In: *EuroGP*. Volume 7831. Lecture Notes in Computer Science. Springer, pages 121–132 (cited on page 34).

Mampaey, Michael, Siegfried Nijssen, Ad Feelders, and Arno J. Knobbe (2012). “Efficient Algorithms for Finding Richer Subgroup Descriptions in Numeric and Nominal Data”. In: *ICDM*. IEEE Computer Society, pages 499–508 (cited on pages 21, 34, 36).

Manolescu, Ioana (2017). “ContentCheck: Content Management Techniques and Tools for Fact-checking”. In: (cited on page 2).

Mathonat, R., D. Nurbakova, J.F. Boulicaut, and M. Kaytoue (2019). “SeqScout: using a bandit algorithm to discover interesting subgroups in Labeled Sequences”. In: *DSAA*. IEEE (cited on page 21).

Meeng, M. and Arno J. Knobbe (2011). “Flexible Enrichment with Cortana – Software Demo”. In: *Proceedings Benelearn*, 117—119 (cited on pages 35, 156).

Meilă, Marina (2007). “Comparing clusterings—an information based distance”. In: *Journal of multivariate analysis* 98.5, pages 873–895 (cited on page 152).

Minato, Shin-ichi, Takeaki Uno, Koji Tsuda, Aika Terada, and Jun Sese (2014). “A Fast Method of Statistical Assessment for Combinatorial Hypotheses Based on Frequent Itemset Enumeration”. In: *ECML/PKDD (2)*. Volume 8725. Lecture Notes in Computer Science. Springer, pages 422–436 (cited on pages 52, 103).

Moens, Sandy and Mario Boley (2014). “Instant exceptional model mining using weighted controlled pattern sampling”. In: *International Symposium on Intelligent Data Analysis*. Springer, pages 203–214 (cited on pages 40, 72).

Moens, Sandy and Bart Goethals (2013). “Randomly sampling maximal itemsets”. In: *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*. ACM, pages 79–86 (cited on pages 34, 71).

Moranges, Maëlle, Marc Plantevit, Arnaud P. Fournel, Moustafa Bensafi, and Céline Robardet (2018). “Exceptional Attributed Subgraph Mining to Understand the Olfactory Percept”. In: *Discovery Science - 21st International Conference, DS 2018, Limassol, Cyprus, October 29-31, 2018, Proceedings*. Volume 11198. Lecture Notes in Computer Science. Springer, pages 276–291 (cited on page 43).

Morishita, Shinichi and Jun Sese (2000). “Traversing Itemset Lattice with Statistical Metric Pruning”. In: *PODS*. ACM, pages 226–236 (cited on page 49).

Mukhopadhyay, Nitis (2000). *Probability and statistical inference*. CRC Press (cited on page 115).

Mullins, Irene M, Mir S Siadaty, Jason Lyman, Ken Scully, Carleton T Garrett, W Greg Miller, Rudy Muller, Barry Robson, Chid Apte, Sholom Weiss, et al. (2006). "Data mining and clinical data repositories: Insights from a 667,000 patient data set". In: *Computers in biology and medicine* 36.12, pages 1351–1377 (cited on page 8).

Murtagh, Fionn and Pedro Contreras (2012). "Algorithms for hierarchical clustering: an overview". In: *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2.1, pages 86–97 (cited on pages 8, 151, 152).

Narendra, Patrenahalli M. and Keinosuke Fukunaga (1977). "A branch and bound algorithm for feature subset selection". In: *IEEE Transactions on computers* 9, pages 917–922 (cited on page 35).

Neter, John, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman (1996). *Applied linear statistical models*. Irwin Chicago (cited on page 41).

Newman, Mark EJ (2004). "Detecting community structure in networks". In: *The European Physical Journal B* 38.2, pages 321–330 (cited on page 8).

Nguyen, An T., Aditya Kharosekar, Matthew Lease, and Byron C. Wallace (2018). "An Interpretable Joint Graphical Model for Fact-Checking From Crowds". In: *AAAI*. AAAI Press, pages 1511–1518 (cited on page 130).

Nijssen, Siegfried and Albrecht Zimmermann (2014). "Constraint-based pattern mining". In: *Frequent pattern mining*. Springer, pages 147–163 (cited on page 35).

Norris, James R (1998). *Markov chains*. 2. Cambridge university press (cited on pages 11, 43).

Omidvar-Tehrani, Behrooz and Sihem Amer-Yahia (2017). "Online Lattice-Based Abstraction of User Groups". In: *DEXA (1)*. Volume 10438. Lecture Notes in Computer Science. Springer, pages 95–110 (cited on page 10).

— (2018). "User Group Analytics: Discovery, Exploration and Visualization". In: *CIKM*. ACM, pages 2307–2308 (cited on page 4).

Omidvar-Tehrani, Behrooz and Sihem Amer-Yahia (2019). "User Group Analytics Survey and Research Opportunities". In: *IEEE Transactions on Knowledge and Data Engineering* (cited on pages 4, 7).

Omidvar-Tehrani, Behrooz, Sihem Amer-Yahia, and Ria Mae Borromeo (2019). "User group analytics: hypothesis generation and exploratory analysis of user data". In: *VLDB J.* 28.2, pages 243–266 (cited on pages 4, 7, 9, 10).

Omidvar-Tehrani, Behrooz, Sihem Amer-Yahia, and Laks V. S. Lakshmanan (2018). "Cohort Representation and Exploration". In: *DSAA*. IEEE, pages 169–178 (cited on page 9).

Omidvar-Tehrani, Behrooz, Sihem Amer-Yahia, and Alexandre Termier (2015). "Interactive User Group Analysis". In: *CIKM*. ACM, pages 403–412 (cited on pages 9, 10).

Omidvar-Tehrani, Behrooz, Sihem Amer-Yahia, Pierre-François Dutot, and Denis Trystram (2016). "Multi-Objective Group Discovery on the Social Web". In: *ECML/PKDD (1)*. Volume 9851. Lecture Notes in Computer Science. Springer, pages 296–312 (cited on pages 8, 10).

Orueta, Juan F, Roberto Nuño-Solinis, Maider Mateos, Itziar Vergara, Gonzalo Grandes, and Santiago Esnaola (2012). "Monitoring the prevalence of chronic conditions: which data should we use?" In: *BMC health services research* 12.1, page 365 (cited on page 84).

Pajala, Antti, Aleks Jakulin, and Wray Buntine (2004). "Parliamentary group and individual voting behaviour in the Finnish parliament in year 2003: a group cohesion and voting similarity analysis". In: (cited on page 82).

Palen, Leysia and Amanda L Hughes (2018). "Social media in disaster communication". In: *Handbook of disaster research*. Springer, pages 497–518 (cited on page 1).

Pasquier, Nicolas, Yves Bastide, Rafik Taouil, and Lotfi Lakhal (1999). "Discovering Frequent Closed Itemsets for Association Rules". In: *ICDT*. Volume 1540. Lecture Notes in Computer Science. Springer, pages 398–416 (cited on pages 27, 59).

Pearson, Karl, Alice Lee, Ernest Warren, Agnes Fry, and Cicely D. Fawcett (1901). "Mathematical Contributions to the Theory of Evolution: IX. On the Principle of Homotyposis and its Relation to Heredity, to Variability of the Individual, and to that of Race. Part I: Homotyposis in the Vegetable Kingdom". In: *Philosophical Transactions of the Royal Society* 197.Series A, pages 285–379 (cited on pages 104, 105).

Pellegrina, Leonardo and Fabio Vandin (2018). "Efficient Mining of the Most Significant Patterns with Permutation Testing". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, pages 2070–2079 (cited on page 164).

Piatetsky-Shapiro, Gregory (1991). "Discovery, Analysis, and Presentation of Strong Rules". In: *Knowledge Discovery in Databases*. AAAI/MIT Press, pages 229–248 (cited on page 32).

Pieters, Barbara FI, Arno Knobbe, and Sašo Dzeroski (2010). "Subgroup discovery in ranked data, with an application to gene set enrichment". In: *Proceedings preference learning workshop (PL 2010) at ECML PKDD*. Volume 10, pages 1–18 (cited on pages 33, 34).

Pool, Simon, Francesco Bonchi, and Matthijs van Leeuwen (2014). "Description-Driven Community Detection". In: *ACM TIST* 5.2, 28:1–28:28 (cited on page 8).

Poole, Keith T and Howard Rosenthal (1985). "A spatial model for legislative roll call analysis". In: *American Journal of Political Science*, pages 357–384 (cited on pages 9, 145, 151).

— (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand (cited on pages 2, 151).

Portet, François, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes (2009). "Automatic generation of textual summaries from neonatal intensive care data". In: *Artificial Intelligence* 173.7-8, pages 789–816 (cited on page 135).

Rajski, C (1961). "A metric space of discrete probability distributions". In: *Information and Control* 4.4, pages 371–377 (cited on page 9).

Riondato, Matteo and Fabio Vandin (2018). "MiSoSouP: Mining Interesting Subgroups with Sampling and Pseudodimension". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. Edited by Yike Guo and Faisal Farooq. ACM, pages 2130–2139 (cited on pages 37, 45).

Roddy, Edward and Michael Doherty (2010). “Gout. Epidemiology of gout”. In: *Arthritis research & therapy* 12.6, page 223 (cited on page 86).

Roman, Steven (2008). *Lattices and ordered sets*. Springer Science & Business Media (cited on pages 25, 27, 58, 73).

Romero, Cristobal and Sebastian Ventura (2013). “Data mining in education”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3.1, pages 12–27 (cited on page 7).

Romero, Cristobal, Sebastian Ventura, Mykola Pechenizkiy, and Ryan SJd Baker (2010). *Handbook of educational data mining*. CRC press (cited on page 7).

Rossetti, Giulio and Rémy Cazabet (2018). “Community Discovery in Dynamic Networks: A Survey”. In: *ACM Comput. Surv.* 51.2, 35:1–35:37 (cited on page 8).

Sá, Cláudio Rebelo de, Wouter Duivesteijn, Carlos Soares, and Arno J. Knobbe (2016). “Exceptional Preferences Mining”. In: *DS*, pages 3–18 (cited on pages 11, 43).

Sá, Cláudio Rebelo de, Wouter Duivesteijn, Paulo Azevedo, Alípio Mário Jorge, Carlos Soares, and Arno Knobbe (2018). “Discovering a taste for the unusual: exceptional models for preference mining”. In: *Machine Learning* 107.11, pages 1775–1807 (cited on pages 11, 43).

Schwartz, Martin A (2008). “The importance of stupidity in scientific research”. In: *Journal of Cell Science* 121.11, pages 1771–1771 (cited on page iii).

Scott, W. A. (1955). “Reliability of Content Analysis: The Case of Nominal Scale Coding”. In: *Public Opinion Quarterly* 19, pages 321–325 (cited on page 105).

Shapiro, Linda G and Robert M Haralick (1985). “A metric for comparing relational descriptions”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1, pages 90–94 (cited on page 43).

Shapiro, Samuel Sanford and Martin B Wilk (1965). “An analysis of variance test for normality (complete samples)”. In: *Biometrika* 52.3/4, pages 591–611 (cited on page 124).

Siebes, Arno (1995). “Data Surveying: Foundations of an Inductive Query Language”. In: *KDD*. AAAI Press, pages 269–274 (cited on pages 18, 19, 21, 51).

Siebes, Arno, Jilles Vreeken, and Matthijs van Leeuwen (2006). “Item Sets that Compress”. In: *SDM*. SIAM, pages 395–406 (cited on page 23).

Silva, Rodrigo Nunes Moni da, Andre Spritzer, and Carla Dal Sasso Freitas (2018). “Visualization of Roll Call Data for Supporting Analyses of Political Profiles”. In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, pages 150–157 (cited on page 9).

Simini, Filippo, Marta C González, Amos Maritan, and Albert-László Barabási (2012). “A universal model for mobility and migration patterns”. In: *Nature* 484.7392, page 96 (cited on page 11).

Singer, Philipp, Denis Helic, Andreas Hotho, and Markus Strohmaier (2015). “HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web”. In: *WWW*. ACM, pages 1003–1013 (cited on page 11).

Smith, Jonathan A and Mike Osborn (2004). “Interpretative phenomenological analysis”. In: *Doing social psychology research*, pages 229–254 (cited on page 7).

Sozio, Mauro and Aristides Gionis (2010). “The community-search problem and how to plan a successful cocktail party”. In: *Proceedings of the 16th ACM SIGKDD international* *conference on Knowledge discovery and data mining*. ACM, pages 939–948 (cited on page 9).

Spearman, C. (1904). “The Proof and Measurement of Association Between Two Things”. In: *American Journal of Psychology* 15.1, pages 72–101 (cited on page 104).

Spiegelman, M., C. Terwilliger, and F. Fearing (1953). “The reliability of agreement in content analysis”. In: *Journal of Social Psychology* 37, pages 175–187 (cited on page 105).

Spyropoulou, Eirini, Tijl De Bie, and Mario Boley (2014). “Interesting pattern mining in multi-relational data”. In: *Data Mining and Knowledge Discovery* 28.3, pages 808–849. ISSN: 1573-756X (cited on pages 34–36, 45).

Steele, J Michael (2004). *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press (cited on page 115).

Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava (2004). “Selecting the right objective measure for association analysis”. In: *Inf. Syst.* 29.4, pages 293–313 (cited on pages 31–33, 45).

Tarone, RE (1990). “A modified Bonferroni method for discrete data”. In: *Biometrics*, pages 515–522 (cited on page 164).

Terada, Aika, David duVerle, and Koji Tsuda (2016). “Significant Pattern Mining with Confounding Variables”. In: *PAKDD (1)*. Volume 9651. Lecture Notes in Computer Science. Springer, pages 277–289 (cited on page 164).

Terada, Aika, Koji Tsuda, and Jun Sese (2013). “Fast Westfall-Young permutation procedure for combinatorial regulation discovery”. In: *2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, December 18-21, 2013*. Edited by Guo-Zheng Li, Sunghoon Kim, Michael Hughes, Geoffrey J. McLachlan, Hongye Sun, Xiaohua Hu, Habtom W. Ressom, Baoyan Liu, and Michael N. Liebman. IEEE Computer Society, pages 153–158 (cited on page 164).

Terada, Aika, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese (2013). “Statistical significance of combinatorial regulations”. In: *Proceedings of the National Academy of Sciences* 110.32, pages 12996–13001 (cited on pages 18, 164).

Todorovski, Ljupčo, Peter Flach, and Nada Lavrač (2000). “Predictive performance of weighted relative accuracy”. In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, pages 255–264 (cited on page 31).

Trajkovski, Igor, Nada Lavrač, and Jakub Tolar (2008). “SEGS: Search for enriched gene sets in microarray data”. In: *Journal of biomedical informatics* 41.4, pages 588–601 (cited on page 33).

Trohidis, Konstantinos, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas (2008). “Multi-Label Classification of Music into Emotions”. In: *ISMIR*, pages 325–330 (cited on page 43).

Tukey, John W (1977). “Exploratory data analysis”. In: (cited on pages 7, 18).

Van Leeuwen, Matthijs (2014). “Interactive data exploration using pattern mining”. In: *Interactive knowledge discovery and data mining in biomedical informatics*. Springer, pages 169–182 (cited on page 9).

Ventura, Sebastián and José María Luna (2018). *Supervised Descriptive Pattern Mining*. Springer (cited on page 41).

Vizzini, Jérémy, Cyril Labbé, and François Portet (Jan. 2017). “Génération automatique de billets journalistiques : singularité et normalité d’une sélection”. In: *Extraction et Gestion des Connaissances (EGC) 2017 Atelier Journalisme Computationnel,* Grenoble, France (cited on page 135).

Vlachos, Andreas and Sebastian Riedel (2014). “Fact Checking: Task definition and dataset construction”. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22 (cited on page 131).

Voeten, Erik (2009). “Enlargement and the’normal’European parliament”. In: *The legitimacy of the European union after enlargement*, pages 93–113 (cited on page 145).

Vreeken, Jilles, Matthijs van Leeuwen, and Arno Siebes (2011). “Krimp: mining itemsets that compress”. In: *Data Min. Knowl. Discov.* 23.1, pages 169–214 (cited on pages 23, 44).

Wang, Clifford and Lawrence M Crapo (1997). “The epidemiology of thyroid disease and implications for screening”. In: *Endocrinology and Metabolism Clinics* 26.1, pages 189–218 (cited on page 84).

Wasserman, Stanley and Katherine Faust (1994). *Social network analysis: Methods and applications*. Volume 8. Cambridge university press (cited on page 7).

Webb, Geoffrey I. (2001). “Discovering associations with numeric variables”. In: *KDD*. ACM, pages 383–388 (cited on page 48).

— (2006). “Discovering significant rules”. In: *KDD*. ACM, pages 434–443 (cited on page 164).

Webb, Geoffrey I (2007). “Discovering significant patterns”. In: *Machine learning* 68.1, pages 1–33 (cited on pages 34, 103, 104, 164).

Webb, Geoffrey I. (2008). “Layered critical values: a powerful direct-adjustment approach to discovering significant patterns”. In: *Machine Learning* 71.2-3, pages 307–323 (cited on page 165).

Westfall, Peter H, S Stanley Young, et al. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Volume 279. John Wiley & Sons (cited on page 164).

Whitley, Darrell (1994). “A genetic algorithm tutorial”. In: *Statistics and computing* 4.2, pages 65–85 (cited on page 36).

Wille, Rudolf (1982). “Restructuring lattice theory: an approach based on hierarchies of concept”. In: *Ordered sets* (cited on page 25).

Wrobel, Stefan (1997). “An Algorithm for Multi-relational Discovery of Subgroups”. In: *PKDD*. Volume 1263. Lecture Notes in Computer Science. Springer, pages 78–87 (cited on pages 2, 10, 13, 18, 22, 28, 31, 34, 44, 51, 100).

— (2001). “Inductive logic programming for knowledge discovery in databases”. In: *Relational data mining*. Springer, pages 74–101 (cited on pages 3, 18).

Wu, You (2015). “Computational Journalism: from Answering Questions to Questioning Answers and Raising Good Questions”. PhD thesis. Duke University (cited on pages 130, 131, 142).

Wu, You, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu (2014). “Toward computational fact-checking”. In: *Proceedings of the VLDB Endowment* 7.7, pages 589–600 (cited on pages 130, 131, 142).

Wu, You, Junyang Gao, Pankaj K Agarwal, and Jun Yang (2017). “Finding diverse, high-value representatives on a surface of answers”. In: *Proceedings of the VLDB Endowment* 10.7, pages 793–804 (cited on pages 130, 142).

Xu, Rui and Donald C. Wunsch II (2005). “Survey of clustering algorithms”. In: *IEEE Trans. Neural Networks* 16.3, pages 645–678 (cited on page 8).

Yan, Xifeng and Jiawei Han (2002). “gSpan: Graph-Based Substructure Pattern Mining”. In: *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, pages 721–724 (cited on page 21).

Yang, Jun, Pankaj K. Agarwal, Sudeepa Roy, Brett Walenz, You Wu, Cong Yu, and Chengkai Li (2018). “Query Perturbation Analysis: An Adventure of Database Researchers in Fact-Checking”. In: *IEEE Data Eng. Bull.* 41.3, pages 28–42 (cited on page 2).

Young, Mary Lynn and Alfred Hermida (2015). “From Mr. and Mrs. outlier to central tendencies: Computational journalism and crime reporting at the Los Angeles Times”. In: *Digital Journalism* 3.3, pages 381–397 (cited on page 2).

Zaki, Mohammed Javeed (2000). “Scalable algorithms for association mining”. In: *IEEE transactions on knowledge and data engineering* 12.3, pages 372–390 (cited on page 35).

Zilberstein, Shlomo (1996). “Using Anytime Algorithms in Intelligent Systems”. In: *AI Magazine* 17.3, pages 73–83 (cited on pages 37, 153).

Zimmermann, Albrecht and Luc De Raedt (2009). “Cluster-grouping: from subgroup discovery to clustering”. In: *Machine Learning* 77.1, pages 125–159 (cited on pages 34, 49).

Zipf, George Kingsley (1946). “The p 1 p 2/d hypothesis: The case of railway express”. In: *The Journal of Psychology* 22.1, pages 3–8 (cited on page 11).

Zitnik, Marinka, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M. Hoffman (2019). “Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities”. In: *Information Fusion* 50, pages 71–91 (cited on page 7).

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : BELFODIL DATE de SOUTENANCE : 21/10/2019
(avec précision du nom de jeune fille, le cas échéant)

Prénoms : Adnene

TITRE: Exceptional Model Mining for Behavioral Data Analysis

NATURE : Doctorat Numéro d'ordre : 2019LYSEI086

Ecole doctorale : InfoMaths (ED 512)

Spécialité : Informatique

RESUME :

Avec la prolifération rapide des plateformes de données qui récoltent des données relatives à plusieurs domaines tels que les données de gouvernements, d'éducation, d'environnement ou les données de notations de produits, plus de données sont disponibles en ligne. Ceci représente une opportunité sans égal pour étudier le comportement des individus et les interactions entre eux. Sur le plan politique, le fait de pouvoir interroger des ensembles de données de votes peut fournir des informations intéressantes pour les journalistes et les analystes politiques. En particulier, ce type de données peut être exploité pour l'investigation des sujet exceptionnellement conflictuels ou consensuels.
Considérons des données décrivant les sessions de votes dans le parlement Européen (PE). Un tel ensemble de données enregistre les votes de chaque député (MPE) dans l'hémicycle en plus des informations relatives aux parlementaires (e.g., genre, parti national, parti européen) et des sessions (e.g., sujet, date). Ces données offrent la possibilité d'étudier les accords et désaccords de sous-groupes cohérents, en particulier pour mettre en évidence des comportements inattendus. Par exemple, il est attendu que sur la majorité des sessions, les députés votent selon la ligne politique de leurs partis politiques respectifs. Cependant, lorsque les sujets sont plutôt d'intérêt d'un pays particulier dans l'Europe, des coalitions peuvent se former ou se dissoudre. À titre d'exemple, quand une procédure législative concernant la pêche est proposée devant les MPE dans l'hémicycle, les MPE des nations insulaires du Royaume-Uni peuvent voter en accord sans être influencés par la différence entre les lignes politiques de leurs alliances respectives, cela peut suggérer un accord exceptionnel comparé à la polarisation observée habituellement. Dans cette thèse, nous nous intéressons à ce type de motifs décrivant des (dés)accords exceptionnels, pas uniquement sur les données de votes mais également sur des données similaires appelées données comportementales. Nous élaborons deux méthodes complémentaires appelées Debunk et Deviant. La première permet la découverte de (dés)accords exceptionnels entre groupes tandis que la seconde permet de mettre en évidence les comportements exceptionnels qui peuvent au sein d'un même groupe. Idéalement, ces deux méthodes ont pour objective de donner un aperçu complet et concis des comportements exceptionnels dans les données comportementales. Dans l'esprit d'évaluer la capacité des deux méthodes à réaliser cet objectif, nous évaluons les performances quantitatives et qualitatives sur plusieurs jeux de données réelles. De plus, nous motivons l'utilisation des méthodes proposées dans le contexte du journalisme computationnel.

MOTS-CLÉS : découverte de sous-groupes intéressants, fouille de modèles exceptionnels, Analyse de Données Comportementales, Journalisme Computationnel.

Laboratoire (s) de recherche : Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)

Directeur de thèse :
Pr. Philippe Lamarre (INSA-Lyon),
Dr. Sylvie Cazalens (INSA-Lyon),
Dr. Marc Plantevit (Université Claude Bernard Lyon 1).

Président de jury : -

Composition du jury :
Sihem Amer-Yahia (Directrice de recherche, CNRS)
Arno Siebes (Professeur des Universités, Université d'Utrecht)
Arno Knobbe (Maître de conférences, Université de Leiden)
Ioana Manolescu (Directrice de recherche, INRIA)
Amedeo Napoli (Directeur de recherche, CNRS)
Philippe Lamarre (Professeur des Universités, INSA-Lyon)
Sylvie Cazalens (Maître de conférences, INSA-Lyon)
Marc Plantevit (Maître de conférences, Université Claude Bernard Lyon 1)