Knowledge Representation with Condensed Set-Valued Attributes

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von Dipl.-Inform. Frank Christopher Rügheimer
geboren am 23. November 1978 in Schwerin

Gutachterinnen und Gutachter:
Prof. Dr. Rudolf Kruse
Prof. Dr. Eyke Hüllermeier
Prof. Dr. Myra Spiliopoulou

Ort und Datum des Promotionskolloquiums:
Magdeburg, 22. Oktober 2012

Zusammenfassung

Die systematische Erhebung komplexer, inhärent mengenwertiger Daten in Wirtschaft und Forschung sowie die Einbindung ontologiebasierter Annotationen zur Darstellung von Kontextwissen stellen statistische Modellierungen zur Wissensrepräsentation vor neue Herausforderungen. Einerseits stehen bei mengenwertigen Daten bereits für vergleichsweise kleine Wertevorräte kombinatorische Effekte der direkten Modellierung durch Random Sets entgegen; andererseits erfordert der Wegfall von Standardannahmen etwa über die Disjunktheit von Attributwerten oder statistische Unabhängigkeiten, eine Neubewertung und Ergänzung des bestehenden Methodenpools für komprimierte Darstellungen.

Der im Rahmen dieser Arbeit entwickelte empirische Modellierungsansatz für Verteilungen über mengenwertigen Attributen unterstützt die Aufbereitung, Verknüpfung und Interpretation derartiger Daten. Durch die Nutzung lokaler Approximationen innerhalb eines probabilistischen Rahmenmodells werden dabei die Vorteile einer kompakten und skalierbaren Repräsentation mit dem interpretierbarer Marginalverteilungen kombiniert.

Erweiterungen des Ansatzes auf multivariate Verteilungen und strukturierte Wertedomänen ergänzen das Modell. Letztere ermöglicht die Integration formal spezifizierten Kontextwissens zur Erschließung umfangreicher inhomogener Datenbestände. Zusammen mit einem in der Arbeit eingeführten mit einer Interpretation im Kontextmodell kompatiblen Aggreationsoperator gestattet sie es hierbei, Beobachtungen unterschiedlicher Granularität in einem gemeinsamen Modell zu verknüpfen und die gesammelte Verteilungsinformation innerhalb einer induzierte Familie gekoppelter Ereignisräume zu projezieren. Das Modell unterstützt so nutzer- und einsatzspezifische Sichten, die durch wahlweises Aggregieren oder Expandieren von Detailinformationen erzeugt werden.

Das vorgestellte Modell wird hinsichtlich seiner Eigenschaften und seines Einsatzbereichs von bestehende Ansätzen wie Dempster-Shafer Theorie, Possibilitätstheorie und einer Codierung unter Nutzung eines probabilistischen Graphischen Models abgrenzt. Die zunächst auf Basis theoretischer Erkenntnisse geführte Argumentation setzt sich in einer experimentellen Evaluierung der implementierten Modelle anhand eines komplexen öffentlich verfügbaren Datensatzes fort.

Obwohl Anwendungen derzeit vorrangig in der Bioinformatik liegen, ist das zugrunde liegende Modell selbst nicht an eine spezifische Interpretationen gebunden. Die im Rahmen der Arbeit erstelle Softwarebibliothek und die dazugehörigen Tools reflektieren dies durch die Unterstützung eine Reihe unterschiedlicher Mengeninterpretationen und Abfragemodi.

Contents

1 Introduction **1**
1.1 Annotated Data 2
1.2 Research Tasks 3
1.3 Outline of this Dissertation 5

2 Foundations **7**
2.1 Perception and Knowledge 7
2.1.1 Related Notions 14
2.2 Mathematical Formulation 15
2.2.1 Attributes and the Frame of Discernment 15
2.2.2 Set-Valued Attributes 17
2.3 Choice of Representation Framework 19
2.3.1 Symbolic Approaches 22
2.3.2 Ordered Facts and Epistemic Entrenchment 24
2.3.3 Probability-Based Representations 26
2.4 Conclusions 32

3 Graphical Models **35**
3.1 Introduction and Principles 35
3.2 Statistical Independence and Decomposition 38
3.3 Node Separation in Graphs 41
3.4 Bayesian Networks 43
3.5 Markov Networks 43
3.6 Reasoning with Graphical Models 45
3.7 Summary and Further Reading 49

4 Probability and Set-Valued Data **51**
4.1 Interpretation of Model Components 51
4.1.1 Interpretations for Sets 52
4.1.2 Interpretation of Probability Distributions 54
4.2 Random Sets 54
4.2.1 Definition 55
4.2.2 Properties of Random Sets 55
4.2.3 Uses and Limitations 56

4.3 Dempster-Shafer Theory of Evidence 57
4.3.1 Upper and Lower Probability 57
4.3.2 Reinterpretation by Shafer 58
4.3.3 Reasoning in the Dempster-Shafer Framework 61
4.3.4 Critical Discussion 64
4.4 Possibility Theory 67
4.4.1 Axiomatic Approach to Possibility 68
4.4.2 Context Model Interpretation of Possibility 71
4.4.3 Formalization 75
4.4.4 Relation to Random Set Properties 76
4.4.5 The Possibilistic Aggregation Problem 78
4.4.6 Maxitivity and Consonance 82
4.4.7 Discussion of Possibility Theory 83
4.5 Conclusions 84

5 Condensed Representation for Set-Attribute Distributions 87
5.1 Conventions and Notation 88
5.2 Condensed Representation 88
5.3 Extension to the Multivariate Case 93
5.3.1 Introduction to Tuple-Based Formalization 94
5.3.2 Multiple Condensed Set-Valued Attributes 97
5.3.3 Marginal Distributions 102
5.3.4 Conditioning and Conditional Distributions 105
5.3.5 Combination with Graphical Models 108
5.4 Application to Tree-Structured Domains 109
5.4.1 Related Frames of Discernment 110
5.4.2 Uncertainty over Hierarchical Attribute Domains 114
5.4.3 Multi-Label Instantiations 118
5.4.4 Summary 125
5.5 Discussion 125

6 Experiments and Evaluation 129
6.1 Saccharomyces Gene Annotation Data 129
6.2 Model Types and Experimental Setup 131
6.3 Data Preparation 132
6.4 Parameter Estimation 133
6.5 Evaluation Measures 135
6.6 Results 137
6.7 Comparison of Computation Time 142
6.8 Conclusions 143

7 Implications and Perspectives 145

7.1 Scientific Results 145
7.2 Software Development 149

List of Symbols **153**

Bibliography **163**

Index **175**

List of Figures

2.1 Two relations corresponding to identical set-instantiations of their component attributes (elements tuples indicated in grey). The relation on the left is a Cartesian product. 18

3.1 Structural component of a Bayesian Network Model for the chest clinic diagnostic problem (example initially proposed by Lauritzen and Spiegelhalter, 1988). 38

3.2 Graph representations, which encode the statements $A \perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp B \mid \emptyset$ (from left to right: using an undirected graph, using directed graphs with diverging edges in C or directed paths from A to B and from B to A respectively). 42

3.3 An undirected graph, its corresponding hypergraph representation and a join tree for propagation on the clique graph 44

3.4 Partitionings of a polytree w.r.t. two different pivot nodes (strong contour); light shading: node contributing to the π-factor, dark shading: nodes contributing to the λ-factor, not shaded: nodes unconnected to pivot node). 46

3.5 Message passing for the polytree propagation algorithm: π-messages travel from parent to child nodes, λ-messages from child to parent nodes. 47

4.1 Three urns containing colored spheres 72

4.2 Two random sets with identical one-point coverages (shading indicates contributions from individual contexts) 76

4.3 Examples of random sets and their respective one-point coverages (from top to bottom: consistent, with consonant focal sets, focal sets are singletons, disjoint focal sets) 77

5.1 Domains of a set-valued attribute A^*, the induced condensed set-valued attribute $A^\diamond$ and underlying basic domain Ω. Arrows indicate the set reduction mapping w.r.t Ω. Shaded elements of $\mathrm{dom}(A^*)$ mark multi-valued outcomes covering ω_2. 91

5.2 Three relations with different projections to the individual attribute domains. The respective tuples assigned by the set reduction mapping are (from left to right): $(a^\diamond, b^\diamond)$; $(a^\diamond, b_3)$ and $(a_2, b^\diamond)$. . . . 98
5.3 Joint distribution with coverage factors for two condensed set-valued attributes and one of its marginal distributions. . . . 100
5.4 Attribute value hierarchy with attached branching probabilities (root label not shown) . . . 115
5.5 Correspondence of subframes and single labels . . . 118
5.6 Possible refinements of an applicable label during sub-frame expansion w.r.t. the label a_1 from Figure 5.4 . . . 120
5.7 Extended attribute value hierarchy as data structure for the condensed representation of distributions over multi-valued instantiations (conditional probabilities and coverage factors indicated by solid and dotted arrows respectively) . . . 121

6.1 Segment of the slim version of the biological process sub-ontology in GO: Edges indicate hyponym relations, that is, “nucleus organization” is represented as a special type of “organelle organization” . . . 130
6.2 Average record Log-Likelihood for Random Set representation on test data as a function of Laplace Correction (log. scale); broken line indicates parameter value used for model RS1 . . . 141
6.3 Comparison of processing time for each distribution model (sum of training and application phases during crossover procedure). . . 142

List of Tables

2.1 Relation between location and leisure activities 24
2.2 Production summary for Example 2.2 30
2.3 Factorized representation using conditional distribution 32

4.1 Mapping used to select urn after die is thrown 73
4.2 Probability distribution for the urn example 73
4.3 Elementary possibility assignments for the modified urn example computed via the probability of compatible contexts 74

6.1 Evaluation results for model using independent binary variables (one-point-coverage) with Laplace correction of 0.5 (INDEP) . . . 138
6.2 Evaluation results for condensed distribution on hierarchically structured domain with Laplace correction of 0.5 (HCDM) 138
6.3 Evaluation results for condensed distribution on unstructured domain with Laplace correction of 0.5 (CDM) 138
6.4 Evaluation results for Bayesian Network Model (Laplace correction of 0.5, independence criterion ≤ 0.1) (BN1) 138
6.5 Evaluation results for Bayesian Network Model (Laplace correction of 0.5, independence criterion ≤ 0.01) (BN2) 139
6.6 Evaluation results for Random Set representation (Laplace correction of $2.5 \cdot 10^{-9}$) (RS1) 139
6.7 Evaluation results for Random Set representation (Laplace correction of $1.0 \cdot 10^{-12}$) (RS2) 139
6.8 Summary of evaluation results over all model classes and criteria . 139
6.9 Cardinality of annotation sets in *Saccharomyces cerevisiae* data . 140

7.1 Software components implementing core functionality of hierarchical distribution model 149

1 Introduction

Owing the rapid development of information technology, economic activities are now supported by an information infrastructure that collects and processes large amounts of data. Technologies such as bar-code readers and transponders permit real-time automated updates of databases to reflect changes in the physical world. In addition to its immediate use to process control and decision making such data can be analyzed to detect error sources or help optimize business processes. The broad coverage of information collected not only entails the creation of large databases but is frequently accompanied by the introduction of additional, less stringent data representation formats and a stronger emphasis on relations between different data.

In research this development is paralleled by the introduction of new instruments and experimental techniques in the natural sciences. It has lead to a previously unmatched rate at which scientific data are being acquired and distributed. These new techniques are complemented with an storage and communication infrastructure, but also provide challenges to data processing. The field of molecular biology, for instance, has undergone a remarkable transformation that saw time-consuming laboratory techniques replaced by high-throughput processes such as genome-wide sequencing or microarray experiments. Even a basic analysis of these data heavily relies on computational methods, e.g., to match observed sequences against those stored in databases, to standardize data from different microarray chips, or to identify protein in a sample by comparing observed mass spectra with the output of *in-silico* models.

However, complex patterns of interdependence, the inhomogeneous nature of data obtained from complementary types of experiments, the need to integrate extensive background knowledge and – not a least – sheer amount of collected data still pose considerable obstacles for researchers that aim to interpret data in the context of the underlying biological systems or harness them for testable predictions. Additionally, data analysis and modeling must deal with biological and technical variance in the measurements, which occurs even in tightly controlled experiments.

In the light of these challenges, methods for data integration, data-driven model induction, as well as hypotheses generation and testing are drawing interest from a growing community. Statistical and Machine Learning approaches address these needs by providing computational tools for the induction of predictive models from empirical data. Such tools draw on a selection of techniques to solve reoccurring tasks, such as the integration of information from multiple sources, quantitative prediction, classification, i.e., the assignment of data to one of several predefined groups, and clustering – understood as the partitioning of a data set into subsets of similar cases.

But regardless of the specific application, any predictive data-driven model relies on on some compact representations of general properties of the data generating process. These dense representations are applied either by themselves or in combination with case-specific observations to make predictions about new cases. The quality of predictions directly depends on the assumptions and methods employed for the internal representations of those models. This is particularly relevant for models in systems biology – an area that focuses on studying complex, evolved systems and their regulation in the living organism. Unfortunately that setting is considerably removed form the idealized scenario of a low number of weakly connected variables for which standard modeling approaches were originally developed. Only through careful reassessment of assumptions in modeling techniques and by supplementing information available from diverse information sources can we hope to construct models that serve current needs and yet withstand the challenge of empirical validation.

1.1 Annotated Data

With the emergence of new types of data sources, requirements for knowledge representations and processing method become more varied. Textbooks on statistics assume that each case is described by a vector of values chosen from a static set of mutually exclusive values per property. However, such data representations no longer constitute a default. Instead, the highly structured data processed in text analysis and biology are typically supplemented with annotations, frequently in the form of *sets of terms* from a restricted vocabulary that are assigned to objects to *collectively* describe one particular property. The advantages of this approach as compared to using mutually exclusive category labels are manifold:

1. Annotation sets are well suited to express structured information such as relations,

2. Annotation sets are easily extensible, i.e. the set of terms used in annotations can be expanded when additional distinctions are required or when new cases are not well covered by concepts described by the existing term

set (it is common practice to release new versions of annotation schemes at intervals to keep them in line with recent research results),

3. Annotation sets tolerate overlap between concepts allowing domain experts to specify facts using the established terminology of their respective field,
4. Annotations sets avoid the artificial, often arbitrary distinctions, when predefined categories force the assignment of intermediate cases to a single descriptor, and
5. Annotators are free to chose concepts on a suitable level of specificity (concept hierarchies).

Unfortunately, automated model induction techniques employed for categorical data are rarely suited to deal with data that features set-annotations. The ambition of this work is the develop a compact representation for statistical properties of set-annotated data that provides access to informative summaries and can serve as an internal knowledge representation for use in predictive models. In particular the work focuses on the aspects of a compact, scalable representation, robustness against overfitting and the capacity to integrate data that is specified in relation to different levels of detail, e.g. due to its origin from diverse sources. The proposed algorithms and models are embedded into a generally applicable formal framework. Nevertheless interpretations for model elements and operations are provided and discussed in relation to application tasks. While tools and concepts introduced in this thesis are general and support applications in other fields, the presentation and the examples given focus on current challenges of computational biology, namely the analysis of expression, proteome or metabolome data and their integration with the increasingly rich body of biological background knowledge that has been formalized in ontologies.

1.2 Research Tasks

The program of providing a modeling framework for set-annotated data was subdivided into six research tasks, which are reflected in the structure of the thesis:

Task 1 – Establishing terms and notations: The objective of this task is to prepare the ground for the subsequent scientific tasks. It involves positioning the research in a broader context, discussing elementary concepts of data and knowledge representations and clarifying notions and terminology. It is furthermore concerned with the development of a notation system that is required to specify model components and operations. This step involves a short review of existing formalizations, which are than adapted to the particular questions at hand and – where necessary – supplemented and extended to support the set-annotation setting.

Task 2 – Review of extant modeling approaches: Distributions over sets are typically defined over large sample spaces. This requirement needs to be reconciled with technological (e.g. number representations, storage capacity) and theoretical limitations (e.g., algorithmic complexity, (un)desirable properties) of modeling techniques for knowledge representation. Conducting an in-depth study of the strengths, application range, inherent assumptions and limitations among the existing approaches allows to identify suitable techniques for modeling set-distributions. The research task therefore consists in investigating these characteristics to identify both documented and implicit assumptions of various model types. Following that, the uncovered model properties are analyzed and assessed with regard to their suitability for the representation of set-distributions.

Task 3 – Model development: Based on the results of the Tasks 1 and 2, this task is concerned with the development of data structures and algorithms for efficient induction, storage and querying of a statistical model for distribution of set-instantiations.

Task 4 – Data integration: To enhance the utility of the model developed in Task 3, data structures and algorithms for the conversion of data between domains of varying underlying variable sets and resolutions are developed. This capability is essential to integrate data from multiple sources or complementary experiments and to adapt the presentation results to the different information needs of users.

Task 5 – Software development: The objective of this task is to implement tools that facilitates the application and evaluation of distribution models for set-valued data. To enable a comparison to alternative representation strategies the task also includes the adaptation or *de-novo* implementation of tools that support several extant approaches and the development of a test framework for assessing model performance. The implementations will be directed at advanced users, who wish to embed the developed approaches into their own data processing pipelines.

Task 6 – Comparative assessment: The final task is to empirically test and evaluate the performance of the developed approaches in comparison to extant methods described in the literature. The goal of this task is to point out advantages and limitations for each method and provide guidelines for the application of the new model.

1.3 Outline of this Dissertation

In Chapter 2 the reader is introduced to notions employed throughout this thesis. I provide a brief exposition of the underlying epistemological theory and its implications for knowledge engineering. On this basis, I proceed to elaborate a terminology and recapitulate suitable formalizations of the related concepts. To facilitate the discussion of advanced representation types the resulting framework is extended by introducing and formalizing new concepts, such as composite and set-valued attributes, along with their associated domains and notation formats.

A subsequent analysis is concerned with the mathematical frameworks that form the basis of several popular approaches to knowledge representations. I comment on the respective objectives, capabilities and limitations of each class of approach. Based on this analysis and the clarified notions of knowledge, I delineate the central problem motivating this thesis, namely the application of statistical modeling to annotation sets.

In Chapter 3 I proceed with a more detailed investigation into Graphical Models. Graphical models have previously been used in knowledge representation as they have comparatively storage requirements and allow for efficient reasoning operations. The decomposition techniques employed in Graphical Models are later applied to multivariate versions of the proposed condensed random set approach. A Probabilistic Graphical Model also provides one of the candidate models for the representation distributions over annotation sets. The chapter furthermore contains a detailed exposition on the notion of statistical independence and the relevance of this notion for modeling. It explains why independence assumptions are essential to achieve compact representations of large distributions including those over sets. Recognizing and understanding how these assumptions are used in any particular model therefore provides insights into the type of tasks to which the model is applicable.

In Chapter 4 I discuss different interpretations of sets and relate them to various applications of the random set concept. Following that, I point out practical limitation of the random set approach and explain why a direct modeling of set distributions is unrealistic for large-scale modeling. Existing frameworks for the representation of information about set-distributions are studied, compared to each other and related to the properties and semantic aspects of random sets that they intended to reproduce. The majority of the discussed limitations arise from a number of subtle assumptions inherent to those frameworks, which nevertheless have to be considered in any earnest attempt to apply these methods. These assumptions are made explicit and their consequences are discussed. From the results of that study I elaborate desirable properties and, eventually, a draft of a new compact and efficient knowledge representation that retains many of the advantages of extant random set-based approaches, without suffering from the critical limitation identified in those models.

The Chapter 5 serves to formulate the ideas developed from the results of the analysis from Chapter 4 into a general model for the condensed representation of distributions over sets. While the model is probability-based, it reuses features previously discussed in conjunction with possibility distributions. It is later extended to the multivariate setting, including a discussion on its integrates with probabilistic Graphical Models and a hierarchical version, which focuses on current needs of structured knowledge representations, e.g., in ontologies is introduced.

Chapter 6 is concerned with the experimental validation of claims and model assumptions. To that end implementations of the proposed knowledge representation and several alternative frameworks are evaluated in an application context using real data. This study reveals advantages and drawbacks of the respective approaches and allows an assessment of prediction quality based on measurable criteria.

In the final Chapter 7 I summarize the results and the main conclusions of my work before discussing recent applications and directions of future development.

2 Foundations

Although the nature of knowledge and the means to obtain it have been as a subject of philosophical discourse since antiquity, the term “knowledge” itself remains surprisingly vague. So far no precise definition has been universally adopted. Instead, a variety of definitions have been proposed and used – often tailored to specific contexts and applications, and although they overlap to a degree, the definitions differ in the use of the term, and the extend of its application. To further complicate matters, the colloquial use of the term “knowledge” has a diffuse boundary with several related concepts.

The objective of 2.1 is to clarify the notions as used throughout this thesis and to point out differences and relations between them. By giving a short exposition of the philosophical background I aim to guard against misunderstandings that may result from the adoption of different vantage points. Based on the concepts introduced in that section, the mathematical foundation for formal knowledge representation using attributes is recapitulated. This is followed by the introduction of the notion of a set-valued attribute. It is then argued that set-valued attributes provide a convenient approach to address many current challenges in knowledge modeling. I then proceed to investigate popular approaches to knowledge representation that rely on the concept of attributes, point out differences and similarities between them, identify the predominant classes of mathematical formulations and discuss the individual capabilities of each class.

2.1 Perception and Knowledge

The capacity to recognize and integrate information, generalize it into recognized patterns of stimuli and provide responses to those stimuli equips living beings with an ability for controlled interaction with their environment (e.g. Ritchie et al., 2008). The way we perceive and experience our environment is studied in the fields of cognitive neuroscience and psychology. While a comprehensive discussion of underlying biological and neurological mechanisms is not within the scope of this work, a brief excursion into models of human perception shall serve as a guide in setting up the framework for investigating knowledge, its acquisition, and ultimately finding suitable formal representations.

Adopting this perception-oriented view leads to the realization, that the majority of sources for human knowledge are subject to limitations – be it due to the construction of sense organs or, in extension, the available instruments. Consequently, knowledge will almost always be imperfect. Besides these limitations, additional factors may diminish the reliability of observation results:

- failure to control/observe variables relevant to a situation,
- limited precision resolution of numerical representations,
- missing or corrupted data,
- systematic biases, e.g. range-dependent sensitivity of instruments.

Moreover, errors may accidentally introduced in early preprocessing, e.g. by an instrument's firmware. Such problematic processsing steps include, for instance

- non-bijective mappings,
- inadequate rounding,
- "outlier removal" heuristics.

Finally, errors may be introduced in the interpretation of collected data. Common examples include:

- detection of spurious correlations (mistaking noise for signal),
- bias towards observers expectations,
- attribution of effects to incorrect causes,
- inadequate null models in statistical tests.

Throughout this work I therefore acknowledge, that absolute certainty about empirical statements is exceptional rather than commonplace. Any result that is based on measurements or observations retains some degree of uncertainty. Because reasoning with conventional, that is truth-functional, formal logic presupposes certainty, its application for real-world problems necessarily constitutes an idealization. Other widely used idealizations in knowledge representations are assumptions about logical or statistical independence of variables. Of course, such idealizations are often necessary or at least very useful to solve real-world problems. Yet, they impose a limit to the accuracy of models and potentially introduce systematic bias.

The task of the knowledge engineer consists in selecting such idealizations, assessing their suitability to a task at hand and devise a representation and operations to model a situation of interest. For instance, most models will provide mechanisms to summarize collected data or to combine different pieces of information to meet information needs of users. If appropriate idealizations are chosen the resulting model is a valuable tool that serves its users in solving complex planning or decision tasks.

Objects The basis of any meaningful discourse is the coherent perception of entities, and their representation as mental concepts. Human beings have in fact evolved a remarkable ability[1] to do so, which they routinely use when interacting with their environment. We call those entities objects and usually take for granted that others arrive at conceptualizations very similar to our own one, so communication is possible. Moreover, we have learned to deal with abstract concepts, such as categories, and to attribute a certain degree of autonomy to them. This allows reasoning about notions independent of an immediate physical environment.

According to the philosopher Karl Popper (Popper and Eccles, 1985, ch. P2), all concepts or entities can be assigned to one or more of three worlds, namely

- The world of physical objects (World 1)
- The world of mental states (World 2)
- The world of products of the mind (World 3)

Popper describes World 1 as a world of physical entities, states of physical entities or physical states. These entities include “physical objects, processes, forces and fields of forces which may interact with material bodies”. World 2 refers to an individuals subjective experience and includes “states of consciousness, psychological dispositions and possibly unconscious states”. Examples of entities belonging to World 3 include stories, myths, language, works of art, institutions, but also scientific theories and problems. In contrast to the entities of World 2 the entities in World 3 are no longer dependent on an individual, but instead represent communicable ideas that are part of a shared culture. Many of those entities have a corresponding representation in World 1. A printed book obviously has a physical existence. The content of the book though is attributed to World 3. Remarkably, with the introduction of World 3 entities, knowledge itself is recognized as an object of study.

A critical reader might point out the lack of agreement among philosophical schools concerning the reality of entities from World 2 and World 3. However, I

[1] That very feat still provides a significant challenge for AI researchers trying to emulate it in artificial systems.

consider that question of no consequence with regard to knowledge representation. It suffices that entities are perceived as having an autonomous identity, so one can mentally refer to and formulate statements about them. The existence of a active market for software and electronic books, that are available as pure data streams clearly demonstrates that this is the case.

Finally, let me point out that many complex objects can be viewed both as composites of more fundamental entities, their relations and their interactions, and as separate entities with emergent properties. This allows observers to choose between levels of detail and provides a strategy to deal with a complex world by means of abstraction. Recently, ontologies have become a prominent tool in computer science to make relations between concepts explicit and available for formal modeling.

Attributes The means to discern, recognize and describe objects are provided by *attributes*. Attributes serve to represent object properties. In order to describe the realization of particular property, the object is assigned an *attribute value* or *instantiation* from a pool of labels. That pool of available labels forms the so-called *attribute domain*. The attribute domain must either be specified explicitly or implicitly by a system of conventions. Such conventions enable compact encodings and are a feature of all natural languages (Lewis, 1969).

attribute	value
color	white
material	ceramic
capacity	50 ml
temperature	60 ℃

As an example, consider the information given on the right about an espresso cup on my desk. Each of the value assignments extends the information about the object. The example also demonstrates different scale types associated with attributes. Nominal attributes, such as “color” and “basic material”, only allow to test for equality. Dichotomies (the cup is either dishwasher-safe or not) are modeled using a special case of a two-valued nominal scale. In contrast, capacity and temperature are measured on metric scales. For scales of the metric type, there is a meaningful interpretation of not only equality, but also of comparisons and differences. It is also possible to specify meaningful ratios of capacities (The cup is half filled.), whereas temperature would have to be converted to an absolute (e.g. Kelvin) scale first. A third type of scale is called ordinal scale and applies, for instance, to academic degrees. It supports equality and comparisons but provides no meaningful differences.

Finally, with respect to admissible attribute values, it is convenient to distinguish between the *single-valued attributes* with mutually exclusive values and *set-valued attributes* that may simultaneously be instantiated with the elements of a subset of their basic attribute domain. In Data-Mining, set-valued attributes are used in association analysis, e.g. to specify which products a customer of an online book store bought on the same purchase. Another example of a set-valued attribute is the indication that a board game is “for 3–5 players” . More recently set-valued attributes have become popular as a means to provide annotations to existing data. For instance, the Gene Ontology (GO) annotation system uses three set-valued attributes (called aspects) to characterize gene products: cellular component, molecular function and biological process (Ashburner et al., 2000). A GO-annotation is considered complete for an organism if each of these aspect attributes is instantiated with one or more values per gene product.

In other cases set-valued characteristics arise from properties with values that were once perceived as mutually exclusive. For instance, many forms treat a persons nationality as a single-valued characteristic, even though it is not uncommon for individuals to hold citizenship of more than one country.

It is remarked that for the above examples set-valued attributes are not the only possible representation. For instance defining a new attribute with all admissible combinations of values in its domain, would always allow for a reduction to the single-valued case. While such a reductionist approach can provide a viable solution for problems on a smaller scale, it is problematic when applied to larger term sets, as the exponential increase of possible value combinations will rapidly overwhelm existing software. It can thus be argued, that the variant using set-valued attributes not only provides a clear intuition of the meaning, but also circumvents a critical representation problem.

Data In order to discuss model induction and reasoning it is convenient to distinguish between data and knowledge. Concerning the definition of the terms I closely follow the terminology presented by Borgelt and Kruse (2002). Data is understood as *specific* information about *individual entities*. Statements like “Oslo is the capital of Norway.” or the “The French revolution started in 1789.” provide data[2]. Data are generated either actively by direct experimentation or collected passively by observation and measurement. In both cases the observation justifies the assignment of values to attributes of specific objects. While data are sometimes available in large quantities they do not by themselves supply explanations or allow to make predictions. Only in concert with information on general patterns, and regularities can data contribute to those tasks. Such patterns and regularities, are provided by context knowledge (see below) about the matter considered. Moreover, such general patterns and regularities can frequently be discovered within large datasets. This inductive approach is used to

[2] As acknowledged by Borgelt and Kruse, such statements are often colloquially referred to as knowledge. Borgelt and Kruse consider this disagreement as unfortunate, but hold on to their stricter distinction of the notions, which will be useful for the upcoming discussions.

obtain new hypotheses, which pending compatibility with new observations, may eventually become part of established knowledge.

Knowledge In contrast to “data”, I will use the the term “knowledge” only to refer to *generic* information, i.e. information that applies to a class of entities. It is expressed in generalizing statements, such as “most mammals have fur”, rules, principles or even complete scientific theories. Knowledge is efficient in the sense that it explains data for a large number of cases. Newtonian Mechanics, for instance, provide a very condensed description of the movement of physical objects. The generic character of knowledge is revealed in its capacity to make predictions about cases that have not been observed yet. This capacity to predict new cases based on acquired knowledge is used in all tasks involving planned action. For instance, a gardener who sows in spring does so in the expectation that the seeds will germinate and grow into plants of a certain species during the following months. This expectation is the result of knowledge about the life circle of plants. Apart from that role, predictions are essential for refining an existing pool of knowledge by means of falsification or repeated validation of new hypotheses.

Classical pre-Aristotelean philosophers distinguished between true, that is absolute, knowledge (epistéme) and mere opinion (doxa). In Plato’s works it is argued that knowledge entails truth (Plato, 1999a). The distinction of knowledge from true opinions (orthai doxai) is further elaborated in (Plato, 1999b):

> ”The difference [. . .] is only that he who has knowledge will always be right; but he who has right opinion will sometimes be right, and sometimes not.”

According to that definition, knowledge is certain. The argumentation in Plato’s dialog continues by stating that in order for a proposition to be certain, it requires justification, which leads to the popular “traditional” definition of knowledge as *justified true opinion*.

The notion that knowledge entails truth and certainty, was adopted in many subsequent definitions (e.g. Reimer, 1991). In the field of knowledge discovery, Fayyad et al. (1996) have used an even more restrictive definition by adding an “interestingness” criterion. How exactly this interestingness is measured follows pragmatic considerations and depends on the application context.

An absolute notion of truth, however, is problematic outside the context of pure logic. Because observations made when learning from data represent only a sample of the full range of possible cases, the means to assess the truth of a statement are limited. The above definitions of knowledge fail to provide a practical mechanism to expand a body of knowledge other than by reduction to statements already contained therein, which we are supposed to accept as true for an unspecified reason.

Furthermore, models may be useful even if a fraction of their predictions are incorrect (e.g., weather forecast). For those reasons the notion of knowledge adopted in this thesis requires neither truth nor interestingness as defining criteria. Instead, individual pieces of knowledge are characterized and assessed w.r.t. gradual qualities. Borgelt and Kruse (2002) suggest a list of assessment criteria, which is repeated below:

- correctness (probability of success in tests)
- generality (range of validity, conditions for validity)
- usefulness (relevance, predictive power)
- comprehensibility (simplicity, clarity, parsimony)
- novelty (previously unknown, unexpected)

The degree of *correctness* determines how often the predictions resulting from the knowledge turn out to be true. Correctness of a theory is often tested experimentally. Established scientific theories have usually been tested through a large number of trials and are assumed to possess a very high degree of correctness. As an example of pieces of knowledge with a lower degree of correctness one may cite heuristics or technical advice like "Restarting the server may solve the problem.". Such pieces of knowledge are useful, despite occasionally supplying incorrect answers. The advantage of the notion of correctness, however is, that it not only applies to a wider range of cases, but also replaces the philosophically cloudy criterion of truth with an empirically assessable gradual property.

Generality refers to the conditions under which a piece of knowledge may be applied. For example, the theory of Newtonian mechanics is quite general, in the sense that the same set of formulae describes a wide range of phenomena. Nevertheless, some limitations to the theory are documented. The "failed" experiments to measure the motion of the then postulated "luminiferous aether" by Michelson (1881) and later Michelson and Morley (1887) pointed to inconsistencies in its explanation of electromagnetic waves. For objects moving at considerable fractions of the speed of light or within strong gravitational fields, predictions from classical mechanics have been found to differ from results obtained from observations and measurements (Le Verrier, 1843; Walsh et al., 1979), nor does the theory apply to processes on sub-molecular scales (Feynman et al., 1965). By quantifying generality, scope is recognized as a criterion for characterizing knowledge.

Usefulness is concerned with the potential relevance of the knowledge to applications. The predictions obtained using Newtonian Mechanics, for instance, are useful due to their applications in engineering.

A high level of *comprehensibility* facilitates the transfer of knowledge. In conjunction with usefulness, comprehensibility also influences how easily knowledge is turned into applications. Comprehensibility depends on the complexity and presentation of the knowledge, but also on previous experiences of potential recipients, i.e., its relation to other pieces of knowledge.

Finally the *novelty* criterion is relevant in knowledge discovery tasks, though, even the affirmation of extant knowledge usually conveys some degree of novelty as it adds to evaluating the degree of correctness of a model or a hypothesis.

2.1.1 Related Notions

To further substantiate and clarify the notion of knowledge and contrast it with related concepts, it is useful to elaborate its relation to *belief* and the idea of *justification*.

Belief The notion of *belief* emphasizes subjectivity and is used to specify which pieces of knowledge a given agent holds for true. Thus, unlike knowledge, belief is always expressed relative to a particular agent and that agent's epistemic state. In quantitative approaches, belief may be seen as a gradual property allowing to express an agents preference w.r.t. (preliminarily) accepting either of two mutually exclusive pieces of knowledge. Belief models aim at providing rules that establish what an intelligent agent's opinion about the truth of statements should be in a certain situation.

Because a rational belief should at least partially be based on facts and knowledge, mathematical tools used for the representation of (partial) knowledge and belief frequently overlap. While it is assumed that rational agents form their opinions in agreement with their knowledge, some authors allow belief to be influenced by preferences or considerations of utility with respect to future decisions or assumptions, e.g. by using of non-empirical priors in a Bayesian approach. Thus not all beliefs necessarily result from immediate evidence. This argument can be used to justify heuristics, such as default assumptions and ambiguity aversion, which many belief models draw upon to deal with insufficient information in decision tasks. Moreover, given the practical difficulty in verifying consistency between all pieces of knowledge, several approaches for belief representation allow to ascribe partially inconsistent belief states to agents.

Justification *Justification* is tied to the idea of knowledge as “justified true belief”. The justification for believing a proposition consists in the reasons for holding it for true. Justification-based approaches to knowledge representation (Doyle, 1979; de Kleer, 1986) aim at supporting dynamic knowledge by keeping track of these reasons. This additional structure allows pieces of knowledge to be retracted in along with any conclusion that depend on them.

2.2 Mathematical Formulation

Of course the practical utility of knowledge results from its application to problems. Problem solving almost always requires some kind of modeling. By focusing on a small set of objects related to the problem at hand, their relevant properties and presumed interaction are more easily understood. In this work I restrict the discussion to reasoning about such ensembles and assume that a finite set $\{A_1, \dots, A_n\}$ of attributes is employed to describe it. For a broader overview of approaches to knowledge representation methods the reader is referred to the dedicated literature (e.g. Brachman and Levesque, 2004; Sowa, 2000).

2.2.1 Attributes and the Frame of Discernment

While so far I focused on providing an intuition to concepts and notions of knowledge representation, is now helpful to supplement that intuition with a framework of formal definitions. Given an a universe of conceivable objects O, a (single-valued) attribute A applicable to the elements in O constitutes a mapping

$$A : O \to \Lambda, \tag{2.1}$$

where Λ is a set of the labels forming the attribute domain (written $\mathrm{dom}(A) = \Lambda$). For an object $o \in O$ the attribute value $A(o)$ provides a partial specification of that object's state. The value $A(o)$ is an *instantiation* of the attribute A. In this dissertation it is generally assumed that $\mathrm{dom}(A)$ is modeled as a finite set.

In a model the choice of attributes determines the possible distinctions between object states. The set of all admissible state descriptions Ω is called *frame of discernment* (Shafer, 1976) or sometimes *universe of discourse* (Zadeh, 1978). Figuratively, the frame of discernment determines the “vocabulary” available for referring to the state of the world, and therefore the type of propositions that can be considered in the knowledge model. A single, precise attribute A only supports the frame of discernment $\Omega_{\{A\}} = \mathrm{dom}(A)$ since state descriptions may only differ with respect to the value of A. Extending the representation to include a set $\{A_1, \ldots, A_n\}$ of such attributes with $\mathrm{dom}(A_i) = \Lambda_i \quad \forall i = 1, \ldots, n$ adds further dimensions to the description and thus permits a more detailed frame of discernment to be formed. When sets of attributes are used, instantiations are n-tuples of attribute values from the respective domains. Since each permutation of the attributes leads to a different way to describe such a tuple, an arbitrary but fixed order $\prec$ is defined over the attributes to enforce a unique representation of elements. Using the implicit order defined by the attribute indices yields

$$\Omega_{\{A_1,\ldots,A_n\}} = \times_{i=1}^{n} \mathrm{dom}(A_i) = \times_{i=1}^{n} \Lambda_i = \Lambda_1 \times \cdots \times \Lambda_n.$$

Obviously, if all attributes A_i have finite domains and are freely combined, the cardinality of the resulting frame is the product of the cardinalities of the individual attribute domains:

$$|\Omega_{\{A_1,\ldots,A_n\}}| = \prod_{i=1}^{n} |\mathrm{dom}(A_i)|.$$

For brevity, I will omit indices if the underlying set of attributes is irrelevant or clear from the context.

In order to convey partial knowledge about Ω one would often focus on selected attributes and their instantiations. A combination of existing attributes that are instantiated simultaneously gives rise to a new, derived attribute. The domain of such a *composite attribute* consist in the product space of the constituent attributes' domains. Accordingly, the composite attribute $\mathrm{catt}(X)$ generated from a set $X = \{A_{j_1}, \ldots, A_{j_k}\} \subseteq \{A_1, \ldots, A_n\}$ of attributes, where the indices are chosen in such a way, that $A_{j_1} \prec \cdots \prec A_{j_k}$, is defined as the function

$$\begin{array}{rrcl} \mathrm{catt}(X): & O & \rightarrow & \displaystyle\mathop{\times}_{l=1}^{k} \mathrm{dom}(A_{j_k}(o) \\ & o & \mapsto & (A_{j_1}(o), \ldots, A_{j_k}(o)). \end{array}$$

In practice, the combined domain may be a true subset of the joint domain due to logical dependencies within the considered attribute set. Although composite attributes rarely translate to a natural language assessment of a situation, they serve as abstractions for investigating interactions between attributes. Specifically, they allow us to treat the elements $\omega \in \Omega$ as instantiations of a single composite attribute, which maps to the frame of discernment:

$$\begin{array}{rrcl} \mathrm{catt}(\{A_1, \ldots, A_n\}): & O & \rightarrow & \Omega \\ & o & \mapsto & (A_1(o), \ldots, A_n(o)). \end{array}$$

Conversely, specifying instantiations for all attributes $A_i, \quad i = 1 \ldots n$ singles out a unique element of Ω. Thus the same information may be stated using values of individual attributes or directly on the higher dimensional combined space. A formalization specifically developed to support operations with composite attributes is introduced in Borgelt and Kruse (2002, page 63ff.)

2.2.2 Set-Valued Attributes

Although single-valued attributes have proven sufficient to represent the relevant information for numerous problems, some concepts are more easily modeled if sets are admitted as attribute values. In addition to their natural interpretation as descriptors for potentially multi-valued properties (cf. page 10), set-attributes can be used, for instance, to refer to *quantities subject to intrinsic variability* (Dubois, 2006) or to imprecision that arises from information deficits (compare page 23).

A formal definition of a set-valued attribute (Equation 2.2) is obtained by extension from the single-valued case given in Equation 2.1:

$$A^* : O \to 2^{\Lambda}. \tag{2.2}$$

To distinguish between conventional, precise attributes and their set-valued counterparts the latter are marked with the superscript "$*$".

In terms of expressiveness, set-valued attributes are an extension of their conventional counterparts. Each conventional attribute can be emulated by a set-valued attribute that is restricted to singleton values. Occasionally, one will find it convenient to modify the definition of set-valued attributes to exclude assignments to the empty set, as the latter may not be meaningful in that a particular interpretation (see discussion of imprecision on page 23):

$$A^* : O \to 2^{\Lambda} \setminus \emptyset. \tag{2.3}$$

The choice of the formalization of a set-valued counterpart of the composite attribute requires some consideration though. A mapping from objects to vectors of set-instantiations w.r.t. the component attributes may initially seem a good candidate for such an extension. However, such a mapping would yield vectors over fixed domains, not sets. This change of representation format is inconvenient for modeling the subsequent combination with with other set-valued attributes. On the other hand, mapping to subsets of the product space of the underlying label sets allows to connect set-descriptions referring to higher dimensional spaces to set-valued instantiations of individual attributes. From a mathematical point of view the images of those mappings are *relations*. By choosing mappings to relations over n-dimensional attribute domains, conventional attributes, set-valued attributes and concurrent instantiation over sets of such attributes seamlessly integrate into a unified framework.

To formalize that idea, consider a subspace $\Lambda = (\Lambda_{j_1} \times \cdots \times \Lambda_{j_k})$, $\{j_1, \ldots, j_k\} \subseteq 1, \ldots, n$ of Ω. The elements of Λ are vectors over instantiations of a selection of variables. A set-valued attribute $A^* : O \to 2^\Lambda$ over that domain induces other set-valued attributes $A^*_{j_l}, \quad l = 1, \ldots, k$ over the component domains via its projections

$$\begin{aligned} A^*_{j_l} : O &\to 2^{\Lambda_{j_l}} \\ o &\mapsto \left\{\lambda' \in \Lambda_{j_l} : \left(\exists \vec{\lambda} \in A^*(o) : \lambda' = \lambda_l\right)\right\}. \end{aligned}$$

The move from instances to relations brings along some changes that are relevant for the choice of the representation. For single-valued attributes, instantiation vectors over the domain of a combined attribute are uniquely described by a combination of the instantiations on each one-dimensional subspace. In contrast, in the multi-valued case, several different relations can induce an identical vector of set-instantiations w.r.t. those elementary subspaces. The largest such relation is the Cartesian product, which can always be formed by listing all combinations of elements from the instantiations of the individual attributes. Figure 2.1 shows two different relations between properties individually described by attribute sets $\{A^*, B^*\}$ with $\mathrm{dom}(A^*) = 2^{\{a_1,a_2,a_3,a_4,a_5\}}$ and $\mathrm{dom}(B^*) = 2^{\{b_1,b_2,b_3,b_4\}}$. Still, both relations correspond to identical set-instantiations $\{a_2, a_3, a_4\}$ and $\{b_2, b_3\}$ on the one-dimensional component domains.

Relations can also be used to reflect information about the *interaction* of attributes. For instance, if the attribute A^* maps to countries that may be visited by a given person o without visa requirements and attribute B^* to the countries

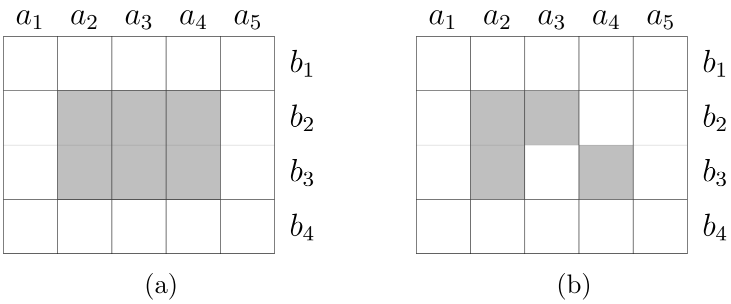

Figure 2.1: Two relations corresponding to identical set-instantiations of their component attributes (elements tuples indicated in grey). The relation on the left is a Cartesian product.

for which the person o holds citizenship, then the relation in Figure 2.1(b) could indicate due to which citizenship visa waivers are granted to o.

While attributes and frames form the common basis of knowledge representations, application tasks often call for emphasis on selected aspects of knowledge. Consequently there is no single universal approach to knowledge representation. Instead the differences in interpretation, inherent assumptions and aims lead to a variety of formalisms and calculi – each with its own benefits and drawbacks. The following section is concerned with some of those representation frameworks as well as their properties, interpretations and uses.

2.3 Choice of Representation Framework

The definition of attributes and the frame of discernment provides basic elements of a formal representations of selected aspects of the world. This permits to express data and knowledge about the modeled world segments in the form of assertions. To construct a knowledge representation we must specify a representation format and corresponding interpretation rules for such assertions. In practice the variety of knowledge representation tasks and their respective sets of requirements, gave rise to a considerable number of different formalisms and interpretations. In order analyze extant approaches and to systematically elaborate the similarities and differences between them I will focus on the following questions:

1. What is the subject of the assertions made?
2. What type of knowledge do the assertions refer to?
3. What is the expressive power of the assertion language?
4. Which the rules apply for drawing inferences?
5. Which assumptions are implicitly made in the approach?

The subject of the assertions reflects a perspective chosen for describing the modeled world. A number of representations are centered around explicitly listing objects within the modeled world (set-of-instances view). In those models attributes are used to directly provide information about particular instances that exist in the considered world segment. The set-of-instances perspective is commonly adopted e.g., for databases.

Alternatively one may consider the modeled world segment as a whole. In this type of model attributes are used to make assertions about the state of that modeled world segment itself, allowing to discerning it from other potential realization. This second view presupposes that the selected set of attributes is at least in principle sufficient to fully discern relevant the state of the model world. A particular, usually unknown *instantiation* $\omega_0 \in \Omega$ of those attributes identifies a unique, true state of the modeled section of the world (state-of-the-world view) and knowledge is given in the form of constraints that aim to help identify the true situation among the alternatives. All statements implicitly refer to the same object, which often allows for a simplified representation. Because of the close correspondence between states and their respective descriptions and for parsimony of expression, I will use "state" also to refer to state descriptions.

Conversely, for the set-of-instances view, the situation of the considered world segment is described by referring to a non-empty subset $W_0 \subseteq \Omega$ of realized instantiations. It is specifically suited to formulate statements about collections of objects or relations. Again, the distinction between that situation itself and its respective formal representation is dropped in favor of linguistic convenience. For the purposes considered in this work it is assumed that the attribute domains – and thus the frame of discernment Ω – are specified in advance.

Depending on the answer to the first question, assertions about different types of knowledge can used. Frequently, the information is conveyed in the form of constraints w.r.t. the value assignments for attributes. Some constraints suitable to the set-of-instances view are:

- restrictions to attribute values for individual objects or classes of objects ("That car is blue"; "None of these books are in English or French"),
- relation between objects and/or classes of objects ("The blue block is on top of the yellow block.","Lions are mammals").
- relations between the values of attributes of an individual object or a class of objects ("Raw potatoes are inedible.")
- relations between the attribute values of different objects or classes of objects ("The car and the bus have the same color"),
- statistical properties about the values assigned to a class of objects, ("40% of the world population have blood type A"),

In the set-of-instances view assertions must specify the object(s) they refer to. Thus assertion languages will usually use conventions (like pre-defined categories) or mechanisms to support such specifications (e.g. predicate logic).

If attributes refer directly to the state of the world, of course, no such specification is required:

- attribute value constants taken or excluded ("The temperature is low.")

- relations between the value of different attributes ("If the taste is sweet, the energy content is high.")
- assessments of the probability of particular attribute-values or attribute-value combinations, ("The chance of precipitation is at 80%.")

The above examples demonstrate very different types of constraints. In particular the last statement in each groups refer a distribution of values rather then the set of possible attribute values for selected subset of objects/object of interest. Moreover the linguistic terms used to make these assertions e.g. 'low' or 'sweet' may themselves be represented as precise labels for scalar values, as intervals or e.g. fuzzy concepts. Knowledge engineers may choose from a number the mathematical formalizations, that emphasize qualitatively different aspects of knowledge. Such a formalization typically consists of a coding method, a matching set of operations (a logic) and a suitable interpretation. Naturally the choice of the framework depends on the desired expressive power of the specification language. Consider for example the two representations of knowledge about Joe's car:

a) a simple attribute vector,

b) lists of possible attribute values for each attribute

Initially all attributes in (a) will be set to "unknown", (b) will simply permit all values for each attribute. Both representations allow to record knowledge about an object that reflects positive evidence ("Joe's car is red."). Negative evidence as in "Joe's car is not yellow.", however, can not be used in (a), as the statement on it's own does not justify the assignment of any of the other values. In contrast, with representation (b) the assertion is handled by simply excluding yellow from the list of remaining values for the color attribute.

Although the simple tuple representation (a) is inferior to (b) in terms of expressiveness, it is also more compact. Moreover, many observation tasks involve only positive information and a compact representation as suggested in (a) can well be an appropriate choice.

Among commonly used formal frameworks two major classes stand out. On one side there are frameworks for dichotomous assessments, which are based on binary logic (e.g. symbolic logic, relational algebra) on the other side those that draw on the probabilistic calculus. Furthermore, a number of ordinal and non-probabilistic numeric frameworks exist. Some of these frameworks forming intermediates, which combine aspects of both the above classes. The members of the main groups emphasize different properties of knowledge reflecting the distinction between non-statistical certain knowledge – or at least knowledge considered certain for the purpose of reasoning – and knowledge, that is uncertain or refers to statistical properties of collections of cases. The difference between those formalisms is explained by looking at the way inferences are drawn.

Inference denotes the process of combining pieces of information to arrive at previously unknown conclusions. Regardless of the calculus used, a successful inference requires premises, as well as an inference mechanism. Only if the pieces of information match, the inference can be drawn. The (potentially repeated) application of that process is called *reasoning*.

Example 2.1. To establish whether a heated piece of metal has reached a desired temperature for further processing the following inference can be used:

$$\begin{array}{l} \text{The metal is glowing orange}(A). \\ \text{If the metal is glowing orange the metal has the right temperature } (A \to B). \\ \hline \text{Therefore, the metal has the right temperature } (B). \end{array}$$

Observing the piece of metal is indeed glowing orange and assuming the specified rule is accepted, the inference leading to the conclusion about the metals temperature can be drawn. In this case A and $A \to B$ serve as premises. The conclusion B is obtained by employing the *modus ponens* as an inference mechanism. □

Example 2.1 illustrated an inference drawn using propositional logic. When deducing specific information by reasoning, the task consists in finding a viable path of inferences leading from the already established pieces of information to other pieces that satisfy current information needs.

2.3.1 Symbolic Approaches

For the symbolic approaches to knowledge representation, it is assumed that a situation can in principle be correctly described in terms of a set of non-contradictory facts, where a fact is a proposition that is true in the described situation. In the context of belief representation that set contains those propositions that are held for true by an agent.

Once the set of facts is established, logical inferences are used to derive additional propositions. Thus reasoning amounts to applying the propositional calculus or – for more advanced models – the predicate calculus on the sentences of a language (Example 2.1). The direct implementation of that idea leads to knowledge representations via belief sets (Levi, 1980) and belief bases.

In the case of belief sets, the set of stored propositions is closed under logical inference and directly equated with the agents current beliefs. Due to the potentially exponential growth of the closure and the correspondingly costly expansion operation, the applications of that approach are limited to very restricted languages or problem settings.

In contrast to that, belief bases distinguish between explicitly accepted sentences and those that are merely derived by inference. Only the originally accepted sentences need to be stored and inferences are drawn only as part of the query processing. The PROLOG programming language considerably contributed to the prominence of the belief base approach in artificial intelligence, as it enabled practitioners to comfortably implement such systems. Unlike with the belief set approach, the distinction between primary and derived pieces of knowledge supports interpretations where acceptance of derived sentences is provisional. This view has implications for belief change operations (Gärdenfors, 1988; Dubois and Prade, 1998b) in the light of new evidence. The recurring problem of dealing with inconsistent information was one of the motivating factors that lead to the development of the ordinal representation frameworks discussed in 2.3.2.

Drawing on the same mathematical foundation of two-valued logic, *relational* knowledge representations are usually employed in connection with the set-of-instances view. Internally the relations are stored as lists of tuples or tuple indices encoding the indicator function of the relation. Relational databases are a prominent application that relies on this interpretation. In that context instantiations W_0 realized in the current situation are identified with tuples stored in the database. To store relations efficiently, the decomposable structure of many high-dimensional relations is used by splitting them into linked sub-relations of lower dimensionality. Decomposition (of relations) forms the basis of the compact data representation and efficient operations achieved with relational databases (compare Chapter 3).

When applied to the state-of-the world approach, relational representations provide the means to deal with situations, where the actual state is underspecified. In that case, we speak of knowledge that is *imprecise* w.r.t. the identity of the true state ω_0. Imprecision occurs when the best possible description of ω_0 under the current knowledge consists in a sets of several candidate states that are compatible with the restrictions imposed by the accepted sentences. Statements like “It will visit either on Wednesday, Friday or Saturday” and “The train will arrive between 8 and 9 o'clock” are imprecise because they do not supply an exact value, but rather limit the “set of alternatives”. In comparison, the statement “The train arrived at 8.37” gives a precise[3] value. Reasoning under imprecision is founded on constraints, which are expressed by compatibility relations that

[3] Time being a continuum, there is some imprecision even in that statement; it is introduced inadvertently due to the limited resolution of the chosen scale (minutes). For all practical purposes, however, imprecision on that level can be neglected if the measurement scale is chosen sufficiently fine-grained.

link the values of different attributes to each other. A restriction of the instantiations of a subset of attributes may then reduce the possible choices for the instantiation of the remaining attributes. Imprecision can be a property of both generic and factual pieces of information.

If imprecision is represented via sets of candidates, that set of candidates should at all times contain the true state ω_0. Thus the empty set is never a valid result. Its appearance as the result of reasoning processes therefore indicates contradictions in the processed information or inadequacies of the modeling.

		location		
		mountains	seaside	urban
activity	climbing	●		
	cycling	●	●	
	sailing		●	
	shopping			●
	sightseeing	●		●
	swimming		●	●

Table 2.1: Relation between location and leisure activities

To illustrate reasoning with relations consider Table 2.1. The table visualizes a body of generic knowledge linking recreation sites with available leisure activities. Each of the marked fields corresponds to a particular tuple in a relation $R \subseteq \mathrm{dom}(\mathrm{activity}) \times \mathrm{dom}(\mathrm{location})$ Suppose, we have not decided on a specific site yet, but have already made some choices for weekend activity; for instance, we would like to go sailing or cycling. That constraint is formalized as a subset $S = \{\mathrm{cycling}, \mathrm{sailing}\}$ of dom(activity). Applying S to R only three tuples remain and the set of choices for location is restricted to $S \circ R = \{\mathrm{seaside}, \mathrm{mountains}\}$

2.3.2 Ordered Facts and Epistemic Entrenchment

With a (binary) symbolic or relational representation, all accepted sentences are treated equally, regardless of the evidence that lead to their acceptance in the first place. This view is founded on the assumption that all included statements are definite and certain. On the modeling level the reliance on the correctness of accepted statements is reflected in the application of truth-preserving inferences.

In practice, however, we must concede that measurements and observations contributing to the evidence are not absolutely reliable. Moreover, even originally correct observations are prone to become outdated in a dynamic world. Hence one must recognize the possibility that the set of assertions held for true becomes inconsistent. The resolution of inconsistency requires operations, such as revision and updating, which can be used withdraw or modify in certain pieces of information, that are deemed as conflicting with the new evidence.

Unfortunately, the existence of a unique solution is not guaranteed. The reason is, that a conflict can arise from a combination of pieces of information, which are each individually compatible with the new evidence, whereas their combination is not. Depending on which of these pieces is retained alternative resolutions of the conflict are obtained. A policy of removing all pieces of information that contribute to a conflict with newly acquired evidence is usually too rigorous. One of the solutions proposed for belief modeling is to introduce a selection function that chooses among the maximal subsets of facts that are still consistent (Alchourrón et al., 1985). However, since the choice of that set would not be based on the available knowledge itself, any specific selection involves a subjective element.

To address these difficulties, Alchourrón and Makinson (1985) suggested to enrich the representation of belief bases by supplying a strict partial order over its elements. That way, a precedence hierarchy for the conflicting sentences is established. This ordering in turn determines how conflicts are resolved. However the approach merely shifts the problem of conflict resolution to the initial specification of the priority of the facts. Other, approaches such as *epistemic entrenchment* (Groove, 1988; Spohn, 1988) or relational belief revision (Lindström and Rabinowicz, 1990) provide interpretable theories for forming rational believes based on dynamic relational knowledge (Spohn, 1990). The underlying idea is that accepted propositions are assigned values from an ordinal scale, which reflect an agents reluctance to abandon those propositions in the light of new evidence. Agents receiving information that is in conflict with a subset of their current beliefs, compare the priority rank assigned to that new information to the level of entrenchment of their current beliefs. Based on this comparison it is decided, which pieces of knowledge are kept and which are discarded to form a new, internally consistent knowledge state. The calculus used for epistemic entrenchment is closely related to the qualitative interpretations of possibility theory (Dubois and Prade, 1988a, 2004), which in turn refines concepts previously presented in the context of modal logic (Lewis and Langford, 1932). Mapping ranks to levels of necessity has even allowed to establish a direct formal correspondence between those two frameworks (Benferhat et al., 1994; Dubois and Prade, 1997; Gérard et al., 2007).

2.3.3 Probability-Based Representations

In Section 2.1 I argued that knowledge is inherently *uncertain*, i.e., there is a risk that a statement does not correctly reflect the actual situation. Uncertainty may be attached to both precise and imprecise statements. The logic-based approaches listed previously relied on the comfortable assumption that for most practical purposes we can treat "sufficiently certain" statements as if they were true. But how should we model knowledge, if this uncertainty is more pronounced?

The extensions introduced in Subsection 2.3.2 addressed some of the limitations of two-valued logic, but do not provide a quantitative approach to modeling uncertainty yet. However, decisions often have to be made on the basis of information that is not fully reliable. In that case a statement's degree of (un)certainty itself becomes a focus of interest. This challenge motivates the introduction of methods for formal reasoning under uncertainty.

Among the mathematical frameworks employed to represent uncertainty, *probability theory* (see, e.g. Feller, 1968) as defined by Kolmogorov's axiomatization (Kolmogorov, 1933) is arguably the most prominent one. For single-valued attributes, the frame of discernment Ω has the properties of a sample space. To reflect the uncertainty w.r.t. the identity of the true state ω_0, the elements $\omega \in \Omega$ are viewed as potential *outcomes* of a random experiment, that is as mutually exclusive, indecomposable *elementary events*. Collected knowledge about the investigated situation is then expressed in terms of a probability measure P, which quantifies the uncertainty about the realization of possible events. In practice the measure is usually encoded via a probability mass (discrete) or probability density (continuous) function

$$\begin{aligned} p \;:\; \Omega &\to [0,1], \\ \omega &\mapsto P(\{\omega\}), \quad \forall \omega \in \Omega \end{aligned}$$

over the frame of discernment. The term *probability distribution* is used generically to refer to the encoded information regardless of its particular formal representation. In the knowledge representation literature, the term "probability distribution" is frequently used to refer specifically to "probability density function" or "probability mass function". In some instances, however, the term has even been applied to probability measures (possibly as a contraction of "cumulative probability distribution function")[4].

[4] This rather confusing practice is discouraged by the author of this thesis.

An additional application of the above formal framework is to model relative frequencies and ratios, i.e. proportions. Proportions are closely related to probabilities, but may be distinguished from them on a semantic level. Proportion are relative frequencies in collections of objects and thus quantify emergent properties of such collections. The explicit representation of proportions accounts for a wide range of applications of the probabilistic framework, for instance in logistics, manufacturing control or other planning and analysis tasks. Additionally, proportions remain important for providing estimates of probabilities on an empirical basis. The idea to define probability directly as the limit of relative frequency for increasingly larger samples, however, had to be discarded as self-referential, because that definition itself is based on convergence in probability (von Mises, 1957).

To illustrate how proportions are linked to probabilities consider the task of modeling the behavior of a six-sided die. The faces of the die are labeled with numbers from 1 to 6 corresponding to 6 elementary outcomes when throwing the die. Suppose we do not want to make any assumptions about the fairness of the die but would still like to model the probability for each outcome. In that case knowing the probability of each of the six elemental events suffices to compute the probability for all other events (such as throwing a number greater than 4). To estimate that probability distribution we could throw the die repeatedly, e.g. $N = 1000$ times, and count for each outcome $\omega \in \Omega$ the respective number N_ω of occurrences in the sequence.

The relative frequencies $\frac{N_\omega}{N}$ with which a given outcome is observed in a series of repeated experiments may be used as an estimate for the probability of the respective realization. While there is no actual guarantee that this procedure will succeed, the probability to obtain an estimate within any given epsilon environment of the true value may be brought arbitrarily close to unity with arbitrarily high probability by choosing a sufficiently large number of trials for the experiment, i.e.

$$\lim_{N\to\infty} P\left(|\frac{N_\omega}{N} - p(\omega)| \geq \epsilon\right) = 0.$$

The above statement is known as Bernoulli's theorem or weak law of large numbers.

Because the relative frequency in a sample is a consistent and unbiased estimator for the probability of an event, it is frequently used in machine learning to fit parameters of probabilistic models. One drawback of this procedure is that the finite size of the sample necessarily precludes guarantees for the coverage of rare events in the training phase. As a result parameters that represent small positive probabilities would often be estimated with a value of 0. If these parameters are subsequently used as factors in computing probability values for more complex events, they force the result to extreme values regardless of the values of other parameters. This behavior is problematic for a number of applications. For instance, when evaluating model fit via the likelihood of test cases (page 135), the zero values produce to undefined or insufficiently discriminating measures.

The above problem can be countered by applying the so-called Laplace correction. The Laplace correction is a small value that is added to the counts of each category when estimating probabilities based on observed relative frequencies[5]. Thus the correction is added to the nominator and once for each of the disjoint categories in the denominator of the fraction. This results in estimates of the form

$$\hat{p}(\omega) = \frac{N_\omega + \mathrm{lcorr}}{\sum\limits_{\omega' \in \Omega} (N_{\omega'} + \mathrm{lcorr})} = \frac{N_\omega + \mathrm{lcorr}}{N + |\Omega| \cdot \mathrm{lcorr}}. \tag{2.4}$$

While heuristic in nature, the modification prevents the parameter estimates from adopting extreme values – namely by introducing a small tendency towards the uniform distribution. For sufficiently large training sets the bias introduced due to the Laplace correction is negligible and the mathematical properties of the model become better suited to the subsequent applications.

If the data generating process is well understood an alternative to the estimation of probability distributions from data may be applied. For simple experiments like throwing dice or drawing cards symmetry arguments may be used to define sets of equiprobable elementary events. For instance, permuting the labels on the faces of a fair die does not change the expected outcomes.

Because the pioneers in probability theory have made extensive use of such assumptions in their the study of games of chance, probability assignments derived from such symmetry arguments are commonly referred to as “classical” probabilities.

Even in the stricter sense of the word, probability is still used in at least two meanings. Depending on the interpretation, postulates chosen to deal with incomplete information may differ in applications. To distinguish these interpretations from each other, they are often called “subjective” and “objective” probabilities. The so called “subjective view”, centers around modeling an agents opinions and beliefs i.e., a mental state. Subjectivists subscribe to the view that the probability assigned to an event merely reflects a subjective assessment of the chance that the event occurs. Subjective probabilities are traditionally associated with the study of betting behavior (de Finetti, 1974, 1975).

In contrast to that, the interpretation called “objective probability” focuses on the description of a data-generating process independent of the observers dispo-

[5] Common choices for the value of the Laplace correction are lcorr = 0.5 and lcorr = 1.

sitions or preferences. They are thus frequently employed for modeling in science and engineering. Theoretical insights about such processes sometimes allow to re-construct a distribution from very few parameters or bypass the estimation step altogether (e.g. emission spectra). Objective probabilities also assume a pivotal role in modern theories of physics (cf. Feynman et al., 1965).

The question of whether probabilities encountered in applications are subjective or objective ones, has been the cause of much confusion. The debate is fed by ambiguity of the terms and unfortunately often interwoven with epistemological issues[6]. To avoid misunderstandings I will use the terms subjective and objective only to denote whether the model attaches probabilities to opinions of observers or to the data-generating process without regard to how their values are determined when applying the model. The advantage of this distinction is that by emphasizing the model aspect it separates the formal representation of a problem from the purely epistemological question of whether or not the exact value of objective probabilities is actually accessible to an external observer.

In the probabilistic approaches, generic information is represented by a probability distribution. Such a distribution, quantifies the chance that a randomly drawn object from a studied population has a given combination of properties. If one is only interested in a specific subset of attributes then the possible value combinations may be grouped by those attributes. Adding the probability of the elements in each group yields so called marginal distributions which refer to the selected attributes only.

Reasoning takes place when a piece of information, allows to associate the object of interest with a specific subgroup, corresponding to the observed values of the measured attributes. In that case the more specific distribution associated with that subgroup provides a more specific description of the situation. It is obtained by conditioning the generic distribution with values of the observed attribute and adapting the distribution for the remaining ones accordingly. Thus conditioning reflects a change of reference class. Another form of conditioning occurs when the individual instantiation is replaced with a marginal distribution. In that case the result reflects a weighted mixture of the distributions that would have been obtained by instantiating with each of the individual values. Together, conditioning and marginalization provide the standard tools for reasoning in the probabilistic framework.

Example 2.2. A manufacturer has set up three production lines to fabricate machine parts. Once finished, the products are tested to separate out defective parts (D) and sort the remaining ones into two qualities. In this case one

[6] Whereas *epistemic* is used to refer to knowledge and its properties in general, epistemology deals with the origin, nature and limits of human knowledge.

would use a probability distribution to model properties of the production process (“objective” interpretation). In practice such distributions are estimated from empirical probabilities, i.e. observed proportions in samples.

To translate this description into a formal notation, we introduce two attributes “production line” and “quality” with respective domains dom(prod. line) $= \{L_1, L_2, L_3\}$ and dom(quality) $= \{Q_1, Q_2, D\}$. Table 2.2 specifies the joint distribution of part properties with respect to both attributes.

		prod. line			
		L_1	L_2	L_3	$\sum$
quality	Q_1	0.08	0.08	0.04	0.2
	Q_2	0.11	0.26	0.32	0.69
	D	0.01	0.06	0.04	0.11
	$\sum$	0.2	0.4	0.4	1

Table 2.2: Production summary for Example 2.2

If one is interested only in product quality the respective marginal distribution given in rightmost column supplies the desired information. Since we are dealing with a discrete distribution, the respective values are computed as sums:

$$\forall q \in \{Q_1, Q_2, D\} :$$
$$P(\text{quality}\,(x) = q) = \sum_{p \in \text{dom}(\text{prod. line})} P\,(\text{quality}(x) = q,\ \text{prod. line}(x) = p)\,.$$

In general, marginal distributions over a subset $Y \subset X$ of attributes are computed from joint distributions over an attribute set $X = \{A_1 \ldots A_n\}$ by “summing out” the variables $X \setminus Y$, i.e.:

$$P\left(\bigwedge_{A_i \in Y} A_i = a_i\right) = \sum_{A_j \in X \setminus Y} \sum_{a_j \in \text{dom}(A_j)} P\left(\bigwedge_{A_k \in X} A_k = a_k\right). \tag{2.5}$$

The shortcut notation $\bigwedge$ shall indicate a conjunction of conditions. It is used here to refer to the simultaneous instantiation of the attributes in a set. Equation 2.5 states, that marginal probabilities are computed by collecting the probability of all compatible instantiations w.r.t in the higher dimensional space. This is achieved by taking the sum over all values of the attributes in $X \setminus Y$, that is, by ignoring any distinctions w.r.t. the attributes outside Y.

In applications one is frequently interested in those cases that realize a specific instantiation of one or more attributes. For instance, to calculate the expected ratio of defective parts in the output of production line L_2 one must compute the conditional probability

$$\begin{aligned}&P\left(\text{quality}(x) = D \mid \text{prod. line}(x) = L_2\right)\\ &\quad = \frac{P\left(\text{quality}(x) = D,\ \text{prod. line}(x) = L_2\right)}{P\left(\text{prod. line}(x) = L_2\right)} = \frac{0.06}{0.4} = 0.15.\end{aligned}$$

□

Note that with the classical definition, conditional probabilities are defined only, if the conditioning event has positive probability[7]. Because conditional distributions are well-suited to compare groups, they are frequently employed to visualize data or summarize it for reports. Moreover, robust generic knowledge about causal dependencies may be expressed using conditional probability. For instance, a statement like “Approximately 1% of the patients experience slight headaches when treated with drug X.” retains its meaning regardless of current prescription levels of that drug or variations in the prevalence of headaches in the general population due to other reasons.

Example 2.3. Recall the situation presented in Example 2.2. This time, it is described by means of a marginal distribution $P(\text{prod. line})$ and a supplementary conditional distribution $P(\text{quality} \mid \text{prod. line})$ (Table 2.3).

From that representation the joint distribution is recovered using the chain rule of probability

$$\begin{aligned}&P(\text{prod. line}(x) = p,\ \text{quality}(x) = q)\\ &\quad = P(\text{quality}(x) = q \mid \text{prod. line}(x) = p) \cdot P(\text{prod. line}(x) = p).\end{aligned}$$

Finally, suppose one is interested in the probability that a product of given quality is produced on a specific production line. The probability that a given defect product originates from line L_2 is computed using Bayes’ rule:

$$\begin{aligned}&P(\text{prod. line}(x) = L_2 \mid \text{quality}(x) = D)\\ &\quad = \frac{P(\text{quality}(x) = D \mid \text{prod. line}(x) = L_2) \cdot P(\text{prod. line}(x) = L_2)}{P(\text{quality}(x) = D)}\\ &\quad = \frac{0.15 \cdot 0.4}{0.11} \approx 0.55.\end{aligned}$$

[7] An alternative proposal, without the restriction to conditioning events with positive probability is presented by Coletti and Scozzafava (2005). Their approach introduces the notion of “coherent conditional probabilities”, and views conditional probability as a fundamental rather than a derived concept.

prod. line		
L_1	L_2	L_3
0.2	0.4	0.4

$P(\text{quality} \mid \text{prod. line})$	quality		
	Q_1	Q_2	D
L_1 :	0.4	0.55	0.05
L_2 :	0.2	0.65	0.15
L_3 :	0.1	0.8	0.1

Table 2.3: Factorized representation using conditional distribution

The value of the denominator is not given directly but can be computed from the provided information:

$$\begin{aligned} & P(\text{quality}(x) = D \\ &= \sum_{p \in \text{dom}(\text{prod. line})} P(\text{quality}(x) = D \mid \text{prod. line}(x) = p) \cdot P(\text{prod. line}(x) = p) \\ &= 0.05 \cdot 0.2 + 0.15 \cdot 0.4 + 0.1 \cdot 0. \quad = \quad 0.11 \end{aligned}$$

□

2.4 Conclusions

The discussion of the predominant formalisms of knowledge representation in Section 2.3 points out their respective properties and their supported range of interpretations. With respect to suitability for applications the advantages of the individual approaches can be summarized as follows:

- Symbolic and relational approaches are adequate for representing set-based concepts such as relations, imprecision and multi-label descriptions (Subsection 2.3.1). They do not support quantitative reasoning though.
- Graded knowledge representation schemes define several discrete levels of evidence an an ordinal scale. They support the representation of epistemic preferences and provide improved capabilities concerning dynamic knowledge and partially inconsistent information (Subsection 2.3.2).
- Probabilistic approaches are used to construct quantitative models of uncertainty or to represent statistical properties of larger collections (Subsection 2.3.3). They use a continuous scale.

These representations already cover a number of practically relevant tasks. For instance, relational models are successfully applied in databases, whereas probabilistic approaches have been shown to provide a powerful tool for solving prediction, classification and planning tasks.

However, a number of applications that are suited for neither of the probabilistic, relational or ordinal knowledge representation frameworks have come to attention. For tasks such as reasoning with vague data and, more recently, the analysis of gene expression and protein level data, statistical information needs to be combined with set-based concepts. Similar difficulties are encountered in connection with ambiguity in natural language processing and the application of statistical methods to multiply-labeled data.

To address that challenge I propose an efficient and scalable knowledge representation for properties of set distributions suitable for these tasks. To achieve that goal I will first discuss means to efficiently represent multivariate distributions (Chapter 3). Following that I compare and analyze existing models that relate to distributions over sets. The insights gained from this comparison are then used to compile desired properties and define requirements for the new model.

3 Graphical Models

The state of any complex system is characterized by a number of variables. To understand and model the behavior of the system the interactions between these variables need to be investigated. Interactions between variables can be represented qualitatively, via relations, or an a quantitative level, e.g. using functions or probability distributions over the frame of discernment formed by the model's variables. For a given set of attributes, that frame of discernment is the product space of the individual attribute domains, that is the set of all *combinations* of values of the individual attributes. Hence, the number of elements in that space equals the product of the cardinality of all attribute domains (see Subsection 2.2.1). Unfortunately, the cardinality of the resulting sample space makes the straightforward approach of building knowledge representations based on relations and distributions over the frame of discernment itself impractical for even moderately sized variable sets.

Graphical Models are powerful tools to represent the interaction of variables within complex systems. They provide compact descriptions of large distributions, and although the structure of a Graphical Model can be learned directly from data, they allow integrating prior knowledge about variable dependencies via structure constraints. This permits to reconstruct likely system states from a comparatively limited number of observations. Finally, Graphical Models provide efficient reasoning operations that allow to simulate how changes in one part of the system affect other elements.

3.1 Introduction and Principles

The defining characteristic of *Graphical Models* is their approach to dealing with the large product spaces in multivariate settings. The central idea consists in finding *decompositions* of high-dimensional distributions into sets of overlapping distributions on lower-dimensional subspaces. Since the cardinality of each individual subspace is considerably lower than the cardinality of the high-dimensional space (see page 16), this may be used to efficiently represent high-dimensional distributions via their decompositions and the respective distributions on the subspaces.

Although Graphical Models have been applied used predominantly in connection with strictly positive probability distributions (Hammersley and Clifford, 1971; Lauritzen and Spiegelhalter, 1988; Pearl, 1988), the decomposition approach has been generalized to a wider class of distribution types (Pearl and Paz, 1985; Studený, 1993; Jiroušek and Vejnarová, 2003)[1].

In de Campos and Huete (1993) the authors have used a calculus for probability intervals to model imprecision in subjective probability assessments. The goal of enriching the knowledge representation in Graphical Models with a notion of imprecision also motivated the development of possibilistic Graphical Models (Dubois and Prade, 1990; Gebhardt, 1997; Borgelt et al., 2000; Vejnarová, 2003; Borgelt and Kruse, 2003). Interpretations and applications of the possibilistic framework are extensively covered by Section 4.4. Since the possibilistic framework may be viewed as an extension of the relational one (compare Borgelt and Kruse, 2002), Graphical Models are also closely connected to database theory. Stated in the language of Graphical Models the database schema describes a decomposition of a high-dimensional relation into lower-dimensional ones. Each table of the relational database constitutes one such lower-dimensional relation, and links between tables are established via the shared key-attributes. It is this common decompositions principle that underlies the compact representation and efficient reasoning methods both and other types of Graphical models. In all cases, the graphical representation is based on analogues of the notion of independence and conditional independence properties for the respective frameworks. Such properties have been studied, e.g. in Hisdal (1978); Studený (1993); Dubois et al. (1994); de Cooman (1997c); Cozman and Walley (2005).

The term Graphical Model itself is derived from the common representation of the decomposition using a graph. A graph as used throughout this dissertation is a pair a $G = (V, E)$ formed from a finite set of nodes V and a set of edges $E \subseteq V \times V$. An edge $(A, B) \in E$ is called undirected, if both (A, B) and (B, A) are in E. Conversely, the edge $(A, B) \in E$ is called directed, $(B, A) \notin E$. In extension, we speak of an *undirected graph* if that graph has only undirected edges and of a *directed graph* if all its edges are directed. Moreover, in context of Graphical Models we are concerned with simple graphs, that is graphs that neither contain loops nor multiple edges. For better distinction I will adopt the notations $G = (V, E)$ and $\vec{G} = (V, \vec{E})$ for undirected and directed graphs respectively (Borgelt and Kruse, 2002, cf.).

In the case of Graphical Models, the set of nodes in the underlying graph is a set of attributes forming a global frame of discernment. The organization of the

[1] As pointed out by Pearl and Paz exact decomposition is based on a trinary relation that satisfies the so called semi-graphoid axioms. For probabilistic decomposition the conditional independence relation is used in this capacity (cf. Section 3.2).

nodes in the graph encodes *independence statements* regarding the respective attributes. Conversely, the set of edges determines paths along which evidence is propagated. The decomposition approach is used, not only for the compact representation of relations and distributions, but also to simplify reasoning. Inferences are drawn using operations on the local distributions only, that is marginal distributions and relations may be computed without actually having to compute the respective joint distribution and relations on the full sample space. Due to their application in reasoning Graphical Models are also known as *Inference Networks* or *Causal Networks*.

An example of such a network – more specifically a Bayesian Network – for reasoning with regard to a hypothetical diagnostic problem in a lung clinic was proposed by Lauritzen and Spiegelhalter (1988). The model is aimed at representing generic knowledge about the interaction between a number of Boolean attributes, linking results of medical tests, risk factors and possible diseases. Information about dependency/independence of attributes is encoded using a directed acyclic graph (Figure 3.1) and supplemented with conditional probability distribution for each of attributes given the set of their parent attributes in the graph. Reasoning with probabilistic graphical models requires the network to be initialized with an prior distribution. This for prior could, for instance, be chosen to reflect the distribution for the variables for the general patient population of the area. Additional knowledge about the a specific case at hand then allows to instantiate variables in the network. This evidence may then be used to update the distribution of other variables resulting in a better assessment of the particular case.

Although all Graphical Models are based on the same fundamental concepts, a large number of variants have been proposed over the time. These variants mainly differ in the the choice of the knowledge representation framework employed, in the representation of local distributions and the way, in which independence statements are expressed. Additionally, implementations may differ w.r.t. the choice of implemented evidence propagation methods. In the literature separate names are sometimes used to refer to subclasses of Graphical Models, which are defined by via constraints to the network architecture (e.g. tree structured networks or chain graphs (Buntine, 1995; Studený, 1996)). The additional constraints usually allow to use modified propagation methods with increased efficiency on that particular type of structure, but come at the cost of a restricted application range. Since the aim of this section is to provide an introduction into the common underlying concepts that are relate to this thesis, these variants will not be covered in detail. For a more extensive study of these variants of Graphical Models and their associated propagation methods, the reader is referred to the specialized literature, e.g. Lauritzen and Spiegelhalter (1988); Pearl (1988); Whittaker (1990); Jensen (1996); Lauritzen (1996); Borgelt and Kruse (2002).

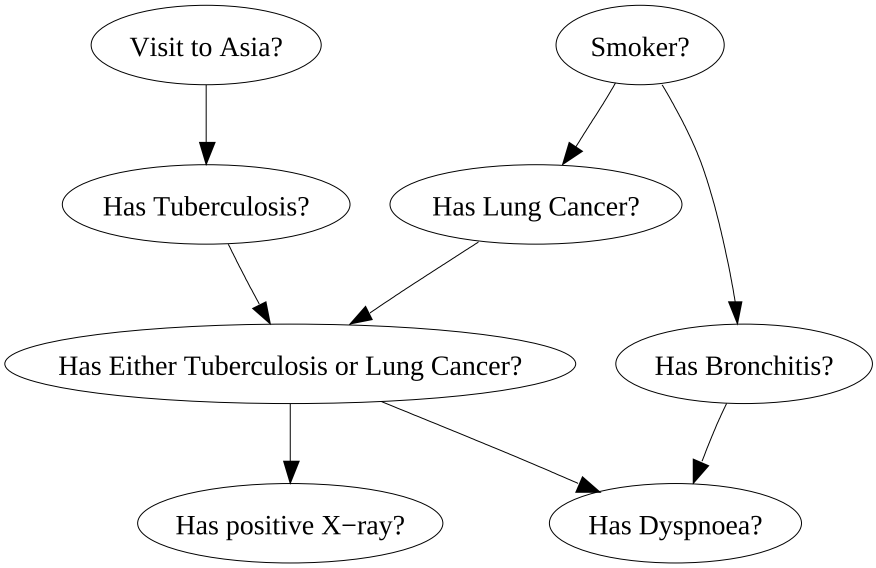

Figure 3.1: Structural component of a Bayesian Network Model for the chest clinic diagnostic problem (example initially proposed by Lauritzen and Spiegelhalter, 1988).

3.2 Statistical Independence and Decomposition

Decomposition in Graphical Models is based on marginal or conditional independence statements between attributes or sets of attributes. Put in simple words, two attributes A and B are independent, if obtaining knowledge about the instantiation of any one of them does not supply additional information about the instantiation of the other one. The exact formulation of the independence condition depends on the particular knowledge representation framework at hand. For the probabilistic framework, decomposition is based on the notion of statistical independence. In order for two attributes or variables A and B to be statistically independent their joint distribution must fulfill the condition:

$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : P(A = a, B = b) = P(A = a) \cdot P(B = b). \quad (3.1)$$

The above formulation of the independence condition immediately provides a mechanism to reconstruct the joint distribution of the independent variables A, and B from their respective marginal distributions. Also, Equation 3.1 shows, that pairwise independence is a symmetric property (exchanging A and B results in an equivalent formula).

Assuming the conditioning event has positive probability, instantiating either of the variables in the joint distributions does not affect the a-posteriori marginal distribution of the other one due to

$$\begin{aligned} P(A=a \mid B=b) &= \tfrac{P(A=a,B=b)}{P(B=b)} = \tfrac{P(A=a)\cdot P(B=b)}{P(B=b)} = P(A=a) \quad \text{and} \\ P(B=b \mid A=a) &= \tfrac{P(A=a,B=b)}{P(A=a)} = \tfrac{P(A=a)\cdot P(B=b)}{P(A=a)} = P(B=b) \quad \text{respectively.} \end{aligned}$$

In their short form $P(A=a \mid B=b) = P(A=a)$ and $P(B=b \mid A=a) = P(B=b)$ the above equations are frequently used as alternative formulations of the independence criterion. That formulation is mainly applied in connection with conditional distributions.

When three or more attributes are considered, more subtle independence statements may be formulated. For instance, a pair of attributes may be statistically dependent when considered alone, yet independent for each fixed value of a third attribute. Such interactions are captured in the notion of conditional independence:

Definition 3.1. *Let X be a set of attributes, $A, B, C \in X$ and p a probability distribution defined over X. With respect to p the attributes A and B are called conditionally independent given C (written $A \perp\!\!\!\perp B \mid C$) iff:*

$$\begin{aligned} &\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) : \\ &\qquad P(A=a, B=b \mid C=c) = P(A=a \mid C=c) \cdot P(B=b \mid C=c) \end{aligned}$$

Like with its unconditional counterpart, the attributes A and B may be exchanged in the equation without changing the meaning, so conditional independence is symmetric. The alternative formulations are given by the equations

$$\begin{aligned} &\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall \in \mathrm{dom}(C) : \\ &\qquad P(A=a \mid B=b, C=c) = P(A=a \mid C=c) \quad \text{and} \\ &\qquad P(B=b \mid A=a, C=c) = P(B=b \mid C=c). \end{aligned}$$

The notion of conditional independence may also be generalized from single attributes to sets of attributes. Any joint instantiation of an attribute set may be viewed as an instantiation of a composite attribute(compare page 16) to which the independence condition may be applied. Drawing on this correspondence, the same notation is used for expressing independence statements w.r.t. both individual attributes and sets of attributes. For attribute sets that only contain a single element, only that attribute will usually be written. Although the term "conditional independence" is normally used only in conjunction with conditioning attributes, the notion also covers the case of marginal independence with marginal dependency/independence statements being expressed using an empty set of conditioning attributes.

So far conditional independence statements have been described as a property of attributes w.r.t. a given distribution. A more intuitive understanding may be obtained by considering typical explanations for conditional independence relations in data:

- Common cause: The value of a single attribute influences the values of two or more otherwise unrelated attributes. But the two dependent attributes are conditionally independent from each other given the common cause.
- Chain structure: An attribute is indirectly affected due to its statistical dependency from the values of another attribute, which in turn is linked to a third one. Yet, for any fixed value of the mediating attribute the first and third attribute are independent.

A textbook example of a common cause relation is an (alleged) statistical dependency observed between ice-cream sales and bathing accidents. The value of both variables plausibly depends on an additional factor, namely the daytime temperature. However, the values of the two attributes are not significantly related if the data is presorted into groups by ranges of outside temperature. A similar relation exist in Lauritzen and Spiegelhalter's chest clinic example (Figure 3.1) between the result of an X-ray test and dyspnoea (shortness of breath) since both may indicate lung cancer or tuberculosis. Moreover, the indirect connection between traveling to Asia and positive X-ray diagnosis provides an example of a chain structure. A recent a visit to Asia is associated with an increased tuberculosis risk, and in the case of an infection it is more likely that an X-ray scan will result in the diagnosis of abnormal lung tissue in the respective patient. Whereas causal explanations are readily understood by humans and often used by domain experts to specify a model structure, the structural component of a Graphical Model merely encodes statistical dependencies and independence relations. The results of an experiment conducted by Borgelt and Kruse with several algorithms for learning the structure of Graphical Models (Chow and Liu, 1968; Cooper and Herskovits, 1992; Chickering et al., 1995, 1997; Borgelt and Kruse, 2002), demonstrate that suitable decompositions need not reflect true causal relations.

In general, a joint distribution fully reflects the statistical interactions within sets of attributes. The presence of (conditional) independence properties allows to write such distributions in terms of their decomposition into local distributions. Ideally, a decomposition expressed by the graph and the associated local distributions would allow to recover the original distribution exactly, though in practice, probabilistic Graphical Models will often represent only close approximations to the original distribution. This discrepancy is tolerated for two reasons: Firstly, applications often benefit from disregarding weak statistical interactions in favor of more efficient decompositions with smaller local distributions. Secondly, because models are trained from finite data sets, sampling error is likely to result in spurious statistical dependencies. Omitting weaker observed dependencies from the model is a strategy to avoid overfitting.

3.3 Node Separation in Graphs

At the beginning of the section it was stated that the graphical representation of decompositions is based on encoding independence statements. To that end, independence statements are reflected via node separation in graphs. In order to explain that idea, it is necessary to make clear the criteria for by which separation in graphs is defined. These criteria differ for directed and undirected graphs (after Borgelt and Kruse, 2002):

Definition 3.2. *Consider an undirected graph $G = (V, E)$ and three disjoint subsets X, Y and Z of the node set V. Z* **u-separates** *X from Y (written $(X \mid Z \mid Y)_G$) if every path from a node in X to a node in Y also contains a node in Z. A path that contains a node in Z is called* **blocked** *by Z. Otherwise it is called* **active**.

For directed graphs the so called d-separation criterion in applied (Pearl, 1988; Pearl and Paz, 1985; Verma and Pearl, 1992):

Definition 3.3. *Consider a directed graph $\vec{G} = (V, \vec{E})$ and the disjoint subsets X, Y and Z of the node set V. A path is called active iff:*

1. *Every node with converging edges (non-terminal node that receives a directed edge from both of its neighbors in the path) is either in Z or has a descendant in Z.*

2. *No other node on the path is in Z.*

In this context, a path is understood as any sequence of nodes connected by edges regardless of the edges direction, that is any sequence of nodes $(V_1, \dots V_n)$, $V_i \in V$, $i = 1, \dots, n$, such that $\forall i : 1 \leq i < n : (V_i, V_{i+1}) \in \vec{E}$ or $(V_{i+1}, V_i) \in \vec{E}$. The set Z **d-separates** *X from Y (written $(X \mid Z \mid Y)_{\vec{G}}$) if all paths from a node in X to a node in Y are blocked by Z, i.e. there is no active path from a node in X to a node in Y.*

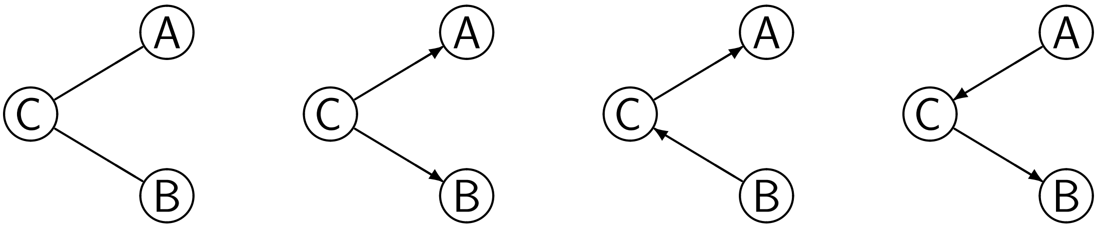

Figure 3.2: Graph representations, which encode the statements $A \perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp B \mid \emptyset$ (from left to right: using an undirected graph, using directed graphs with diverging edges in C or directed paths from A to B and from B to A respectively).

Figure 3.2 shows a simple example for the representation of independence statements with graphs. Applying the d-separation criterion to directed graphs there are several possibilities for the direction of the edges of an active path from A to B. However, a structure with converging edges in C would represent a different set of independence relations since the path from A to B would become active only if C is given (corresponding to the statements $A \perp\!\!\!\perp B \mid \emptyset$ and $A \not\perp\!\!\!\perp B \mid C$) in that case.

When an undirected or a directed graph is used as the structural component of a Graphical Model, the respective separation criterion encodes (conditional) independence statements that hold for the represented distribution. For two disjoint attribute sets that are separated by a third attribute set Z in the graph, their respective elements are conditionally independent given Z. For $Z = \emptyset$ the elements are called marginally independent. This connection between separation in graphs and conditional independence statements leads to the notion of a conditional independence graph (Pearl, 1988):

Definition 3.4. *Let $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ be a tree-place relation representing the set of conditional independence statements that hold in a given distribution δ over a set of attributes V. An undirected graph G is called a* **conditional independence graph** *or* **independence map** *w.r.t. δ iff for all disjoint subsets $X, Y, Z \subseteq V$ of attributes*

$$(X \mid Z \mid Y)_G \implies X \perp\!\!\!\perp_\delta Y \mid Z.$$

An analog notion has been defined for directed graph and the d-separation criterion. If the implication also holds in the converse direction, i.e. $X \perp\!\!\!\perp_\delta Y \mid Z \implies (X \mid Z \mid Y)_G$, the graph is called a *perfect map*. Because neither graph representation is capable of capturing all possible conditional independence statements that may hold in distributions, such a perfect map does not always exist however (compare Borgelt and Kruse, 2002, ch. 4 for counterexamples).

3.4 Bayesian Networks

For probabilistic Bayesian Networks Pearl (1988), the joint distribution over the global frame of discernment is written as a product of marginal and conditional distributions. Any multivariate probability distribution can be factorized by applying the chain rule of probability (see page 31). By adding knowledge about conditional independence statements these products can then be further simplified. Typically independence statements are given via a directed acyclic graph $G(X, \vec{E})$, $X = \{A_1, \ldots, A_n\}$, where each attribute is independent of its non-descendants given its parents (d-separation criterion). The graph corresponds to a factorization into a set of conditional probability distributions (Castillo et al., 1997; Borgelt and Kruse, 2002):

$$\begin{aligned}&\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) : \\ &\qquad P_X\left(\bigwedge_{i=1}^{n} A_i = a_i\right) = \prod_{i=1}^{n} P\left(A_i = a_i \,\middle|\, \bigwedge_{A_j \in \mathrm{pred}(A_i)} A_j = a_j\right),\end{aligned} \tag{3.2}$$

where again $\bigwedge$ denotes the conjunction of the conditions regarding the instantiation of the attributes and $\mathrm{pred}(A_i) = \{A_j \in X \mid (A_j, A_i) \in E\}$ is the set of the attribute A_i's direct predecessors in the graph $\vec{G}$. This type of factorization is called *chain rule factorization*.

In addition to the graph structure, Bayesian Networks include specifications of the factor distributions that appear in the above formula. These distributions form the qualitative component of the Bayesian Network Model. For each of the attribute $A_i \in X$ that do not have predecessors in the graph, the marginal distribution w.r.t. that attribute appears as a factor distribution in Equation 3.2. For the remaining attributes the conditional distribution given their respective predecessors $\mathrm{pred}(A_i)$ in $\vec{G}$ is stored instead.

3.5 Markov Networks

In contrast to Bayesian Networks, the independence structure for Markov networks (Lauritzen and Spiegelhalter, 1988; Guyon, 1995) – also called Markov Random fields (Kindermann and Snell, 1980) is represented using an undirected graph. For the construction of Markov networks, preprocessing may include a triangulation step, in which edges are inserted into a dependency graph learned from data or given by experts to break circles of length ≥ 4. Triangulating cycles leads to graph structures that are better suited to reasoning using the hypertree propagation method (compare Section 3.6). Although a triangulated graph may fail to express some of the original independence statements, all dependencies are preserved, so the accuracy of the decomposition is not compromised by that procedure.

In a Markov Network the joint distribution for the considered attributes is expressed as a product of so called *factor potentials* assigned to the maximal cliques $\mathbf{C} = \{C_1, \ldots, C_k\}$ of G Lauritzen and Spiegelhalter (1988).

$$\begin{aligned} &\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) : \\ &P\left(\bigwedge_{i=1}^{n} A_i = a_i\right) = \prod_{C_j \in \mathbf{C}} \phi_{C_j}\left(\bigwedge_{A_i \in C_j} A_i = a_i\right) \end{aligned} \tag{3.3}$$

These factor potentials are closely related to the marginal clique distributions.. To illustrate the above statement consider the very simple undirected graph previously shown in Figure 3.2. The corresponding clique graph only contains the cliques $\{A, C\}$ and $\{B, C\}$, which are connected via the shared attribute set $\{C\}$ (Figure 3.3). In order to represent the quantitative aspect of the models the factor potentials may be distributed to the cliques in different ways. For instance, factor potentials can be computed from the marginal distributions w.r.t. the attributes in each clique.

Since separator variables are common to two or more cliques, the contribution of these separator sets is divided between the cliques sharing them. For the decomposition shown in Figure 3.3 this strategy would yield two factors for the marginal distributions w.r.t. $\{A, C\}$ and $\{B, C\}$ respectively, divided by the marginal distribution w.r.t. the separator set $\{C\}$:

$$\begin{aligned} &\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(c) : \\ &P(A = a, B = b, C = c) = \tfrac{P(A=a,C=c) \cdot P(B=b,C=c)}{P(C=c)}. \end{aligned}$$

Different ways to distribute the factor $\frac{1}{P(C=c)}$ correspond to different solutions for assigning the clique potentials (see Borgelt and Kruse (2002) for further

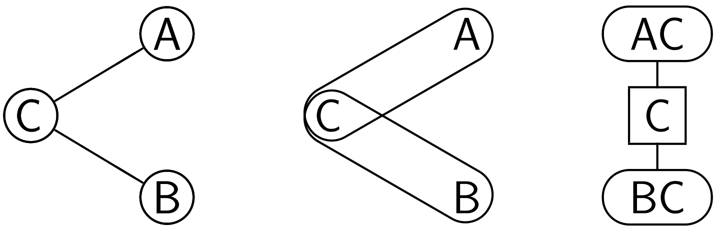

Figure 3.3: An undirected graph, its corresponding hypergraph representation and a join tree for propagation on the clique graph

examples). For an application of probabilistic Markov Models to a real-world planing problem see Gebhardt et al. (2004); Kruse et al. (2006); Steinbrecher et al. (2008).

3.6 Reasoning with Graphical Models

As stated in the introduction of this chapter, Graphical Models allow for efficient reasoning on large distributions. To elaborate that statement, is is appropriate to review evidence propagation methods for probabilistic Graphical Models. A list of popular propagation methods along with a list of references is given below:

- polytree propagation (Pearl, 1986, 1988)
- join tree propagation (also called clique tree propagation) (Lauritzen and Spiegelhalter, 1988; Castillo et al., 1997)
- iterative proportional fitting (Whittaker, 1990)
- bucket elimination (Dechter, 1996; Zhang and Poole, 1996)

The most well-known among these methods are polytree propagation (Pearl, 1986) and join tree propagation (Lauritzen and Spiegelhalter, 1988). Given the introductory character of this chapter I will not discuss the individual propagation algorithms in detail, but confine myself to present the underlying ideas. For proofs and implementation details the reader is referred to the referenced original publications.

As suggested by its name, the polytree propagation method (Pearl, 1986) presupposes a special structure of the given independence graphs. A directed graph $\vec{G}(V, \vec{E})$ is called a polytree, if for all pairs of two nodes from V there is at most one path connecting these nodes in the underlying undirected graph G'. The underlying undirected graph G' is obtained by replacing all directed edges in $\vec{G}$ by undirected ones. Since there is only one path between any two nodes in the graph, any fixed node $A \in V$ corresponds to one particular partitioning of V (Figure 3.4), with the individual partitions are defined as follows:

- the set of nodes consisting of A, the predecessors of A in $\vec{G}$ and nodes that are connected to A in G' via such a predecessor node (light shading in Figure 3.4),
- the set of nodes that are successors of A in $\vec{G}$ or are connected to A in G' via such a successor node (dark shading in Figure 3.4), and
- the set of nodes that are not connected to A in G'.

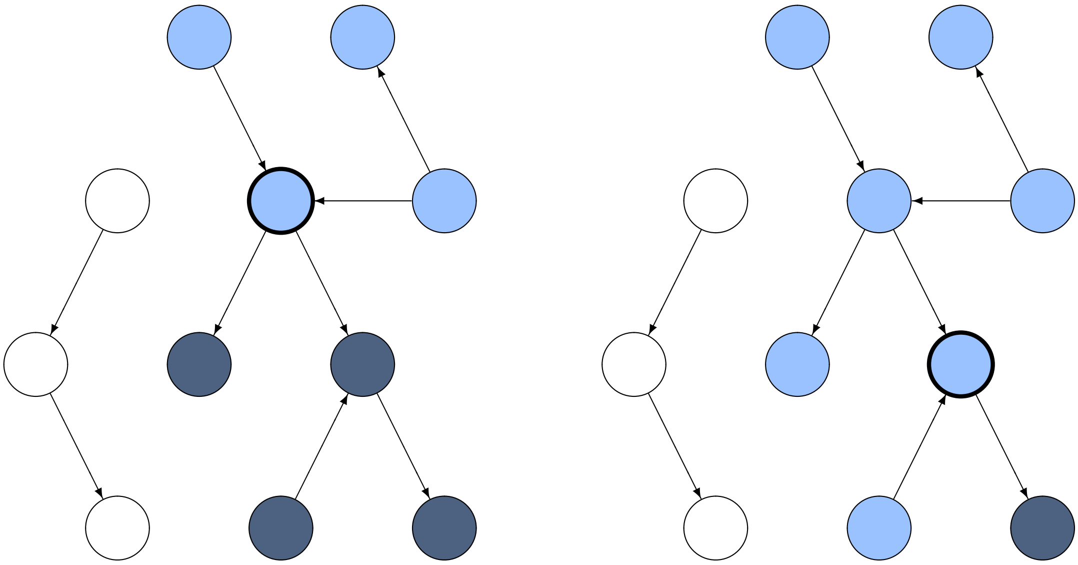

Figure 3.4: Partitionings of a polytree w.r.t. two different pivot nodes (strong contour); light shading: node contributing to the π-factor, dark shading: nodes contributing to the λ-factor, not shaded: nodes unconnected to pivot node).

Obviously, the pivot node A d-separates any pair of nodes from two different partitions (compare Definition 3.3 on page 41). Thus the respective variables are modeled as conditionally independent of each other given the pivot node variable.

This property may be used to to construct a local propagation algorithm for polytrees. Given the partitioning of V w.r.t. any particular attribute A, consider the decomposition of the joint distribution achieved by accumulating the conditional probability distributions for node instantiations given their respective parents for each of the partitions separately (cf. Equation 3.2).

Using Equation 2.5, the marginal distribution for the attribute A may be computed by summing out all other attributes from the joint distribution. Summing out the attributes in $V \setminus \{A\}$ from the above factorization yields a so-called π-*value* and the λ-*value*. The π-value represents the influence of the information reaching A via its parent attributes whereas the λ-value represents the information obtained via child-attributes (cf. Borgelt and Kruse, 2002, ch 4). For the variables not connected to A in G' a constant of 1 is obtained, regardless of their instantiation, reflecting their independence of A. If A has more than one parent in the graph, the π-value itself may be further decomposed into a products with each factor referring to the subgraph of G connected to A via one particular parent attribute. Similarly, the influence of the attributes "below" A may be split into one factor per child node. For static distributions, the product of the π-values for each non-root variable in the graph mirrors the univariate distribution of the values of that attribute, whereas λ-value(s) are one. However changing the distribution of variables in the respective set affects the distribution of A. The local propagation in the polytree algorithm is based on restoring that invariant when either value is altered.

Due to the polytree property, changing the distribution of a variable in either of the subsets contributing to the π- or λ-value may affect variables in the respective other subset only via the attribute A. Since this reasoning applies w.r.t. any attribute in the graph, information about by new evidence may be propagated through the graph by sending local update messages between neighboring nodes in the graph.

These messages are generated whenever the distribution of an attribute is altered due to the integration of new evidence. To that end messages in the form of factor distributions are sent from the corresponding node to its respective parent (λ-messages) and child nodes (π-messages)in the graph (Figure 3.6. For each receiving node the incoming messages are processed to generate new update messages that are then be passed on to the remaining neighbors of the receiving node.

If several variables are instantiated concurrently the message passing scheme is slightly altered reflecting the fact that instantiated variables d-separate certain subsets of nodes, thereby preventing the propagation of evidence between those sets. As a result instantiated variables do not generate messages for incoming λ-messages. The incoming π-messages, however, are still processed and trigger λ-messages to other parent attributes.

After all messages have been propagated, the now updated $\pi-$ and $\lambda-$factors associated with each node allow to compute the new distribution for all modeled variables.

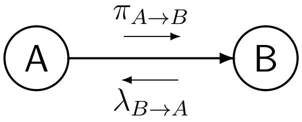

Figure 3.5: Message passing for the polytree propagation algorithm: π-messages travel from parent to child nodes, λ-messages from child to parent nodes.

While based on the same principles as the polytree propagation algorithm, the clique tree propagation approach associates the node processors not with single attributes, but with groups of closely interrelated attributes – namely the maximal cliques of an underlying independence graph. The quantitative part of the model is defined by multivariate distributions for these groups of attributes. Local distributions interact by means of so called separator sets, that is sets of attributes shared between two or more cliques. Conditioning a local distribution may also cause the distribution of a separator set to be changed. The modified distribution on the separator set is in turn used to condition the local distribution of other cliques covering that separator set. Since these cliques can in turn share variables with further cliques, this approach results in a mechanism for iteratively propagating evidence through the network. Whereas the original version of the algorithm proposed by Lauritzen and Spiegelhalter operates on cliques of the junction tree, the version implemented in the HUGIN Software package (Andersen et al., 1989) provides additional storage for explicitly representing the distributions for separator sets. To extend the algorithms to non-probabilistic settings Shenoy and Shafer (1986) have formulated a version of clique tree propagation that does not require an operation for division.

In analogy to the polytree condition, many clique graph propagation algorithms require that the clique graph associated with the underlying independence graph has tree structure (hypertree property) to ensure a unique path of evidence propagation. The reason is, that the presence of circular connections would allow information about the same evidence to travel on different paths leading to duplicate updates. For the clique tree approach that difficulty can be avoided by triangulating the underlying independence graph. Triangulation entails that the associated clique graph has tree structure and only one path between two given cliques exists (e.g. Andersen et al., 1989; Lauritzen, 1996). Additionally Castillo et al. have proposed a method for converting directed independence graphs into a Markov hypertree by triangulating their moral graph (Castillo et al., 1997). In order to facilitate propagation, this hypertree is then further transformed into a join tree, that is a graph denoting the cliques and paths for evidence propagation via separator sets. The transition from graphs based on single attributes to clique based approaches is an example of the so called node merging approach to resolve circular evidence propagation paths.

An alternative approach to dealing with circles consists in instantiating a selection of attributes in the graph, thus restricting evidence propagation via the respective nodes. The additional restrictions introduced by those instantiations give rise to a set of partial solutions. The unrestricted solution is then recovered from a weighted superposition of partial solutions for all possible combinations of attribute value assignments. Unfortunately, as the number of instantiations to be considered grows exponentially in the number of additionally instantiated attributes this approach is only applicable on a small scale.

Finally the variable elimination algorithm may be applied to replace the need for propagation altogether(Dechter, 1996; Zhang and Poole, 1996). To that end, Equation 2.5 is applied directly to the factor potentials to sum out individual variables. This elimination process is repeated for all variables that are neither observables $\mathbf{O}$ or query variables $\mathbf{Q}$ using heuristics to determine an efficient elimination order. During this process, variables that are no longer connected[2] to any of the variables in $\mathbf{O} \cup \mathbf{Q}$ may be removed without calculation, as they have no influence on the output.

The variable elimination algorithms effectively collapses the model in such a way, that only variables relevant to the given query are represented. Due to this initial reduction in model complexity the algorithm may be used to process queries to very large Graphical Models more efficiently than the propagation-based methods. Moreover, variable elimination permits to project a complex graphical model to a subspace. The resulting auxiliary model can then be used for processing large numbers of queries involving a limited subset of variables in an more efficient manner. This strategy is used in chapter 6 to compute marginal one point coverages from set-instantiations represented using Graphical models. Since the variable elimination approach processes queries one-by-one, it is less suited, however, to dealing with dynamic states of knowledge, where the value and instantiation state of nodes may change between queries time, or if the set of variables eventually to be instantiated is not known in advance.

3.7 Summary and Further Reading

Graphical Models allow for the decompositions of relations or distributions on high-dimensional spaces into overlapping relations or distributions on lower-dimensional ones (decomposition) and provide efficient reasoning techniques for large multivariate problems. By means of these decompositions they help to overcome limitations of knowledge representation frameworks that arise from the computational complexity of operations with large frames of discernment.

Although the algorithms outlined in this chapter are used in most practical applications of Graphical Models, several modifications and alternatives have been suggested to improve performance for specific applications. Buntine and Studený

[2] The notion of connected in this contexts pays no regard to edge direction, i.e. two nodes in a directed graphs $\vec{G}$ are considered connected, if there exists a path between those nodes in the *underlying undirected* graph. The underlying graph is obtained from $\vec{G}$ by replacing all directed edges with undirected ones.

have investigated properties and fast propagation methods for so-called chain graphs (Buntine, 1995; Studený, 1996) – a subclass of Graphical Models with constrained architecture. In contrast, Gaussian Graphical Models (Whittaker, 1990) heavily restrict the distribution types to be modeled and assume linear dependence of attributes, but in return draw on simplified structure learning methods previously described in the context of covariance selection (Dempster, 1972).

Causal Bayesian Networks (Heckerman, 1993; Pearl, 2000) add functionality to model the effects of external interventions on the modeled systems. The interventions artificially fix the value of variables decoupling it from its normal causal influences in the system. In contrast to the passive observation of statistical interactions between variables, which does not allow to distinguish between causes and effects, the effects of such interventions propagate according to in the direction of causality. For this reason Causal Bayesian Networks make use of two separate modes of propagation – one for reasoning about observations, and another one for inferring effects of interventions. More recently their capacity to model high-dimensional dependency structures and to incorporate information from several sources has motivated a number of applications of Graphical Models in Computational Linguistics (e.g. De Luca and Rügheimer, 2007) and Bioinformatics (e.g. Lauritzen and Sheehan, 2003; Friedman, 2004; Segal et al., 2005; Listgarten et al., 2007; Peña, 2008).

4 Probability and Set-Valued Data

The majority of commonly applied knowledge representations fall into one of two categories: those suited to deal with qualitative, relational pieces of information and those adapted for modeling quantitative aspects, such as value distributions. This division is sensible as long as the problem setting is dominated by either aspect of knowledge. But with the developments in data collection and experimental techniques scientific questions are now frequently addressed by a combination of data from multiple different sources. Within this setting the problem of applying probabilistic methods in conjunction with set-based concepts has become an important challenge, e.g. for investigating statistical properties over imprecise data, relations and candidate sets or modeling an agents subjective knowledge state. The following sections are concerned with approaches for representing statistical information about sets and set-interactions. After discussing a straightforward approach based on random sets (Section 4.2), I highlight difficulties related to the implementation and application of such a representation with large-scale data. Following that, two popular extant approaches to address these challenges – the Dempster-Shafer theory and the possibilistic framework – are discussed (the latter with a strong emphasis on the so called context model interpretation). In a detailed analysis of these frameworks their advantages and limitations are pointed out. Based on the results of that analysis, I proceed to draft a consistent frame-spanning representation for information about set-concepts in multivariate modeling.

4.1 Interpretation of Model Components

In the previous chapters several applications for knowledge representations using sets were mentioned. Although these applications use a variety of interpretation of the sets, they are very similar on the level formal representation. Similar observation can be made about other mathematical concepts, as the abstract mathematical notions are flexible enough to support applications in a wide range of contexts. Yet, to ensure meaningful results, the interpretations often have to impose constraints on the choice of operations permitted. For this reason it is useful to investigate the interpretation level of knowledge representations separately from the mathematical tools that are used to formalize them. This approach allows to select and adapt operation in such a way that the alignment of the formal model with a meaningful empirical interpretation is preserved.

4.1.1 Interpretations for Sets

In order to point out the sometimes subtle differences between knowledge representations I distinguish between four different uses of sets:

Classical Interpretation I apply the term classical interpretation if the set represents a collection of physical objects (compare page 9). This approach is often used to provide a detailed model of a small section of the world. For instance software systems for warehouse management would employ a database that represents palettes with goods as data objects. Each palette is linked to a physical location, content description, transport history, storage requirements and similar information. Sets are natural means to represent groups of such palettes e.g. for further shipping. Moreover the database itself models the set of objects that constitute the current inventory of the warehouse. In modern computational biology sets are particularly relevant due to their role in expressing relations. For instance Protein-DNA interaction experiments identify sets of putative transcription factors (regulatory proteins) that bind to the promoter region of a gene. Because functional units in biological systems are frequently protein complexes formed from several interacting components, sets constitute a natural elements in the description of such processes.

Applicable Attribute Values With a slightly higher level of abstraction sets are used to specify realizations of properties applicable to an object. The distinctive difference to the classical interpretation is that elements no longer correspond to representations of physical objects themselves but rather values of their attributes. Specifying sets of applicable values for one or more attributes characterizes classes of objects without explicitly listing their members. The indirect reference to objects via characteristic properties is particularly advantageous when knowledge about the properties of individual objects or the possible realizations of those properties in general is yet incomplete. In particular the set of known applicable values can be expanded if additional information becomes available. For this reason the interpretation is often used in conjunctions of set-valued attributes in annotation databases, which are updated at intervals to account for insights from recent publications.

Imprecise Specification The third interpretation refers to an incomplete state of knowledge that manifests itself as *imprecision*. It is often associated with the state-of-the-world view, but may also be applied to states of individual objects or object attributes. In both cases the goal is to discern the current true state of the world (or of the object under consideration) among a number of alternatives in a state space. It is assumed that this true state is unique and in principle precisely defined. However the available information is not necessarily sufficient to identify this "true state". Instead a partial specification is provided via a *set of candidates*. Alternatives that are inconsistent with observations are excluded from the candidate set. In the special case, where *precise* knowledge about the targeted state is available, the candidate set collapses to a singleton. In contrast to the applicable attribute values interpretation imprecise specifications require fully known attribute domains (knowledge is represented by successively eliminating incompatible alternatives).

Non-Quantified Intrinsic Variability Finally sets may be used to express non-quantified intrinsic variability, that is to refer variable do not have a fixed value, or for the specification of perception intervals when the variable can be observed with limited resolution only (Dubois, 2006). For example a coastline varies to the effect of tides and waves. Above a given spacial resolution a precise specification of a coastline could only represent a snapshot with questionable utility. Instead the notion is extended to the more convenient set-based concept of a coastal zone[1]. The main difference between intrinsic variability and imprecision is that imprecision refers to an epistemic state only, whereas the former is used to express an intrinsic property of the described objects. To illustrate this difference consider a series of photographs of a section of guitar string in three-dimensional space, that are take from several known locations. After one photograph is obtained, knowledge about the position of a string at rest is imprecise, as only a projection was observed. However combining photos from multiple angles allows to reconstruct the strings exact position via triangulation. In contrast a string oscillating at a sufficiently high frequency, does not have a precise discernible position when viewed at the typical temporal resolution of the camera shutter. Thus additional photos will not resolve the location of the string segment beyond its intrinsic variability within the amplitude of the oscillation.

Due to the close relation of the set-based concepts the applicable interpretation sometimes depends on the focus of interest and the resulting point of view taken. For instance, the set of train stations in a larger city is used with the classical

[1] An even more flexible way to formally describe such vague concepts is by means of a fuzzy set Zadeh (1965). In a fuzzy set representation each geographical location is assigned a degree of membership to the concept of being "coast".

interpretation when speaking about that city in general, yet, when referring to a particular location only given as “at the station” the same set is used in the role of an imprecise specification resulting from that ambiguous description.

4.1.2 Interpretation of Probability Distributions

Regardless of the interpretation applied, the knowledge represented can be subject to uncertainty. Thus uncertainty may originate from different sources, e.g.:

- Uncertainty about properties of individual objects due the variability within a reference class (frequentist view),
- Uncertainty about the state of a system due to stochastic evolution since the last observation or
- Uncertainty due to limited accuracy of the observation process itself, e.g. measurement error (see Subsection 2.3.3 of the previous Chapter).

For the purpose of formal modeling uncertainty can be quantified using probability distributions. Apart from this direct description of random processes, probability distributions have also been used in belief representation, where they are interpreted as subjective assessments of the uncertainty w.r.t. propositions or states.

When probability distributions are applied to set-based representations they give rise to *random sets*. The notion of random sets establishes a general framework that subsumes a number of popular approaches such as upper and lower probabilities (Dempster, 1967) or the Dempster-Shafer theory of evidence (Shafer, 1976).

4.2 Random Sets

With the semantics for sets and distributions outlined in the previous sections, it is now possible to descend to the level of formal representations for distributions over sets. First we notice, that the definition of a probability space does not restrict the type of the mathematical objects that are used to describe experimental outcomes. Choosing sets for the elements of the sample space is a straightforward way to extend probabilistic reasoning to sets-based concepts. This idea gives rise to random sets.

4.2.1 Definition

Random sets were brought to the attention of the knowledge representation community by Nguyen (1978a, 2005) who coined the term and investigated their relation to to belief functions, fuzzy sets and several uncertainty calculi. A random set is introduced as the a set-valued mapping from a probability space:

Definition 4.1. *Consider a probability space* $(C, 2^C, P)$ *and a nonempty set* U*. Given that probability space, a mapping* $\Gamma\colon C \to 2^U$ *is called a* **random set***. The sets* $\Gamma(c), \quad c \in C$ *with* $P(c) > 0$ *are called* **focal sets** *of* Γ*.*

Random sets may be used to extend probabilistic reasoning to set-valued attributes. To that end the sets O and Λ in Equation 2.2 are respectively identified with C and U in the definition of the random set (Definition 4.1). Accordingly, the attribute A takes the role of the set-valued mapping Γ.

Some authors have used definitions that refer not to the mapping Γ itself, but to the tuple (P, Γ) as a random set. However, definition 4.1 clearly points out the relation to random variables. A random set may simply be viewed as a set-valued random variable with each focal set forming a possible set-valued outcome. Provided that the number of potential set-outcomes is low, random sets by themselves already constitute effective knowledge representations.

4.2.2 Properties of Random Sets

In knowledge representation one is often interested in random sets that exhibit specific properties. Two of the properties frequently discussed in the context of knowledge representation frameworks are *consistency* and *consonance*. The consistency criterion requires that all focal sets overlap, so that some elements are contained in all focal sets (after Kruse et al., 1994):

Definition 4.2. *A random set is called* **consistent** *iff*

$$\bigcap_{c \in C} \Gamma(c) \neq \emptyset.$$

Clearly, a random set Γ can only be consistent if the empty set is not among the focal sets of Γ. Under the interpretation of focal sets as representing possible states under hypotheses based on individual piece of evidence, consistency is equivalent to the absence of contradiction between those pieces of evidence.

The other subclass of random sets that is useful to discuss in the context of knowledge representation are those with *consonant* focal sets (Kruse et al., 1994; Dubois and Prade, 1999; Borgelt and Kruse, 2002):

Definition 4.3. *Let $\Gamma : C \to 2^\Omega$ be a random set with $C = \{c_1, \dots, c_n\}$. The focal sets $\Gamma(c_i)$, $1 \leq i \leq n$, are called* **consonant** *iff there exists a sequence $c_{i_1}, c_{i_2}, \dots, c_{i_n}$, $1 \leq i_1, \dots, i_n \leq n$, $\forall 1 \leq j < k \leq n : i_j \neq i_k$, so that*

$$\Gamma(c_{i_1}) \subseteq \Gamma(c_{i_2}) \subseteq \cdots \subseteq \Gamma(c_{i_n}).$$

In other words, consonance describes the property that focal sets are nested. Unless Γ maps to the empty set for some $c \in C$, consonant random sets are also consistent. In that case the common intersection is simply the smallest focal set, that is the leftmost element in the sequence denoting the subset relations from Definition 4.3.

4.2.3 Uses and Limitations

Whereas the random set concept is useful in modeling imprecision and experiments with a limited number of multi-valued outcomes there are also some unfavorable aspects to the approach. Firstly, switching from elementary descriptions to sets considerably increases the number of potential alternative outcomes to be considered. In the absence of external information restricting the focal sets, a full modeling calls for the representation of a probability distribution defined over the power set of Ω. As a result, the number of potential outcomes grows exponentially with the cardinality of the base domain. Because Ω itself is often formed as a product space, the number of values to be stored reaches adverse levels even for comparatively small problems. Apart from storage requirements, larger distributions are increasingly expensive to determine from data. The reason is, that in order to obtain reliable estimates for marginal and conditional probabilities, the required sample needs to supply a sufficient number of supporting cases per class. Hence, with a large number of classes, data availability becomes a limiting factor. Therefore, the random set approach is rarely applied directly. Nevertheless it provides a useful reference model that helps to illustrate the benefits and limitations of alternative frameworks.

The practical challenges listed above lead several authors to suggest frameworks that emphasize applicability rather than attempting to exactly represent potentially inassessable probability distributions. Although these approaches usually give up some of the accuracy and representation power of the direct approach using random sets, some of them have turned out to be very successful. One such approach the Dempster-Shafer theory of evidence has been received with considerable interest (e.g. Kohlas and Monney, 1995; Lee and Zhu, 1995; Lalmas, 1997; Nakamura et al., 2007), in particular within the field of economics.

4.3 Dempster-Shafer Theory of Evidence

In 1967 Dempster published an article, in which he investigated upper and lower probability measures induced by set-valued mappings from a probability space to a frame of discernment Ω. To that end, he depicted a scenario, where each element of a probability space is mapped to a subset of Ω. In conjunction with the probability distribution on its domain, such a set-valued mapping constitutes a random set.

4.3.1 Upper and Lower Probability

The random set approach emphasizes focal sets as central elements of the representations. But considering that question often relate to occurrences of certain $\omega \in \Omega$ that can be elements of more than one focal set, an element-wise perspective is sometimes desirable for querying the model. To summarize such information, Dempster defined upper and lower probability measures $P^*\colon 2^\Omega \to [0,1]$ and $P_*\colon 2^\Omega \to [0,1]$, which reflect the a rescaled aggregated probability from all non-empty focal sets that overlap with, or are fully contained within a subset E of Ω respectively. For a discussion of Dempster's idea, it is advantageous to initially concentrate on the cases that do not require rescaling. Assuming $\forall c \in C : (\Gamma(c) = \emptyset) \implies P(\{c\}) = 0$, $P_*(E)$ and $P^*(E)$ may be computed $\forall E \subseteq \Omega$ as

$$P_*(E) = \sum_{\substack{c \in C \\ \emptyset \neq \Gamma(c) \subseteq E}} P(\{c\}) = P\left(\{c \in C \colon \emptyset \neq \Gamma(c) \subseteq E\}\right) \tag{4.1}$$

$$P^*(E) = \sum_{\substack{c \in C \\ E \cap \Gamma(c) \neq \emptyset}} P(\{c\}) = P\left(\{c \in C \colon E \cap \Gamma(c) \neq \emptyset\}\right). \tag{4.2}$$

Dempster then turned his attention to the family of probability measures over Ω that are compatible with those bounds and provided two further, equivalent definitions. One of these definitions characterizes compatible measures as those, which may be obtained by freely distributing probability mass within focal sets. Thus Dempster's approach can be understood as an attempt to trace probability mass for underspecified distributions. Every focal set marks out the most specific subset of the frame of discernment Ω, to which a distinct portion of probability mass is assigned. Beyond that, nothing is known about its distribution to the elementary events within the focal set. The class of compatible probability distributions is generated by the different ways of shifting the probability mass within their respective focal set. One of the advantages of this representation is the capacity to express states of partial knowledge by restricting the set of admissible distributions. In a state of total ignorance all probability distributions over Ω would be admissible, which is represented by assigning the complete probability mass to the whole of Ω. The other extreme occurs when all focal sets $\Gamma(c)$ are singletons. For such cases the index set used for computing P_* and P^* in Equations 4.1 and 4.2 is identical and the boundaries determine a unique probability measure.

The interpretation of set-valuedness applied in the previous paragraph slightly differs from imprecision (see page 23), in that probability mass associated with a focal set may be distributed freely among the elementary events within the focal set, rather than being assigned "en bloc" to an unknown, yet specific element. However, the imprecision interpretation does lead to the subclass of so called *extremal* measures. Dempster pointed out, that his compatible measures can also be described as the closure of those extremal solutions under mixing. In a brief example he further suggested, that random sets be used to present degrees of belief that quantify the partial knowledge available from information sources, such as human experts. That idea was further elaborated by Shafer and is treated in the following subsection.

It is also noteworthy that the function Γ in Dempster's formalization may map to the empty set even though such an assignment would be meaningless under the suggested interpretation. According to Dempster, that case was included in order to arrive at a "more general" theory. Yet, this effectively allows to attach a positive probability to the impossible event giving rise to a problem of "lost probability mass" – an anomaly that occurs because the calculation of $P^*(\Omega)$ and $P_*(\Omega)$ does not account for probability mass linked to the empty set. To compensate for that loss of probability mass, the definitions suggested by Dempster include a renormalization constant for rescaling the probability associated with non-empty focal sets to unity. The required renormalization factor is

$$\frac{1}{\sum_{\substack{c\in C\\ \Gamma(c)\neq\emptyset}} P(\{c\})}.$$

Because the denominator collects the probability mass assigned to non-empty subsets of Ω, the definition assumes, that $P\left(\{c\colon c\in C \wedge \Gamma(c)=\emptyset\}\right) < 1$ holds. Otherwise the denominator would become zero, so both upper and lower probability remain undefined.

4.3.2 Reinterpretation by Shafer

Although Dempster already suggested to apply his approach to represent evidential support, it was Shafer, who reformulated, reinterpreted and extended these ideas for a theory of evidence Shafer (1976). Using a state-of-the world type representation (cf. page 20), Shafer modeled rational beliefs of intelligent agents with only partial knowledge about a situation. That partial knowledge is due to pieces of *evidence*, including prior knowledge and possibly preconceptions. However, it is usually not known, whether a certain piece of evidence is relevant to the question at hand or not. To express the agents commitments regarding these alternative interpretations of the pieces of evidence an assessment of their relative relevance is given in terms of a subjective probability assignment. The agents use these assessment to attribute an amount of belief to propositions. In this context, a proposition is the statement, that the true state of the world is described by an element of a certain subset E of Ω. When referring to a fixed frame of discernment, I will apply the term proposition for both the statement itself, and the corresponding subset of the frame that represents that statement in the model.

Different propositions can be related to each other in that they refer to overlapping or even nested subsets of Ω. For instance, a tourist may attach belief to the proposition that the Luxembourg Palace is somewhere in Europe but also to the more specific proposition that it is actually situated in the French capital of Paris. Of course with Paris being located in Europe, belief attributed to the second hypothesis should also be attributed to the more general one expressed by the first proposition. Formally, considering two nonempty sets E_1 and E_2 the implication

$$E_1 \subset E_2 \implies \mathrm{Bel}(E_1) \leq \mathrm{Bel}(E_2) \tag{4.3}$$

should hold.

In this modeling, a piece of evidence justifies attributing an amount of belief to a whole class of hypotheses, but Shafer noticed that the same information can be represented by considering only the most specific propositions supported by each interpretation of the given evidence. These most specific hypotheses are described by the focal sets of a random set Γ.

The probability distribution over C reflects the relative amount of belief attached to the different interpretations of the evidence. The belief in any given proposition is determined from the amount of belief attributed to the supporting interpretations of the evidence. A focal set reflects the most specific proposition supported by some interpretation of the evidence. Equipped with these semantics, belief is more comfortably quantified in terms of belief mass attributed to focal sets.

The belief mass $\mathrm{m}(H)$ of a set $H \subseteq \Omega$ measures the belief attributed to evidence that points exactly to H but none of the more specific hypotheses reflected by subsets of H. Formally $\mathrm{m}(H)$ can be defined as the probability assigned to the preimage of a set H, so $\mathrm{m}(H) = P(\Gamma^{-1}(H))$. Since only the focal sets belong to the range of Γ, the function m describes a distribution of weight to focal sets. For all other sets, the assigned value is $\mathrm{m}(H) = P(\Gamma^{-1}(H)) = P(\emptyset) = 0$. This so called *basic probability assignment* forms the fundament of Shafer's formalization (adapted from Shafer, 1976):

Definition 4.4. *A function*

$$\begin{aligned} &\mathrm{m} : 2^{\Omega} \to [0,1], && \textit{with} \\ &\mathrm{m}(\emptyset) = 0, && \textit{and} \\ &\sum_{H \subseteq \Omega} \mathrm{m}(H) = 1. \end{aligned}$$

is called a basic probability assignment.

In comparison to Dempster's suggestion, two additional criteria of that definition reflect the adopted interpretation. Since the true state must be some element of Ω, the empty set encodes an unsatisfiable hypothesis. Moreover, the total belief mass distributed to the focal sets must be one.

Having assigned the belief mass to a focal set, it is now possible to compute a measure of belief for arbitrary subsets of Ω. Any given proposition $E \subseteq \Omega$ is supported by exactly those interpretations of the evidence that correspond to a focal set fully contained in that proposition. With that interpretation the lower probability of Dempster's approach is viewed as a measure of belief, i.e.

$$P_*(E) = \sum_{\substack{c \in C \\ \Gamma(c) \subseteq E}} P(\{c\}) = \sum_{H \subseteq E} \mathrm{m}(H) = \mathrm{Bel}(E). \tag{4.4}$$

Conversely the upper probability

$$P^*(E) = \sum_{\substack{c \in C \\ \Gamma(c) \cap E \neq \emptyset}} P(\{c\}) = \sum_{H \cap E \neq \emptyset} \mathrm{m}(H) = \mathrm{Pl}(E). \tag{4.5}$$

measures the plausibility of proposition E. Although the above equations are applicable for the original interpretation of evidence, Dempster (1967) and Shafer (1981) realized that the operations they had proposed for their approach do not generally preserve the probability bounds. Thus, results acquired by reasoning within the framework are usually not interpretable as probability bounds any more.

4.3.3 Reasoning in the Dempster-Shafer Framework

Dempster-Shafer theory provides mechanisms for combining information sources (Dempster, 1967, 1968) now commonly referred to as "Dempster's rule of conditioning" and "Dempster's rule of combination", which Dempster considered as an extension of the conditioning rule.

Dempster's Rule of Conditioning The term "conditioning" refers to an operation, in which the representation of an epistemic state is supplemented with a new piece of information, e.g., due to an assumption or an observation. The new information is expressed as a proposition $F \subseteq \Omega$. It is then asserted, that outcomes outside F are excluded. In the framework suggested by Dempster, computing the conditional measures $P^*(E \mid F)$ and $P_*(E \mid F)$ is based on the idea of restricting all focal sets to F. For conditioning with proposition $F \subseteq \Omega$, images of the set-valued mapping are updated such that $\forall c \in C : \Gamma'(c) = \Gamma(c) \cap F$. If F and some of the original focal sets of Γ are disjoint, however, this operation will assign a positive probability to the empty set, resulting in an inconsistent intermediate state. Consistency is restored by subsequent renormalization, that is a rescaling of probability/belief mass attributed to the remaining focal sets. The approach can be viewed as a proportional redistribution of belief after ruling out interpretations of the evidence that conflict with F, i.e. as a form of Bayesian conditioning. Indeed, if all images under Γ are singletons the operation is equivalent to Bayesian conditioning of a probability distribution. At the same time, the restriction of the focal sets directly corresponds to an application of an expansion operation (Alchourrón et al., 1985) from the relational framework for each of the remaining interpretations of the evidence.

The conditioning rule is frequently expressed in terms of the upper and lower probability measures P^* and P_*, which are derived in Dempster (1967):

$$P^*(E \mid F) = \frac{P^*(E \cap F)}{P^*(F)} \quad \text{and} \tag{4.6}$$

$$P_*(E \mid F) = 1 - P^*(\overline{E} \mid F) = 1 - \frac{P^*(\overline{E} \cap F)}{P^*(F)}. \tag{4.7}$$

However, renormalization produces counterintuitive results when other interpretations of the framework are applied. For that reason, some variants and extensions of Dempster-Shafer theory use alternative rules that skip the renormalization step (e.g. Smets, 1990; Smets and Kennes, 1994). The normality criterion from Definition 4.4 is relaxed to $\sum_{H \subseteq \Omega} \mathrm{m}(H) \leq 1$ in those frameworks. A broader discussion of semantic problems related to renormalization will be given in Subsection 4.3.4.

Dempster's Rule of Combination Among the most notable results of Dempster's work is the formulation of a mechanism for combining evidence now commonly referred to as Dempster's Rule of combination (Dempster, 1967; Shafer, 1976). The rule was originally proposed as a method to summarize information from several independent information sources. Each individual source provides a random set representation that reflects an assessment of the situation using only the evidence from that source, which is called a *body of evidence*. The rule may be used to produce a combined assessment from several bodies of evidence, which again is expressed as a random set representation.

The rule of combination is more easily understood by referring to basic probability assignments. Each focal set corresponds to a certain interpretation of the evidence, and belief in any particular interpretation is quantified via the belief/probability mass assigned to its corresponding focal set. The rule aims at providing new basic probability assignments to distribute belief to interpretations of the combined evidence.

The combination of basic probability assignments conducted by that rule follows the suggestion, that an an interpretation for the combined evidence is formed by selecting an interpretation for each of the input representations and combining the constraints by intersecting the corresponding focal sets. The complete representation resulting from the combination of basic probability assignments then contains all possible (non-empty) intersections that can be generated from such combinations.

The belief/probability mass associated with a focal set in the new representation is computed from aggregated contributions for each combination of focal sets in the input representation that produce it as their intersection, were mass contributed by a combination of focal sets is the product of its elements' belief/probability masses in the respective in the input representations.

Again, one drawback of this idea is revealed when it is applied to selections of focal sets, which have an empty intersection. According to the above calculation, such as situation would lead to belief mass being assigned to the empty set. To remain compatible with Definition 4.4, the belief mass assigned to the empty set is discarded and the remainder rescaled to unity (Equation 4.9).

Although that combination rule was originally presented for only two information sources (e.g. Shafer, 1976), it is easily extended to a more general form that is suitable for the simultaneous combination of several information sources:

Definition 4.5. *Suppose* $\mathrm{m}_1, \ldots, \mathrm{m}_k$ *are basic probability assignments to subsets of the same frame* Ω. *Let* $\mathcal{F}_1, \ldots, \mathcal{F}_k$ *denote the respective sets of focal sets, i.e,*

$$\mathcal{F}_i = \{H \colon H \subseteq \Omega \wedge \mathrm{m}_i(H) > 0\} .$$

Under the condition, that

$$\sum_{\substack{(H_1,\ldots,H_k)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_k\\ \bigcap_{i=1}^k H_i=\emptyset}} \prod_{i=1}^{k} \mathrm{m}_i(H_i) < 1 \tag{4.8}$$

the function $\mathrm{m} : 2^\Omega \to [0,1], \quad \emptyset \mapsto 0$ *and for* $H : \emptyset \neq H \subseteq \Omega$*:*

$$H \quad\mapsto\quad \frac{\sum\limits_{\substack{(H_1,\ldots,H_k)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_k\\ \bigcap_{i=1}^k H_i=H}} \prod\limits_{i=1}^{k} \mathrm{m}_i(H_i)}{\sum\limits_{\substack{(H_1,\ldots,H_k)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_k\\ \bigcap_{i=1}^k H_i\neq\emptyset}} \prod\limits_{i=1}^{k} \mathrm{m}_i(H_i)} \tag{4.9}$$

is a basic probability assignment.

This combined representation only exists if the condition expressed in Equation 4.8 holds. Shafer noted that, since the focal sets refer to mutually exclusive interpretations of the combined evidence, the term

$$\sum_{\substack{(H_1,\ldots,H_k)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_k\\ \bigcap_{i=1}^k H_i=\emptyset}} \prod_{i=1}^{k} \mathrm{m}_i(H_i)$$

measures a degree of conflict between the information sources. After rescaling, the non-empty intersections from the set of focal sets $\mathcal{F} = \{H : \emptyset \neq H \subseteq \Omega : (\exists (H_1,\ldots,H_k) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k : \bigcap_{i=1}^k H_i = H)\}$ for the unified representation.

Obviously, the rule of combination is as an extension of the rule of conditioning, where the conditioning information itself is provided via a random set representation. Conversely the rule of conditioning results can be viewed as a special case where a random set representation is combined with a piece of information that has a unique interpretation.

Combination via Common Refinements So far, it was assumed that information acquired from different sources was expressed with respect to the same frame of discernment. However, such a common frame may not always have been established in advance. If, for instance, two information sources rely on different sets of attributes to supply interpretations of their respective evidence, the propositions reflecting those interpretations of the evidence may well refer to different frames of discernment. In order to combine these propositions, their corresponding set-representations must first be converted to a common reference. In Shafer's notion such a reference frame is provided by a *common refinement* of the original frames of discernment. Assuming what Shafer calls independent frames, i.e. that "no proposition on one of [the original frames] non-trivially implies a proposition discerned by the other"[2], that common refinement is the Cartesian product of the original frames of discernment. Shafer argued that since propositions merely reflect constraints on the frames of discernment, any given proposition may be expressed using a semantically equivalent restriction w.r.t. a finer frame, namely the Cartesian product of that proposition and the frame used by of respective other representation. With the subsequent application of Dempster's rule the intersections obtained as focal sets the in the combined representations are then the Cartesian products of focal sets from the parent frames.

4.3.4 Critical Discussion

One of the controversial aspects of using Dempster's rule of combination in the framework refers to the way, it deals with conflict. Zadeh (1984) proposed an example, where the combination of two strongly conflicting information sources using Dempster's rule resulted in a representation that suggested high belief in a proposition that is only weakly supported by each of the individual information sources. While some authors have argued that conflict merely indicates an incomplete definition of the frame of discernment (Smets, 1990), others have connected the paradoxical result to the reassignment of conflict mass to focal sets only (Yager, 1987) or to inherent, overly optimistic assumptions about the reliability of information sources (Dubois and Prade, 1988b). For each of these views, alternative combination rules have been proposed, which are extensively discussed and compared in Lefevre et al. (2002).

The normalization problem can also be viewed a symptom of a more fundamental discrepancy between the mechanism applied for combining propositions or theories linked to specific interpretations of evidence from individual sources, and the assumptions used to integrate information from different sources. The former operations are based on a set-theoretic mechanism using a binary logic, whereas the latter model the relative relevance of individual interpretations using a probability-based approach with corresponding conditioning and combination rules. In Dempster (1967) it is remarked that the mechanism assumes "independence of sources", which is later explained as "independence of errors". The

[2] Relations between logically dependent frames are investigated in Chapter 5. For a discussion of special cases see chapter 6 of Shafer (1976).

meaning of this informal statement, becomes clearer when the combination rule itself is investigated more closely: From the multiplicative combination of belief mass, one can infer that the independence assumption not just refers to the acquisition of the pieces of evidence, but to statistical independence of their interpretations (given as the focal hypotheses) w.r.t. the basic probability assignment m (the subjective probability measure used to assesses the relevance of the interpretations). Yet, with Shafer's interpretation the focal sets obtained from the different sources are still formed over the same domain. This situation can lead to logical dependencies that conflict with an independent combination of interpretations, so the above assumption does not generally hold:

Example 4.1. Consider two basic probability assignments m_1 and m_2 that describe belief structures obtained using two pieces of evidence from different information sources w.r.t. a frame of discernment $\Omega = \{\omega_1, \omega_2, \omega_3\, \omega_4\}$. Each assignments assigns probability mass to the focal sets H_1 and H_2. Let

$$\begin{aligned} H_1 &= \{\omega_1\}, & H_2 &= \{\omega_2, \omega_3\}, \\ \mathrm{m}_1(H_1) &= 0.8, & \mathrm{m}_1(\Omega) &= 0.2, \\ \mathrm{m}_2(H_2) &= 0.5, & \mathrm{m}_2(\Omega) &= 0.5. \end{aligned}$$

In this case H_1 and H_2 are disjoint, so at most one of the proposition may cover the unknown true state of the world. Thus, provided both pieces of evidence were reliable, at least one of these pieces must be irrelevant and the interpretation that both pieces of evidence apply is not admissible. Nevertheless the subjective probability of $m_1(H_1) \cdot \mathrm{m}_2(H_2) = 0.4$ assigned to that interpretation under the independence assumption is greater than 0 and therefore inconsistent with logical restrictions already given due to background knowledge about the intended semantics. □

The example demonstrates, that inherent assumptions of the combination rule used in the model may result in an inappropriate approximation of the unknown true interaction structure. It disregards possible logical dependencies between admissible interpretations of subsets of the evidence. These dependencies arise from compatible or incompatible interpretations of these pieces of evidence within the calculus for propositional logic. Indeed, the problem of positive probability mass assignments to the empty set could not even occur, if the statistical independence assumption w.r.t. interpretations of disjoint subsets of the evidence really applied. Even in the absence of such logical constraints, the assumption of independence would require further justification because statistical interactions between interpretations of evidence from different information sources may not be ruled out (absence of logical dependency does not imply statistical independence).

The lack of means to represent probabilistic interaction structures, such as conditional information, extends to other versions of the belief function formalism and has lead Pearl (1990) to point out, that the belief framework is too limited to serve as a general representation of partial knowledge. In their approach to this problem, Kruse et al. (1991a), Kruse et al. (1991b, ch. 14) have applied separate data structures (specialization matrices) to model information about interactions when reasoning with mass distributions.

Other limitations of the Dempster-Shafer approach have been overcome by applying relatively minor modifications to the framework. For instance the basic version of the Dempster-Shafer framework calls for repeated renormalization, to update the internal belief state whenever new evidence is encountered. In a multiple update scenario this has the undesired effect of giving to each new piece of evidence the same weight as to the entire prior evidence encountered (weight of evidence problem). More advanced variants of the framework, however, avoid that problem by either keeping track of the weight of the accumulated evidence or removing the renormalization step from the actual belief representation altogether.

Regardless of its limitations, the framework has successfully been applied to random set representations as a heuristic to obtain coarse numerical assessments, if detailed information on the interaction were unavailable. An unnormalized version of Dempster's rule has also been employed for the pooling of evidence in the transferable belief model (Smets, 1990, 1993; Smets and Kennes, 1994). That model uses a subjectivist interpretation of probability and focuses on supporting decision tasks. The transferable belief model (TBM) maintains a strict separation between the epistemic or credal level encompassing the parameters that encode the actual belief state, and a so-called pignistic level used for decision making. Normalization, in the form of the pignistic transformation operation is shifted into the decision stage of the model to be performed on-demand whenever a decision needs to be made. While this delayed renormalization remedies the weight of evidence problem, the criticism with respect to the independence assumptions still applies (see Snow (1998) for an example that leads to a counter-intuitive belief state with the TBM.) In addition the TBM relies on storing belief functions over the power set of the base domain for encoding a belief state at the credal level, limiting its applicability on large domains such as those typically encountered in computational biology.

4.4 Possibility Theory

Besides belief functions, other frameworks may be subsumed under the random set formalism – among them possibility theory. The notion of possibility that underlies this theory was suggested by Zadeh (1978) as means to express certain aspects of natural language. It may be used to refer to statements or pieces of information that are both imprecise and uncertain. In contrast to the colloquial meaning of the term "possible", which is based on a binary distinction, Zadeh proposed *degrees of possibility* assigned to elements of a frame of discernment[3].

The original formulation presented by Zadeh closely relates possibility to the theory of fuzzy sets by the same author (Zadeh, 1965). Zadeh's initial consideration was that information is often provided in the form of statements that relate attributes to fuzzy concepts (Zadeh had previously suggested to use such fuzzy concepts to reflect the intrinsic vagueness of expressions in natural language). This induces a fuzzy restriction of the attribute domain, which reflects the *compatibility* of individual instantiations with that concept. He argued, that such restrictions are not well represented in probabilistic terms since they merely quantify if an instantiation can occur *in principle*, but not the likelihood of its actual realization. Nevertheless, Zadeh acknowledges a *heuristic connection between possibilities and probabilities* in that any probable event must at least be possible.

Zadeh's suggestions were subsequently taken up and expanded by several other authors (Hisdal, 1978; Nguyen, 1978b; Yager, 1981; Higashi and Klir, 1983; Shafer, 1986; Dubois and Prade, 1988a), who worked towards establishing the semantics of possibility, added further interpretations and provided views on operations and constraints vital to making testable predictions in the framework. Many of the contributions published in the 1980's investigated the formal similarities of possibility theory and the belief function framework to develop information measures to quantify evidence. While this development induced a fruitful discussion on the meaning of possibility it also added to the existing confusion w.r.t. the term, so that several interpretations and formalizations coexist to this day.

[3] Unfortunately, "possibility" is not used consistently, even within the field of knowledge representation itself. For instance, in modal logic the term is applied in its original dichotomous meaning.

4.4.1 Axiomatic Approach to Possibility

The arguably best developed formalization of possibility theory is the axiomatic approach (Dubois and Prade, 1988a). With that axiomatic approach, events from a sample space Ω are assigned a degree of possibility from a totally ordered set V. That set contains at least a smallest element 0 and a largest element 1. For $V = \{0, 1\}$ this leads called to so called binary possibility measures, which reflect the concept of possibility as it is understood in modal logic, but typically the complete interval $[0, 1]$ is chosen The assignment is formally represented as a *possibility distribution*:

Definition 4.6. *A possibility distribution is a function*

$$\pi : \Omega \to [0, 1]. \tag{4.10}$$

A possibility distribution is called **normalized***, iff*

$$\exists \omega \in \Omega : \pi(\omega) = 1. \tag{4.11}$$

Like with the probabilistic framework, possibility distributions allow to construct a function on subsets of Ω. For standard possibility theory this function is given by the following definition (Borgelt and Kruse, 2002), (cf. Dubois and Prade, 1988a; Zadeh, 1978):

Definition 4.7. *Let Ω be a sample space. A (general)* **possibility measure** *is a function $\Pi : 2^\Omega \to [0, 1]$ satisfying*

1. $\Pi(\emptyset) = 0$ *and*
2. $\forall E_1, E_2 \subseteq \Omega : \ \Pi(E_1 \cup E_2) = \max\{\Pi(E_1), \Pi(E_2)\}$ *(Maxitivity)*

It is appropriate to remark that Π is not required to be σ-additive. Therefore it does not generally constitute a measure in the sense of measure theory. The name "possibility measure" was merely coined in analogy to the concept of a probability measure. Subsection 4.4.5 will discuss the consequences of this difference in terms of operations within the possibilistic framework.

Each possibility distribution π leads to a unique possibility measure Π via $\Pi(\{\omega\}) = \pi(\omega)$ and therefore

$$\Pi(E) = \begin{cases} 0 & \text{if } E = \emptyset, \\ \max_{\omega \in E} \pi(\omega) & \text{otherwise.} \end{cases} \tag{4.12}$$

An interesting property of possibility measures is that the possibility for the union of two events may be computed from their respective possibility measures only, even if these events are not disjoint. This contrasts with the probabilistic case, were such an operation is only applicable to disjoint events. Moreover, from the second axiom in Definition 4.7 it may be concluded $\forall E_1, E_2 : E_1 \subseteq E_2 \subseteq \Omega :$

$$\Pi(E_1) \leq \max(\Pi(E_1), \Pi(E_2)) = \Pi(E_1 \cup E_2) = \Pi(E_2). \tag{4.13}$$

This result is in analogous to the monotonicity condition for belief functions given on page 59.

With the axiom stating how to compute the possibility of unions of events $E_1, E_2 \in \Omega$, it would be convenient if a similar rule could be formulated for the intersection of events. However this is not possible in general, because the attribution of the degree of possibility of a set to its subsets is ambiguous for all but trivial cases (see Dubois and Prade (1998a) for a brief discussion of this problem). Yet, a simple consideration permits to compute an upper bound for the possibility of the intersection of events. Let $E_1, E_2 \subseteq \Omega$. Obviously, $(E_1 \cap E_2) \subseteq E_1$, and $(E_1 \cap E_2) \subseteq E_2$. Then substituting into our result about the union of events from Equation 4.13 yields two conditions:

$$\begin{aligned}
\Pi(E_1 \cap E_2) &\leq \max\{\Pi(E_1 \cap E_2), \Pi(E_1)\} = \Pi((E_1 \cap E_2) \cup E_1) = \Pi(E_1) \\
\Pi(E_1 \cap E_2) &\leq \max\{\Pi(E_1 \cap E_2), \Pi(E_2)\} = \Pi((E_1 \cap E_2) \cup E_2) = \Pi(E_2).
\end{aligned}$$

Combined, these two conditions can be summarized as

$$\Pi(E_1 \cap E_2) \leq \min\{\Pi(E_1), \Pi(E_2)\}. \tag{4.14}$$

Indeed, the intersection $E_1 \cap E_2$ may have a lower possibility than each of the events E_1 and E_2, for the maximally possible elements of E_1 and E_2 need not be in the common intersection of the two events. On the other hand, the possibility of $E_1 \cap E_2$ cannot be higher than that of either E_1 or E_2 because that would require the existence of an element ω' in $E_1 \cap E_2$ such that $\forall \omega \in (E_1 \cup E_2) : \pi(\omega') > \pi(\omega)$. However, since as $E_1 \cap E_2 \subseteq E_1 \cup E_2$, any maximally possible element ω' would be contained in at least one the events E_1 and E_2 as well.

Among the possibility measures demarcated by Definition 4.7, the subclass of measures corresponding to normalized possibility distributions has received particular attention by researchers. Because for normalized possibility distributions at least one element in Ω must be fully possible, $\Pi(\Omega) = \max_{\omega \in E} \pi(\omega) = 1$ holds for the measures in that class.

Such normalized measures may be complemented with necessity measures, which are usually defined via the equation $N(E) = 1 - \Pi(\overline{E})$. This reconciles possibility theory, with the understanding of possibility adopted in philosophy. On account of the normalization condition, these versions of possibility theory modify Definition 4.6 by introducing $\Pi(\Omega) = 1$ as an additional axiom (e.g. Dubois and Prade, 1988a). Unless stated otherwise, I will apply the term possibility theory in this dissertation to refer to the axiomatization without the normality axiom. For a discussion and comparison of other notions of possibility theory the reader is referred to Dubois (2006) and the very extensive papers of De Cooman (de Cooman, 1997a,b,c).

The practical value of the axiomatic approach to possibility theory is in its standardization of several operations with possibility measures. However, it leaves open the essential question of how possibility degrees are determined in the first place. In order to apply the formal model to an empirical problem so testable and practically relevant conclusions can be derived from it, it is necessary to find a suitable interpretation of possibility values.

The fuzzy-set-based epistemic interpretation suggested in Zadeh (1978) merely shifts this problem to the semantically appropriate definition of fuzzy concepts, which would, for instance, have to be supplied by experts. While it can be questioned whether users can be expected to specify such fuzzy sets in accordance with the given axiomatization, the approach has been used with predefined fuzzy restrictions to define possibilistic databases (e.g. Bosc et al., 2006). A non-quantitative version of possibility degrees (on a finite range) is applied in decision support to pre-order goals and rank decision alternatives (Grabisch, 1995; Benferhat et al., 2001; Gérard et al., 2007). Similarly, Spohn's representation of epistemic entrenchment can be reformulated in terms of possibility theory (Spohn, 1990; Gebhardt and Kruse, 1998). Commonly employed quantitative notions of possibility distributions are those used for dealing with imperfect statistical information, e.g. on the basis of likelihoods (Dubois et al., 1997; Dubois, 2006), that of an upper bound for probability distributions for reasoning with imprecise probabilities (Delgado and Moral, 1987; Dubois and Prade, 1992) and the related interpretation as a plausibility measure on a consonant body of evidence[4] (e.g. Higashi and Klir, 1983; Dubois and Prade, 1999; Masson and Denœux, 2006). The latter interpretations belongs to a larger class of proposals that draw on the random sets framework to provide a meaning for possibility measures. Other representatives of that class are the theory of large deviations explored in Nguyen and Bouchon-Meunier (2003) and the context model (Gebhardt and Kruse, 1993, 1998; Borgelt and Kruse, 1998). The idea of reasoning with a consonant body of evidence is also taken up for the knowledge representation using mass distribution (Baldwin et al., 1995). However, the consonance requirement is a very strict condition, which considerably limits the applicability of those frameworks.

[4] compare Definition 4.3

Under the assumption that focal sets represent imprecise specifications, Yager (1983) has developed an information-theoretic approach, to compare random sets with regard to their compliance with the consistency/consonance assumptions. Additionally, Yager (1983) and Higashi and Klir (1983) proposed measures of specificity which allow to numerically assess the information content of possibility distributions.

4.4.2 Context Model Interpretation of Possibility

In the following I will adopt the so called *context model* interpretation of possibility described in Gebhardt and Kruse (1993, 1998); Borgelt and Kruse (1998). Using a small example I will explain the differences between the probabilistic and the possibility-based approach to modeling a partial knowledge state and explore the relation between possibility measures and random sets.

The context model aims at combining both uncertainty and imprecision about a facts into a single function. That function assigns a degree of possibility to every element a sample space Ω and is called *elementary possibility assignment* (Borgelt and Kruse, 2002). The construction of possibility assignments is directly derived from a random set representation.

With the elementary possibility assignment directly applied to individual elements of the frame of discernment, the number of parameters required to represent the possibility distribution is small in comparison to a full random set representation. Thus they are more convenient to store and operate upon than random sets. Of course that benefit does not come without a reduction in representation power. It is therefore appropriate to investigate what information may be represented using possibility assignments.

The context model is used for instance to express partial knowledge of a given process. Consider an observer, who can assign probability distributions to describe value distributions for a subset of the variables governing the process, but whose relevant knowledge about the remaining variables is limited to simple constraints for the range of values, with no preference for any of the remaining realizations. The probability distribution about the state of the first set of variables reflects uncertainty in the distinction between so called contexts, and may be an objective probabilities from a theoretical model, reflect a subjective assessment, or have been determined empirically. The contexts themselves represent, for instance, sets of physical frame conditions, that cannot be distinguished or controlled for in that particular experiment, though their general distribution is known from other types of experiments. Each context is characterized by specific constraints on the values of the remaining variables, determining a set of value combinations that are possible under that context. The existence of more than one admissible combination of values under a given context reflect imprecision, either from inability to discern occurrences or due irrelevance of further distinctions. The idea of the model is to combine the uncertainty w.r.t. *context selection* with the imprecise specification of variable values under each context.

With these interpretations the context model is employed as a basis for a quantitative assessment of possibility in terms of a *probability of logical possibility*. In a broader application of the mathematical framework, the probability measure may also be used reflect a relative weight of importance or reliability (Gebhardt and Kruse, 1998).

In order to further clarify the view of possibility proposed by the context model consider the experiment in Example 4.2.

Example 4.2. Consider three identical opaque urns (A through C) as pictured in Figure 4.1. Each of the urns contains spheres of up to tree different colors (represented by different shades), that are otherwise identical. The experiment consist of two steps. First a die is thrown by an experimenter. Depending on the number shown, an urn is selected in accordance to the mapping given by Table 4.1. In a second step one sphere is drawn from the selected urn. The color/shade of the drawn sphere determines the outcome of the experiment, so a sample space may be defined as a set $\Omega = \{\text{light,medium,dark}\}$. □

In the above experiment the result of the first step set the conditions – a context – for the second part of the experiment. Before the die is thrown, all of the outcomes in Ω are possible. But because some urns do not contain spheres of all colors, certain results may be excluded once an urn is selected. To model the experiment it is not necessary to consider all numbers that may be rolled as separate contexts. Since several outcomes lead to the same urn being selected it suffices to group them into three mutually exclusive intermediate events {⚀,⚂,⚄},{⚁,⚃} and {⚅}. Each of these events is associated with one element

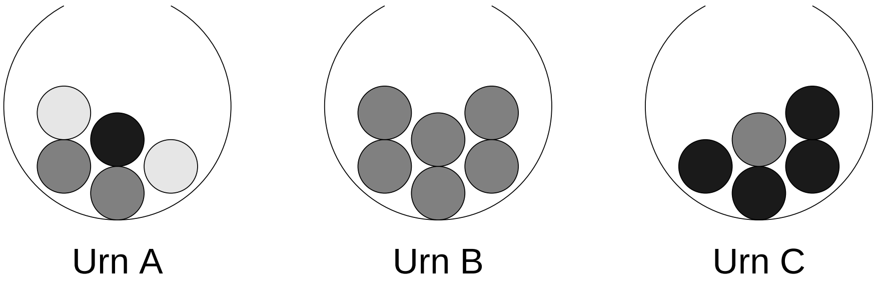

Figure 4.1: Three urns containing colored spheres

⚀, ⚂, ⚄	⚁, ⚃	⚅
urn A	urn B	urn C

Table 4.1: Mapping used to select urn after die is thrown

in a set of contexts $C = \{c_A, c_B, c_C\}$. Let us now consider an agent familiar with that experimental setup. Suppose that the die has already been thrown, but the intermediate result is not accessible to the agent, so it remains ignorant about the identity of the urn selected. Can the agent assess the situation using its limited information about the context?

To reconstruct a probabilistic assessment of the uncertainty w.r.t the outcome of the experiment as a whole, the agent would require information regarding context selection (Table 4.1) and – for each of the urns – the respective conditional probabilities of drawing a given color. Assuming a fair die and interpreting the proportions of colors as pictured in Figure 4.1, as conditional probabilities, the chain rule $P(\omega = \Omega) = \sum_{c \in C} P(\omega = \Omega \mid C = c)P(C = c)$ allows to compute a probability assignment for the final outcomes of the experiment (Table 4.2).

Picture a slightly altered situation now. This time the agent gets to know which types of spheres are contained in any the urns A–C only, but remains ignorant of their actual ratios. In this case, the information is insufficient to provide a well-founded probability assessment. Nevertheless some useful constraints are specified. For instance, whenever urn B is selected (context c_B), the outcome is guaranteed to be "‘medium”. Similarly, in the context of urn C being selected (c_C), "light" can be ruled out as a result of the experiment. For every context a set of logically possible candidate outcomes is generated. According to the context model interpretation the possibility of an outcome in any given situation reflects the *probability that a context applies, in which the outcome is logically possible.* Applying this notion to the experiment from Example 4.2 yields the elementary possibility assignment given in Table 4.3.

shade ω	elementary probability $p(\omega)$
medium	$\frac{1}{2} \cdot \frac{2}{5} + \frac{1}{3} \cdot 1 + \frac{1}{6} \cdot \frac{1}{5} = \frac{17}{30}$
dark	$\frac{1}{2} \cdot \frac{1}{5} + \frac{1}{6} \cdot \frac{4}{5} = \frac{7}{30}$
light	$\frac{1}{2} \cdot \frac{2}{5} = \frac{6}{30}$

Table 4.2: Probability distribution for the urn example

shade ω	elementary possibility $\pi(\omega)$
medium	$\frac{1}{3} + \frac{1}{2} + \frac{1}{6} = 1$
dark	$\frac{1}{2} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$
light	$\frac{1}{2} = \frac{1}{2}$

Table 4.3: Elementary possibility assignments for the modified urn example computed via the probability of compatible contexts

In the axiomatic approach to possibility discussed in the previous section a possibility distribution π defines a possibility measure Π Equation 4.12. That measure allows to assign possibility degrees to the events from 2^{Ω}. For the example it can easily be verified that the computed degrees of possibility are consistent with the context model interpretation as a probability of logical possibility. However this is due to special way, in which the example was constructed, and does not hold in general (compare Subsection 4.4.5).

It should also be remarked that possibility assignments in the context model depend on the set of contexts distinguished for their construction. That set of contexts is usually determined by the agents ability to make a probability assessment for the individual contexts. The contexts are in turn identified by means of their associated candidate sets which reflect the constraints derived from frame conditions, but also the agent's background knowledge and the latter's capacity of observation. Thus, even for identical physical frame conditions two agents may obtain different sets of candidates and consequently different elementary possibility assignments. Depending on the agents capacity, coarser of finer distinctions between contexts may be applied. For instance, suppose that instead of the mapping from Table 4.1 a second agent's knowledge consist merely of the statements

- If an even number faces up then the outcome is medium or dark.
- If an odd number faces up any of the three outcomes light,medium or dark is possible.

That agent might distinguish between the context even/odd number faces up only, which nets the less informative possibility assignment.

$$\pi_2 : \quad \text{light} \mapsto 0.5, \quad \text{medium} \mapsto 1, \quad \text{dark} \mapsto 1.$$

Since a possibility degree in the context model refers to (the absence of) weighted constraints of possible attribute values, an elementary possibility assignment reflects *negative information.* The second agent, being ignorant of the distinction between contexts c_B and c_C, may not exploit the additional restriction in context c_B leading to the less informative possibility assessment π_2. In general, if a distribution π_1 is at least as specific as another possibility distribution π_2 if the condition $\forall\omega \in \Omega : \pi_2(\omega) \geq \pi_1(\omega)$ holds. Interestingly, unlike with the probabilistic framework, the context model interpretation remain applicable even when the outcome is a set-valued attribute, that may take several values from their domain concurrently.

4.4.3 Formalization

In the previous subsection it was explained, how in the context model a possibility degree is viewed as a result of uncertainty w.r.t. context selection (modeled by means of probability theory) and context-dependent logical constraints expressed via candidate sets. Obviously such a structure lends itself well to a random set representation.

Indeed, the logically possible outcomes in any given context may be seen as focal sets of a random set. The underlying probability measure P functions as a description of the agents uncertainty about which of the contexts applies to the situation at hand. The basic possibility assignment induced by that random set may then be defined according to (Kruse et al., 1994; Borgelt and Kruse, 2002):

Definition 4.8. *Consider a set of contexts C that is the carrier of a probability space $(C, 2^C, P)$ associated with a random set $\Gamma : C \to 2^\Omega$, such that the focal sets $\Gamma(c)$ reflects possible outcomes of an experiment under the uncertain contexts $c \in C$. The possibility distribution induced by Γ is the mapping*

$$\begin{array}{rrcl} \pi: & \Omega & \to & [0,1] \\ \forall\omega \in \Omega: & \omega & \mapsto & P\left(\{c \in C \colon \omega \in \Gamma(c)\}\right). \end{array}$$

This definition formalizes the interpretation of possibility given in the previous subsection and permits to view an elementary possibility assignment in the context model as the *one-point coverage of a random set* (compare also Dubois and Prade, 1982). The main difference compared to the direct representation using a random set is, that the basic possibility assignment disregards the connection of candidates to their original context. For this reason different random-set representations may induce identical possibility assignments (Figure 4.2). In the following I will only distinguish between contexts that are mapped to different focal sets. This merger is justified by the identical effect of these contexts within the modeled piece of the world. Due to their comparatively small storage requirements, possibility distributions are applied as information-compressed representations of random sets when the association with the original contexts is not relevant to the application (Gebhardt and Kruse, 1998).

context	prob.	focal set				
		ω_1	ω_2	ω_3	ω_4	ω_5
c_1	$\frac{1}{3}$	●	●			
c_2	$\frac{1}{2}$			●	●	
c_3	$\frac{1}{6}$		●		●	●

context	prob.	focal set				
		ω_1	ω_2	ω_3	ω_4	ω_5
c_1	$\frac{1}{3}$	●		●		
c_2	$\frac{1}{2}$		●		●	
c_3	$\frac{1}{6}$			●	●	●

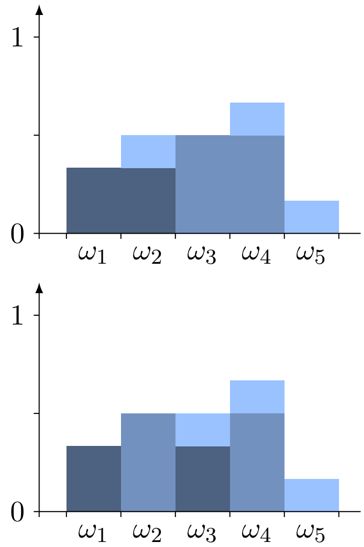

Figure 4.2: Two random sets with identical one-point coverages (shading indicates contributions from individual contexts)

4.4.4 Relation to Random Set Properties

Definition 4.8 permits to link some properties of possibility distributions to properties of the random sets from which they were generated. In order for a possibility distribution to be normalized in the context model at least one element of its domain Ω must exhibit a possibility degree of 1. This means that the element has to be logically possible in all contexts with positive probability. Therefore, the underlying random set must fulfill the consistency criterion previously given by Definition 4.2. Conversely, every consistent random sets induces a normalized possibility distribution. Figure 4.3 demonstrates the connection between properties of random sets and induced one-point coverages.

If we further restrict the admissible random sets to those with consonant focal sets, then the elementary possibility assignment suffices to reconstruct the original random set representation. Because the focal sets are nested, the subset relation defines a total ordering on those sets (compare Definition 4.3). Consequently, for each element $\omega \in \Omega$ with $\pi(\omega) > 0$ there is a minimal focal F set such that the element is covered by all focal sets at least as large as F but none

context	prob.	focal set				
		ω_1	ω_2	ω_3	ω_4	ω_5
c_1	$\frac{1}{3}$	●	●		●	
c_2	$\frac{1}{2}$		●	●	●	
c_3	$\frac{1}{6}$		●			●

context	prob.	focal set				
		ω_1	ω_2	ω_3	ω_4	ω_5
c_1	$\frac{1}{3}$	●	●	●	●	
c_2	$\frac{1}{2}$		●		●	
c_3	$\frac{1}{6}$		●			

context	prob.	focal set				
		ω_1	ω_2	ω_3	ω_4	ω_5
c_1	$\frac{1}{3}$	●				
c_2	$\frac{1}{2}$			●		
c_3	$\frac{1}{6}$				●	

context	prob.	focal set				
		ω_1	ω_2	ω_3	ω_4	ω_5
c_1	$\frac{1}{3}$	●	●			
c_2	$\frac{1}{2}$			●	●	
c_3	$\frac{1}{6}$					●

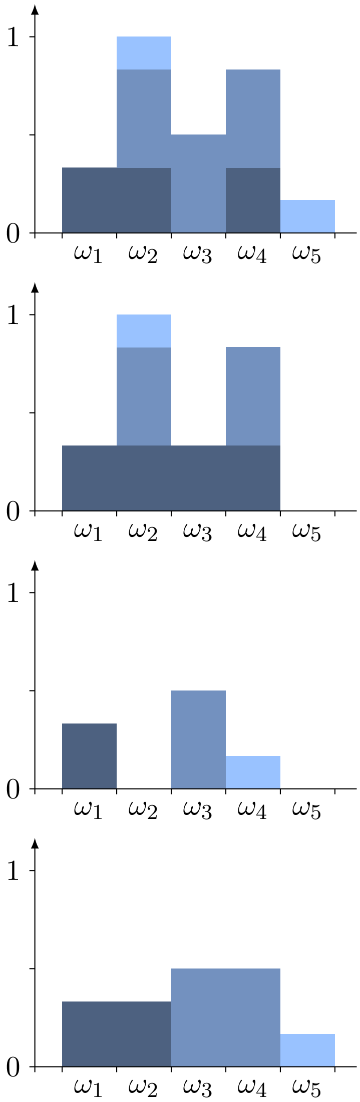

Figure 4.3: Examples of random sets and their respective one-point coverages (from top to bottom: consistent, with consonant focal sets, focal sets are singletons, disjoint focal sets)

of the smaller ones. Moreover, contexts are required to have distinct focal sets, so every particular focal set F contains at least one element ω_F not covered by any smaller focal set.

The probability for the largest focal set may be recovered directly as the minimum of the individual possibility degrees of its elements. Once the probability of the contexts associated with all larger focal sets is known, the same argument can be applied to deduce the context probability of the second largest and then increasingly smaller focal sets. To compute the probability $P(\Gamma^{-1}(F))$ of the context related too a focal set F the aggregated probability of the contexts $P(\{c \in C\colon \Gamma(c) \supset F\})$ is deducted from the possibility $\pi(\omega_F) = \min_{\omega \in F} \pi(\omega) = P(\{c \in C\colon \Gamma(c) \supseteq F\})$ for one such least possible element $\omega_F \in F$.

Finally it is remarked that if all focal sets are singletons, than the resulting one-point coverage is a probability distribution. Likewise, the probabilistic approach is applicable whenever focal sets are mutually disjoint. In that case the underlying frame of discernment can simply be replaced with a coarsened version that drops the distinction between elements of the same focal set. Unfortunately, some operations of the axiomatic possibilistic framework are inconsistent with the probabilistic view adopted under a random set-based interpretation, as will be demonstrated below.

4.4.5 The Possibilistic Aggregation Problem

Although the random set-based interpretation applied in the context model provides semantics for elementary possibility assignments, possibility measures over subsets of a frame of discernment must still be addressed. Applying the intuition proposed by the context model, the possibility of an event E should reflect the combined probability of contexts, under which an event $E \subseteq \Omega$ is logically possible. A formalization of that idea results in the re-definition of a possibility measure in the sense of the context model (Borgelt and Kruse, 2002):

Definition 4.9. *Consider a set of contexts C that is the carrier of a probability space $(C, 2^C, P)$ associated with a random set $\Gamma : C \to 2^\Omega$, such that the focal sets $\Gamma(c)$ reflects possible outcomes of an experiment under the uncertain contexts $c \in C$. The possibility measure induced by Γ is the mapping*

$$\begin{array}{llcl} \Pi : & 2^\Omega & \to & [0,1] \\ \forall E \subseteq \Omega : & E & \mapsto & P\left(\{c \in C\colon E \cap \Gamma(c) \neq \emptyset\}\right). \end{array}$$

In order for an event E to be possible under the a given context it suffices if one of the elementary outcomes in E is logically possible. Thus E will be possible under *any* context that enables at least one of the elements $\omega \in E$. This view formally coincides with the concept of a plausibility measure (Equation 4.5), which was discussed earlier.

The advantage of using Definition 4.9 as the basis of a possibility measure is that the possibility assignments reflect upper bounds for probabilities, yet do not require observations of the same quality as would be required for a full probabilistic assessment. This property is useful, whenever variables are difficult or costly to observe, become available only at a late stage of an ongoing process under observation or rely on extensive collections of sample cases for acceptable distribution estimates. For the same reason it is sometimes appropriate to drop the distinction with respect to a subset of variables, that are irrelevant to the investigated situation. The the latter case one chooses to operate with statements that refer to a coarsened frame of discernment.

Example 4.3. During admission to hospital for a planned procedure a patient tests positive for a bacterial pathogen. The infection is at an early stage and the initial analysis does not reveal the actual strain involved. Results of a detailed analysis will not be available for several days. Patients will almost never be infected with more than one strain at a time and the distribution of strains attributed to recent infections in the area is as follows:

Strain	Strain A	Strain B	Strain C
proportion of infections	0.75	0.2	0.05

The medical literature lists possible progressions and associated transmission risks for each of the strains:

Strain A:

infectious-	symptoms		
ness	none	mild	serious
low	●	●	
high	●		

Strain B:

infectious-	symptoms		
ness	none	mild	serious
low			
high	●	●	●

Strain C:

infectious-	symptoms		
ness	none	mild	serious
low		●	●
high			

Setting the unknown strain as a context attribute and applying the context model to summarize the above pieces of information yields a possibilistic assessment:

infectious-	symptoms		
ness	none	mild	serious
low	0.75	0.80	0.05
high	0.95	0.20	0.20

Because the patient was isolated before reaching an infective stage, infectiousness is not relevant for the assessment. Dropping the distinction by that attribute, we observe that asymptomatic progression is possible for infections with strain A and B. Moreover, while mild symptoms are possible for all strains, only the strains B and C have potential to cause serious symptoms in the patient. The application of Equation 4.9 with the now coarsened frame of discernment yields the one-dimensional possibility distribution:

symptoms		
none	mild	serious
0.95	1.0	0.25

□

Example 4.3 demonstrates the effect of coarsening a frame of discernment on a possibility assignment in the context model. If some attributes defining a frame of discernment are unobserved, irrelevant to the question at hand or have a domain that is insufficiently discerned by the observation a modeler can choose to merge subsets of an original frame of discernment and view them as single states. With the distinction of the original elements abandoned, that new states are logically possible in *any* of the contexts that permit *at least one* of its constituent element (those distinguished only w.r.t. the original frame of discernment). Hence, the possibility degree induced w.r.t. an element of the resulting, coarser frame is equivalent to assigning a degree of plausibility to their corresponding subsets of the original, finer frame (cf. Equation 4.5. Any event discerned in the coarser frame will also be discernible in the finer one. The resulting possibility assignments that are compatible with the formal requirements for plausibility measures in the random set framework.

Unfortunately, the axiomatic approach to possibility and its mechanism for extending a possibility distributions to compute a possibility measure do not meet this compatibility criterion. According to Equation 4.12 the possibility of any nonempty subset E of Ω is simply fixed as the largest possibility degree assigned to any individual element in E. In the diagnosis example 4.3 this would result in underestimating the possibility of symptomatic progression.

Phrased in the terminology of the context model, the aggregation rule of the axiomatic approach, attributes the probability that the context in effect allows for the maximal possible element in E as a degree of possibility $\Pi(E)$ to the whole subset $E \subseteq \Omega$ claiming:

$$\Pi(E) = \max_{\omega \in E} \pi(\omega) = \max_{\omega \in E} P(\{c \in C \colon \omega \in \Gamma(c)\}). \tag{4.15}$$

Yet, possibility measures in the sense of the context model do not in general exhibit the maxitivity property from Definition 4.6.

The difference is highlighted when looking at the sequence of operations that leads form the random set representations reflecting context selection probabilities and their associated permissible outcomes to the marginal possibility assignment: The context model interpretation requires to project the sets of outcomes to the target attribute set first and only then aggregate context probabilities for elements covered by the projection. In contrast computation via the maximum involves aggregation of context probabilities on the *original* domain, followed by a projection of the resulting possibility assignments. Yet, as demonstrated by Example 4.3 the two operations are not distributive. Hence, Equation 4.15 is not applicable.

This presents us with the choice to restrict the model to the case of consonant focal sets, for which aggregation is consistent with the maxitivity axiom, or to dismiss the maxitivity axiom altogether in order to retain a meaningful interpretation of numerical possibility degrees in the more general setting.

First we remark that the consonance assumption for an application context requires justification in the form of background knowledge about the relation of contexts. If such constraints are available, however, a more direct approach using a parametric model based on probability theory would usually be preferable to reasoning with possibility distributions. We will therefore pursue the second option.

As a consequence of our dismissal of the maxitivity criterion, a possibility measure may not be computed from the elementary possibility assignments alone. As a one-point coverage of a random set, the possibility assignment accounts for the probability mass of the contexts compatible to individual elements in a subset $E \subseteq \Omega$ only. But unlike the full random set representation, the more compact one-point-coverage does not retain the separation of contributions from different contexts when $|E| > 1$. If the possibility of an event cannot be computed using the maximum operator, can we at least determine bounds for this possibility from the elementary possibility distribution? Obviously, the elementary possibility assigned to a maximally possible element $\omega_i \in E\colon \forall \omega \in E \pi(\omega_j) \geq \pi(\omega)$ constitutes a lower bound for the possibility of E. Any context that permits the "most possible" element of E also permits the event E as a whole. On the other hand, $\Pi(E)$ potentially includes additional probability mass from contexts associated with focal sets that cover elements of E other than ω_i, so $\Pi(E) \geq \max_{\omega \in E} \pi(\omega)$. To find an upper boundary of $\Pi(E)$ one must consider a compatible random set such that the probability mass associated with focal sets with E is maximal for a given one point coverage. A straightforward approach to construct such a random set is to distribute possibility assignments to the elements of E, to separate context whenever possible, permitting overlap only to ensure the total mass of all involved contexts remains limited to unity.

Summarizing these considerations, the context model interpretation demands that Equation 4.15 is replaced by the weaker condition

$$\max_{\omega \in E} \pi(\omega) \leq \Pi(E) \leq \min \left\{ \sum_{\omega \in E} \pi(\omega), 1 \right\}. \tag{4.16}$$

With only the weak link between joint and marginal possibility distributions provided by the above inequality the capacity of joint possibility distributions to serve as knowledge representations is limited. I will refer to this limitation as the *possibilistic aggregation problem*.

4.4.6 Maxitivity and Consonance

The problem of possibilistic aggregation may be resolved for the case of consonant, i.e. nested focal sets. For random sets with the consonance property all focal sets overlapping with an arbitrary set $E \subseteq \Omega$ are guaranteed to have a common intersection. It is this intersection, that necessarily contains the maximally possible elements of E. This argument allows to reconcile Equation 4.15 with the idea of a plausibility measure:

$$\begin{aligned} \max_{\omega \in E} \pi(\omega) &= \max_{\omega \in E} P(\{c \in C \colon \omega \in \Gamma(c)\}) \\ &= P(\{c \in C \colon \Gamma(c) \cap E \neq \emptyset\}) = \Pi(E) = \mathrm{Pl}(E). \end{aligned} \tag{4.17}$$

It can also be demonstrated, that in the maximum of elementary possibility degrees are not necessarily plausibility measure: For instance, using the possibility distribution induced from the consistent random set from Figure 4.3 (topmost diagrams) and $E = \{\omega_4, \omega_5\}$ the maximal value of the elementary possibility assignment $\max_{\omega \in E} \pi(\omega)$ is only $\frac{2}{3}$ whereas the plausibility measure yields $\mathrm{Pl}(E) = 1$.

More generally: for any random set with non-consonant focal sets there exists a pair of contexts $c_1, c_2 \in C, P(\{c_1\}) > 0, P(\{c_2\}) > 0$, such that neither $\Gamma(c_1) \subseteq \Gamma(c_2)$ nor $\Gamma(c_2) \subseteq \Gamma(c_1)$ (Equation 4.3). If we choose $E = (\Gamma(c_1) \cup \Gamma(c_2)) \setminus (\Gamma(c_1) \cap \Gamma(c_2))$, then the resulting set contains elements of both $\Gamma(c_1)$ and $\Gamma(c_2)$. Thus $\Pi(E)$ draws on the probability mass of both contexts. Yet, since the intersection $\Gamma(c_1) \cap \Gamma(c_2)$ is excluded, there is no single element in $\omega \in E$ that is enabled under both contexts. Because any other contexts under which an individual element of $\omega \in E$ is logically possible also contribute to $\Pi(E)$, it follows that $\max_{\omega \in E} \pi(\omega) < \Pi(E)$.

This argument limits the maximum-based aggregation operator in the axiomatic possibility theory to distributions induced from consonant random sets. Although this limitation was already recognized in (Dubois and Prade, 1988a), the restrictive consonance requirements still present mayor limitation to proposed applications of the framework.

4.4.7 Discussion of Possibility Theory

Possibility distributions in the sense of the context model provide easily computed, interpretable and compact summary of random sets. Even though possibility distributions may not reflect the relation between contexts, the representation preserves practically relevant aspects of random sets. For that reason the concept of possibility proves useful for the presentation or visualization of static distributions over sets. The arguably most substantial limitation of possibilistic information representations using the context model, however, is due to the lack of a well-defined aggregation operator. This effectively restricts purely possibilistic models to elementary events and a single fixed frame of discernment at a time.

Although restricting the application of possibilistic models to consonant focal sets, would ensure the semantically meaningful aggregation of possibility measures via the maximum operator, the assumption, that unrelated contexts should produce nested focal sets cannot be reconciled with the goal of developing a generally applicable information-compressed representation of random sets. For instance with random sets based on empirical probability, it is hard to see why the observed sets should exhibit the consonance property. If, however the application context provides such constraints the possibilistic approach permits simple and efficient tools for reasoning.

While it might be suggested to employ the possibility bounds identified in Equation 4.16 as a basis of a knowledge model, at least two arguments may be brought up against this proposal. Firstly, the possibility degree of any event already constitutes an upper boundary of a probability degree. A lower bound of such a possibility degree would be difficult to interpret and hardly convey useful information about the occurrence of an event. Secondly, unless almost all focal sets associated with common contexts are singletons, the upper boundary based on the bounded sum of elementary possibility degrees would not be very informative, almost always taking the value 1. In other words, the transfer of information between different frames would not be sufficiently information-efficient to be of practical utility in a knowledge model.

4.5 Conclusions

Having investigated extant techniques to represent distributions over set-valued concepts, we may now assess those approaches w.r.t their suitability to the development of an extended framework that allows to construct such models from inhomogeneous data sources and convert the resulting distributions between different frames of discernment.

The most direct approach using random set applies the conventional probabilistic framework to set-concepts by considering probability distributions over power sets of an underlying set-universe (i.e., the frame of discernment). The framework may draw on the well-founded probabilistic reasoning methods to combine information from different sources. As outlined in Section 4.2 its main drawback when applied to modeling knowledge is the insufficient scalability of the approach due to the exponential growth of the sample space, when the resolution of the frame of discernment is increased. The large number of degrees of freedom for the resulting distributions is detrimental both in regard to its storage requirements and, more importantly, due to the difficulties of reliably estimating the numerous model parameters from limited data.

In contrast to that, an approach based on the Dempster-Shafer framework (Section 4.3) lacks the means to encode the statistical interaction between attributes or information sources. While the Dempster-Shafer theory complies with the more general random set framework with regard to the transfer of information from finer to coarser frames of discernment, its simple rule for combining information must be rejected for the knowledge representation setting at hand due to the lack of justification for the postulated independence assumptions. Moreover, as with random sets, the need to explicitly represent focal sets seriously restricts the application of the approach for larger attribute domains, as the number of potential focal sets grows exponentially with the cardinality of the underlying set domain.

Finally, the possibilistic framework, which was investigated in Section 4.4, permits to capture some relevant pieces of information about random sets in a compact representation. Yet, unless consonance is presumed, the maximum-based aggregation method of as standard possibility theory fails to conform with the semantics of the underlying context model interpretation. Due to the lack of a generally applicable aggregation operator, possibility theory as discussed in the preceding section may not serve as the basis of a frame-spanning representation for set-attributes.

Nevertheless, each of the above approaches can contributes to the development of a frame-spanning knowledge representation for set-based concepts. The random set model, although limited in its practical applicability due to scalability issues, provides a reference for comparison via test cases and a starting point for the development of a simplified representation framework.

The inspiration that may be drawn from Shafer's approach is the introduction a heuristic component, which may in part takes over the role of the potentially inassessable probabilistic interaction structure of a direct random set representation. However, to avoid inconsistencies due to unjustified assumptions, the application of heuristics in models should be limited. For instance, it may be admissible to supplement missing detail to a model otherwise built upon a solid foundation of empirical estimates, if the consequences of these assumption do not propagate to pollute empirically accessible parameters.

The arguably most interesting feature of the possibilistic framework for the context model is its use of one-point coverages for summarizing random-set representations. Depending on the particular concepts formalized using the focal sets, the one-point coverage has an intuitive interpretation, e.g. as a degree of evidential support or as a measure of an elements compatibility with soft constraints (Hüllermeier, 2003). The key to their successful application in a knowledge model is to overcome the limitations of the possibilistic aggregation operator. To that end it is necessary to provide algorithms and data structures for relating one-point coverages induced w.r.t. different frames of discernment to each other.

In the following chapter I will introduce a condensed representation for multivariate distributions obtained from set-valued data, which was developed based on these conclusions. It adopts one-point coverages and the context model interpretation as its main tools for presenting information, but anchors their distributions on a probabilistic framework and a partitioning of the power set of a frame of discernment. Specialized data structures serve to connect joint and marginal distributions and can be used to to implement reasoning on structured domains. The proposed representation allows to recover the probability of singleton events and one-point-coverages for each element of the frame of discernment.

5 Condensed Representation for Set-Attribute Distributions

In the previous chapters I have argued for developing a new frame-spanning knowledge representation for set-valued attributes, discussed desired properties such a framework and elaborated an outline for implementing it. The present chapter serves to further elaborate the mathematical foundations, data structures and formal mechanisms that enabled the realization of these ideas.

Following the introduction of some conventions and helpful notations, Section 5.2 presents a condensed representation for set-valued attributes (Rügheimer, 2007). While conceptually drawing on the random-set approach, the proposed representations avoids the difficulties linked to distributions on power sets by grouping multi-valued outcomes. This permits to strike a balance between practical considerations, such as scalability and the availability of data for parameter estimation on one hand, and accuracy on various levels of detail on the other hand. The approach is later extended to multi-dimensional domains in Section 5.3 before being complemented with an investigation into operations for conditioning condensed set-valued distributions and for computing their marginal distributions. The results are then integrated into the general framework of Graphical Models to achieve a compact representation that supports efficient reasoning.

Section 5.4 deals with the application of the condensed distributions to hierarchically structured attribute domains. It is shown that condensed distributions provide the means to extend such structures to adapt them to the case of multi-label instantiations. The resulting data structures and associated mechanisms are well suited to modern knowledge representation and data analysis tasks. When projecting onto ontologies used in information retrieval or bioinformatics, for instance, condensed distributions provide meaningful operations for relating and mapping annotations or compare their distributions on multiple levels of detail while taking into account the structure of underlying term relations. Because the model uses a decomposition of distributions, it permits the combination of scarce case-specific observation with a lager body of generic distribution information about reference cases.

5.1 Conventions and Notation

To enhance the presentation of subsequent sections and facilitate their reception, it is useful to stipulate some conventions. I will designate attributes with capital letters from the beginning of the alphabet, whereas letters from the second half are used for sets. Resuming the notation introduced in chapter 2, the superscripts $*$ and $\diamond$ indicate set-valued attributes and condensed set-valued attributes[1] respectively. The same superscripts are used to mark the associated distributions. When applied to symbols that denote a distribution, subscripts indicate the set of attributes or base attributes over whose joint domain the distribution is defined. The same scheme is applied to symbols for frames of discernment. Furthermore, subscripts are used for indices, e.g. when enumerating the elements of an attribute domain.

Although, in a strict sense, proportions and empirical probability refer to an underlying set of objects O, I will treat the induced distributions over attribute domains as results of random processes in their own right. Thus, when dealing with distributions, attributes are represented as random variables. For instance, I write $P(\omega)$ or $P(A = \omega)$ as a shortcut for $P(\{o \in O\colon A(o) = \omega\})$. The justification for this abstraction is that the original set of objects is not part of the model itself and of concern only on the interpretation level. For similar reasons, I will not exclude the empty set from the admissible values of a set-valued attribute, as the question of whether it should be permitted or not is related to the applied interpretation rather than the formal representation itself. Although for the proposed versions of the framework, no special precautions are taken to ensure high precision for recovering the probability mass assigned to the empty set outcome[2] from the condensed representation, the proposed data structures are easily modified by an additional field to better support interpretations that attribute a special role to the empty-set outcome.

5.2 Condensed Representation

As demonstrated in Subsection 2.3.3 the probability (density) distribution p over an attribute domain $\Omega_{\{A\}}$ carries all the information required to recover a probability measure P over that domain. Obviously, this property is independent of the interpretation or type of elements forming the sample space. Hence the above argument fully applies to a set-valued attribute A^* taking values from $2^{\Omega_{\{A\}}}$, in which the underlying probability distribution p^* is defined over the power set

[1] to be introduced in Section 5.2

[2] Note the difference between the empty-set outcome $\{\emptyset\}$ and the impossible event $\emptyset$

(i.e., the set of all subsets) of $\Omega_{\{A\}}$. Each individual subset corresponds to a combination of the two possible states (presence or absence) for each element of the carrier set. Thus the number of elements in $2^{\Omega_{\{A\}}}$ is exponential in the size of $\Omega_{\{A\}}$.

Given that the "base vocabulary" of even the smallest ontologies rarely comprises less than 50 terms, (e.g., Gene Ontology: $>$ 120 in slim version, $>$ 32000 in full version) several issues must be addressed to develop useable computational models for distributions over sets. Because the space of possible set instantiations grows exponentially in the number of admissible elements, memory requirements become an obvious obstacle to a direct representation (Gigabyte range being reached for underlying sets with around 30 items). The same considerations apply to computational resources required for operating on such distributions. But even if all values can be represented in memory and sufficient computational capacity is available, the problem of providing estimates of these often very small probabilities from observed data remains. This however would requires unrealistically large samples (see, e.g. Wasserman, 2006).

Fortunately, probability estimates are rarely required on the level of individual set-valued outcomes. Instead, applications may draw on certain summaries of value distributions to provide useful decision criteria or even comparisons between sets of observations. Probabilities of singletons and the probability of coverage of elements by set-valued events (one-point coverages) are of particular interest in that respect. In fact, for application, such as mass spectrometry, one point coverages are correspond to only measurements directly obtained in experiments with other variables being inferred by comparison with results from to in-silico models. By focusing on a small set of relevant parameters the proposed condensed representation achieves a compact, scalable summary of statistical information regarding set attributes.

The approach is based on partitioning the set of the subsets of a sample space Ω and a mapping of set-distributions to a probability/possibility distribution over the condensed domains. In the formalization of that approach a special attribute value is introduced to label outcomes that are multi-valued w.r.t. a frame of discernment Ω. In the standard variant of the model the same label is applied to empty set outcomes if it is enabled for the setting. For simplicity, let us initially consider the case of a frame of discernment based on a single attribute:

Definition 5.1. *Let Ω denote a frame of discernment. Furthermore let $\omega^{\diamond}$ be a special symbol uniquely associated with and not already contained in Ω. Consider* *a mapping σ from the set of subsets 2^Ω to the* **extended set universe** $\Omega \cup \{\omega^\diamond\}$

$$\begin{aligned} \sigma: \quad 2^\Omega &\rightarrow \Omega \cup \{\omega^\diamond\} \\ \forall S \subseteq \Omega: \quad \sigma(S) &= \begin{cases} \omega & \textit{if } S = \{\omega\}, \omega \in \Omega, \\ \omega^\diamond & \textit{otherwise.} \end{cases} \end{aligned}$$

I call σ the **set reduction mapping** *w.r.t. Ω.*

It is easily verified that σ preserves the distinction between singleton elements of 2^Ω, but groups the multi-valued outcomes in a separate class. Consider now a set-valued attribute A^* taking values from 2^Ω. Using the the set reduction mapping, it is now possible to define an condensed set-valued attribute $A^\diamond$ that is linked to the values of A^*:

Definition 5.2. *Let A^* be a set-valued attribute $A^* : O \rightarrow 2^\Omega$. Additionally let $\sigma : 2^\Omega \rightarrow \Omega \cup \{\omega^\diamond\}$ denote the set reduction mapping w.r.t. Ω. The* **condensed set-valued attribute $A^\diamond$ induced by** A^* *is a mapping:*

$$\begin{aligned} A^\diamond: \quad O &\rightarrow \Omega \cup \{\omega^\diamond\} \\ \forall o \in O: \quad o &\mapsto \sigma(A^*(o)). \end{aligned}$$

The relation between the attribute domain conveyed by the set reduction mapping is illustrated in Figure 5.1. The underlying set Ω is referred to as the *basic domain* of the condensed set-valued attribute $A^\diamond$ (written $\Omega = \mathrm{bdom}(A^\diamond)$).

Per Definition 5.2 the values of $A^\diamond$ depend directly on the values of A^*. Consequently a probability distribution p^* over $\mathrm{dom}(A^*)$ induces a probability distribution $p^\diamond$ over $\mathrm{dom}(A^\diamond)$, which summarizes p^*.

$$\begin{aligned} p^\diamond(\omega) &= P^*(\{S\colon \sigma(S) = \omega\}) \\ &= P^*(\sigma^{-1}(\omega)) \\ &= \begin{cases} p^*(\{\omega\}) & \text{if } \omega \in \mathrm{bdom}(A^\diamond), \\ \sum\limits_{S \in \mathrm{dom}(A^*), |S| \neq 1} p^*(S) & \text{if } \omega = \omega^\diamond. \end{cases} \end{aligned} \tag{5.1}$$

The function $p^\diamond$ is called a *condensed probability distribution* One advantage of defining P^* via the pre-image of ω under σ is the abstraction from the underlying set of objects. This allows to separate the interpretation layer of the model from the mathematical tools for reasoning. Moreover, the abstraction permits to view

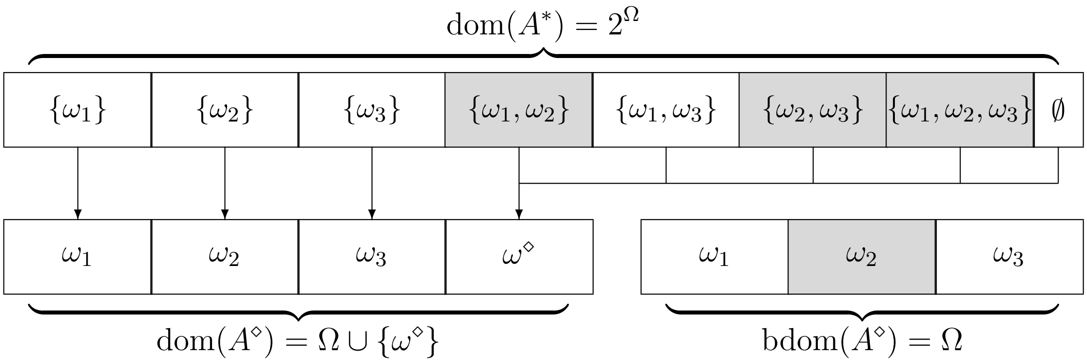

Figure 5.1: Domains of a set-valued attribute A^*, the induced condensed set-valued attribute $A^\diamond$ and underlying basic domain Ω. Arrows indicate the set reduction mapping w.r.t Ω. Shaded elements of $\mathrm{dom}(A^*)$ mark multi-valued outcomes covering ω_2.

attributes as random variables, when the focus is on distributions of attribute values rather than on individual objects in O. Specifically, a set-valued attribute A^* will be viewed as a random set.

It is remarked, that for any element $\omega \in \mathrm{bdom}(A^\diamond)$, the value $p^\diamond(\omega)$ quantifies not a probability of occurrence, but rather the probability of observing ω as the *only* element in a set-valued outcome of $A^*(\omega)$. The probability mass originally associated with multi-valued outcomes $S\colon S \in \mathrm{dom}(A^*), |S| > 1$ or with the empty-set outcome is assigned to a surrogate attribute value $\omega^\diamond$ in the condensed probability distribution. This approach leads to two immediate benefits: Firstly, since $p^\diamond$ is still a probability distribution, well established operations of the probabilistic framework like conditioning and marginalization can be employed with this representation. In addition to that, Definition 5.2 can be applied to estimate the condensed probability distributions directly from observations of set-outcomes, i.e. without prior computation of the distribution p^*. It is also remarked any that probability distribution over a single-valued attribute $A, \mathrm{dom}(A) = \mathrm{bdom}(A^\diamond) = \Omega$ is subsumed in the representation without loss of information. In that case the corresponding condensed probability distribution is simply given by $p^\diamond(\omega) = p(\omega)$ and $p^\diamond(\omega^\diamond) = 0$.

So far the discussion on a compact representation of distributions related to set-valued attributes was focused on the probabilities of singleton outcomes. Due to the equality $p^*(\{\omega\}) = p^\diamond(\omega)$, the probability of such outcomes may be read directly from the distributions for the corresponding condensed set-valued attributes. To support the reconstruction of one-point coverages, however, a richer representation is required.

Given a probability distribution p^* for a set-valued attribute A^* taking values from 2^Ω, the one-point coverage of individual elements $\omega \in \Omega$ is computed as follows:

$$\forall \omega \in \Omega : \mathrm{opc}(\omega) = P^*(S : S \subseteq \Omega \wedge \omega \in S) = \sum_{\substack{S \subseteq \Omega \\ \omega \in S}} p^*(S). \tag{5.2}$$

For each $\omega \in \Omega$ one element of the sum in the right-hand expression of Equation 5.2 is obtained directly from the distribution $p^\diamond$ of the induced condensed attribute $A^\diamond$. For $S = \{\omega\}$ the summand is recovered due to the equality $p^*(S) = p^*(\{\omega\}) = p^\diamond(\omega)$. Moreover $S = \emptyset$ needs not be considered, when computing one-point coverages (no element can be covered by the empty set). To represent the joint contributions from all other subsets of Ω, the latter are encoded as proportions relative to $p^\diamond(\omega^\diamond)$ (called coverage factors):

Definition 5.3. *Let p^* denote a distribution linked to a set-valued attribute (A^*) over 2^Ω and $p^\diamond$ the distribution over the domain $\mathrm{dom}(A^\diamond)$ of an induced condensed set-valued attribute $A^\diamond$ obtained by applying equation 5.1. Then the* **coverage function** *$c^\diamond$ relative to multi-valued outcomes of A^* is defined as a function*

$$\begin{aligned} c^\diamond : \quad \Omega &\rightarrow [0,1] \\ \omega &\mapsto \begin{cases} \frac{\sum_{S \subseteq \Omega, \omega \in S, |S|>1} p^*(S)}{p^\diamond(\omega^\diamond)} & \text{if } p^\diamond(\omega^\diamond) > 0, \\ 1 & \text{otherwise.} \end{cases} \end{aligned}$$

For $p^\diamond(\omega^\diamond)$ the value $c^\diamond(\omega)$ denotes the conditional probability for ω being *contained* in an outcome, given that outcome is not a singleton. Although the contributions to the one-point coverage could have been stored directly, the representation via relative coverage factors was chosen to better support probabilistic conditioning and marginalization operations discussed in Section 5.3.

It is remarked that for $p^\diamond(\omega^\diamond) = 0$, the conditional probability expressed in the coverage factor remains undefined. The reference implementation handles this situation by assigning a fixed positive value of 1. This has no adverse effects on subsequent calculations, as those values will always be weighted with a zero probability $p^\diamond(\omega^\diamond)$. Once assigned, a zero-value is not affected by subsequent conditioning operations, so the is guaranteed to work even in a dynamic setting. An additional benefit of the above technique is that it permits to recover contributions to the one-point coverage simply by multiplying with $p^\diamond(\omega^\diamond)$.

Like the distribution $p^\diamond$, the *relative coverage factors* assigned by $c^\diamond$ can be computed directly from data. Replacing the sum in Equation 5.2 the one-point coverage may now be rewritten as

$$\forall \omega \in \Omega : \operatorname{opc}(\omega) = \sum_{\substack{S \subseteq \Omega \\ \omega \in S}} p^*(S) = p^\diamond(\omega) + p^\diamond(\omega^\diamond) \cdot c^\diamond(\omega) \tag{5.3}$$

In the following, the term *condensed distribution* is understood to refer to a tuple $(p^\diamond, c^\diamond)$ that is formed by a condensed probability distribution and the corresponding coverage function.

The advantage of the condensed set-valued attribute $A^\diamond$ and the function $p^\diamond$ and $c^\diamond$ over the full random set representation is the reduction of the number of parameters. For each attribute value only the probability for the singleton outcome and the coverage factor, that is $2|\Omega|$ values, need to be stored. For practical reasons, it is also advantageous to explicitly represent the total probability mass of multi-valued outcomes (though in principle this value could be recovered by deducting the combined probability of all singleton outcomes from unity). Thus the number of distribution parameters grows linearly in the size of the underlying base domain Ω. In contrast, a full distribution over sets considers $2^{|\Omega|}$ elements, where one parameter is redundant due to the normality condition for the probability distribution p^*.

Besides its capacity to provide information summaries, the condensed representation is well suited to model the interactions of set distributions and one point coverages between joint and marginal domains. This property allows to substitute condensed distributions for random sets in multivariate or structured domains.

5.3 Extension to the Multivariate Case

So far only distributions over single attributes have been considered. This is usually sufficient for tasks centered around processing static information, such as the comparison between of gene expression patterns between conditions or between different populations. For reasoning or dependency analysis however, it is necessary to examine the interactions across attributes. Attribute interactions are captured directly by modeling joint distributions (contingency tables) or indirectly by estimating derived properties of such distributions (e.g. correlation analysis). This section extends the condensed distribution framework to the multivariate setting. Furthermore, it proposes an analogue of the contingency table approach for obtaining multivariate versions of condensed distributions from empirical data. Following that, operations for reasoning with condensed distribution are elaborated.

5.3.1 Introduction to Tuple-Based Formalization

For clarity the condensed distribution was introduced for the one-dimensional case first. To that end a simple representation on the basis of attributes as functions defined on sets of objects was used. When working with multivariate distributions however, observations and propositions may refer not only to the complete variable set under investigation, but also to lower-dimensional projections thereof. Therefore the specification language and connected operations must clearly identify the set of variables an observation or statement refers to. Moreover the integration of data from different sources and goal-oriented presentation of results call for operations to convert information about observed instances and distributions between different sets of reference variables.

A tuple-oriented formalization of the sample space provides a convenient tool to address these issues. Although the additional layer of abstraction introduced by tuples initially calls for a slightly more complicated formalization of object properties, this is more than made up for by the unique advantages this formalization offers for the discussion of multivariate distributions. The formalism allows to approach the projection and aggregation operations used for computing and exchanging information via marginal distributions in an intuitive manner. Those operations have a central role when dealing with high-dimensional multivariate data because they allow to consider distributions on lower-dimensional subspaces rather than the variable set as a whole. The tuple-based formalization was previously used in Borgelt and Kruse (2002), where it is explicated in detail and applied to the discussion of problems in reasoning with probabilistic and possibilistic Graphical Models. This subsection recapitulates some essential definitions adapted from that source and supplements them with the extensions for the additional distribution types proposed in this thesis.

The central idea of the alternative formalization suggested by Borgelt and Kruse is to provide object descriptions not via attributes themselves, but rather indirectly by mapping to so called tuples or instantiations. The instantiations in turn are represented as functions that assign values to the elements of a selection of attributes. The attributes are thus reduced to the role of variables that are connected to a range of possible values. The tuples themselves are represent combinations of constraints on a set of entities, that is, they delineate subsets or categories of such entities.

This indirect approach allows to make explicit the set of attributes considered in the model. More importantly, it permits to define partial instantiations w.r.t. different subsets of attributes. While all these functions could also be achieved using a more “conventional’ formalization that encodes each attribute as a position in a value vector for a specified reference set, the convenience of conducting operation without the need of index transformations provides a powerful argument for switching to the function-based formalization (compare Borgelt and Kruse, 2002, p. 64).

Since the tuple-based formalization of joint attribute domains was originally used for single-valued attributes. I will start by re-iterating the definitions for that case (after Borgelt and Kruse (2002)):

Definition 5.4. *Consider a finite set X of attributes. An* **instantiation** *of the attributes in X or a* **tuple** *over X is a mapping*

$$\begin{aligned} t_X : \quad X \rightarrow & \textstyle\bigcup_{A \in X} \operatorname{dom}(A) \\ & \textit{such that} \\ \forall A \in X : \quad t_X(A) \quad & \in \operatorname{dom}(A). \end{aligned}$$

The set of all tuples over X is denoted T_X.

The definition ensures that each attribute $A \in X$ may only be mapped to elements of its original domain $\operatorname{dom}(A)$. The notation $\operatorname{dom}(t) = X$ is used to indicate that t is a tuple over X. A tuple t over $\{A, B, C\}$, which maps attribute A to a_1, B to b_2 and C to c_2 is written $t = (A \mapsto a_1, B \mapsto b_2, C \mapsto c_2)$. This is shortened to $t = (a_1, b_2, c_2)$, if an implicit order is fixed.

As remarked earlier, the formalization is particularly useful for discussing the projection operation. The practical utility of the projections is that information need not be presented with respect to a fixed frame of discernment. Projections allows to "disregard" distinctions based on attributes that are unavailable or deemed irrelevant to particular information needs. Thus the presentation of information can be adapted to suit specific tasks. To project a single tuple over an attribute set X to an attribute set $Y \subseteq X$ it suffices to apply the mappings for the attributes in Y from the tuple t_X, or formally (from Borgelt and Kruse, 2002, p. 64):

Definition 5.5. *If t_X is a tuple over a set X of attributes and $Y \subseteq X$, then $t_{X|Y}$ denotes the* **restriction** *or* **projection** *of the tuple t_X to Y. That is, the mapping $t_{X|Y}$ assigns values only to the attributes in Y. Hence $\operatorname{dom}(t_{X|Y}) = Y$, i.e. $t_{X|Y}$ is a tuple over the attribute set Y.*

In other words, the operation allows to select a relevant subset of attributes. In particular, $t_{X|\emptyset}$ yields the empty tuple $()$. For application to sets of tuples over an attribute set (i.e. relations), that definition is extended (after Borgelt and Kruse, 2002, p. 64):

Definition 5.6. *Let R_X be a relation over a set X of attributes and $Y \subseteq X$. The* **projection** $\mathrm{proj}_Y^X(R_X)$ *of the relation R_X from X to Y is defined as*

$$\mathrm{proj}_Y^X(R_X) \overset{\text{def}}{=} \{t_Y \in T_Y \mid \exists t_X \in R_X \colon t_Y \equiv t_{X|Y}\}.$$

The relevance of the projection operation results from its capability to relate and convert between set-valued instantiations with respect to different subsets of underlying attribute set X. This may be used, for instance, to simplify the notation of the decomposition rules given in Chapter 3. In the new notation, the chain rule decomposition formula applied in Bayesian networks (Equation 3.2) can be written as

$$p_X(t_X) = \prod_{A_i \in X} p_{\{A_i\}|\mathrm{pred}(A_i)}(t_{X|\{A_i\}} \mid t_{X|\mathrm{pred}(A_i)}). \tag{5.4}$$

Similarly, a modified formulation of the decomposition rule for undirected graphs used in Markov Networks may be given (compare Equation 3.3):

$$p_X(t_X) = \prod_{C_i \in \mathbf{C}} \phi_{C_i}(t_{X|C_i}). \tag{5.5}$$

As argued in Subsection 2.2.2 the set-valued descriptions with respect to groups of attributes X are not limited to Cartesian products, but may well be a more general relation over the domain of a combined attribute[3]. In the tuple-based formalism such a relation is expressed directly as a subset $R_X \subseteq T_X$. It is noted that by switching to the tuple-based formalization the observation-style attributes that directly differentiated between classes of objects in a population have been superseded by a new type of attribute. In the advanced formalization the elements of the attribute set X discern the tuples in T_X, and those tuples now serve as the exclusive interface between formal model and modeled world segment.

If the only cases considered are instantiations yielding a singleton, one trivially represent this situation by using conventional attribute with $\mathrm{dom}(A) = \mathrm{bdom}(A^\diamond)$. We call such an attribute A the *base attribute* of a condensed set valued attribute $A^\diamond$.

[3] In this point I deviate from Borgelt and Kruse (2002). The reason is, that Borgelt and Kruse employed set-instantiations in connection to a database of sample cases, that contained imprecise specifications for individual attribute values only, making a the limitation to Cartesian product a convenient choice. Yet, their equations concerning the mechanism of projection transfer to the more general case considered here.

5.3.2 Multiple Condensed Set-Valued Attributes

Having generalized the notion of set-outcomes to combined attribute domains, we can now formalize the proposed approach for generating, interpreting and and operating with the condensed representation on in a multivariate setting. Like with the one-dimensional case the condensed representation is based on mapping *sets of tuples* from T_X to instantiations $t_{X^\diamond} \in T_{X^\diamond}$ w.r.t. condensed set-valued attributes. On the formal level, the tuples $T_{X^\diamond}$ are treated just like instantiations of single-valued attributes. Each of the tuples in turn specifies a combination of the instantiations of component attributes taking values from extensions of the domains of the attributes in X.

In the one-dimensional case, the conversion from of the set-valued attribute to a regular one was achieved by grouping the non-singleton set-outcomes using the set reduction mapping (Equation 5.1). In the multivariate case, however, individual condensed set-valued attributes refer to projections of higher-dimensional relations and not necessarily to those relations themselves. To keep track of singleton elements in marginal distribution w.r.t. arbitrary subsets Y of the underlying attribute set X, a finer partitioning based on the projections of the multi-dimensional relation to the domains of individual attributes is employed. To that end the simple set reduction mapping introduced in Definition 5.1 is replaced with a more general mapping σ_X from set-valued outcomes to tuples over of condensed set-valued attributes:

Definition 5.7. *Consider a set $X = \{A_1, \dots, A_n\}$ of attributes and an associated set of condensed set-valued attributes $X^\diamond = \{A_1^\diamond, \dots, A_n^\diamond\}$ with domains* $\mathrm{dom}(A_i^\diamond) = \mathrm{dom}(A_i) \cup \{a_i^\diamond\}$ *for $i = 1, \dots, n$. The tuple-based* **set reduction mapping** *is the function*

$$\sigma_X : \quad 2^{T_X} \quad \to \quad T_{X^\diamond}$$

that assigns to a relation $S \subseteq T_X$ the tuple

$$\begin{aligned} \sigma_X(S) : \quad X^\diamond \quad &\to \quad \bigcup_{A^\diamond \in X^\diamond} \mathrm{dom}(A^\diamond) \\ \forall A_i^\diamond \in X^\diamond : \quad A_i^\diamond \quad &\mapsto \quad \begin{cases} t_{\{A_i\}}(A_i), & \text{\textit{if} } \mathrm{proj}^X_{\{A_i\}}(S) = \{t_{\{A_i\}}\}, \\ a_i^\diamond, & \text{\textit{if} } |\mathrm{proj}^X_{\{A_i\}}(S)| \neq 1. \end{cases} \end{aligned}$$

If S is a singleton $\{t_X\}$, so are all its projections and $\sigma_X(\{t_X\})(A^\diamond) = t_X(A)$ for all attributes $A \in X$. Conversely, if the cardinality of S differs from one, $\sigma_X(S)$ maps at least one of the attributes to the special symbol denoting a non-singleton outcome. That mapping σ_X is now used to relate a distribution p^* over the power set of T_X to a corresponding distribution over $T_{X^\diamond}$.

Definition 5.8. *Let* $X^\diamond = \{A_1^\diamond, \dots, A_n^\diamond\}$ *be a finite nonempty set of condensed set-valued attributes with respective domains* $\mathrm{dom}(A_1^\diamond), \dots, \mathrm{dom}(A_n^\diamond)$ *and let* $X = \{A_1, \dots, A_n\}$ *denote the set of associated base attributes, such that* $\mathrm{dom}(A_i) = \mathrm{bdom}(A_i^\diamond)$*. A probability distribution* p^* *over* 2^{T_X} *induces a condensed probability distribution* $p^\diamond : T_{X^\diamond} \to [0,1]$ *with*

$$p^\diamond(t_{X^\diamond}) = P^*(\{S \in T_X : \sigma(S) = t_{X^\diamond}\}) = \sum_{S \in \sigma_X^{-1}(t_{X^\diamond})} p^*(S) \tag{5.6}$$

For each tuple $t_{X^\diamond}$, the inverse image under σ_X identifies the set-valued outcomes that contribute to its probability mass in the condensed joint distribution. The definition assigns exactly those set-outcomes S to $t_{X^\diamond}$ that fulfill either of the following conditions for all attributes $A \in X$:

- $t_{X^\diamond}(A^\diamond) \in \mathrm{bdom}(A^\diamond)$ and S contains only tuples that assign the same value to A as $t_{X^\diamond}$ does to $A^\diamond$ (the outcome is single-valued w.r.t. A)
- $t_{X^\diamond}(A^\diamond) \notin \mathrm{bdom}(A^\diamond)$ and S has a projection to $\{A\}$ with cardinality different from one, i.e., the outcome is either multi-valued w.r.t. A or the empty-set outcome.

Thus, a tuple w.r.t. condensed set-valued attributes represents a class of set-valued outcomes, defined by the cardinality class (singleton vs. non-singleton) of an the outcome's projections to each individual attribute's domain and, if it projects to a singleton, the particular instantiation of that attribute. Due to this distinction, the representation retains all necessary information for computing marginal distribution w.r.t. to any chosen subset of attributes (Figure 5.2).

The same distinction is applied for coverage functions. Whereas in the one-dimensional case only one set of coverage factors was required, the multivariate

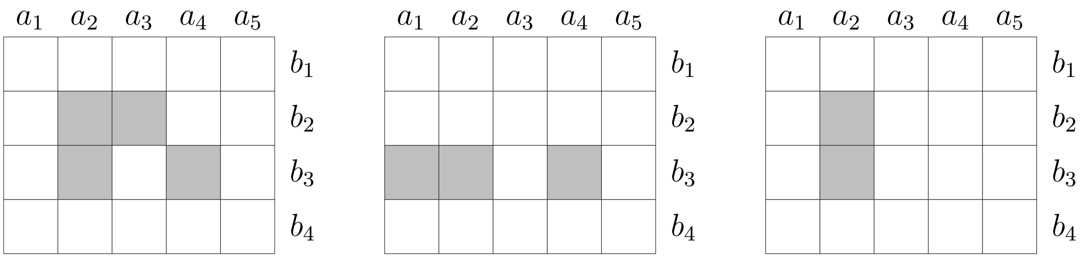
a_1 a_2 a_3 a_4 a_5
b_1
b_2
b_3
b_4
a_1 a_2 a_3 a_4 a_5
b_1
b_2
b_3
b_4
a_1 a_2 a_3 a_4 a_5
b_1
b_2
b_3
b_4

Figure 5.2: Three relations with different projections to the individual attribute domains. The respective tuples assigned by the set reduction mapping are (from left to right): $(a^\diamond, b^\diamond)$; $(a^\diamond, b_3)$ and $(a_2, b^\diamond)$.

approach defines separate coverage functions with respect to all tuples over condensed set-valued attributes. For any fixed tuple $t_{X^\diamond}$ the attribute set X can be partitioned into subsets:

$$\begin{aligned} Y^\diamond &= \{A^\diamond \in X^\diamond \colon t_{X^\diamond}(A^\diamond) \notin \mathrm{bdom}(A^\diamond)\} \text{ and} && (5.7)\\ Z^\diamond &= \{A^\diamond \in X^\diamond \colon t_{Y^\diamond}(A^\diamond) \in \mathrm{bdom}(A^\diamond)\} = X^\diamond \setminus Y^\diamond. && (5.8) \end{aligned}$$

By Definition 5.7 the relations of the class represented via $t_{X^\diamond}$ contain only tuples for which the instantiation of base attributes $Z^\diamond$ have values corresponding to the images of the respective condensed set-valued attributes under $t_{X^\diamond}$. Furthermore, these relations are set-valued w.r.t any of the attributes in Y. To enable the reconstruction of the contribution of those relations to the one point-coverage w.r.t. any set of tuples $T_W, W \subseteq X$, is suffices to store the coverage of tuples in T_Y by the projections of the represented relations from 2^{T_X}.

This observation motivates a definition of the coverage factors with respect to the classes induced by σ. To specify these factors a coverage function $c^\diamond_{X,[Y],t_{Z^\diamond}} : T_Y \to [0,1]$ is defined for any tuple $T_{X^\diamond}$. The reference set, and thus $T_{X^\diamond}$, is indicated via the 2nd and 3rd subscripts. Assuming that the coverage function refers to those relations that are mapped to a tuple $t_{X^\diamond}$ by the function σ, Y specifies the set of attributes for which the relations have non-singleton projections, whereas the tuple $t_{Z^\diamond} \equiv t_{X^\diamond|Z^\diamond}$ serves to identify the common singleton projections to the remaining attributes (compare Equations 5.7 and Equations 5.8).

Definition 5.9. *Let $X^\diamond = \{A^\diamond_1, \ldots, A^\diamond_n\}$ be a finite nonempty set of condensed set-valued attributes and X the set of respective base attributes. Furthermore let p^* denote a distribution over 2^{T_X} and $p^\diamond$ the condensed set-valued distribution induced by p^*. Let $t_{X^\diamond}$ be a tuple from $T_{X^\diamond}$ and the subsets $Y^\diamond$ and $Z^\diamond$ of $X^\diamond$ determined by applying Equations 5.7 and 5.8 w.r.t the tuple $t_{X^\diamond}$. The coverage function associated with the tuple $t_{X^\diamond}$ is a mapping $c^\diamond_{X,[Y],t_{Z^\diamond}} : T_Y \to [0,1]$*

$$\begin{aligned} c^\diamond_{X,[Y],t_{Z^\diamond}}(t_Y) &\overset{\text{def}}{=} \begin{cases} \frac{P^*(\{S:\sigma(S)=t_{X^\diamond} \wedge t_Y \in \mathrm{proj}^X_Y(S)\})}{P^*(\{S:\sigma(S)=t_{X^\diamond}\})} & \text{if } P^*(\{S : \sigma(S) = t_{X^\diamond}\}) > 0 \\ 1 & \text{otherwise,} \end{cases} \\ &= \begin{cases} \frac{\sum_{\substack{S \in \sigma^{-1}(t_{X^\diamond}) \\ t_Y \in \mathrm{proj}^X_Y(S)}} p^*(S)}{p^\diamond(t_{X^\diamond})} & \text{if } p^\diamond(t_{X^\diamond}) > 0 \\ 1 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.9}$$

where $t_{Z^\diamond} \equiv t_{X^\diamond|Z^\diamond}$.

Intuitively $c^\diamond_{X,[Y],t_{Z^\diamond}}(t_{X|Y})$ may be interpreted as the proportion of cases, in which a particular tuple $t_X \in T_X$ is covered by a relation, known to be multi-valued w.r.t. all attributes in Y and none of the other attributes.

The function $c^\diamond_{X,[X],()}$ specifies the relative coverage due to relations that under no projection other than the one to the empty set appear as singletons. Its second argument is always the empty tuple and coverage factors are given for the original space T_X. On the other end of the scale are the functions $c^\diamond_{X,[\emptyset],t_{X^\diamond}}$. For these, Equation 5.9 may be simplified as the inverse image of the respective tuples $t_{X^\diamond} = t_{Z^\diamond}$ under σ_X contains only a singleton $\{t_X\}$. Furthermore, the projection condition in the numerator may be dropped. This is seen by computing the projection $\mathrm{proj}^X_\emptyset(\{t_X\}) = \{()\}$, which is also the only possible tuple over the empty variable set. Together with Equation 5.6 this yields unit value coverage factors

$$\begin{aligned} c^\diamond_{X,[\emptyset],t_{X^\diamond}}(()) &= \begin{cases} \frac{\sum_{S\in\sigma_X^{-1}(t_{X^\diamond})} p^*(S)}{p^\diamond(t_{X^\diamond})} & \text{if } p^\diamond(t_{X^\diamond}) > 0 \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{p^*(\{t_X\})}{p^*(\{t_X\})} & \text{if } p^\diamond(t_{X^\diamond}) > 0 \\ 1 & \text{otherwise} \end{cases} = 1. \end{aligned}$$

Of course, these constant values need not be represented explicitly. A schematic representation of a two-dimensional distribution over condensed set-valued attributes with associated coverage factors is given in Figure 5.3.

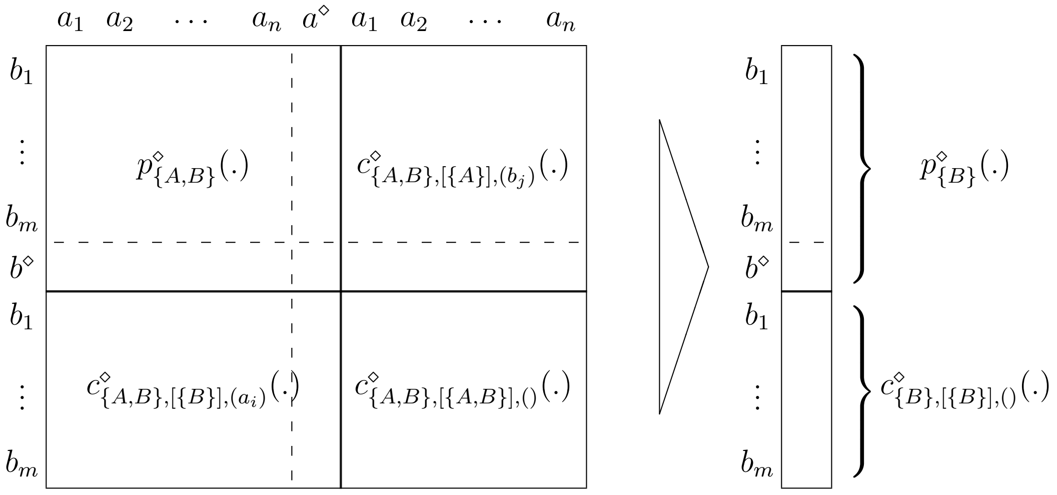
a_1 a_2 $\cdots$ a_n $a^\diamond$ a_1 a_2 $\cdots$ a_n
b_1
$\vdots$
b_m
$b^\diamond$
b_1
$\vdots$
b_m
$p^\diamond_{\{A,B\}}(\cdot)$
$c^\diamond_{\{A,B\},[\{A\}],(b_j)}(\cdot)$
$c^\diamond_{\{A,B\},[\{B\}],(a_i)}(\cdot)$
$c^\diamond_{\{A,B\},[\{A,B\}],()}(\cdot)$
b_1
$\vdots$
b_m
$b^\diamond$
b_1
$\vdots$
b_m
$p^\diamond_{\{B\}}(\cdot)$
$c^\diamond_{\{B\},[\{B\}],()}(\cdot)$

Figure 5.3: Joint distribution with coverage factors for two condensed set-valued attributes and one of its marginal distributions.

The one-point coverage of a tuple $t_Y \in T_Y$ from that representation, is computed from the relative coverage functions for each the partitions introduced by the set reduction mapping. For iterating over these coverage functions, it is convenient to introduce an auxiliary function:

Definition 5.10. *Let $X = \{A_1 \dots A_n\}$ be a set of attributes. Furthermore, let $X^\diamond = \{A_1^\diamond \dots A_n^\diamond\}$ be an associated set of condensed set-valued attributes, such that* $\operatorname{dom}(A_i^\diamond) = \operatorname{dom}(A_i) \cup \{a_i^\diamond\}$, *$\forall i = 1 \dots n$. Given a tuple $t_X \in T_X$ and an attribute set $Y \subseteq X$ The mapping* mv_X *is the two-place function $T_X \times 2^X \to T_{X^\diamond}$ satisfying*

$$\operatorname{mv}_X(t_X, Y)(A_i^\diamond) = \begin{cases} t_X(A_i), & \text{if } A_i \notin Y, \\ a_i^\diamond, & \text{if } A_i \in Y, \end{cases} \tag{5.10}$$

where $Y^\diamond \subseteq X^\diamond$ denotes those condensed-set valued attributes, that correspond to the attributes in Y.

For convenience, the index indicating the reference set will be dropped if it is clear from to the first argument. For instance $\operatorname{mv}(t_X, Y)$ will be used as a shorthand for $\operatorname{mv}_X(t_X, Y)$. When working with condensed set-valued attributes, the function mv_X can be used to mark out tuples that are images of relations containing a given tuple t_X in the corresponding full random set representation. Together with the set of attributes Y w.r.t which the relations have non-singleton projections, this allows to identify the cases that contribute to a particular coverage factor. Based on this consideration we can now express one-point coverages $\operatorname{opc}(t_X)$ in terms of relative coverage factors and of probability degrees for tuples of condensed attributes:

$$\begin{aligned} \operatorname{opc}(t_X) &= P^*(\{S \in 2^{T_X} \colon t_X \in S\}) \\ &= \textstyle\sum_{t_{X^\diamond} \in T_{X^\diamond}} P^*(\{S \colon \sigma(S) = t_{X^\diamond} \wedge t_X \in S\}) \end{aligned}$$

Assuming that $P^*(\{S \colon \sigma(S) = t_{X^\diamond}\}) > 0$, this is further expanded to

$$\begin{aligned} \operatorname{opc}(t_X) &= \textstyle\sum_{t_{X^\diamond} \in T_{X^\diamond}} P^*(\{S \colon \sigma(S) = t_{X^\diamond}\}) \cdot \frac{P^*(\{S : \sigma(S) = t_{X^\diamond} \wedge t_X \in S\})}{P^*(\{S : \sigma(S) = t_{X^\diamond}\})} \end{aligned}$$

$$\begin{aligned} &= \textstyle\sum_{\substack{Y \subseteq X \\ Z = X \setminus Y}} P^*(\{S \colon \sigma(S) = \operatorname{mv}(t_X, Y)\}) \cdot \frac{P^*(\{S : \sigma(S) = \operatorname{mv}(t_X, Y) \wedge t_{X|Y} \in \operatorname{proj}_Y^X(S)\})}{P^*(\{S : \sigma(S) = \operatorname{mv}(t_X, Y)\})} \\ &= \textstyle\sum_{\substack{Y \subseteq X \\ Z = X \setminus Y}} p^\diamond(\operatorname{mv}(t_X, Y)) \cdot \frac{\sum_{\substack{S \in \sigma^{-1}(\operatorname{mv}(t_X, Y)) \\ t_{X|Y} \in \operatorname{proj}_Y^X(S)}} p^*(S)}{p^\diamond(\operatorname{mv}(t_X, Y))} \\ &= \textstyle\sum_{\substack{Y \subseteq X \\ Z = X \setminus Y}} p^\diamond(\operatorname{mv}(t_X, Y)) \cdot c^\diamond_{X,[Y],\operatorname{mv}(t_{X|Z}, \emptyset)}(t_{X|Y}). \end{aligned}$$

The tuples $t_{X^\diamond}$ compatible with t_X are constructed by applying mv with all selections of attributes Y from X. To compute the one-point coverage of a tuple $t_X \in T_X$, probabilities assigned to those compatible tuples are weighted with the respective relative coverage factors. Conveniently, the above result also applies if $p^\diamond(t_{X^\diamond}) = P^*(\sigma(S) = t_{X^\diamond}) = 0$ for some $t_{X^\diamond}$ as the correct contribution $p^\diamond(t_{X^\diamond}) \cdot c^\diamond_{X,[Y],(t_{X^\diamond})}(t_Y) = 0 \cdot 1 = 0$ will be returned in that case.

5.3.3 Marginal Distributions

Whereas joint distributions provide detailed information about dependencies and statistical interactions between attributes, they are impractical when attention is turned to specific aspects of a complex system. Marginal distributions permit to discern events according to a chosen subset of the attributes rather than the complete attribute set. Thus they extract relevant information for specific tasks, summarize information from more complex models, provide interfaces to link between global and specialized local models or data sources. To obtain marginals for condensed distributions both marginal probability distributions and coverage factors need to be computed.

For the probabilistic component $p^\diamond$ of the condensed set-valued distribution over a set of attributes $X^\diamond$ the marginal distribution w.r.t. any subset $W^\diamond \subseteq X^\diamond$ is computed by aggregating the probability mass of tuples t_X that have common projections into the lower dimensional space T_W.

$$\begin{aligned} p^\diamond_W(t_{W^\diamond}) &= P^\diamond_W(\{t_{W^\diamond}\}) = P^\diamond_X(\{t_{X^\diamond} \mid t_{X^\diamond|W^\diamond} \equiv t_{W^\diamond}\}) \\ &= \sum_{t_{X^\diamond} \in T_{X^\diamond}, t_{X^\diamond|W^\diamond} \equiv t_{W^\diamond}} p^\diamond_X(t_{X^\diamond}). \end{aligned} \tag{5.11}$$

This computation yields the same result as applying the set reduction mapping σ_W to projections of the original relations from X to W, as is pointed out by the following argument:

$$\begin{aligned} p^\diamond_W(t_{W^\diamond}) &= P^*_W(\{R\colon \sigma_W(R) = t_{W^\diamond}\}) \\ &= P^*_X\left(\{S\colon \sigma_W\left(\mathrm{proj}^X_W(S)\right) = t_{W^\diamond}\}\right) \\ &= P^*_X\left(\{S\colon \sigma_X(S) = t_{X^\diamond} \wedge t_{X^\diamond|W^\diamond} \equiv t_{W^\diamond}\}\right) \\ &= \sum_{\substack{t_{X^\diamond} \in T_{X^\diamond} \\ t_{X^\diamond|W^\diamond} \equiv t_{W^\diamond}}} p^\diamond_X(t_{X^\diamond}) \end{aligned} \tag{5.12}$$

The third equality in the above argument follows from the definition of the set reduction mapping. Since the images of the individual attributes under the tuples are determined from the one-dimensional projections only, $\sigma_X(S)$ assigns the same values to the attributes in $W^\diamond$ as $\sigma_W\left(\mathrm{proj}^X_W(S)\right)$ for every $S \subseteq T_X$.

In a similar manner, coverage factors w.r.t. a set of attributes $W \subseteq X$ should be defined to be consistent with an application of Equation 5.9 to the projection of a distributions over 2^{T_X} to 2^{T_W}. To that end the coverage functions of the condensed representation for distributions over that projected space are defined with respect to tuples $T_{W^\diamond}$. For each tuple $t_{W^\diamond} \in T_{W^\diamond}$ the attribute set $W^\diamond$ is composed of the two disjoint subsets $Y^\diamond = \{A^\diamond \in W^\diamond\colon t_{W^\diamond}(A^\diamond) \notin \mathrm{bdom}(A^\diamond)\}$ and

$Z^\diamond = \{A^\diamond \in W^\diamond \colon t_{W^\diamond}(A^\diamond) \in \mathrm{bdom}(A^\diamond)\}$. The former set contains the attribute that with non-singleton values, whereas the latter contains the attributes for which the value is a singleton. As before, the set Y contains the base attributes corresponding to the condensed set valued attributes in $Y^\diamond$ and $t_{Z^\diamond} \equiv t_{W^\diamond|Z^\diamond}$. Given the condensed distribution and coverage factors referring to tuples $T_{X^\diamond}$ the coverage functions associated with $T_{W^\diamond}$ are computed as follows:

$$c^\diamond_{W,[Y],t_{Z^\diamond}}(t_Y) = \begin{cases} \frac{P^*_W(\{R:\sigma_W(R)=t_{W^\diamond} \wedge t_Y \in \mathrm{proj}^W_Y(R)\})}{P^*_W(\{R:\sigma_W(R)=t_{W^\diamond}\})} & \text{if } P^*_W(\{R:\sigma_W(R)=t_{W^\diamond}\})>0 \\ 1 & \text{otherwise} \end{cases} \tag{5.13}$$

By applying Equation 5.12 to the denominator of the fraction and substituting the projections we obtain:

$$\begin{aligned} c^\diamond_{W,[Y],t_{Z^\diamond}}(t_Y) &= \begin{cases} \frac{P^*_X\left(\{S:\sigma_W\left(\mathrm{proj}^X_W(S)\right)=t_{W^\diamond} \wedge t_Y \in \mathrm{proj}^W_Y\left(\mathrm{proj}^X_W(S)\right)\}\right)}{p^\diamond(t_{W^\diamond})} & \text{if } p^\diamond(t_{W^\diamond}) > 0 \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{P^*(\{S:\sigma_X(S)=t_{X^\diamond} \wedge t_{X^\diamond|W^\diamond} \equiv t_{W^\diamond} \wedge t_Y \in \mathrm{proj}^X_Y(S)\})}{p^\diamond(t_{W^\diamond})} & \text{if } p^\diamond(t_{W^\diamond}) > 0 \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{\sum\limits_{t_{X^\diamond|W^\diamond} \equiv t_{W^\diamond}} P^*(\{S:\sigma_X(S)=t_{X^\diamond} \wedge t_Y \in \mathrm{proj}^X_Y(S)\})}{p^\diamond(t_{W^\diamond})} & \text{if } p^\diamond(t_{W^\diamond}) > 0 \\ 1 & \text{otherwise} \end{cases} \end{aligned}$$

The expression in the numerator computes a one-point coverage. As before, probability-weighted relative coverage factors compatible with the restrictions imposed by t_Y and $t_{Z^\diamond}$ are aggregated. However, since these restrictions only apply to attributes in $W^\diamond$, combinations with all possible instantiations for the remaining attributes in $X^\diamond$ need to be considered:

$$\begin{aligned} &c^\diamond_{W,[Y],t_{Z^\diamond}}(t_Y) \\ &= \begin{cases} \frac{\sum\limits_{\substack{t_X \in T_X, t_{X|Y} \equiv t_Y, \mathrm{mv}(t_{X|Z},\emptyset)=t_{Z^\diamond} \\ Y' \subseteq X \setminus W, t_{X^\diamond}=\mathrm{mv}(t_X, Y \cup Y')}} p^\diamond(t_{X^\diamond}) \cdot c^\diamond_{X,[Y \cup Y'],\mathrm{mv}(t_X, Y \cup Y')}\left(t_{X|(Y \cup Y')}\right)}{p^\diamond(t_{W^\diamond})} & \text{if } p^\diamond(t_{W^\diamond})>0 \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} \sum\limits_{\substack{t_X \in T_X \\ t_{X|Y} \equiv t_Y \\ \mathrm{mv}(t_{X|Z},\emptyset)=t_{Z^\diamond} \\ Y' \subseteq X \setminus W \\ t_{X^\diamond}=\mathrm{mv}(t_X, Y \cup Y')}} \frac{p^\diamond(t_{X^\diamond})}{p^\diamond(t_{W^\diamond})} \cdot c^\diamond_{X,[Y \cup Y'],t_{X^\diamond}}\left(t_{X|(Y \cup Y')}\right) & \text{if } p^\diamond(t_{W^\diamond})>0 \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{5.14}$$

Because the above equation requires operations on the parameters of the sum to identify the correct attribute sets and coverage functions, this result is still difficult to apply. Using an auxiliary function τ the transformation of the parameters can be separated from the remaining computations:
Let τ denote a function that maps its arguments $(W, Y, Z^\diamond, t_{Z^\diamond}, t_Y)$ to the set of all tuples $(Y_+, t_{W^\diamond}, t_X, t_{X^\diamond}, t_Y)$ that fulfill the following criteria:

$$\begin{array}{ll}
t_X \in T_X & \text{(iterate over instantiations)} \\
t_{X|Y} \equiv t_Y & \text{(compatible with } t_Y\text{)} \\
\mathrm{mv}(t_{X|Z}, \emptyset) = t_{Z^\diamond} & \text{(compatible with singleton instantiations} \\
 & \text{ for attributes in } Z^\diamond\text{)} \\
Y' \subseteq X \setminus W & \text{(iterate over attributes with possible} \\
Y_+ = Y' \cup Y & \text{ set instantiations)} \\
t_{X^\diamond} = \mathrm{mv}(t_X, Y_+) & \text{(generate compatible tuples over } T_{X^\diamond}\text{)} \\
t_{W^\diamond} \equiv t_{X^\diamond|W^\diamond} & \text{(from definition of attribute sets)}
\end{array}$$

On the implementation side the function τ is reflected by an iterator that is used to traverse the instantiations of the non-constrained attributes $X^\diamond \setminus W^\diamond$. For each of the resulting instantiations in the marginal distribution it supplies the respective higher-dimensional coverage factors allowing to compute their contribution to the marginal coverage factor.

With the above definition of the function τ the relation between the joint and marginal coverage factors can be expressed in a more convenient manner:

$$c^\diamond_{W,[Y],t_{Z^\diamond}}(t_Y) \quad = \begin{cases} \displaystyle\sum_{\substack{(Y_+,t_{W^\diamond},t_X,t_{X^\diamond},t_{Y^\diamond}) \\ \in\tau(W,Y,Z^\diamond,t_{Z^\diamond},t_Y)}} \frac{p^\diamond(t_{X^\diamond})}{p^\diamond(t_{W^\diamond})} \cdot c^\diamond_{X,[Y_+],t_{X^\diamond}}\left(t_{X|(Y_+)}\right) & \text{if } p^\diamond(t_{W^\diamond})>0 \\ 1 & \text{otherwise} \end{cases} \tag{5.15}$$

The short summary of the above results is that marginal coverage factors are obtained by computing a weighted average of coverage factors over a higher dimensional reference set, with the weights given by the tuple probabilities. To better understand this result, it is helpful to to remember that coverage values specify a proportion relative to the a tuple $t^\diamond_X$. Each tuple in turn stands for a class of set-valued outcomes. The projection, makes those classes of set-valued outcomes that are reflected by identical tuples $t_{W^\diamond}$ in the target space indistinguishable from each other. This results in the formation of to a new combined group. The weighted average reflects the contributions from each original class to the newly formed group.

Alternatively one may choose to view each individual coverage factors as a parameter of a binary probability distribution. According to this interpretation the coverage factors in the target space reflect the parameters of mixtures of such distributions. To distinguish coverage factors referring to the full attribute set of a model from their counterparts for marginal attribute domains, the former will be called *elementary coverage factors*.

An important property of condensed distributions is, that the marginal distributions coincide with the condensed distributions that are computed directly from projections of the original set-valued outcomes. This semantic consistency is essential to relate observations that are made with respect to different but overlapping sets of attributes. Moreover, because the coarsened distribution $p^{\diamond}$ is a regular probability distribution. Thus the representation is fully embedded within the probabilistic framework. Specifically, if only singleton outcomes occur, the condensed set-valued distribution is equivalent to a representation via a conventional probability distribution. Therefore set-valued attributes can interface with other probabilistic model when used along with them in applications.

5.3.4 Conditioning and Conditional Distributions

Like marginal distributions and the operations connected with them, conditional distributions and conditioning are basic tools of probabilistic and relational reasoning. Conditioning serves to combine case-specific observations with generic (a-priori) knowledge or to simulate the effects of changes to the modeled system. In bioinformatics it can be applied to predict function for expression when annotations on an organism are still limited. In that case conditioning permits to adapt detailed statistical models available for better studied model organisms to observations on the target organism.

Under an empirical interpretation of probability, conditioning reflects a change of the reference set. The particular reference set to which a conditional distribution refers is specified in one of three modes, namely

- by choosing a fixed instantiation,
- by excluding instantiations incompatible with observations or
- by providing a specific distribution over the values of so called conditioning variables (attributes).

The result of conditioning is a conditional distribution for the instantiations of the modeled systems variables w.r.t. the selected reference set.

If the conditioning variables are known in advance, knowledge representations can provide pre-calculated families of conditional distributions given the individual instantiations of these variables. Once the actual instantiations of the conditioning variables become available, the corresponding conditional distribution is selected. If the conditioning information is given as a marginal distribution, the conditional distribution is computed as a mixture of the conditional distributions w.r.t. the individual instantiations, with the weights given by the provided marginal distribution. A similar procedure is applied to exclude incompatible instantiations, though an intermediate step is required to re-distribute probability mass on marginal distributions.

It is remarked that in the relational framework conditioning simply amounts to a restriction of the tuples in a relation to those, which exhibit the required values for the conditioning attributes. In the context of databases the equivalent of the conditioning operation would be called a selection.

Probabilistic Component

When applying the notion of conditioning to condensed distributions it is desirable to maintain consistency with the random set framework. This constraint immediately determines the definition of most conditioning operations. For the probabilistic part $p^\diamond$ of the condensed distribution, conditioning with an event $E \subseteq 2^{T_{X^\diamond}}$ with $P^\diamond(E) > 0$ is straightforward:

$$\begin{aligned} p^\diamond(t_{X^\diamond} \mid E) &= P^\diamond(\{t_{X^\diamond}\} \mid E) = \frac{P^\diamond(\{t_{X^\diamond}\} \cap E)}{P^\diamond(E)} \\ &= \begin{cases} \frac{p^\diamond(t_{X^\diamond})}{\sum\limits_{t'_{X^\diamond} \in E} p^\diamond(t'_{X^\diamond})} & \text{if } t_{X^\diamond} \in E \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{5.16}$$

A conditioning event may alternatively be specified using tuples over a subset $U^\diamond \subseteq X^\diamond$ of attributes. Such a specification $E_{U^\diamond}$ is considered a shortcut notation for the event $E = \{t_{X^\diamond} \mid t_{X^\diamond} \in T_{X^\diamond} \wedge t_{X^\diamond|U^\diamond} \in E_{U^\diamond}\}$ i.e., the most general event over $T_{X^\diamond}$ matching the restrictions expressed in $E_{U^\diamond}$.

To support conditioning with distributions over tuples, it is useful to formulate a rule for conditioning with individual tuples w.r.t. arbitrary reference sets, that is to compute conditional probabilities of the form $p^\diamond_{W^\diamond}(t_{W^\diamond} \mid t_{U^\diamond})$. The conditional distributions for elementary tuples can subsequently be combined to achieve conditioning with distributions. In order to adapt the probabilistic conditioning rules to the tuple-based notation, it is important to remember, that tuples reflect sets of restrictions w.r.t. the instantiations of attributes. When conditioning in the tuple-based notation the intersection of events is expressed by joining tuples, with the result being a tuple over the combined set of attributes.

Obviously, if the set of attributes $U^\diamond$ and $W^\diamond$ have a nonempty intersection, then $t_{W^\diamond|(U^\diamond \cap W^\diamond)} \equiv t_{U^\diamond|(U^\diamond \cap W^\diamond)}$ must hold in order for the conditioned and the conditioning events to compatible. The expressed restrictions would otherwise require mutually exclusive instantiations for at least one of the shared attributes in $U^\diamond \cap W^\diamond$, resulting in a conditional probability of 0. Assuming $t_{U^\diamond|U^\diamond \cap W^\diamond} \equiv t_{W^\diamond|U^\diamond \cap W^\diamond}$ as well as $p^\diamond_{U^\diamond}(t_{U^\diamond}) > 0$, the formula for computing a conditional probability distributions for condensed set-valued attributes may be written

$$p^\diamond_{W^\diamond}(t_{W^\diamond} \mid t_{U^\diamond}) = \frac{p^\diamond_{U^\diamond \cup W^\diamond}(t_{U^\diamond \cup W^\diamond})}{p^\diamond_{W^\diamond}(t_{W^\diamond})}, \tag{5.17}$$

where $U^\diamond, W^\diamond \subseteq X^\diamond$, $t_{(U^\diamond \cup W^\diamond)|U^\diamond} \equiv t_{U^\diamond}$ and $t_{(U^\diamond \cup W^\diamond)|W^\diamond} \equiv t_{W^\diamond}$. For $p^\diamond_{W^\diamond}(t_{W^\diamond}) = 0$ the conditional distribution is neither defined, nor required to reconstruct the joint distribution. However, to simplify evidence propagation when working with distributions estimated from empirical data, one may choose $p^\diamond_{W^\diamond|U^\diamond}(t_{W^\diamond} \mid t_{U^\diamond}) \overset{\text{def}}{=} p^\diamond_{W^\diamond}(t_{W^\diamond})$ for that case, thereby substituting the known unconditioned distribution for an unknown conditional one.

Coverage Factors and Conditioning

Having discussed the effect of conditioning on the probabilistic part of the model we can now turn to the coverage factors . To that end, it is worthwhile to differentiate between two cases: (a) new information is supplied only w.r.t. the condensed probability distribution $p^\diamond$ only (all three modes), and (b) the conditioning pieces information specify both a condensed probability distribution and coverage factors (including conditioning with individual set-outcomes). In comparison to a random-set representation, however, conditioning on the condensed distribution is subject to certain limitations. In a full random set representation (as with any other probability distribution), conditioning redistributes probability mass of individual outcomes to those that are more compatible to the conditioning information. In contrast, coverage factors of the condensed representation refer to a group of set-outcomes, in which contributions from individual outcomes are no longer tracked separately.

In case (a) the elementary coverage factors are left unchanged taking advantage of the fact that the elementary coverage functions are already defined in relation to fixed tuples in $T_{X^\diamond}$. The assumption behind this approach is that the average coverage ratios of elements by the multi-valued outcomes are more robust than the joint distribution (compare page 31). Conditioning a joint distribution via observations on a subspace, requires some additional consideration. The key is to recall that the model represents the marginal coverage factor as a probability-weighted mixture of conditional coverage factors. Because the weights are determined by the probabilistic part of the distribution, the marginal coverage factors have to be adapted under conditioning. To achieve this, it suffices to replace all references to the condensed probability distribution $p^{\diamond}$ in Equation 5.13 with the respective values of the new conditional distribution.

Another possible scenario is that conditioning information has been obtained empirically from data on observed set-instantiations. In that case both the probabilistic part of the model and the coverage factors may be specified (b). If the specified coverage factors are defined with respect to the full attribute set X of the model, this new information takes precedence over the respective elementary coverage factors of the unconditioned distribution. Marginal coverage factors, however, are modeled as mixtures of elementary coverage factors. Thus the differences between the observed marginal coverage factors $c^{\diamond}_{W,[Y],t_{W^{\diamond}\setminus Y^{\diamond}}\,\mathrm{obs}}(t_Y)$ and the computed marginal coverage factor need to be resolved. The above problem can be approached heuristically, e.g., by assigning to each of the coverage factors $c^{\diamond}_{X,[Y_+],t_{X^{\diamond}}}\left(t_{X|(Y_+)}(Y_+, t_{W^{\diamond}}, t_X, t_{X^{\diamond}}, t_{Y^{\diamond}})\right)(t_{Y_+})$, with $(Y_+, t_{W^{\diamond}}, t_X, t_{X^{\diamond}}, t_{Y^{\diamond}}) \in \tau(W, Y, W^{\diamond} \setminus Y, t_{W^{\diamond}\setminus Y^{\diamond}}, t_Y)$ the value $c^{\diamond}_{W,[Y],t_{W^{\diamond}\setminus Y^{\diamond}}\,\mathrm{obs}}(t_Y)$ in the conditioned distribution. An alternative heuristic is to apply a proportional fitting strategy, that rescales the values of the coverage factors. To ensure valid results, however, the individual coverage factor need to be bounded by the interval $(0, 1]$. Therefore, such a fitting procedure should be conducted in an iterative manner with correction applied in between runs to enforce boundary constraints.

5.3.5 Combination with Graphical Models

On the previous pages I have discussed, how the basic operations of conditioning and marginalization are adapted to joined distributions of condensed set-valued attributes. With those results it is possible to combine the decomposition approach of Graphical Models to probability distributions over condensed set-valued attributes. To that end, consider the reconstruction of a joint distribution from independent marginal ones. The probabilistic part of the joint distribution is computed using the standard probabilistic operation (Section 3.2). Moreover, the conditional coverage functions of an attribute given its parents in the graph are identical to those of the joint distribution of these attributes. Thus, the task is reduced to combining sets of marginal coverage functions. This combination of marginal coverage functions is achieved by computing their products (see Equation 5.18 below).

The above recombination procedure implicitly assumes a property parallel to the notion of independence for condensed set-valued attributes $A^{\diamond}$ and $B^{\diamond}$. Due to the reuse of the probabilistic operation, a minimum requirement is independence w.r.t. the condensed probability distribution. In addition to that, changing marginal distributions w.r.t. one attribute must not affect the coverage factors for the other attribute, so all coverage factors $c^{\diamond}_{\{A,B\},[\{A\}],(b)}((a))$ and $c^{\diamond}_{\{A,B\},[\{B\}],(a)}((b))$ must coincide with the respective marginal coverage factors $c^{\diamond}_{\{A\},[\{A\}],()}((a))$ and $c^{\diamond}_{\{B\},[\{B\}],()}((b))$. For the higher dimensional coverage functions there are several possible solutions. In the absence of more specific information about the focal sets we may choose

$$c^{\diamond}_{\{A,B\},[\{A,B\}],()}((a,b)) = c^{\diamond}_{\{A\},[\{A\}],()}((a)) \cdot c^{\diamond}_{\{B\},[\{B\}],()}((b)) \,. \qquad (5.18)$$

An interesting observation is, that the factors $c^{\diamond}_{\{A,B\},[\{A\}],()}((a,b))$ are required for computation of one-point coverages in the joint distribution only. If the joint distribution merely serves to mediate the interaction between the distributions of the attributes $A^{\diamond}$ and $B^{\diamond}$, the respective criterion can be omitted for a weaker definition. The analogue extension is applied to conditional independence (w.r.t. the conditional coverage factors). When conditioning the joint distribution with marginal probability distributions for the separator sets the coverage factors defined w.r.t. conditioning attributes are considered as fixed. Nevertheless, other marginal coverage factors have to be corrected for the altered distributions and coverage functions for higher dimensional tuples have to be re-computed.

Although Graphical Models provide a very general approach to model interactions between variables with set-instantiations, it is remarked that the predominant source of complex data involving set-valued attributes in bioinformatics are gene, protein, or pathway annotation databases. To the authors knowledge all major publicly available data sets use annotations based on ontologies with tree-based relation structures. In the following section I present adapted models that achieve efficient operations on such structured attribute domains.

5.4 Application to Tree-Structured Domains

As previously stated, the amount of information provided in the description of objects can be controlled via the choice of the attribute set used to describe their properties. In addition to that, many properties can be presented with variable resolution. For instance, the birthplace of a person may be specified in terms of a country, but also on the finer scales of region or community. It is convenient to organize such different layers of detail into hierarchies that allow to easily summarize detail information. When enriched with statistical information about value distributions, hierarchically structured attribute domains support the transfer of knowledge between alternative frames of discernment. Thus they provide flexibly in serving various information needs and facilitate the processing of inhomogeneous data.

By applying the condensed representation of set-valued data to hierarchically structured attribute domains I aim to provide a robust and interpretable approach to problems such as the incomplete classification of objects encountered when processing/mining annotation sets formed from hierarchically organized terms. I doing so I restrict myself to finite frames of discernment, which cover all cases of set-valued data annotations currently known to the author and can be applied to discretized continuous attributes as well. Nevertheless, most considerations directly transfer to the continuous case using parametric descriptions.

Using probability trees as a starting point, I will initially investigate a data representation with attributes exhibiting a hierarchical value domain structure. After shortly recapitulating the concept of frames of discernment (cf. 2.2.1), I proceed to discuss relations between frames and their relevance to structured attribute domains (Subsection reference frames). From this discussion a method to efficiently model parallel frames and their interactions is developed (Subsection 5.4.2). In Subsection 5.4.3 the approach is extended to set-valued attributes, which represent, e.g. parallel instantiations, sets of alternatives or imprecise data. To that end condensed distributions are integrated into the hierarchical model leading to a scalable non-parametric distribution model for structured annotation data.

5.4.1 Related Frames of Discernment

In 2.2.1 the single-valued attribute was introduced via a universe O of objects and a set Λ of labels – the *frame of discernment*, which supplies attribute values for characterizing individual objects w. r. t. a particular property. In that representation and attribute A is identified with a (potentially unknown) function $A : O \to \Lambda$ that assigns the appropriate label to describe each of the considered objects. This definition presumes, that there exists a generally accepted set of mutually exclusive attribute values suitable for recording, processing and presenting any information about the considered property. Yet, it is sometimes useful to define several frames of discernment that complement each other. Such multi-frame representations are used, for instance:

- to combine information sources with differing observation capabilities,
- to meet restrictions of certain data analysis or processing methods (discretization, binning, data conversion for use in existing models)
- to tend to user- and task-specific information requirements
- to enable informative summaries

Adapting the level of detail in a description to a particular user's level of prior knowledge can contribute to better reception of relevant pieces of information.

To support the integrated discernment of a property on multiple frames, an attribute must be re-imagined as a collection of frame-specific mappings that assign labels to objects. But in contrast to the more general interaction between different attributes, label assignments for frames that refer to the same attribute are closely coupled. Labels that are shared between frames have identical instantiation states and a conversion of a statement to a coarser, that is, less informative, *frame of discernment* (Shafer, 1976) must be reflected in a compatible mapping of instantiation states and distributions. That change of resolution defines a partial ordering on the set of frames. The concept of refinement captures these ideas in a formal notion:

Definition 5.11. (Shafer 1976) *A set* Λ' *is a refinement of* Λ *if there is a mapping* $\mathrm{ref} : 2^{\Lambda} \to 2^{\Lambda'}$ *such that* $\forall \lambda_1, \lambda_2 \in \Lambda$:

1. $\forall \lambda \in \Lambda : \mathrm{ref}(\{\lambda\}) \neq \emptyset$

2. $(\lambda_1 \neq \lambda_2) \Rightarrow \mathrm{ref}(\{\lambda_1\}) \cap \mathrm{ref}(\{\lambda_2\}) = \emptyset$

3. $\bigcup \{\mathrm{ref}(\{\lambda\}) \mid \lambda \in \Lambda\} = \Lambda'$

4. $\mathrm{ref}(L) = \bigcup \{\mathrm{ref}(\{\lambda\}) \mid \lambda \in L\}$

Condition (1) ensures that any label in the original frame is still represented by at least one label in the refined frame, whereas condition (2) guarantees the preservation of any existing distinctions. Conditions (3) and (4) ensure correspondence of the considered attribute domains and provide a set extension for mapping operations, respectively. A set Λ is called a *coarsening* of a set Λ' if there is a refinement ref, such that $\mathrm{ref}(\Lambda) = \Lambda'$. By this property the refinement relation defines a partial ordering on the set of reference frames.

Although there are usually several meaningful ways to subdivide the equivalence class associated with a label during refinement, a hierarchical organization of the attribute domain is advantageous because it permits to easily find, summarize or arrange objects and information (the same reasoning is behind the organization of libraries according to a fixed topic hierarchies, even though several equally suited taxonomies can be conceived).

To formalize the hierarchical organization of the attribute labels consider the a nonempty set $\mathcal{L}$ together with a binary relation $\mathrm{R}_{\mathrm{parent}}$ over $\mathcal{L}$, where $(\lambda_2, \lambda_1) \in \mathrm{R}_{\mathrm{parent}}$ indicates that λ_1 is a direct parent (superior) of λ_2 in the hierarchy.

Definition 5.12. *Let* $\mathcal{L}$ *be a set of labels and* $\mathrm{R}_{\mathrm{parent}}$ *a binary relation over the elements of* $\mathcal{L}$*. Moreover let* $\mathrm{R}_{\mathrm{anc}}$ *denote the transitive closure of* $\mathrm{R}_{\mathrm{parent}}$*. The tuple* $(\mathrm{R}_{\mathrm{parent}},\mathcal{L})$ *defines a hierarchy, iff*

1. $\forall \lambda, \lambda', \lambda'' \in \mathcal{L} : (\lambda, \lambda') \in \mathrm{R}_{\mathrm{parent}} \wedge (\lambda, \lambda'') \in \mathrm{R}_{\mathrm{parent}} \implies \lambda' = \lambda''$

2. $\exists \lambda_0 \in \mathcal{L} : \forall \lambda \in \mathcal{L} \setminus \{\lambda_0\} : (\lambda, \lambda_0) \in \mathrm{R}_{\mathrm{anc}}$

3. $\forall \lambda \in \mathcal{L} : (\lambda, \lambda) \notin \mathrm{R}_{\mathrm{anc}}$

The element λ_0 forms the root of the attribute value hierarchy. Due to the first two conditions all other elements of $\mathcal{L}$ must have a parent element (2), which is also unique (1). Thus for hierarchies, $\mathrm{R}_{\mathrm{parent}}$ gives rise to a function over the values of $\mathcal{L} \setminus \{\lambda_0\}$. The third condition ensures an acyclic structure. Moreover – together with the second one – it guarantees the uniqueness of the root element. To show that this is indeed the case, assume for a moment that condition (2) would hold w.r.t. two different root labels $\lambda_0 \neq \lambda_0'$. Then, by application of condition (2), both $(\lambda_0, \lambda_0') \in \mathrm{R}_{\mathrm{anc}}$ and $(\lambda_0', \lambda_0) \in \mathrm{R}_{\mathrm{anc}}$. However, due to the transitivity of $\mathrm{R}_{\mathrm{anc}}$ this entails $(\lambda_0, \lambda_0) \in \mathrm{R}_{\mathrm{anc}}$ violating the third condition.

With the above properties specified, it is useful to introduce notions for expressing relations in the hierarchy H. I write $\mathrm{parent}_H(\lambda)$ to denote the unique parent element of label $\lambda \in \mathcal{L} \setminus \{\lambda_0\}$ w.r.t. the hierarchy H. Similarly $\mathrm{anc}_H(\lambda)$ is used to denote the set of a label's ancestors in H (including indirect ones). In a similar manner I write $\mathrm{children}_H(\lambda)$ for the set of a label λ's direct children $\{\lambda' : \mathrm{R}_{\mathrm{parent}}(\lambda', \lambda)\}$ in the hierarchy H. In extension of that, the set of all descendants, including indirect ones, of a label λ in H is denoted by $\mathrm{desc}_H(\lambda)$.

Any label hierarchy H generates a family of related frames, $(\Lambda)_H$, which differ only with regard to the level of detail discerned in their respective sub-frames. Starting with $\{\lambda_0\} = \Lambda_0 \in (\Lambda)_H$, new frames are iteratively generated by replacing non-leaf elements of a previously found frame in $(\Lambda)_H$ with their respective sets of child labels in H *(sub-frame expansion)*. Whenever this set of child labels contains more than one element, the resulting frame allows for more specific descriptions than its parent frame.

Definition 5.13. *Let H be a label hierarchy with root label λ_0. H generates a family of frames $(\Lambda)_H$ via the following rules:*

1. $\Lambda_0 = \{\lambda_0\} \in (\Lambda)_H$

2. $$\mathrm{children}_H(\lambda_r) \neq \emptyset \wedge \lambda_r \in \Lambda \wedge \Lambda \in (\Lambda)_H \implies \Lambda' = \mathrm{children}_H(\lambda_r) \cup \Lambda \setminus \{\lambda_r\} \in (\Lambda)_H$$

Theorem 5.1. *For any frame of discernment $\Lambda \in (\Lambda)_H$ and for any element $\lambda_r : \lambda_r \in \Lambda \wedge \mathrm{children}_H(\lambda_r) \neq \emptyset$ the frame $\Lambda' = \mathrm{children}_H(\lambda_r) \cup \Lambda \setminus \{\lambda_r\}$ is a refinement of Λ.*

The particular sub-frame $\mathrm{children}_H(\lambda_r) \subseteq \Lambda'$ is called *direct refinement* of λ_r w.r.t. H, and the replacement of labels that generates Λ' from Λ an *elementary refinement operation.*

To prove the above theorem, we need to find a mapping $\mathrm{ref} : 2^{\Lambda} \to 2^{\Lambda'}$ and verify, that it possesses the properties of a refinement mapping as stated in Definition 5.11. Indeed, using $\forall L \subseteq \Lambda$

$$\mathrm{ref}(L) = \begin{cases} L & \text{if } \lambda_r \notin L \\ \mathrm{children}_H(\lambda_r) \cup L \setminus \{\lambda_r\} & \text{otherwise,} \end{cases} \tag{5.19}$$

and recalling that the theorem is restricted to non-leaf λ_r, one concludes for the refinements of the singleton subframes:

$$\mathrm{ref}(\{\lambda\}) = \begin{cases} \{\lambda\} \neq \emptyset & \text{if } \lambda \neq \lambda_r \\ \mathrm{children}_H(\lambda_r) \neq \emptyset & \text{otherwise.} \end{cases} \tag{5.20}$$

To check whether refinements of singleton subsets of Λ are disjoint, we note that except for $\{\lambda_r\}$ all singletons are mapped to themselves. Hence, $\forall \lambda_1, \lambda_2 \in \Lambda$, $\lambda_1 \neq \lambda_2$ and $\lambda_1 \neq \lambda_r \neq \lambda_2$:

$$\mathrm{ref}(\{\lambda_1\}) \cap \mathrm{ref}(\{\lambda_2\}) = \{\lambda_1\} \cap \{\lambda_2\} = \emptyset.$$

If, however, one of the two elements is replaced by λ_r the application of the refinement mapping only yields

$$\mathrm{ref}(\{\lambda_1\}) \cap \mathrm{ref}(\{\lambda_r\} = \{\lambda_1\} \cap \mathrm{children}_H(\lambda_r),$$

and it remains to verified that the set $\mathrm{children}_H(\lambda_r)$ does not contain any element from $\Lambda \setminus \{\lambda_r\}$. This condition can be further transformed:

$$\begin{aligned} & \forall \Lambda \in (\Lambda)_H : \forall \lambda \in \Lambda \setminus \{\lambda_r\} : \lambda \notin \mathrm{children}_H(\lambda_r) \\ \Leftrightarrow \quad & \forall \Lambda \in (\Lambda)_H : \forall \lambda \in \Lambda \setminus \{\lambda_r\} : (\lambda, \lambda_r) \notin \mathrm{R}_{\mathrm{parent}_H}. \end{aligned} \tag{5.21}$$

The statement obviously holds for $\lambda = \lambda_0$, as the root node does not have any parent. Moreover, by Definition 5.12, for any other choice for λ, $\Lambda_0 = \{\lambda_0\}$ contains exactly one ancestor. Due to the unique parent property, each application of the replacement operation (Equation 5.19) on that ancestor replaces it with a sub-frame containing either a single closer ancestor of λ or λ itself in the expanded frame. On the other hand, expanding any label that is not already an ancestor of λ may neither produce λ nor ancestors of λ. Consequently, for any frame generated by this procedure the presence of a label in a frame excludes all ancestors of that element. In particular, this entails the weaker statement expressed in Equation 5.21.

Finally the third and forth conditions of Definition 5.11 are checked by directly substituting the chosen refinement mapping from Equation 5.19. This respectively yields

$$\begin{aligned}
\bigcup\{\mathrm{ref}(\{\lambda\}) \mid \lambda \in \Lambda\} &= \mathrm{ref}(\{\lambda_r\}) \cup \bigcup\{\mathrm{ref}(\{\lambda\}) \mid \lambda \in \Lambda \setminus \{\lambda_r\}\} \\
&= \mathrm{children}_H(\lambda_r) \cup \bigcup\{\lambda \mid \lambda \in \Lambda \setminus \{\lambda_r\}\} \\
&= \mathrm{children}_H(\lambda_r) \cup \Lambda \setminus \{\lambda_r\} = \Lambda'
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{ref}(L) &\overset{(5.19)}{=} \begin{cases} L & \text{if } \lambda_r \notin L \\ \mathrm{children}_H(\lambda_r) \cup L \setminus \{\lambda_r\} & \text{otherwise,} \end{cases} \\
&= \begin{cases} \bigcup\{\{\lambda\} \mid \lambda \in L\} & \text{if } \lambda_r \notin L \\ \mathrm{children}_H(\lambda_r) \cup \bigcup\{\{\lambda\} \mid \lambda \in L \setminus \{\lambda_r\}\} & \text{otherwise,} \end{cases} \\
&\overset{(5.19)}{=} \begin{cases} \bigcup\{\mathrm{ref}(\{\lambda\}) \mid \lambda \in L\} & \text{if } \lambda_r \notin L \\ \mathrm{ref}(\{\lambda_r\}) \cup \bigcup\{\mathrm{ref}(\{\lambda\}) \mid \lambda \in L \setminus \{\lambda_r\}\} & \text{otherwise,} \end{cases} \\
&= \bigcup\{\mathrm{ref}(\{\lambda\}) \mid \lambda \in L\},
\end{aligned}$$

concluding our argument that Λ' is indeed a refinement of Λ.

Apart from the obvious restriction, that the elementary refinement operation can be applied to a given label only after all of that label's ancestors in the hierarchy have been expanded, the order in which the label expansions are carried out is irrelevant to the composition of the generated frame. The finest frame that can be generated for a given label hierarchy is $\{\lambda \in \mathcal{L} \mid \mathrm{children}_H(\lambda) = \emptyset\}$, which consists of the leaves in the label hierarchy.

An example of a label hierarchy is shown in Figure 5.4. The attribute value hierarchy is obtained by subdividing the values a_1 and a_3 according to

$$\begin{aligned}
\mathrm{desc}_H(a_1) = \mathrm{children}_H(a_1) = \{a_{11}, a_{21}, a_{31}\}, \\
\mathrm{desc}_H(a_3) = \mathrm{children}_H(a_3) = \{a_{31}, a_{32}\}.
\end{aligned}$$

Starting from the coarsest non-trivial frame $\{a_1, a_2, a_3\}$ (repeatedly) replacing labels with their direct refinements H produces three new frames of discernment $\{a_{11}, a_{12}, a_{13}, a_2, a_3\}$, $\{a_{11}, a_{12}, a_{13}, a_2, a_{31}, a_{32}\}$ and $\{a_1, a_2, a_{31}, a_{32}\}$.

5.4.2 Uncertainty over Hierarchical Attribute Domains

Having discussed the formalization of attributes and their domains, let us now consider the representation of uncertain knowledge on a family of frames. One factor that gives rise to uncertainty about label assignments are limitations to

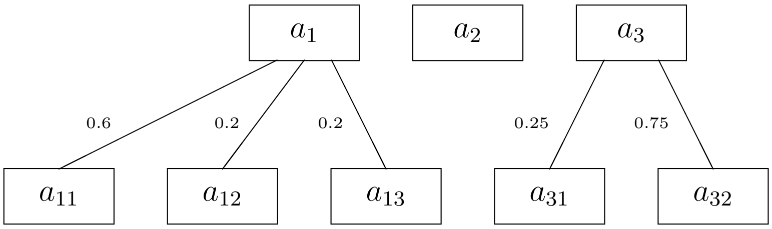

Figure 5.4: Attribute value hierarchy with attached branching probabilities (root label not shown)

observation capabilities. Although the observed label should be linked to the true one at least in a statistical sense, the probability to actually observe the correct label in measurements is usually lower than one. In practice, quantitative models for this type of uncertainty are based on background knowledge about the measuring process or on estimates from representative reference data. They can be formalized by a family of conditional probability distributions characterized by $P_\Lambda(A_\Lambda = \lambda_i \mid A_{\text{obs},\Lambda} = \lambda_j)$, which model the statistical relation of the unknown matching labels A_Λ on a given frame Λ to observed ones $A_{\text{obs},\Lambda}$. In combination with observed instantiations, this determines a distribution $p_\Lambda : \Lambda \to [0,1]$ over Λ and the associated probability measure P_Λ.

When switching between alternative frames of discernment further information deficits arise due to non-corresponding labels. Unlike the previously discussed one, this uncertainty component is not inherent to the measurement process but due to the frame conversion itself. Adequate choices for the frames of discernment help minimize uncertainty in frame conversion. To model this uncertainty component, the representation needs to be supplemented with information about the pairwise statistical interaction between distributions on pairs of frames.

In the absence of a logical dependency structure, knowing the correct label in one frame does not allow to exclude any of the labels of the other frame with certainty. Fortunately, if the admissible frames are restricted to those generated from a single hierarchically structured attribute domain, logical dependencies considerably simplify the representation. In particular a given label λ will always have the same (conditional) probability regardless of the frame it appears in. The (conditional) prior distributions over all frames can therefore be represented using probability assignment functions $p_H(\lambda)$ and $p_H(\lambda_1 \mid \lambda_2)$ on the labels in $\mathcal{L}$, with the distribution on any particular frame Λ given by the restriction of the respective assignment function to the elements of Λ.

Moving the discussion to the determination of the probability values, it is serviceable to start with the conditional assignments. In the most simple case, no label conversion is required at all. This trivially leads to

$$\forall \lambda \in \mathcal{L}: \quad p_H(\lambda \mid \lambda) = 1. \tag{5.22}$$

For an object that is correctly described by a label λ, the branch probabilities $p_H(\lambda' \mid \lambda)$ quantify the uncertainty regarding which of the sub-labels $\lambda' \in \mathrm{children}_H(\lambda)$ provides the matching description on a frame, where label λ is expanded. Since a label λ in a frame Λ correctly describes a situation whenever any one of its descendants does so on a refinement of Λ, we also have:

$$\forall \lambda_1, \lambda_2 \in \mathcal{L},\ \lambda_2 \in \mathrm{desc}_H(\lambda_1): \quad p_H(\lambda_1 \mid \lambda_2) = 1. \tag{5.23}$$

To go from general labels to more specific ones, the branch probabilities for each elementary refinement step on the path from λ_2 to λ_1 are combined:

$$\begin{aligned} &\forall \lambda_1, \lambda_2 \in \mathcal{L},\ \lambda_2 \in \mathrm{anc}_H(\lambda_1): \\ &\qquad p_H(\lambda_1 \mid \lambda_2) \quad = \prod_{\lambda' \in \{\lambda_1\} \cup (\mathrm{anc}_H(\lambda_1) \cap \mathrm{desc}_H(\lambda_2))} (\lambda' \mid \mathrm{parent}_H(\lambda')). \end{aligned} \tag{5.24}$$

In the remaining cases, the labels that are not located on the same refinement path through H. Such labels are mutually exclusive:

$$\forall \lambda_1, \lambda_2 \in \mathcal{L},\ \lambda_1 \neq \lambda_2,\ \lambda_1 \notin \mathrm{anc}_H(\lambda_2),\ \lambda_2 \notin \mathrm{anc}_H(\lambda_1): \quad p_H(\lambda_1 \mid \lambda_2) = 0 \tag{5.25}$$

To clarify the last statement, consider the frame from Λ that is generated by consecutive application of the elementary frame refinement operation w.r.t. the labels $\{\mathrm{anc}_H(\lambda_1) \cup \mathrm{anc}_H(\lambda_2)$. Although both labels are specializations of λ_0 there is some point in the refinement process where their last common ancestor is split into a new sub-frame, thereby discerning the cases described by λ_1 from those described by λ_2.

The marginal prior probability assignments for the labels in $\mathcal{L}$ can be obtained directly from the conditional ones. Because the coarsest possible frame Λ_0 only contains one element, there is only one valid probability distribution over that frame. It immediately follows that $p_H(\lambda_0) = P_H(\{\lambda_0\}) = P_H(\Lambda_0) = 1$. The prior probability distributions over frames are then reconstructed by multiplying branch probabilities along a path of serial refinements, i.e.:

$$\begin{aligned} \forall \lambda \in \mathcal{L}: \ p_H(\lambda) &= p_H(\lambda_0) \cdot \prod_{\lambda' \in (\{\lambda\} \cup \mathrm{anc}_H(\lambda)) \setminus \lambda_0} p_H(\lambda' \mid \mathrm{parent}_H(\lambda')) \\ &= \prod_{\lambda' \in (\{\lambda\} \cup \mathrm{anc}_H(\lambda)) \setminus \lambda_0} p_H(\lambda' \mid \mathrm{parent}_H(\lambda')). \end{aligned} \tag{5.26}$$

Because the marginal and conditional prior probability assignments can be assembled from the branch probabilities $p_H(\lambda \mid \mathrm{parent}_H(\lambda))$, $\lambda \neq \lambda_0$ alone, the situation lends itself well to a probability tree representation. In this representation the value assigned to a non-leaf label λ summarizes the probability mass p_H of the leaves in the subtree rooted at λ (illustrated in Figure 5.4). When computing the distributions p_H the values $p_H(\lambda)$ can be reused as common factors in the decomposition of the probability function for all descendants of λ. Moreover, uncertainty due to frame conversion is contained locally in the respective subtrees.

With the conditional prior distributions over the labels in $\mathcal{L}$ we now have the means to convert distributions between frames of the same family. As a result of the imposed restrictions, only three cases have to be considered when mapping an element λ_1 from a frame Λ_1 to a frame Λ_2 generated by the same hierarchy H of attribute values:

- λ_1 is an element of Λ_2 as well,
- λ_1 summarizes a sub-frame $L \subseteq \Lambda_2$ consisting only of its (possibly indirect) descendants in the hierarchy
- λ_1 is itself part of a sub-frame associated with a unique element of Λ_2.

It is remarked, that none of the frames is marked out so Λ_1 and Λ_2 can be interchanged in that statement (Figure 5.5). Instantiating with observations on the source frame Λ_1 and applying the Equations 5.22–5.25 to infer probability values for the elements of Λ_2 results in a rule for converting distributions between related frames:

Definition 5.14. *Let Λ_1 and Λ_2 be two frames of discernment generated from the same hierarchy H and p_{Λ_1} a probability function over Λ_1. The mapping* $\mathrm{conv}_{\Lambda_1 \to \Lambda_2} : \mathrm{Prob}(\Lambda_1) \to \mathrm{Prob}(\Lambda_2)$ *that converts a probability distribution over a frame Λ_1 to a probability distribution over Λ_2 is computed as:*

$$\begin{aligned} p_{\Lambda_2}(\lambda_2) &= \mathrm{conv}_{\Lambda_1 \to \Lambda_2}(\lambda_2) \\ &= \begin{cases} p_{\Lambda_1}(\lambda_2) & \text{if } \lambda_2 \in \Lambda_1, \\ \displaystyle\sum_{\lambda \in \mathrm{desc}_H(\lambda_2) \cap \Lambda_1} p_{\Lambda_1}(\lambda) & \text{if } \mathrm{desc}_H(\lambda_2) \cap \Lambda_1 \neq \emptyset, \\ p_H(\lambda_2 \mid \lambda) \cdot p_{\Lambda_1}(\lambda) & \text{if } \exists \lambda \in \Lambda_1 : \lambda_2 \in \mathrm{desc}_H(\lambda), \end{cases} \end{aligned} \tag{5.27}$$

where $\mathrm{Prob}(\Lambda)$ *denotes the set of all conceivable probability distributions on a frame Λ.*

Note, that the probability assigned to a given label does not depend on the composition of the particular frame Λ_2 under consideration. Indeed, equation 5.27

may be used to condition the probabilities for all labels in the hierarchy with new information from observations on a specific frame. Moreover, the function p_H itself encodes the a-priori probabilities for all labels and – via its restrictions – the a-priori distributions over the frames of the hierarchy reflecting the generic component of this knowledge model.

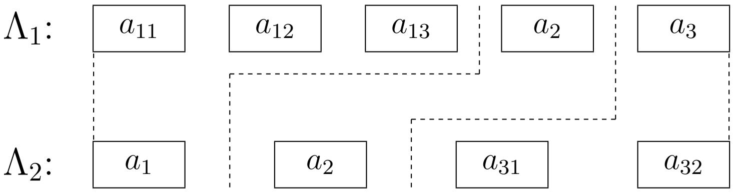

Figure 5.5: Correspondence of subframes and single labels

Example 5.1. Consider a conversion from Λ_1 to Λ_2 as given by figure 5.5 with an original distribution on Λ_1 given by

$$p_{\Lambda_1}(a_{11}) = p_{\Lambda_1}(a_2) = 0.2, \quad p_{\Lambda_1}(a_{12}) = p_{\Lambda_1}(a_{13}) = 0.1 \quad \text{and} \quad p_{\Lambda_1}(a_3) = 0.4.$$

We apply Equation 5.27 to compute the distribution $p_{\Lambda_2} = \mathrm{conv}_{\Lambda_1 \to \Lambda_2}(p_{\Lambda_1})$ induced over the new frame. The sub-frame $\{a_{11}, a_{12}, a_{13}\} = L \subseteq \Lambda_1$ is represented in Λ_2 by the single attribute value a_1. Thus the probability assigned to the elements of L in Λ_1 is attributed to a_1 in Λ_2, i.e

$$p_{\Lambda_2}(a_1) = p_{\Lambda_1}(a_{11}) + p_{\Lambda_1}(a_{12}) + p_{\Lambda_1}(a_{13}) = 0.4.$$

Label a_2 appears in both frames, so

$$p_{\Lambda_2}(a_2) = p_{\Lambda_1}(a_2).$$

The two remaining elementary probability values are computed using the estimated sub-label distribution:

$$\begin{aligned} p_{\Lambda_2}(a_{31}) &= p_H(a_{31} \mid a_3) \cdot p_{\Lambda_1}(a_3) = 0.1 \quad \text{and} \\ p_{\Lambda_2}(a_{32}) &= p_H(a_{32} \mid a_3) \cdot p_{\Lambda_1}(a_3) = 0.3. \end{aligned}$$

This fully determines the probability function p_{Λ_2}. □

5.4.3 Multi-Label Instantiations

So far it was assumed that all objects in O could be described using no more than one label per object. But given that only a subset of those objects would actually have been observed by the time the attribute hierarchy is chosen, new cases may not always be suitably characterized by any of the predefined categories. Furthermore, if composite objects are considered (e.g. texts) or objects that interact within a complex system (genes) a single label per attribute does not provide the best possible specification. In such cases, it is more realistic to have an attribute A^* assign a set of *applicable* labels to each object, that is, $A^*_\Lambda : O \to 2^\Lambda \setminus \{\emptyset\}$, where 2^Λ denotes the power set of Λ. An object is thus described by the set of all labels applicable to it. The downside of that solution is, that the labels are not mutually exclusive, so that uncertainty may no longer be represented by a simple probability distribution over the label domain.

The introduction of random sets formally reduces the problem of uncertainty representation for set-valued attributes to the probabilistic case. The main difference to the approach for disjoint labels discussed in the previous subsection is, that the distributions are now defined on the power set of the frames. When investigating corresponding multi-label instantiations on frames from the same family $(\Lambda)_H$, the constraints previously discussed for instantiations with individual labels still apply: Since the frames of discernment generated by the same hierarchy may have labels in common, any labels that are applicable in one frame also apply in all other frames containing those labels. Moreover, if a label applies to a given case or object, so do all its ancestors in H. Finally we know that at least one of the children of an applicable non-leaf label must apply when using a description in a refined frame. In the light of these considerations it makes sense to define multi-label instantiations globally for all frames in $(\Lambda)_H$.

Definition 5.15. *Let H be a hierarchy of labels $\mathcal{L}$. A set of labels $S_H \subseteq \mathcal{L}$ is a* **multi-label instantiation** *w.r.t. H, if it fulfills the following properties:*

1. $\forall \lambda \in S_H : \forall \lambda' \in \mathcal{L} : \lambda' \in \mathrm{anc}_H(\lambda) \implies \lambda' \in S_H$,
2. $\forall \lambda \in S_H : \mathrm{children}_H(\lambda) \neq \emptyset \implies \exists \lambda' \in \mathrm{children}_H(\lambda) : \lambda' \in S_H$.

With these constraint each S_H specifies a sub-tree of the labels in H that contains exactly those elements of $\mathcal{L}$ that apply to a given situation. The multi-label instantiation for a particular frame Λ is obtained by intersecting S_H with that frame. It is remarked that due to the first condition the maximally refined frame contains all the information to recover instantiations for all other frames. In practice however, one would expect that for at least some of the cases observations are only available on one of the coarser frames, so that knowledge about the elements of S_H remains incomplete.

The above strategy for the consistent representation of set-instantiations in a hierarchy is easily adapted for use with distributions. To model a family of probability distributions p^*_Λ, $\Lambda \in (\Lambda)_H$ over set-instantiations on frames in a hierarchy, the latter are viewed as induced manifestations of a single probability distribution p^*_H over the permissible multi-label instantiations $S_H \subseteq \mathcal{L}$. In order to reconstruct that global distribution p^*_H from observations, it is assumed that the distributions on sub-frames that arise from an elementary refinement operations are (statistically) independent of each other given their respective parent labels. This means that direct interactions w.r.t. to the applicability of labels of a frame are considered only between the labels that are generated from the same refinement operation. With this assumption the distribution over set instantiations can be decomposed into conditional branch probabilities

$$p^*_{H,\lambda_r}(S) = P^*_H(\{S_H \mid S = S_H \cap \mathrm{children}_H(\lambda_r)\} \mid \{S_H \mid \lambda_r \in S_H\}), \qquad (5.28)$$

with $\lambda_r \in \mathcal{L}$, $S \subseteq \mathrm{children}_H(\lambda_r)$ and $P^*_H(\{S_H \mid \lambda_r \in S_H\}) > 0$. For each λ_r the values $p^*_{H,\lambda_r}(S)$ define a probability distribution over the selections of children of λ_r providing information about how the the applicable labels of an instantiation are split when switching to a finer frame of discernment (Figure 5.6). An advantage of the representation strategy via branch distributions is that even partial specifications of S_H contribute to the estimation of $p^*_{H,\lambda_r}(S)$ as long as the respective label λ_r has been expanded in the frame on which the observation is based. If nevertheless $P^*_H(\{S_H \mid \lambda_r \in S_H\}) = 0$ for some λ_r then the branch probabilities for the respective sub-frame instantiations remain undefined, and the modeled probability for instantiations containing labels from that subtree is 0. To counter practical difficulties connected to undefined branch distributions in reasoning, one would often choose to set those branch distributions heuristically, e.g. using the Laplace correction, thereby ensuring that results for previously unseen instantiations can still be computed (see Algorithm 1).

Although in order to convert distributions between frames the branch distributions would need to be normalized to

$$\sum_{S \subseteq \mathrm{children}_H(\lambda_r)} p^*_{H,\lambda_r}(S) = 1,$$

it is remarked that values may also be stored in their unnormalized form, e.g. to encode the relative frequency of observations of the unrefined sub-frame in empirical data. Unless explicitly stated otherwise, all explanations in this section refer

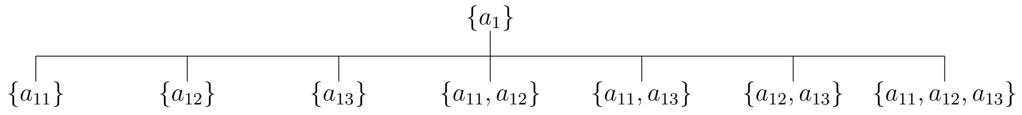

Figure 5.6: Possible refinements of an applicable label during sub-frame expansion w.r.t. the label a_1 from Figure 5.4

to normalized sub-frame distributions though. A further layer of compression is introduced by switching to a condensed representation for the subset branch distributions in the refinement of applicable labels. Applied to all expandable labels of the hierarchy, this leads to the data structure depicted in Figure 5.7. For each non-leaf label λ_r an additional label $\lambda_r^\diamond$ is introduced. In the condensed representation the conditional probability assigned to that label reflects the probability that the applicable label λ_r is split into more than one applicable child labels during the next stage of sub-frame expansion. This is complemented with conditional coverage factors, which are stored for each element in the direct refinement of λ_r.

To estimate the model parameters from empirical data the Equations 5.1 and 5.3 are applied to the branch distributions of non-leaf labels λ_r. The respective reference set is formed by those observations, for which λ_r is both applicable and has been expanded on the observed frame. In that case information on the applicability of the individual child labels of λ_r is available too. By viewing the respective intersections $S_H \cap \mathrm{children}_H(\lambda_r)$ as instantiations of a multi-valued attribute with basic domain $\mathrm{children}_H(\lambda)$ the empirical branch distribution is calculated from the relative frequencies of these instantiation. An algorithm to calculate the branch distributions for a given label hierarchy H is given below (Algorithms 1 and 2).

The branch distribution on the originally set-valued selections of applicable labels from elementary refinement of λ_r is represented using the condensed distribution $(p^\diamond_{H,\lambda_r}, c^\diamond_{H,\lambda_r})$, with the new element $\lambda_r^\diamond$ representing the multi-valued outcomes. The lcorr parameter denotes a user-defined constant for an optional Laplace correction, which is applied for both the induction of branch probabilities and

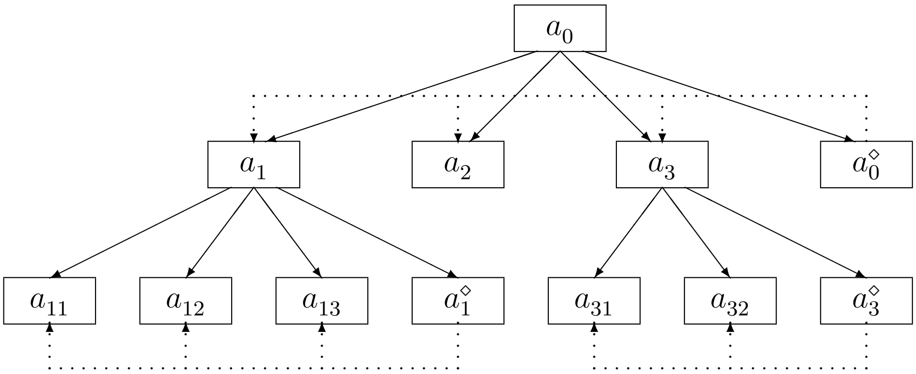

Figure 5.7: Extended attribute value hierarchy as data structure for the condensed representation of distributions over multi-valued instantiations (conditional probabilities and coverage factors indicated by solid and dotted arrows respectively)

Algorithm 1 Inducing branch distributions

```
procedure getFrequencies(H,Observations,lcorr)
    ResetCounters
    for currentObs ∈ Observations do
        for $\lambda$ ∈ currentObs do
            markLabel($H, \lambda$)
            markAncestors($H, \lambda$)
        end for
        updateCounters($H$)
        obsCnt ← obsCnt + 1
    end for
    $\lambda_r$ ← rootLabel($H$)
    while validLabel($\lambda_r$) do
        nrm_p ← $\max\{1, \text{GetNumSgltExp}(\lambda_r) + \text{GetNumMltvExp}(\lambda_r)$
                         $+ (1 + |\text{children}_H(H, \lambda_r)|) \cdot \text{lcorr}\}$
        nrm_c ← $\max\{1, \text{GetNumMltvExp}(\lambda_r) + 2 \cdot \text{lcorr}\}$
        for $\lambda$ ∈ childrenH($\lambda_r$) do
            $p^{\diamond}_{H,\lambda_r}(\lambda) \leftarrow \frac{(\text{GetNumAsSglt}(\lambda) + \text{lcorr})}{\text{nrm\_p}}$
            $c^{\diamond}_{H,\lambda_r}(\lambda) \leftarrow \frac{(\text{GetNumAsCvrd}(\lambda) + \text{lcorr})}{\text{nrm\_c}}$
        end for
        $p^{\diamond}_{H,\lambda_r}(\lambda_r^{\diamond}) \leftarrow \frac{(\text{GetNumMltvExp}(\lambda_r) + \text{lcorr})}{\text{nrm\_p}}$
        $\lambda_r$ ← nextLabelInDepthFirstSearchOrder($H$)
    end while
end procedure
```

conditional coverage factors (the latter being instances of a two class problem). The bounding of the normalization factors ensures that all marginal probabilities will be defined, even if the Laplace correction is not applied. This guarantee does not extend to conditional branch probabilities though. By altering the normalization factors the algorithm is easily adapted to alternative interpretations of the non-expanded values in the training data set. An experimental evaluation of these algorithms is provided in Chapter 6.

To facilitate the use of the above representation in models, let us now turn to the recovery the stored information. To recover a set-distribution from the representation, the conditional branch distributions on the hierarchy are recombined into corresponding distributions on the frames. For for singleton outcomes this is analogous to the case with single-label instantiations and achieved by multiplying

Algorithm 2 Inducing condensed branch distributions (counting)

```
procedure updateCounters(H)
    $p \leftarrow$ rootLabel($H$)
    while validLabel($p$) do
        nMarkedChildren $\leftarrow 0$
        for $c \in$ ChildrenH($p$) do
            if isMarked(c) then
                if nMarkedChildren $= 0$ then
                    firstChild $\leftarrow c$
                    nMarkedchildren $\leftarrow 1$
                else
                    countAsCovered($c$)
                    nMarkedChildren $\leftarrow$ nMarkedChildren $+ 1$
                end if
            end if
        end for
        if nMarkedChildren $= 1$ then
            countAsSingleton(firstChild)
            countSgltExpansion($p$)
        else if nMarkedChildren $> 1$ then
            countAsCovered(firstChild)
            countMltvExpansion($p$)
        end if
        clearMark($H, p$)
        $p \leftarrow$ nextMarkedLabelInDepthFirstSearchOrder($H$)
    end while
end procedure
```

branch probabilities along a path of label refinement, i.e.

$$\forall \lambda \in \mathcal{L}: \quad p^{*\prime}_{H}(\{\lambda\}) = \prod_{\lambda' \in (\{\lambda\} \cup \mathrm{anc}_H(\lambda)) \setminus \lambda_0} p^{\diamond}_{H,\mathrm{parent}_H(\lambda')}(\lambda'). \tag{5.29}$$

In general the approximation will be imperfect. In addition to the unavoidable sampling error, the branch distributions do not distinguish between real singletons and cases where a label is merely the only applicable element in the local sub-frame. If required and provided sufficient training data is available, additional precision in the probability approximation for singletons can be obtained by adding a separate set of branch distributions though.

The one point coverages of individual labels are retrieved by recursively accumulating conditional probabilities and coverage factors for each elementary refinement leading to the label in question. For a single recursion step the reconstructed one-point coverage of a given label is obtained by application of Equation 5.3. Because each branch distribution refers only to those cases where the respective ancestor labels are applicable, the result is than multiplied with the respective one-point coverage for the ancestors: $\forall \lambda \in \mathcal{L} \neq \lambda_0,\ \lambda_r \overset{\text{def}}{=} \text{parent}_H(\lambda)$:

$$\text{opc}'_H(\lambda) \quad = \quad \text{opc}'_H(\lambda_r) \cdot \left(p^{\diamond}_{H,\lambda_r}(\lambda) + p^{\diamond}_{H,\lambda_r}(\lambda^{\diamond}_r) \cdot c_{H,\lambda_r}(\lambda)\right), \tag{5.30}$$

where λ_r is used as a shorthand notation for the parent label of λ in the hierarchy and $\lambda^{\diamond}_r$ the corresponding surrogate label that indicates multiple applicable elements in the extension of λ_r. Each level in the hierarchy adds another factor until the root label λ_0 is reached. If the empty instantiations are excluded the one-point coverage of that label is always one, as it is the only element in its frame[4] To efficiently compute one-point coverages for several elements of a frame an implementation would reuse partial results whenever the recursion runs over shared ancestors in the hierarchy. Under the assumption that applicability of the individual labels within an elementary refinement is independent for non-singleton instantiations, the one-point coverages can also be used to approximate probability values for multi-valued instantiations, though the approximation quality is lower than for the singletons (consult Section 6.6 for detailed evaluation results).

Finally case-specific information on one-point coverages and probabilities can be integrated to allow reasoning. This is achieved by temporarily fixing conditional branch distributions to externally supplied inputs. In the next step the distributions on the target frame(s) is recomputed with the provided values taking precedence over those supplied by the model. Recursions are broken early whenever one of the externally provided values is encountered and only the missing conditional branch probabilities are supplemented by the model.

While the proposed representation strategy leads to a considerable reduction in the information content, it focuses on preserving local interactions between closely related labels. To justify this emphasis one can argue that interactions between labels that are separated by longer paths in the hierarchy are often indirect. Such indirect interactions are difficult to observe in isolation though and the superposition of many influences is likely to mask the signal of genuine interactions, making it indistinguishable from sample variation in empirical data. The emphasis on the local sub-frames is therefore in line with the goal of obtaining a good generalization. Equations 5.29 and 5.30 assume the local distribution

[4] Otherwise, empty instantiations can easily be represented by inserting a "virtual" root label with an unnormalized branch distribution at the top of the hierarchy. In that case, the one point coverage of the original root label is computed using Equation 5.30, whereas the one-point coverage of the new root label is set to one.

within direct refinements to be invariant w.r.t. presence of alternative labels on coarser frames. Depending on the interpretation of set-valuedness, this assumption may require justification. It can be avoided though, by introducing separate sets of conditional probabilities.

5.4.4 Summary

The hierarchically structured attribute domain permits to combine observations from sources that differ w.r.t. resolution, reliability and focus, into a single coherent probabilistic model. In particular the representation makes use of a family of interrelated frames of discernment that are generated from the same attribute value hierarchy. The integration of information from different sources is due operations that map distributions between frames. By projection of distribution information onto any of the linked frames, the same operations enable application task, such as enrichment analysis, clustering, classification and identification, the quantitative assessment of the (dis)agreement of experiments to control conditions, or simply the presentation of data on multiple scales. If the hierarchical model is used to represent prior knowledge on the general distribution of sublabels, the operations can to support decisions under uncertainty by integrating case-specific observations w.r.t. specific frames. Finally, the model can serve as a component in larger probability-based models.

For use with annotation sets, the approach was adapted to use a condensed representation of uncertain set-valued information (introduced in Subsection 5.4.3). This approach is loosely related to the more general framework of random sets but trades storage efficiency for some representation capabilities. It was argued, that this the reduction in representational power is justified when models have to be induced from data as the detailed interaction structure potentially available with a full representation is often masked by sampling effects. A demonstration of the developed model including an evaluation on a large, publicly available scientific dataset will be discussed in the next chapter.

5.5 Discussion

The condensed set-valued attributes and their associated distributions introduced in this work provide a compact, yet informative summary of probability distributions of over sets. Due to its combination of a coarsened frame of discernment and the use of coverage factors the approach takes an intermediate positions between a full probabilistic modeling (using random sets) and approximations via independent binary attributes for each element in the underlying set universe.

The representation allows for high-quality approximations of distributions over sets in empirical data (see Chapter 6 for evaluation results).

One of the most distinguishing features of the suggested representation format is the separate representation of singleton outcomes in conjunction with additional model parameters that permit to reconstruct one-point coverages. This property addresses elements that are important for the interpretation of many knowledge representations. For instance, singleton rate and one-point coverage correspond to lower and upper probability bounds when interpreting set-valued outcomes to model imprecision. The high incidence of singleton annotations in the datasets studied in the evaluation part of this work indicates that the focus on singletons is also advantageous in other contexts where variables are permitted simultaneously adopt several values of their domain, such as with annotation databases. Because annotation databases are one of the most widely available resources in computational biology the model has a number of applications in that field.

The principle behind the condensed representation is a partitioning of a power set an the basis of element cardinality. These partitions are then used to group set outcomes and calculate frequency statistics and coverage factors for each group. Of course the same principle could be applied to further subdivide the group of non-singleton subsets leading to a complete family of distribution models that form intermediates between with full random sets and possibility distributions in the sense of the context model at the other end of the spectrum. However restricting the approach to the separate representation of singleton subsets has distinct advantages:

- Singleton outcomes are abundant in many datasets featuring set-valued attributes.
- The separate representation of singletons is sufficient to subsume the probabilistic models as a special case, thus establishing compatibility with extant tools and approaches.
- Singletons have a special role in the interpretation of knowledge models (see previous paragraph).
- The separate representation of singletons entails only a moderate increase in storage requirements over probability and possibility distributions. Because a base domain with n elements has $\binom{n}{m}$ different m-elementary subsets, the restriction to singletons ($m = 1$) allows the model size to remain linear in the cardinality of the underlying base domain.

It is also remarked that due to the capability of the model to reconstruct one-point coverages, non-normalized possibility distributions are included as special cases.

It was shown, how the approach is transferred to multivariate case and how marginal and conditional distributions are computed. These operations allow to employ Graphical Models to summarize distributions for high-dimensional domains and to propagate the effects of local modifications. Like with the probabilistic framework, marginal and conditional distributions always retain their empirical interpretation w.r.t. the modeled world. In particular one-point coverages of all marginal distribution are consistent with the context model interpretation of possibility. Owing to this property generic knowledge represented in the distributions can be combined with partial case-specific information from observations. In this respect the model exceeds the capabilities of possibility models, for which marginal distributions retain their empirical interpretation only if the underlying random set has the consonance property.

The increasing relevance of ontologies as knowledge representations motivated the development of a hierarchical version of the framework. In that variant hierarchically structured relations specified in those ontologies are enriched with empirical distributions obtained from data. When operating with annotations that draw on labels from a structured term set, the models capability to "sum out" variables is employed to present empirical distribution on multiple levels of detail. The resulting combination of domain knowledge reflected in the ontology with meaningful statistics obtained via the model supports the analysis of large complex datasets.

The next chapter provides an experimental validation of the introduced concepts and algorithms for a prototypical application case. The analysis is based on a widely-known publicly available gene-annotation data set and uses the associated ontology to compute and represent a functional profile of the genome of a model organism. Such profiles can be applied to highlight and assess similarities and differences between populations grown under different conditions or between phylogenetically related species.

6 Experiments and Evaluation

The elaboration of desired model properties and the study of limitations of extant approaches given in Chapter 4 provided a theoretical basis for assessing the suitability of particular model assumptions and mathematical formalisms. These results were reflected in the design decisions for the condensed representation for set distributions introduced in Chapter 5.

But it remains to be investigated how the proposed model and its inherent set of assumptions affect results when processing empirical data under realistic conditions. This chapter provides this complementary picture, pointing out the strengths and weaknesses of the suggested representation scheme in a prototypical application context. The performance of the condensed random set models is evaluated and and compared to popular alternative approaches. All models were tested on a biological dataset from a publicly available research database.

Section 6.1 briefly introduces the dataset and discusses the purpose of the annotations provided therein. The outline of the experiment and a list of investigated model types are given in Section 6.2. This is followed by details on data preparation and preprocessing (Section 6.3), an exposition of the respective methods to estimate the model parameters from training data and some information on how these parameters are used to predict target values (Section 6.4). Section 6.5 is concerned with a discussion of the evaluation measures used. The results of the experiments are presented and discussed in Section 6.6. An abbreviated version of this evaluation was presented in Rügheimer (2010).

6.1 Saccharomyces Gene Annotation Data

The study has been conducted on an annotated genome dataset released to the public via the Saccharomyces Genome Database project (SGD Curators, b). The SGD-project maintains a curated database that summarizes published results about the function of the genes and gene products of the baker's and brewer's yeast Saccharomyces cerevisiae, as well as their respective roles in biological processes and their intracellular activity sites. Annotation follows a domain-wide standard defined in the gene-ontology (GO, Ashburner et al., 2000). The ontology provides a controlled vocabulary and defines term relations that allow to relate annotations on different levels of specificity to each other. Terms are organized into three non-overlapping term hierarchies provided for the tree aspects of annotation: (biological) process, (molecular) function, and cellular component. The process aspect describes *what* general cell-level functionality a gene product is contributing to, e.g., "carbohydrate metabolism". The molecular function aspect focuses on *how* the gene product is involved in biochemical reactions relevant for that process. The term "hydrolase activity", for instance, marks the gene product as a member of an particular enzyme class. Finally the "cellular component" aspect describes *where* the activity of the gene product takes place (example: cytoplasm). The three term hierarchies form separate branches of the ontology and are connected to each other only via the common root node and some supplementary relations that provide cross references are not relevant to definition of the ontology structure itself.

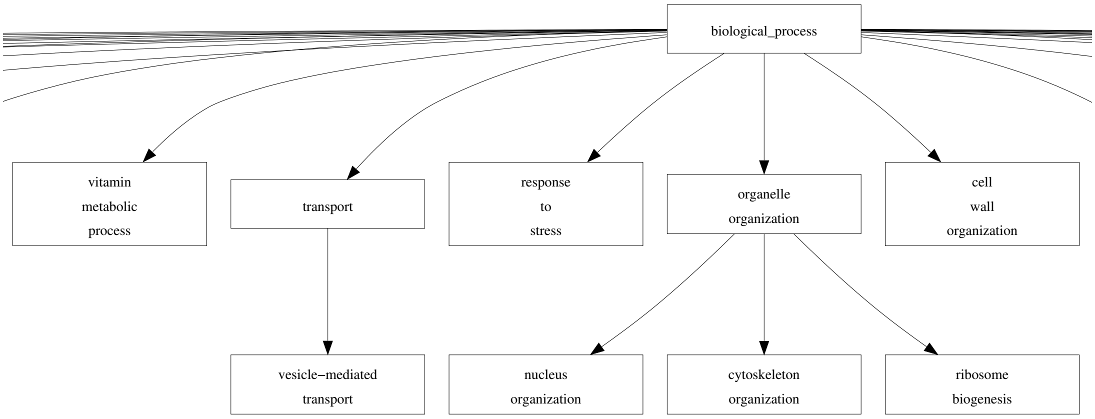

Figure 6.1: Segment of the slim version of the biological process sub-ontology in GO: Edges indicate hyponym relations, that is, “nucleus organization” is represented as a special type of “organelle organization”

Because the full annotation is very detailed, a considerable fraction of the annotation terms is only applied to a very small subset of the database. Due to their extremely low term coverage they do not lend themselves to a statistical analysis. To provide a standardized broader view of the represented knowledge, less specific versions of the ontology have been released by the consortium. These so-called “slim ontologies” define species-specific subsets using comparatively general Gene Ontology terms. They are usually released together with the annotation data collected in coordinated efforts to analyze the genome and proteome of the respective organism. The dataset used in this study was based on a projection of the full SGD annotations to a subset of relatively broad gene-ontology terms – the GO-Slim terms for yeast (SGD Curators, a). Terms that were not included in the in the slim version of the ontology were mapped to their most specific remaining ancestor in the original term hierarchy set. Current and archived versions of both that mapping and the GO-Slim itself are maintained at the SGD website.

6.2 Model Types and Experimental Setup

To evaluate the proposed framework, test its underlying assumptions and compare its predictions with those of alternative frameworks, several distribution models were implemented and evaluated on the *S. cerevisiae* dataset. In particular, this comparison included the following model types:

- A model in which presence or absence of elements in a set are encoded using binary variables (INDEP). The latter variables are treated as independent, so the distribution of set-instantiations is obtained as a product of binary distributions for the state of the elements of the underlying carrier set. The set-distribution is described via its one-point-coverage.
- A condensed distribution model using an unstructured attribute domain (CDM) as described in Section 5.2;
- The hierarchical version of the condensed distribution model (HCDM) as described in Section 5.4);
- Two Bayesian Models (BN1 and BN2) induced from the training data using conditional independence tests (cf. Section 3.4). The models differ in that BN2 applies more stringent conditional independence tests than BN1. Thus the model variants represent different trade-offs between efficiency gained from the decomposition and accuracy on the training data.
- Two models based on a full Random Set representation (RS1 and RS2) using Laplace corrections of 10^{-9} and 10^{-12} respectively.

The Random Set representation serve as a reference for the assessment of the evaluation scores. Their practical use, however, is limited due to the large number of possible set-instantiations. Even when only the focal sets are stored the model effectively consists of a slightly compressed representation of the training data set[1].

[1] The same argument applies to all other models that rely on the direct assignment of probability or belief mass to the focal sets (see Section 4.3).

For the experiment the models for each of the listed types were used to represent the distribution of annotation sets for a randomly selected subset of the yeast genome. The resulting distribution models were then compared with the distribution of the annotation term combinations on the remaining genes. To that end approximation quality and generalization were evaluated based on measures that respectively emphasize overall quality of fit, the representation of singleton outcomes and the prediction of element coverage.

Because the distributions are learned from data, it is worthwhile to dedicate some attention to the so called *variance error*. If the observed data is considered as a sample obtained by drawing from the unknown distribution to be approximated, then the learning methods have no means to distinguish between general patterns due to regularities in that distribution and mere coincidences that arose as a result of the sampling procedure. Thus, models induced from data exhibit a tendency to adapt to such random, non-generalizable patterns in the data – an effect known as *overfitting*. This undesirable effect is revealed by evaluating data on a test data set, which is separate from the one used in the training. To increase the robustness of the evaluation and avoid biases due to sampling effects, a cross-validation strategy was employed for all experiments.

6.3 Data Preparation

Due to the structure of the database maintained in the SGD project, each assignment of an annotation term to a gene is represented as separate database record. Apart form the gene name and annotation term these records contain supplementary information, such as alternative gene names, the annotation aspect class, types of information sources used to assign the annotation, references to the location of the gene within the genome or connected publications.

Because the content of the annotation databases has been compiled from various publications spanning several years, preprocessing had to ensure that unique identifiers were used for each gene. For historical reasons some genes were given several alternative names. For instance a gene may have been assigned a name at the time of its initial sequencing and been renamed once the gene was linked to a specific function. In other cases gene homonyms arose from their association with homologue genes in other species, or due to independent discovery by different research groups. In order to ensure that annotations can be correctly attributed, the first step of preprocessing consisted in mapping all alternative gene names to unique standard identifiers which are used throughout all subsequent processes.

Following the mapping to standard identifiers, the records where filtered according to the annotation aspect given. For the purpose of this evaluation only annotation w.r.t. the “biological process” aspect were chosen. The annotations on the biological processes provide a comparatively reliable and extensive higher-level description of the role of the gene product in the organism. In the remaining part of the database, annotations for individual genes are still spread over several database records. To better support a gene-based view on the data annotations where grouped by the genes they refer to. The resulting file summarizes the known biological function for each of 6849 genes using 909 distinct annotation sets.

In addition to compiling the annotations-sets from the database, the preprocessing stage included routines to assemble information about the annotation scheme itself: The term hierarchy structure was extracted from the ontology and converted into a domain specification for the hierarchical version of the condensed distribution models. In a similar manner domain specifications were prepared for the non-hierarchical version, the model based on independent binary variables, the Graphical Models and the Random Set representations. The domain specifications for those models, however, were simple lists of annotation terms, that is the information on term organization was disregarded. The generated domain specifications were later used in the training phase to configure the learning algorithms for the respective distribution models .

The above preprocessing resulted in a database of annotation sets for 6849 genes. For the assessment the database randomly divided into five disjoint partitions (4 partition with 1370 genes each and one partition with 1369 genes). To limit sampling effects, the evaluation measures were computed in a 5-fold cross-validation process (Mosier, 1951; Kohavi, 1995; Mitchell, 1997) with a different partition serving as a test data set and the remaining partitions providing training data in each run. Because the implementations of the models require different input formats all training and tests sets were converted to an alternative file format based on a binary encoding of the annotation sets.

6.4 Parameter Estimation

Using the model configuration files prepared in the preprocessing step and the training data for each validation run, the different model types were trained for the distribution of gene annotation sets. In the case of the model with independent binary variables the parameter set consists of one value per element in the carrier set. Each value describes the probability of an instantiation containing its associated element. The modeled probability $\hat{p}_{\mathrm{INDEP}}(S)$ of any given set-instantiation $S \subseteq \Omega$ is obtained by computing the products

$$\hat{p}_{\text{INDEP}}(S) = \left(\prod_{\omega \in S} \operatorname{opc}(\omega)\right) \cdot \left(\prod_{\omega \in \Omega \setminus S} (1 - \operatorname{opc}(\omega))\right), \tag{6.1}$$

with the model parameters $\operatorname{opc}(\omega)$ denoting the (estimated) probability of ω to be an element of the outcome. Coverage rates for the elements of the carrier set are estimated from the observed frequencies of the two possible outcomes “element is present in the instantiation” and “element is absent in the instantiation” in training data.

For the condensed distribution the parameters are singleton probabilities and conditional coverage factors for the distribution. The hierarchy-based condensed distribution model coverage factors refer to subtrees of the label hierarchy instead. For a detailed description of parameters and the model induction procedures see Sections 5.2 and 5.4 respectively.

The Bayesian Network models were induced from the binary encoding of the training files using INeS Borgelt and Kruse (2002) – an open source implementation of several algorithms for the induction of Graphical Models. To determine network structure the program was configured to use the conditional independence test method with information gain as evaluation measure (applied for the independence tests during network induction) and indepence tresholds of 0.1 and 0.01 respectively. For the evaluation of the induced models against the test datasets the program “neval” of INeS package was substituted by a modified version, which accesses the existing Bayesian Network structure for efficient computation of predicted one-point coverages and implements the additional evaluation measures introduced in Section 6.5. In all of the above cases, the parameters were estimated from the observed frequencies in the training data applying a Laplace correction of 0.5 (compare page 28).

Due to the size of the distributions over the power set, the random sets where not explicitly represented in memory. Instead all required evaluation measures were computed directly from the size of the training and test sets, the number of tuples in the overlap of both sets and the respective absolute frequencies, of set-outcomes in this overlap with Laplace corrections applied according to the settings for the evaluation run. This procedure avoids the iteration over a large number of tuples (for which the pre-set computation time cap was exceeded already for base domains of as few as 20 elements). The fact that the distribution over the power set is more efficiently modeled by the training data set itself than by its parametric form underlines the lack of generalization inherent to approaches that rely on direct modeling over the power set of the base domain.

For the Random Set Model a Laplace correction of 0.5 dominates the results because the correction is applied to each element of the power set. To prevent this undesired effect a much lower Laplace correction of $2.5 \cdot 10^{-9}$ was used in the evaluation. With this choice the representation achieves considerably better evaluation results (Figure 6.2) and the fraction of probability mass distributed due to the Laplace correction is approximately the same as for the other models. Nevertheless the large number of parameters used for storing the distribution over the power set in comparison to the available sample size makes the estimates obtained from Random Set representations sensitive to the choice of the Laplace correction value. In the experiment results for a run using a much lower Laplace correction of $10 \cdot 10^{-12}$ were added to the evaluation. The comparison of results for those different parameters is used to illustrate the effects of the Laplace correction on the performance of the Random Set model.

6.5 Evaluation Measures

Having discussed the model classes, their respective training procedures and the general evaluation method, we shall now investigate evaluation measures. The measures where chosen to provide complementary information on how well different aspects of the set-distribution are captured by each model type.

Log-Likelihood To describe those measures we consider a process were distribution models are evaluated against a test database $D_{\text{tst}} = (d_1, d_2, \ldots, d_m)$. Each record d_i formed by the set of annotations applicable to one particular gene. A common way to evaluate the fit of a probability-based model M is to consider the likelihood of the observed test data D_{tst} under the model, that is, the conditional probability estimate $P(D_{\text{tst}} \mid M)$. The closer the agreement between test data and model, the higher that likelihood will be. An advantage of using the likelihood-based approach over distribution-oriented measures, such as the χ^2 measure, is the possibility to compute the probability assessments using the instantiations actually found in the training and test sets only rather than the complete power-set.

Of course the likelihood can also be used to test model generalization. Models that overfit the training data predict low likelihoods for independent test datasets drawn from the same background distribution as the training data. To circumvent technical limitations concerning the representation of multiplication with small numbers in the computer, the actual measure used is based on the logarithm of the likelihood:

$$\log L(D_{\text{tst}}) = \log \prod_{d \in D_{\text{tst}}} P(d \mid M) \tag{6.2}$$

$$= \sum_{d \in D_{\text{tst}}} \log P(d \mid M). \tag{6.3}$$

Of course the particular term used to estimate the probabilities $P(d \mid M)$ of the records in D depends on the model type and its parameters. The measure builds on the idea that the individual cases (genes) in both the training and the test set are considered as independent samples of a multi-valued random variable drawn from the same distribution. Having modeled that distribution by estimating parameters form the training data, the likelihood of a particular test database of size m is computed as the product of the likelihoods of its m records. However, due to the low likelihood of individual sample realizations even for good model approximation, the formula is almost always implemented using the mathematically equivalent formulation given in Equation 6.3, which yields intermediate results within the bounds of standard floating point format number representations.

One particular difficulty connected with the Log-Likelihood, relates to the treatment of previously unobserved cases in the test data set. If such values were simply assigned a likelihood of zero by the model then the whole database would have to be considered impossible and the Log-Likelihood becomes undefined. In the experiment this undesired effect was countered by applying a Laplace correction during the training phase. This modification ensures that conceivable events are assigned non-zero probability estimates even if they have not been covered in the training data and enables the measures to discriminate between databases containing such records.

Average Record Log-Likelihood: The main idea of the Log-Likelihood measure is to separately evaluate the likelihood of each record in the test database with respect to the model and consider the database construction process a sequence of a finite number of independent trials. This makes it difficult to compare measures obtained for test databases of different size. By correcting for the size of the test database one may obtain an average record (Log-)Likelihood as a more suitable measure for such tasks:

$$\text{arLL}(D_{\text{tst}}) = \frac{\log L(D_{\text{tst}})}{|D_{\text{tst}}|} \tag{6.4}$$

Note that in the untransformed domain the mean of the log-likelihoods corresponds to the geometric mean of the likelihoods, and is thus consistent with the construction of the measure from a product of evaluations for independently generated records. The arLL is used to measure the overall fit of the models to test data sets in the experiments and takes values from the range $(-\inf, 0]$. Values closer to 0 indicate better fits.

Singleton and Coverage Rate Errors: In addition to the overall fit between model and data, it is desirable to characterize how well particular properties of a set-distribution are represented. It has previously been pointed out that the condensed distribution emphasizes the approximation of both singleton probabilities and the values of the element coverage[2]. To assess the quality of the approximations from an application-oriented viewpoint and compare it to results achieved using by other methods, two additional measures – $\mathrm{d_{sglt}}$ and $\mathrm{d_{cov}}$ – have been employed. These measures are based on the sum of squared errors for the respective statistics over all elements of the base domain:

$$\begin{aligned} \mathrm{d_{sglt}} &= \sum_{\omega\in\Omega} \left(p'(\omega) - p(\omega)\right)^2 , &(6.5)\\ \mathrm{d_{cov}} &= \sum_{\omega\in\Omega} \left(\mathrm{opc}'(\omega) - \mathrm{opc}(\omega)\right)^2 . &(6.6)\end{aligned}$$

For both measures smaller values indicate better reconstructions.

6.6 Results

In order to increase the robustness of the results the evaluation was conducted using 5-fold cross-validation. In each of the five runs the models were trained using a Laplace correction of 0.5. For the assessment and comparison of the different methods, the evaluation results of the individual runs were collected and – with the exception of the logL measure[3] – averaged. These results are summarized in the Tables 6.1–6.5. For comparison, the same procedure was applied to two full Random Set representations using an adapted Laplace correction of $2.5 \cdot 10^{-9}$ and a reduced value of 10^{-12} respectively (Tables 6.6 and 6.7).

As demonstrated by the arLL measure (logarithmic scale!) both versions of the condensed distribution model provide a better overall fit to the test data than the one with independent modeling of term appearance using one-point coverage

[2] Compare Section 4.1 for the role of singleton probabilities and coverage in different interpretations

[3] See the discussion on the arLL measure on page 136 to review the argument why averaging Log-Likelihoods is not meaningful here

$\log L$	arLL	d_{sglt}	d_{cov}
-9039.60	-6.60	0.067856	0.001324
-8957.19	-6.54	0.064273	0.001524
-9132.09	-6.67	0.060619	0.001851
-8935.82	-6.52	0.074337	0.001906
-9193.44	-6.72	0.059949	0.001321
	-6.61	0.065406	0.001585

Table 6.1: Evaluation results for model using independent binary variables (one-point-coverage) with Laplace correction of 0.5 (INDEP)

$\log L$	arLL	d_{sglt}	d_{cov}
-7629.66	-5.57	0.000539	0.008293
-7559.38	-5.52	0.000457	0.011652
-7752.21	-5.66	0.000857	0.006998
-7529.83	-5.50	0.001014	0.004767
-7828.44	-5.72	0.000567	0.009961
	-5.59	0.000686	0.008334

Table 6.2: Evaluation results for condensed distribution on hierarchically structured domain with Laplace correction of 0.5 (HCDM)

$\log L$	arLL	d_{sglt}	d_{cov}
-7992.76	-5.83	0.000241	0.001342
-7885.19	-5.76	0.000222	0.001531
-8045.31	-5.87	0.000411	0.001838
-7839.16	-5.72	0.000612	0.001895
-8195.49	-5.99	0.000268	0.001316
	-5.83	0.000350	0.001584

Table 6.3: Evaluation results for condensed distribution on unstructured domain with Laplace correction of 0.5 (CDM)

$\log L$	arLL	d_{sglt}	d_{cov}
-7305.45	-5.33	0.005469	0.000928
-7385.09	-5.39	0.005544	0.000946
-7276.01	-5.31	0.006352	0.001218
-7514.56	-5.49	0.004684	0.000625
	-5.38	0.005512	0.000929

Table 6.4: Evaluation results for Bayesian Network Model (Laplace correction of 0.5, independence criterion ≤ 0.1) (BN1)

$\log L$	arLL	d_{sglt}	d_{cov}
-7349.38	-5.36	0.005877	0.001324
-7324.65	-5.35	0.005638	0.001524
-7455.47	-5.44	0.005360	0.001851
-7333.04	-5.35	0.008129	0.001906
-7562.74	-5.52	0.005980	0.001321
	-5.41	0.006196	0.001585

Table 6.5: Evaluation results for Bayesian Network Model (Laplace correction of 0.5, independence criterion ≤ 0.01) (BN2)

$\log L$	arLL	d_{sglt}	d_{cov}
-8259.66	-6.03	0.001823	0.098462
-8346.13	-6.09	0.001311	0.100860
-8651.12	-6.31	0.000964	0.095850
-8288.68	-6.05	0.003105	0.103200
-8534.30	-6.23	0.000671	0.094305
	-6.14	0.001574	0.098536

Table 6.6: Evaluation results for Random Set representation (Laplace correction of $2.5 \cdot 10^{-9}$) (RS1)

$\log L$	arLL	d_{sglt}	d_{cov}
-8974.15	-6.55	0.000226	0.001324
-9107.57	-6.65	0.000228	0.001521
-9490.80	-6.93	0.000433	0.001854
-9050.11	-6.61	0.000575	0.001901
-9319.35	-6.81	0.000289	0.001326
	-6.71	0.000350	0.001585

Table 6.7: Evaluation results for Random Set representation (Laplace correction of $1.0 \cdot 10^{-12}$) (RS2)

model	arLL	d_{sglt}	d_{cov}
Rand. Set (RS1)	-6.14	0.001574	0.098536
Rand. Set (RS2)	-6.71	0.000350	0.001585
indep. bin. var.	-6.61	0.065406	0.001585
CDM (flat term set)	-5.83	0.000350	0.001584
HCDM (enriched ontol.)	-5.59	0.000686	0.008334
Bayesian Network (BN1)	-5.38	0.005512	0.000929
Bayesian Network (BN2)	-5.41	0.006196	0.001585

Table 6.8: Summary of evaluation results over all model classes and criteria

alone. Among the two condensed distribution models the variant integrating the term hierarchy exhibits a consistently better fit. The advantage of the condensed distribution models over the independent modeling can be explained by their accurate representations of annotations with singletons or terms from a single path of refinements. This particular class of annotations is frequent in the dataset (compare Table 6.9). In fact, only the probabilistic Bayesian Networks achieve a better assessment with respect to this criterion. Indeed, Graphical Models are well-known for finding accurate, yet compact approximations of high dimensional distributions. Although the complexity of the conditional independence tests required to construct the models increases with the number of variables, use of heuristics allows to find good model structures at acceptable computational costs. Interestingly, the approximation results with the network based on more stringent independence test achieves a lower fit to the test data than the more tolerant model. Again, this effect can be attributed to overfitting as the inverse relation is found for the training data.

cardinality	1	2	3	4	5	6	7	8	9	10	>10
abs. frequency	3896	1186	947	427	228	74	54	21	9	3	4

Table 6.9: Cardinality of annotation sets in *Saccharomyces cerevisiae* data

In spite of using the largest number of parameters the Random Set representations are clearly behind other model types in the comparison of general approximation quality ($arTT$ measure). This is a result of overfitting as the even the considerable training sample appears small in comparison to the number of possible elements in the power set of the domain. Laplace correction (RS1) mitigates this effect to some degree but also introduces biases that interfere with the accurate reconstruction of coverage factors. The effect of different choices for the Laplace correction value on the accuracy of the Random Set representation is shown in Figure 6.2.

With regard to the accuracy of the predicted rates of single-element annotations the independent modeling of elements leads to a high error rate. In contrast, the condensed distribution models benefit from their separate representations of the probability associated with singleton annotation sets (the latter representing 56.9% of all database entries) to produce very low error measures. The accuracy of the hierarchical variant of the condensed distribution model is marginally lower, which is explained by its reliance on local branch distributions for determining singleton rates[4]. Even the Graphical Models, with their excellent overall fit to the distribution cannot compete with the specialized models for this task. For the Random Set representation it becomes apparent that the very Laplace

[4] see page 125 for a discussion of and a proposed solution to that problem

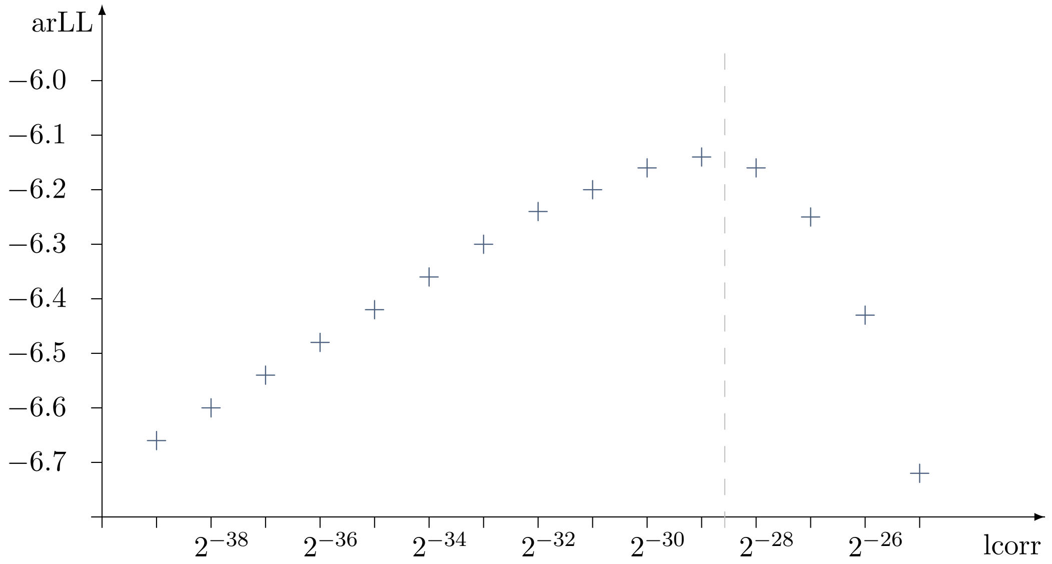

Figure 6.2: Average record Log-Likelihood for Random Set representation on test data as a function of Laplace Correction (log. scale); broken line indicates parameter value used for model RS1

correction that allowed to increase the overall fit of the distribution also increases the error in the estimates of singleton rates (Table 6.6). As already explained on page 134, this results from the sensitivity of the the Random Set approach to the bias introduced by the correction. That bias is avoided by choosing low values for the Laplace correction (Table 6.7) allowing to match the low error rate of the condensed representation model for the unstructured domain. However, this reduced Laplace correction is not effective in alleviating the overfitting problem (see comparison of RS1 and RS2 in Table 6.8).

The adverse effect of the Laplace correction in the Random Set representation is even more pronounced w.r.t. the error in one point coverages ($\mathrm{d}_{\mathrm{cov}}$ column in Table 6.6. Again, that bias can be neglected when the Laplace correction is set sufficiently low to balance the number of available observations with the cumulative contributions of applying the correction to the elements from the power set (Table 6.7, Laplace correction 10^{-12}). By construction, the independence model (Table 6.1) as well as the non-hierarchical condensed distribution model accurately reflect the one point coverages in the training data. They achieve essentially identical prediction errors (negligible differences between the tables occur due to numerical effects in the calculations to recover one-point coverage values from several parameters in the condensed representation). As before, the error in the hierarchical version is slightly larger due to the conditional independence assumptions in the construction of the local distribution. Whereas the Bayesian Network Model trained with stringent independence tests predicts the one-point coverages (corresponding to the marginal distributions in the encoding using binary variables) with error comparable to the error using the accurate representations, the conditional independence assumption in the less constrained model allows even better predictions. This indicates that good generalization properties the less stringent condition.

6.7 Comparison of Computation Time

Its potential for processing of large annotation data sets was one of the motivating factors for the development of the condensed distribution model. Obviously its suitability for this tasks critically depends on the efficiency and scalability of the approach. It is therefore appropriate to study the computational resources required for running analyses using the condensed distribution in comparison to other models for set distributions. In the experiment on the yeast dataset processing times for each of the model types were measured and recorded in a logfile. Figure 6.3 visualizes that data.

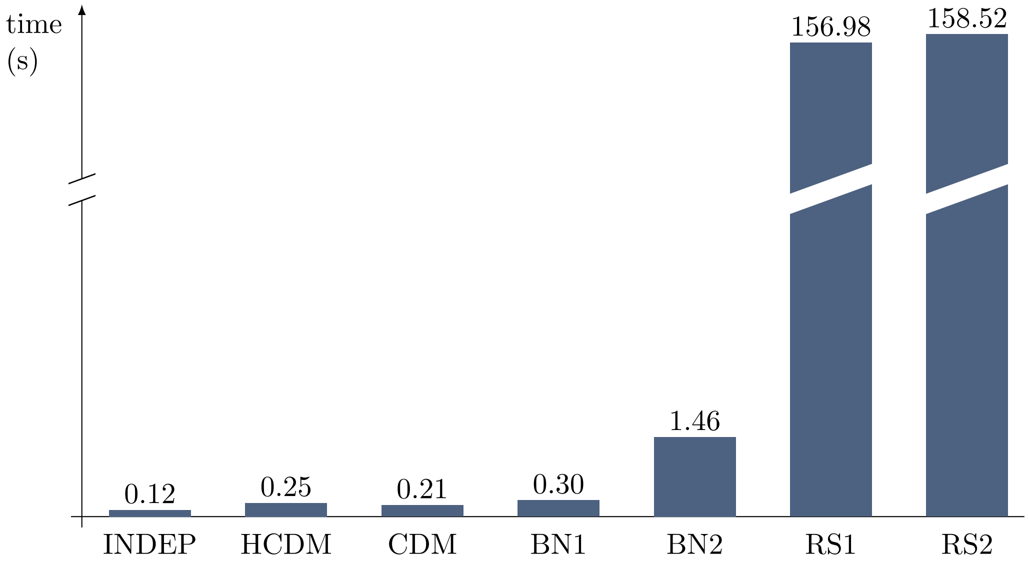

Figure 6.3: Comparison of processing time for each distribution model (sum of training and application phases during crossover procedure).

6.8 Conclusions

The results of the experimental evaluation of the tested models agree with the predicted properties of underlying frameworks that were elaborated in (Chapters 3–5). Although their memory and processing time requirements by far exceed those of all other tested methods, Random Set representations achieved low accuracy of predictions in comparison to the other approaches. This is largely attributed to overfitting. On the other hand the independent treatment of term occurrences in the annotation sets also led to low evaluation results, due to is its disregard for the logical and statistical relations between terms. In particular, the difference between the observed frequency of single-term annotations and the respective predicted frequency under the independence assumptions in the data indicates that such assumptions do not generally hold. Conversely both Graphical Models and the newly introduced condensed representations of set distributions were shown to provide compact yet accurate models for distributions over sets. The most important difference between these two classes of models is the origin of the structural component of the models. Whereas in the Graphical Models the structure is directly induced from the training data. The hierarchical version of the condensed representation constructs the model around a given hierarchy representing prior relational knowledge.

In the structure learning step of the Bayesian Network construction algorithm independence assumptions expressed via the graph component are tested on the training data before the final model structure is fixed. Whereas this strategy leads to a better overall fit to the data, the best results were achieved when the criteria used in the independence test were relaxed. The reason is that the distinction of genuine but weak statistical interactions against a background of sampling effects requires very large sets of training data. Allowing the learning process to accept additional independence assumption results in a lower performance on the training data, but improves generalization due to the reduced risk of overfitting the model structure.

For the condensed representation of distributions over sets the model induction process is restricted by the fixed model structure. This results in lower overall fit to the data (see arLL-measure) as compared to the Bayesian Network approach. But due to the separate representation of the singleton refinements in the (branch) distribution(s), the error rates for single-label descriptions are lower. The evaluation of the experiments indicates, that the enforced structural correspondence with the knowledge on term relations does not substantially limit the overall accuracy of predictions in comparison to the Graphical Models. The practical implications of these results are discussed in the final chapter of this dissertation.

7 Implications and Perspectives

The study of representations for statistical models of annotations and other set-valued data types, demonstrated that the existing methods can summarily be grouped into a small number of formal approaches. Details and assumptions that allow to reduce the model complexity for individual frameworks are commonly derived from particular interpretations, but prove difficult to transfer to new application contexts.

After the requirements and desired properties for such an approach had been elaborated, major classes of knowledge representations have been investigated and evaluated w.r.t. their suitability to these requirements. While many properties of the extant representations are documented in the literature, the detailed investigation also points out lesser known, often undesirable consequences of the subtle independence assumptions in different model types. Although none of the studied extant approaches fully met the desired model requirements, the detailed analysis of the existing approaches and their capabilities provided valuable lessons, which were integrated into the concept of a new knowledge representation for empirical distributions over set-valued data. This concept was later implemented and tested using a publicly available compilation of current biological knowledge about gene function in *S. cerevisiae.*

7.1 Scientific Results

The central result of the present dissertation was the development of a knowledge representation framework that is suited to data integration and modeling tasks with annotation data. These capabilities enable the model to address critical challenges in fields, such as systems biology.

In the order of their discussion in this dissertation the main results of this work are:

- A survey of interpretations for distributions over sets and the extant approaches to represent such distributions. Major model classes were identified, then analyzed with respect to their properties, inherent assumptions and the aspects of knowledge emphasized by them. For each class of models I discussed comparative strengths and limitations. This analysis allowed to outline the aggregation problem in possibility theory, which impedes the direct interpretation of marginal possibility distributions in the multivariate case (Chapter 4).

- The introduction of the condensed representation to efficiently model distributions over sets. The condensed representation applies ideas inspired by knowledge representations via possibility distributions, but embeds them in a probabilistic framework to maintain interpretable aggregation operations (Section 5.2).
- The extensions of said model to the multivariate case (Section 5.3) and for attributes with hierarchically structured domains (Section 5.4). I proposed data structures and operations for converting between information representations relative to descriptions via different subsets of attributes, that is, different frames of discernment. Because operations for changing the frame of discernment are based on the probabilistic part of the model, it circumvents the possibilistic aggregation problem and allows to obtain interpretable results for marginal or coarsened domains. This information consists of one-point coverages in the sense of the context model, and detailed probability assignments for all singleton outcomes in the considered frame.
- The experimental validation and comparative assessment of the proposed models using publicly available biological datasets (Chapter 6). In particular it was demonstrated that the separate representation of singleton outcomes is well-suited to the properties of annotation sets and allows for improved approximations of their distributions as compared to other representations with similar storage and processing time requirements.

Among the features of the condensed representation for random sets introduced in this work, some aspects are highlighted due their implications for applications in modeling and data analysis:

Empirical Foundation The condensed representation of random sets offers empirical interpretations of all its marginal distributions. This empirical interpretation allows models utilizing the representation to be induced from data on objective observations or measurements and permits to generate empirically testable predictions. Whereas the interpretation of singleton probabilities is identical to the that of the probabilistic framework, represented one-point coverages are consistent with the interpretation offered by the context model. This makes the context model interpretation applicable in a frame-spanning manner, as opposed to having to specify all inputs with respect to a fixed reference set of variables (see 4.4.5).

Integration with Probabilistic Models Due to the use of a coarsened probability space as a starting point of the representations, the effects resulting from independence assumptions made to reduce model complexity are limited to the reconstruction of set frequencies in local distributions. In contrast, marginal distributions are computed using a the probabilistic aggregation method. If all sets-outcomes are singletons for some selection of variables then the model is equivalent to a probabilistic knowledge representation. Together with the probabilistic aggregation operator this property can be used to extend the capabilities of existing probability-based models by interfacing with the multivariate or hierarchical versions of the models' proposed condensed representations, with data being exchanged via shared probabilistic marginal or coarsened distributions.

Data Representation and Ontology Enrichment By offering a method to switch between frames of discernment within an efficient information-compressed representation for distributions over sets, the proposed representation strikes a compromise between full random sets, and simpler representations e.g., based on term frequency. This makes the model suitable for tasks where the former approaches are prone to overfitting due to a high number of focal sets, whereas the accuracy requirements or and application context do not admit the additional independence assumptions made by simpler approaches. Moreover, in the one-dimensional version of the representation, the number of parameters is linear in the size of the base domain, allowing for a compact representation. In particular the condensed representations are well-suited to annotations and can be used to enrich associated ontologies. For example, the hierarchical model variant used in the evaluation (in Chapter 6) enriches the Gene Ontology with species-specific quantitative information about the relation between biological processes and the composition of the genome.

Significance for Research Set-based concepts and annotations are extensively used in emerging research fields, in particular in Computational Biology. Although this research is fueled by rapid improvements in experimental techniques and the wide availability of computational resources the a lack of suitable evaluated algorithms, models, and model components often limits to progress in application projects. In Computational Biology ontologies and complementary annotations in genome and proteome databases from publicly available resources have acquired a pivotal role for the interpretation experimental data and the formulation of new hypotheses. They are routinely used, for instance, to integrate mRNA level data for thousands gene loci into a higher level biological context, integrate such context information into graph-based models of gene-interaction (Maere et al., 2005) or select candidate hypotheses for further investigation in targeted experiments (Buescher et al., 2012); furthermore to study the emergence of new capabilities during evolution by analyzing enrichment and differences of annotations or annotation groups in phylogenetic trees. The central role of annotations and ontologies is illustrated by the number of large international collaborations that aim to create, maintain and extend such resources.

The main incentive for the development of such automated tools is the shift from a small number of experiments focused on a particular pathway to large-scale, high-throughput experimental techniques. Due to the amount the collected data, the analysis and interpretation of these results can no longer be conducted unassisted. Providing adequate statistical models and tools that are suitable to support operating with set-based data, they aid researchers to automate tasks in the integration of information from different sources as well as the planning and analysis of experiments.

Possible Use in Clustering and Classification Because the condensed representation of Random Sets provides a very compact summary representation for a group of annotated objects, it may itself serve as a representation of cluster or class prototypes in machine learning. In particular the evaluation measures applied in Chapter 6 already measure distances in a data and model space for clustering or classification tasks. However, in their current form the measures are not well adapted to such tasks, and modifications would have to be applied to achieve suitable value ranges and better comparability between distances measured. Given the continuing interest in multi-clustering by the machine learning community, the development of this approach and comparison to alternative strategies are an option for a future line of investigation.

Applications in Combination with Graphical Models While one of the inspirations for the development of the condensed representation came from the potential integration of set-based data into graphical models, at the time of this writing the author is not aware of any extensive publicly available collections of multivariate data sets using set-valued data. The comparison with other approaches was thus limited to the more readily available hierarchical data-sets. Nevertheless an application to more general Graphical models would build on the same principles and was considered in the method design.

7.2 Software Development

The condensed representation of random sets as well as additional data structures and operations to support hierarchical models for imported ontologies have been implemented as C libraries, which are available from the author. A software tool that allows the induction of and operations on condensed distributions for random sets has been implemented on the basis of these libraries. The software supports several modeling approaches and can be adapted for different interpretations of set distributions. Configuration options are accessed via a command line interface, that facilitates integration into scripted workflows.

package/tool	Description
hdist.c, hdist.h	C library implementing (hierarchical) distribution over sets and support functionality, permits programmatic access to all distribution parameters, conversion between frames, etc.
psvmodel	integrated model induction and evaluation tool for both flat and hierarchical versions of distribution model over sets; also implements model based on independent binary variables (accessed via command line parameters)
crsinduce	command line interface for inducing (hierarchical) models for set-valued data from case database
crsapply	command line interface for querying (hierarchical) model for set-value data and assessing the likelihood of sample databases under the model

Table 7.1: Software components implementing core functionality of hierarchical distribution model

An introduction to the software interface is given via a demonstration script (output abbreviated): The first commands show small sample files for a hierarchy and set instantiations compatible with that hierarchy. From this data a hierarchical model is induced and saved to file (`crsinduce`). The model is subsequently reloaded and applied to a sample database that lists all possible set-instantiations compatible with the chosen hierarchy (`crsapply`). This evaluation provides statistics for the fit of the sample database to the model and, optionally, assessments of the likelihood of each individual instantiation under the induced model. Converting these likelihood assessments back to sample probabilities demonstrates that the condensed representation indeed expands into a probability distribution over the sets.

```
> cat ex/demo.htr
digraph demo {
  A -> B;
  A [label="nodeA"] -> C;
  C -> D;
}

> cat ex/demo.tab
A
A
A
B
A
B
B C
B C
D
B D

> ./crsinduce -s -10.0 -mincl ex/demo.htr ex/demo.tab ex/demo.mdl
...reading training data from file ex/demo.tab

> cat ex/demoqry.tab
A
B
B C
B D
C
D

> ./crsapply -s ex/demo.mdl ex/demoqry.tab -

...assessing data from file ex/demoqry.tab
{ A }   -0.916291
{ B }   -1.609438
{ B C } -1.897120
{ B D } -1.897120
{ C }   -2.995732
{ D }   -2.995732

-------------------------------------------------------------
model  logL(D_test) logL(d_avg) sum err_frq^2 sum err_cov^2
-------------------------------------------------------------
hdist -12.311433 -2.051906 0.082778 0.090000
```

```
> ./crsapply -s ex/demo.mdl  ex/demoqry.tab - |head -n6 | \
   awk -vOFS="\t" -vFS="\t" '{ print $0, exp($2)}'

...assessing data from file ex/demoqry.tab
{ A }   -0.916291       0.4
{ B }   -1.609438       0.2
{ B C } -1.897120       0.15
{ B D } -1.897120       0.15
{ C }   -2.995732       0.05
{ D }   -2.995732       0.05
```

Alternatively `crsapply` can be configured to show the probability mass of the complete subtree rooted at a node. In that mode the probability of any compatible refinement will count towards the assessment. This mode can be used for conducting enrichment studies based on sample sets with inhomogeneous granularity. The combination of both modes can be used for assessing instantiations for a particular selected frame, depending on the expansion level of elements in the queried instantiation relative to the frame resolution hierarchy, and the chosen model settings regarding the assignment of probability mass to instantiations with non-leaf elements (allowed by default).

Passing the option "-?" to any of the programs will open a screen with command-line help.

Additional software implemented includes:

- An emulation of the Random Set model via database operations on training and test databases
- Modified components of the INES software package for network induction with Graphical Models, which were adapted to evaluation measures defined and applied in chapter 6
- Various support scripts used for format conversion, identifier mapping, data import and pre-processing of Gene Ontology and Gene Ontology Annotation files, automated evaluation and aggregation of results (many of the reusable ones are available on the authors website).
- Supporting tools for frame conversion in structured domains.

All software is available on request or via the authors website www.ruegheimer.org.

List of Symbols

$A \perp\!\!\!\perp B \mid Z$ Variables A and B are conditionally independent for every fixed instantiation of the attributes in the set Z

$A \not\perp\!\!\!\perp B \mid Z$ Variables A and B are not independent for every fixed instantiation of the attributes in the set Z

$X \perp\!\!\!\perp_{\delta} Y \mid Z$ Variable sets X and Y are conditionally independent for every fixed instantiation of the attributes in the set Z under the distribution δ

$\lambda_{B \to A}$ A λ-message sent from attribute B to attribute A, used in polytree propagation algorithm for Bayesian Network

$\pi_{A \to B}$ A π-message sent from attribute A to attribute B, used in polytree propagation algorithm for Bayesian Network

$(X \mid Z \mid Y)_G$ Variable sets X and Y are u-separated by the variable set Z in the undirected graph G

$(X \mid Z \mid Y)_{\vec{G}}$ Variable sets X and Y are d-separated by the variable set Z in the directed graph $\vec{G}$

$X \setminus Y$ The set difference of set X and set Y

$(\Lambda)_H$ The set of related frames generated by the label hierarchy H

$(p^{\diamond}, c^{\diamond})$ A condensed distribution composed of a condensed probability distribution on the coarsened power domain and an associated coverage function

$(p^{\diamond}_{H,\lambda_r}, c^{\diamond}_{H,\lambda_r})$ A condensed conditional distribution over power set of subframe formed by children of lable λ_r in the hierarchy H. Used to refine distributions over multi-label description.

2^{Λ} The power set (set of subsets) of the label set Λ

$\bigwedge$ Extension of the conjunction operator to a set of conditions. The aggregate condition is true if and only if all conditions of the indexed family hold.

$|X|$ The cardinality of the set X

$\overline{X}$ The complement of the set X with respect to its set universe

A Mostly: attribute (single-valued)
Also: proposition (Example 2.1)

a_1, a_2, b_1, b_2 etc. Labels – usually associated with domain or base domain of attribute designated by corresponding capital letter

$A^{\diamond}$ A condensed set-valued attribute

A_i Attributes

A_{Λ} An attribute that takes values from the frame Λ

A^{*}_{Λ} Set-valued attribute taking values from the power set of the frame Λ

$A_{\mathrm{obs},\Lambda}$ Attribute that represents an observed value w.r.t. the frame Λ in a hierarchy.

A^{*}, B^{*} etc. Set-valued attributes

$\mathrm{anc}_H(\lambda)$ The set of ancestors of a label λ in a hierarchy H

arLL average record Log-Likelihood (see page 136)

B Mostly: attribute (single-valued)
Also: proposition (Example 2.1)

$\mathrm{bdom}(A^{\diamond})$ The basic domain of the condensed set-valued attribute (set of labels from which set-outcomes can be formed)

Bel Belief measure

$B^{\diamond}$ A condensed set-valued attribute

$\mathbf{C}$	Set of Cliques in graph component of a Markov Network
C	Attribute
C	Mostly: a sample space – used in random set definition (Definition 4.1), interpreted as set of contexts in the context model Also: attribute (Chapter 3)
$c^{\diamond}$	A coverage function
$c^{\diamond}$	The coverage function of a condensed representation for set-valued distributions
$c^{\diamond}(\omega)$	The relative coverage factor associated with an element of the basic domain of a condensed set-valued attribute
$c^{\diamond}_{H,\lambda_r}$	A conditional coverage function for children of λ_r in hierarchy H (used for mapping distributions over multi-label instantiations to finer frames)
$c^{\diamond}_{X,[Y],t_{Z^{\diamond}}}$	A coverage function for the multivariate version of the condensed representation of set-valued attributes, parameters identify a set of reference attribute $X^{\diamond}$, the subset $Y^{\diamond}$ of $X^{\diamond}$ with attributes that have non-singleton instantiations and the (precise) instantiation for all attributes $Z^{\diamond} = X^{\diamond} \setminus Y^{\diamond}$ (subscripts indicate respective base attributes)
C_i	Clique in the graph component of a Markov Network
$\mathrm{catt}(X)$	A composite attribute constructed from attributes in set X – instantiations are defined as combinations of instantiations of component attributes
$\mathrm{children}_H(\lambda)$	The set of children (direct descendants) of a label λ in the hierarchy H
$\mathrm{conv}_{\Lambda_1 \to \Lambda_2}$	Mapping to convert probability distribution over frame of discernment Λ_1 to probability distribution over frame Λ_2 (Λ_1 and Λ_2 must be generated by the same hierarchy)
δ	A distribution, e.g., probability distribution

$\mathrm{dom}(A^\diamond)$	Domain of a condensed set-valued attribute (probabilistic component)
$\mathrm{d}_{\mathrm{cov}}$	Singleton coverage rate error sum (see page 137)
$\mathrm{d}_{\mathrm{sglt}}$	Singleton probability error sum (see page 137)
$\mathrm{desc}_H(\lambda)$	The set of descendants of a label λ in a hierarchy H
$\mathrm{dom}(A)$	The range of values (domain) of attribute A
E	A proposition, expressed as subset of Ω
E	The edge set for the graph component of a Graphical Model (with undirected graphs)
$\vec{E}$	The edge set for the graph component of a Graphical Model (with directed graphs)
$E_{U^\diamond}$	Set of tuples over attribute set specifying an event in terms of the condensed set-valued attribute set $E_{U^\diamond}$ (tuple-based notation)
F	General use: a proposition, expressed as subset of Ω Specifically: a focal set
$\mathcal{F}$	The set of focal sets for a random set-based representation
$\mathcal{F}_i$	A focal set
G	An undirected graph, used to represent set of independence relations for a Graphical Model
G'	The underlying undirected graph w.r.t. the directed graph $\vec{G}$ of a Bayesian Network, obtained by replacing directed edges in $\vec{G}$ by undirected ones
$\vec{G}$	A directed graph, used to represent set of independence relations for a Graphical Model
Γ	General use: set-valued mapping Specifically: a random set
Γ'	A random set

H	A hierarchy
L	A set of labels
$\mathcal{L}$	The set of labels in a hierarchy
Λ	General use: A set of labels Specifically: A frame of discernment
Λ_0	Coarsest frame of a label hierarchy, contains root label only ($\Lambda_0 = \{\lambda_0\}$)
λ_0	The root label of a hierarchy
Λ'	General use: A set of labels Specifically: A frame of discernment
λ_r	Used to indicate non-leaf element of a hierachically structured label set $\mathcal{L}$ that is replaced by the set of its children under an elementary refinement operation
$\lambda_r^\diamond$	Surrogate label to represent non-singleton outcomes in condesed version of distribution over subsets of the direct refinement of λ_r
lcorr	Value of the Laplace correction
$\mathrm{m}(H)$	Mass assignment to $H \subseteq \Omega$, also called basic probability assignment
$\mathrm{mv}_X(t_X, Y)$	A tuple over the set-valued attributes $X^\diamond$ – values for the attribute subsets $X^\diamond \setminus Y^\diamond$ corresponds to precise values specified in t_X, other variables are set to symbol for non-singleton outcomes
N (function)	Necessity measure
N (variable)	The total number of trials in a series of experiments
N_ω	The number of trials, in which outcome ω was observed in a series of experiments
$\mathbf{O}$	A set of observed variables (used in variable elimination algorithm for reasoning with Graphical Models)

O A set of objects or entities in the modeled world

o An object or entity in the modeled world

Ω A frame of discernment (general)

ω An element of a frame of discernment

ω_0 In modeling approaches based on the state-of-the-world view: the element of a frame of discernment representing the single true state of a modeled world (uncertainty about the identity of this state may be reduced due to the integration of new facts)

$\omega^\diamond$ Surrogate label used to summarize set-outcomes in probabilistic part of the condensed representation for random sets.

ω_F An element of focal set F, not contained in any focal subset of F

Ω_X A frame of discernment providing distinctions w.r.t. the values of the attributes in the attribute set X

$\mathrm{opc}(\omega)$ The one-point coverage for element ω by a random set

$\mathrm{opc}'_H(\lambda)$ Estimate of one-point coverage (probability of applicability) of label λ in label hierarchy H – reconstructed from model using condensed distribution to represent branch probabilities in H.

P A probability measure

p A probability distribution (probability density function)

$P(A = a \mid B = b)$ Conditional probability for instantiation a of attribute A, given that attribute B is instantiated with b.

$p^\diamond$ A condensed probability distribution

$p^\diamond_{H,\lambda_r}$ A condensed conditional probability distribution over the power set of sub-frame formed by the children of λ_r in the hierarchy H

p_H	Function providing probability assignments for each label of a hierarchy H, used to construct probability distributions for individual frames
$\hat{p}(\omega)$	A point estimate for the probability of observing outcome ω in a random experiment
$\hat{p}_{\text{INDEP}}$	Estimated probability distribution for over set-instantiations according to a model with independnent element membership provided by model)
P_Λ	A probability measure over the frame of discernment Λ
P^*	Probability measure over a domain of a set-valued attribute
p^*	A probability distributions over a power set
$P^*(E)$	Upper probability of event E induced by a set-valued mapping from a probability space (Chapter 4 only)
$P_*(E)$	Lower probability of event E induced by a set-valued mapping from a probability space (Chapter 4 only)
P^*_H	Probability measure for multi-label instantiantions w.r.t. a hierarchy H (see underlying distribution p^*_H)
p^*_H	Probability distribution over the space of multi-label instantiations w.r.t. a hierarchy H.
$p^*_{H,\lambda_r}(S)$	Conditional probability distribution over power set of the subframe defined by the children of label λ_r in hierarchy H – used for mapping set-distributions in hierarchies to finer frames.
$\text{parent}_H\lambda$	The direct parent of a label λ in the hierarchy H
ϕ_{C_i}	The factor potential associated with clique C_i in a Markov Model
Π	The set extension of a possibility distribution function (possibility measure)
π	A possibility distribution

Pl	Plausibility measure
$\mathrm{pred}(A)$	Set of predecessors of node A in a directed graph
$\mathrm{proj}_Y^X(R_X)$	The projection of the set or relation R_X to the attribute set Y
$\mathbf{Q}$	A set of query variables (used in variable elimination algorithm for reasoning with Graphical Models)
R	A relation linking activities to locations (used in Chapter 2)
$\mathrm{R}_{\mathrm{anc}}$	Ancestor relation in a hierarchy, transitive closure of $\mathrm{R}_{\mathrm{parent}}$
$\mathrm{R}_{\mathrm{parent}}$	A relation: $(\lambda_2, \lambda_1) \in \mathrm{R}_{\mathrm{parent}}$ indicates that λ_1 is a direct parent (superior) of λ_2 a hierarchy
ref	General use: A refinement function Specifically: The elementary refinement operation on a hierarchy
S	General use: a set or relation Specifically: a set of holiday activities (Chapter 2) a set-valued outcome of a random experiment (Chapter 5)
σ	Set reduction mapping (used for condensed representation of random sets)
σ_X	Set reduction mapping (extension to multivariate case using attributes from X as base attributes)
S_H	A multi-label instantiation with respect to a hierarchy H
τ	A parameter conversion function – used for iteration over coverage groups when computing marginal and conditional coverage functions in the multivariate case
T_X	The set of all tuples (instantiations) over the attribute set X

t_X	A tuple or instantiation represented as a mapping from the attributes in an attribute set X to their respective attribute values
$T_{X^\diamond}$	The set of tuples over a set $X^\diamond$ of condensed set-valued attributes
$t_{X^\diamond}$	A tuple over a set $X^\diamond$ of condensed set-valued attributes
$t_{X\|Y}$	A tuple over the attribute set Y, obtained as a restriction of a tuple t_X over the attribute set X, with $X \supseteq Y$ (all attributes of Y are mapped to the same values as under t_X)
$U^\diamond$	General use: set of condensed set-valued attributes Specifically: used in specification of conditioning events or distributions
V	The set of variables referred to by a Graphical Model, constitutes the set of vertices in the model's graph component
W_0	A subset of a frame of discernment that comprises instances realized in modeled situation, applied in set-of-instances view
$W^\diamond, X^\diamond, Y^\diamond, Z^\diamond$	Sets of condensed set-valued attributes
X, Y, Z	Attribute sets

Bibliography

Carlos E. Alchourrón and David Makinson. On the logic of theory change: Safe contractions. *Studia Logica*, 44:405–422, 1985.

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, June 1985.

Stig K. Andersen, Kristian G. Olesen, Finn V. Jensen, and Frank Jensen. HUGIN—a shell for building Bayesian belief universes for expert systems. In *Proc. 11th International Joint Conference on Artificial Intelligence (IJCAI'89)*, pages 1080–1085, 1989.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

James F. Baldwin, Trevor P. Martin, and Bruce W. Pilsworth. *FRIL – Fuzzy and Evidential Reasoning in Artificial Intelligence.* Research Studies Press / Wiley & Sons, Taunton / Chichester, UK, 1995.

Salem Benferhat, Didier Dubois, and Henri Prade. Expressing independence in a possibilisitic framework and its application to default reasoning. In *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI 94)*, 1994.

Salem Benferhat, Didier Dubois, and Henri Prade. Towards a possibilistic logic handling of preferences. *Applied Intelligence*, 14(3):303–317, 2001.

Christian Borgelt and Rudolf Kruse. Possibilistic networks with local structure. In *Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98, Aachen, Germany)*, volume 1, pages 634–638, Aachen, Germany, 1998. Verlag Mainz.

Christian Borgelt and Rudolf Kruse. *Graphical Models—Methods for Data Analysis and Mining.* J. Wiley & Sons, Chichester, 2002.

Christian Borgelt and Rudolf Kruse. Operations and evaluation measures for learning possibilistic graphical models. *Artificial Intelligence*, 148:385–418, 2003.

Christian Borgelt, Jörg Gebhardt, and Rudolf Kruse. Possibilistic graphical models. In Giacomo Della Riccia, Rudolf Kruse, and Hans-Joachim Lenz, editors, *Computational Intelligence in Data Mining (Proc. 3rd Int. Workshop, Udine, Italy 1998)*, number 408 in CISM Courses and Lectures. Springer, Wien, Austria, 2000.

Patric Bosc, Nadia Liétard, and Oliver Pivert. About the processing of possibilistic queries involving a difference operation. *Fuzzy Sets and Systems*, 157: 1622–1640, 2006.

Ronald J. Brachman and Hector J. Levesque. *Knowledge Representation and Reasoning*. Morgan Kaufmann, 2004.

Joerg Martin Buescher, Wolfram Liebermeister, Matthieu Jules, Markus Uhr, Jan Muntel, Eric Botella, Bernd Hessling, Roelco Jacobus Kleijn, Ludovic Le Chat, François Lecointe, Ulrike Mäder, Pierre Nicolas, Sjouke Piersma, Frank Rügheimer, Dörte Becher, Philippe Bessieres, Elena Bidnenko, Emma L. Denham, Entienne Dervyn, Kevin M. Devine, Geoff Doherty, Samuel Drulhe, Liza Felicori, Mark J. Fogg, Anne Goelzer, Annette Hansen, Colin R. Harwood, Michael Hecker, Sebastian Hubner, Claus Hultschig, Hanne Jarmer, Edda Klipp, Aurélie Leduc, Peter Lewis, Frank Molina, Philippe Noirot, Sabine Peres, Nathalie Pigeonneau, Susanne Pohl, Simon Rasmussen, Bernd Rinn, Marc Schaffer, Julian Schnidder, Benno Schwikowski, Jan Maarten van Dijl, Patrick Veiga, Sean Walsh, Anthony J. Wilkinson, Jörg Stelling, Stéphane Aymerich, and Uwe Sauer. Global network reorganization during dynamic adaptation of *Bacillus subtilis* metabolism. *Science*, 335(6072):1099–1103, March 2012. doi: 10.1126/science.1206871.

Wray Buntine. Chain graphs for learning. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 46–54, San Francisco, CA, USA, 1995. Morgan Kaufmann.

Enrique Castillo, José M. Guitérrez, and Ali S. Hadi. *Expert Systems and Probabilistic Network Models*. Springer-Verlag, New York, USA, 1997.

David M. Chickering, Dan Geiger, and David Heckerman. Learning Bayesian networks from data. *Machine Learning*, 20(3):197–243, 1995.

David M. Chickering, David Heckerman, and Christopher Meek. A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of* *the 13th Conference on Uncertainty in Artificial Intelligence (UAI'97)*, pages 80–89, 1997.

C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Giulianella Coletti and Romano Scozzafava. Conditioning in a coherent setting: Theory and applications. *Fuzzy Sets and Systems*, 155(1):26–49, 2005.

Gregory F. Cooper and Edward Herskovits. A Bayesian method for induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

Fabio Gagliardi Cozman and Peter Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1–2):173–195, 2005.

Luis M. de Campos and Juan F. Huete. Learning non probabilistic belief networks. In Michael Clarke, Rudolf Kruse, and Serafín Moral, editors, *Symbolic and Quantitative Approaches to Reasoning and Uncertainty (LNCS 747)*, pages 57–64. Springer, 1993.

Gert de Cooman. Possibility theory I: the measure- and integral-theoretic groundwork. *International Journal of General Systems*, 25:291–323, 1997a.

Gert de Cooman. Possibility theory II: conditional possibility. *International Journal of General Systems*, 25:325–351, 1997b.

Gert de Cooman. Possibility theory III: possibilistic independence. *International Journal of General Systems*, 25:353–371, 1997c.

Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*, volume 1. John Wiley, New York, 1974.

Bruno de Finetti. *Theory of Probability: A Critical Introductory Treatment*, volume 2. John Wiley, New York, 1975.

Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

Ernesto William De Luca and Frank Rügheimer. Discovering linguistic depencencies with graphical models. In *LWA 2007 Workshop Proceedings*, pages 119–125, Germany, September 2007. Martin-Luther-University Halle-Wittenberg.

Rina Dechter. Bucket elimination: A unifying framework for probabilistic inference. *Uncertainty in Artificial Intelligence*, 196:211–219., 1996.

Miguel Delgado and Serafín Moral. On the concept of possibility-probability consistency. *Fuzzy Sets and Systems*, 21:311–318, 1987.

Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.*, 38:325–339, 1967.

Arthur P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30:205–247, 1968.

Arthur P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, March 1972.

Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979.

Didier Dubois. Possibility theory and statistical reasoning. *Computational Statistics & Data Analysis*, 51:47–69, 2006.

Didier Dubois and Henri Prade. On several representations of an uncertain body of evidence. In M.M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-Holland, Amsterdam, Netherlands, 1982.

Didier Dubois and Henri Prade. *Possibility Theory.* Plenum Press, New York, New York, 1988a. Translation of: Théorie des possibilités.

Didier Dubois and Henri Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4 (3):244–264, 1988b.

Didier Dubois and Henri Prade. Inference in possibilistic hypergraphs. In *Proceedings of the 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 1990)*, pages 250–259, 1990.

Didier Dubois and Henri Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

Didier Dubois and Henri Prade. A synthetic view of belief revision with uncertain inputs in the framework of possibility theory. *International Journal of Approximate Reasoning*, 17:295–324, 1997.

Didier Dubois and Henri Prade. Possibility theory is not fully compositional! a comment on a short note by H.J. Greenberg. *Fuzzy Sets and Systems*, 95: 131–134, 1998a.

Didier Dubois and Henri Prade, editors. *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Belief Change*, volume 3. Kluwer, Dordrecht, 1998b. ISBN 0-7923-5162-2.

Didier Dubois and Henri Prade. Properties of measures of information in evidence and possibility theory. *Fuzzy Sets and Systems*, 100:35–49, 1999. reprinted from Fuzzy Sets and Systems 24 (1987) 161–182.

Didier Dubois and Henri Prade. Possibilistic logic: a retrospective and prospective view. *Fuzzy Sets and Systems*, 144:3–23, 2004.

Didier Dubois, Luis Fariñas del Cerro, Andreas Herzig, and Henri Prade. An ordinal view of independence with application to plausible reasoning. In Ramon López de Mántaras and David Poole, editors, *Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence*, volume 1, pages 195–203, Seattle, WA, July 1994.

Didier Dubois, Serafín Moral, and Henri Prade. A semantics for possibility theory based on likelihoods. *Journal of Mathematical Analysis and Applications*, 205: 359–380, 1997.

Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* AAAI Press and MIT Press, Menlo Park and Cambridge, USA, 1996.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, 3rd edition, 1968.

Richard P. Feynman, Robert B. Leighton, and Matthew Sands. *The Feynman Lectures on Physics: Quantum Mechanics*, volume 3. Addison-Wesley, Reading, MA, USA, 1965.

Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

Peter Gärdenfors. *Knowledge in flux: modeling the dynamics of epistemic states.* MIT Press, Cambridge, Mass., 1988.

Jörg Gebhardt. Learning from data: Possibilisitic graphical models. Habilitation Thesis, University of Braunschweig, Germany, 1997.

Jörg Gebhardt and Rudolf Kruse. The context model – an integrating view of vagueness and uncertainty. *International Journal of Approximate Reasoning*, 9:283–314, 1993.

Jörg Gebhardt and Rudolf Kruse. *Handbook of Defeasible Reasoning and Uncertainty Management Systems: Belief Change*, volume 3, chapter Parallel Combination of Information Sources, pages 393–439. Kluwer, 1998.

Jörg Gebhardt, Frank Rügheimer, Heinz Detmer, and Rudolf Kruse. Adaptable markov models in industrial planning. In *Proceedings of the 2004 IEEE International Conference on Fuzzy Systems (Budapest)*, Piscataway, NJ, USA, 2004. IEEE Press. URL http://ruegheimer.org/publications/fieee_04.pdf.

Romain Gérard, Souhila Kaci, and Henri Prade. Ranking alternatives on the basis of generic constraints and examples – a possibilisitic approach. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.

Michel Grabisch. Fuzzy integral in multicriteria decision making. *Fuzzy Sets and Systems*, 69(3):279–298, 1995.

Adam Groove. Two modellings of theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

Xavier Guyon. *Random Fields on a Network: Modelling, Statistics, and Applications*. Probability and its Applications. Springer-Verlag, New York, Berlin, Heidelberg, 1995.

John M. Hammersley and Peter E. Clifford. Markov fields on finite graphs and lattices. Cited in Isham (1981), 1971.

David Heckerman. Causal independence for knowledge acquisition and inference. In *Proceedings of the 9th Conference on Artificial Intelligence*, pages 122–127, 1993.

Masahiko Higashi and George J. Klir. Measures of uncertainty and information based possibility distributions. *International Journal of General Systems*, 9: 43–58, 1983.

Ellen Hisdal. Conditional possibilities, independence and noninteraction. *Fuzzy Sets and Systems*, 1:283–297, 1978.

Eyke Hüllermeier. Possibilistic instance-based learning. *Artificial Intelligence*, 148:335–383, 2003.

Valerie Isham. An introduction to spatial point processes and markov random fields. *Int. Statistical Review*, 49:21–43, 1981.

Finn V. Jensen. *An Introduction to Bayesian Networks*. University College London Press, London, UK, 1996.

Radim Jiroušek and Jiřina Vejnarová. Construction of multidimensional models by operators of composition: Current state of art. *Soft Computing*, 7(5):328–335, 2003.

Ross Kindermann and J. Laurie Snell. *Markov Random Fields and Their Applications*. Contemporary Mathematics. American Mathematical Society, Providence, Rhode Island, USA, 1980.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the 14th Int. Joint Conference on Artificial Intellligence (IJCAI 95)*, pages 1137–1145, 1995.

Jürg Kohlas and Paul-André Monney. *A Mathematical Theory of Hints: An Approach to the Dempster-Shafer Theory of Evidence*. Springer, 1995.

Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, Heidelberg, 1933. English edition: *Foundations of the theory of Probability*. Chelsea, New York, NY, USA, 1987.

Rudolf Kruse, Detlef Nauck, and Frank Klawonn. Reasoning with mass distributions. In B. D. D'Ambrosio, P. Smets, and P. P. Bonissone, editors, *Uncertainty in Artificial Intelligence*, San Mateo, California, 1991a. Morgan Kaufmann.

Rudolf Kruse, Erhard Schwenke, and Jochen Heinsohn. *Uncertainty and Vagueness in Knowledge Based Systems*. Springer, Berlin, Heidelberg, New York, 1991b.

Rudolf Kruse, Jörg Gebhardt, and Frank Klawonn. *Foundations of Fuzzy Systems*. Wiley & Sons, Chichester, UK, 1994.

Rudolf Kruse, Jörg Gebhardt, Frank Rügheimer, and Heinz Detmer. Planning with graphical models. In *Proc. of the 2006 Conference on COGnitive systems with Interactive Sensors (COGIS'06), Paris, France*, 2006.

Mounia Lalmas. Dempster-Shafer's theory of evidence applied to structured documents: modelling uncertainty. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 110–118, New York, NY, USA, 1997. ACM.

Steffen L. Lauritzen. *Graphical Models*. Oxford, New York, USA, 1996.

Steffen L. Lauritzen and Nuala A. Sheehan. Graphical models for genetic analyses. *Statistical Science*, 18(4):489–514, 2003.

Steffen L. Lauritzen and David J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 2(50):157–224, 1988.

Urbain Jean Joseph Le Verrier. *Recherches sur l'orbite de Mercure et sur ses perturbations ; Détermination de la masse de Venus et du diamétre du soleil.* Bachelier, Paris, France, 1843.

Eugene Stanley Lee and Qing Zhu. *Fuzzy and evidence reasoning.* Physica, Heidelberg, 1995.

Eric Lefevre, Olivier Colot, and Patrick Vannoorenberghe. Belief function combination and conflict management. *Information Fusion*, 3:149–163, 2002.

Isaak Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability and Chance.* MIT Press, Cambridge, Mass. and London, 1980.

Clarence Irving Lewis and Cooper Harold Langford. *Symbolic Logic.* The Century co, New York, 1932.

David K. Lewis. *Convention: A Philosophical Study.* Harvard University Press, 1969.

Sten Lindström and Wlodzimierz Rabinowicz. Epistemic entrenchment with incomparabilities and relational belief revision. In A. Fuhrmann and M. Morreau, editors, *The Logic of Theory Change, Proc. of the Konstanz 1989 Workshop*, pages 93–126. Springer-Verlag, 1990.

Jennifer Listgarten, Nicole Frahm, Carl Kadie, Christian Brander, and David Heckerman. A statistical framework for modeling hla-dependent t cell response data. *PLoS Computational Biology*, 3(10):e188, 2007. doi: 10.1371/journal.pcbi.0030188. published online.

Steven Maere, Karel Heymans, and Martin Kuiper. *BiNGO*: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005. doi: 10.1093/bioinformatics/bti551.

Marie-Hélène Masson and Thierry Denœux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3):319–340, 2006.

Albert Abraham Michelson. The relative motion of the earth and the luminiferous aether. *American Journal of Science*, 3(22):120–129, 1881.

Albert Abraham Michelson and Edward W. Morley. On the relative motion of the earth and the luminiferous aether. *American Journal of Science*, 3(34): 333–345, 1887.

Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

Charles J. Mosier. Problems and designs of cross-validation. *Educational and Psychological Measurement*, 11(1):5–11, 1951.

Eduardo F. Nakamura, Antonio A. F. Loureiro, and Alejandro C. Frery. Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Comput. Surv.*, 39(3), 2007.

Hung T. Nguyen. On random sets and belief functions. *Journal Math. Anal. Appl.*, 65:531–542, 1978a.

Hung T. Nguyen. On conditional possibility distributions. *Fuzzy Sets and Systems*, 1(4):299–309, 1978b.

Hung T. Nguyen. Fuzzy and random sets. *Fuzzy Sets and Systems*, 156:349–356, 2005.

Hung T. Nguyen and Bernadette Bouchon-Meunier. Random sets and large deviations principle as a foundation for possibility measure. *Soft Computing*, 8:61–70, 2003.

Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, USA, 1988.

Judea Pearl. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 4:363–389, 1990.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000.

Judea Pearl and Azaria Paz. Graphoids – a graph based logic for reasoning about relevance relations. Technical Report 850038 (R-53-L), Cognitive Systems Laboratory, University of California, Los Angeles, California, CA, 1985.

Jose M. Peña. Learning gaussian graphical models of gene networks with false discovery rate control. In *Proceedings of the 6th European Conference on Machine Learning and Data Mining in Bioinformatics (EvoBIO 2008)*, pages 165–176, 2008.

Plato. Gorgias (dialog). Published online, March 1999a. URL http://www.gutenberg.org/etext/1672. transl. by Benjamin Jowett.

Plato. Meno (dialog). Published online, February 1999b. URL http://www.gutenberg.org/etext/1643. transl. by Benjamin Jowett.

Karl R. Popper and John C. Eccles. *The Self and Its Brain: An Argument for Interactionism*. Springer-Verlag, New York London Heidelberg Berlin, corrected 2nd printing edition, 1985.

Ulrich Reimer. *Einführung in die Wissensrepräsentation*. Teubner, Stuttgart, Germany, 1991.

Allyson V. Ritchie, Saskia van Es, Celine Fouquet, and Pauline Schaap. From drought sensing to developmental control: evolution of cyclic amp signalling in social amoebas. *Molecular Biology and Evolution*, 25(10):2109–2118, 2008.

Frank Rügheimer. A condensed representation for distributions over set-valued attributes. In *Proc. 17. Workshop Computational Intelligence*, Karlsruhe, Germany, 2007. Universitätsverlag Karlsruhe.

Frank Rügheimer. *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, volume 81 of *Communications in Computer and Information Science*, chapter Using Enriched Ontology Structure for Improving Statistical Models of Gene Annotation Sets, pages 55–64. Springer, Heidelberg, 2010. ISBN 978-3-642-14057-0. doi: 10.1007/978-3-642-14058-7.

Eran Segal, Dana Pe'er, Aviv Regev, Daphne Koller, and Nir Friedman. Learning module networks. *Journal of Machine Learning Research*, 6:557–588, 2005.

SGD Curators. SGD yeast gene annotation dataset (slim ontology version). via Saccharomyces Genome Database Project (SGD Curators, b), a. URL ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/go_slim_mapping.tab. (accessed 2008/11/16).

SGD Curators. Saccharomyces genome database, b. URL http://www.yeastgenome.org. (accessed 2008/11/16).

Ross D. Shachter, Tod S. Levitt, Laveen N. Kanal, and John F. Lemmer, editors. *Uncertainty in Artificial Intelligence 4*. North Holland, Amsterdam, Netherlands, 1990.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

Glenn Shafer. Constructive probability. *Synthese*, 48:1–60, 1981. Reprinted in Classic Works of the Dempster-Shafer Theory of Belief Functions, edited by Ronald R. Yager and Liping Liu, Springer, 2007.

Glenn Shafer. *The analysis of fuzzy Information*, volume 1, chapter Belief functions and possibility measures. CRC Press, Boca Raton, FL, 1986.

Prakash P. Shenoy and Glenn Shafer. Propagating belief functions using local computation. *IEEE Expert*, 1(3):43–52, 1986.

Philippe Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

Philippe Smets. Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

Paul Snow. The vulnerability of the transferable belief model to dutch books. *Artificial Intelligence*, 105:345–354, 1998.

John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, Pacific Grove, CA, USA, 2000.

Wolfgang Spohn. Ordinal conditional functions. a dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, pages 105–134. Kluwer, Dordrecht, 1988.

Wolfgang Spohn. A general non-probabilistic theory of inductive reasoning. In *Shachter et al. (1990)*. North Holland, 1990.

Matthias Steinbrecher, Frank Rügheimer, and Rudolf Kruse. *Computational Intelligence in Automotive Applications*, volume 132/2008 of *Studies in Computational Intelligence*, chapter Application of Graphical Models in the Automotive Industry, pages 79–88. Springer, Berlin / Heidelberg, 2008. ISBN 978-3-540-79256-7. doi: 10.1007/978-3-540-79257-4_5. URL http://www.springerlink.com/content/v2g556403238145v/.

Milan Studený. Formal properties of conditional independence in different calculi of AI. In Michael Clarke, Rudolf Kruse, and Serafín Moral, editors, *Proceeding of the 2nd European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU1993)*, Berlin, Heidelberg, Germany, 1993. Springer.

Milan Studený. On separation criterion and recovery algorithm for chain graphs. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI-96)*, 1996.

Jiřina Vejnarová. Design of an iterative proportional fitting procedure for possibility distributions. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications (ISIPTA '03)*, pages 577–592. Carleton Scientific, 2003.

Thomas Verma and Judea Pearl. An algorithm for deciding whether a set of observed independencies has a causual explanation. In *Proc. of the 8th UAI Conference*, pages 323–330, 1992.

Richard von Mises. *Probability, Statistics and Truth.* Allen and Unwin, Woking and London, UK, 2nd edition, 1957. Originally published in German by Springer 1928.

Dennis Walsh, Robert F. Carswell, and Ray J. Weymann. 0957 + 561 A, B - twin quasistellar objects or gravitational lens. *Nature*, 279:381–384, May 1979.

Larry Wasserman. *All of Nonparametric Statistics.* Springer Texts in Statistics. Springer, New York, 2006.

Joe Whittaker. *Graphical Models in Applied Multivariate Statistics.* J. Wiley & Sons, Chichester, UK, 1990.

Ronald R. Yager. Measurements of properties on fuzzy sets and possibility distributions. In E. P. Klement, editor, *Proceedings 3rd Inter. Seminar on Fuzzy Set Theory*, Linz, 1981. Johannes Kepler Universität.

Ronald R. Yager. Entropy and specifity in the mathematical theory of evidence. *International Journal of General Systems*, 9:249–260, 1983.

Ronald R. Yager. On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41:93–137, 1987.

Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

Lotfi A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978. Reprinted in Fuzzy Sets and Systems 100 Supplement (1999) 9–24.

Lotfi A. Zadeh. Review of Shafer's *A Mathematical Theory of Evidence. AI Magazine*, 5(3):81–83, 1984.

Nevin Lianwen Zhang and David Poole. Exploiting causal independence in bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.

Index

λ-message, 47
λ-value, 46
π-message, 47
π-value, 46

active path, 41
annotations, 2
applicable label, 119
attribute, 10–11, 15–17
  composite, 16, 39
  set-valued, 17–19
  tuple-based formalization, 94
average record Log-Likelihood, 136

base attributes, 96
basic domain, 90
basic probability assignment, 60
Bayes' rule, 31
Bayesian Network, 43
belief, 14
belief base, 22
belief function, 59
belief measure, 60
belief set, 22
Bernoulli's theorem, 27
blocked path, 41
body of evidence, 62

Causal Networks, *see* Graphical Model
chain rule decomposition, 96
chain rule factorization, 43
chain rule of probability, 31
circles
  removal of, 48
  resolution by superposition, 48
  resolution by variable elimination, 49
classical interpretation of sets, 52
classical probability, 28
clique tree propagation, 47
coarsening, 111
coherent conditional probability, 31
common cause, 40
common refinement, 64
compatibility, 23
composite attribute, 16, 39
conclusion (inference), 22
condensed distribution, 93
  application to trees, 118
  conditioning, 106, 107
  marginal, 102
  scalability, 93
condensed probability distribution, 90
condensed set-valued attribute, 90
conditional distribution, 105
conditional independence graph, 42
conditional probability, 31
conditioning, 105
  condensed distribution, 106, 107
  probabilistic, 29
consistent, 55
consonant, 56, 70
covariance selection, 50
coverage factor, 92
  conditional, 107
  elementary, 105
  multivalued case, 99
coverage function, 92

multivalued case, 99
coverage rate
error, 137

d-separation, 41
data, 11–12
decomposition, 35
Dempster's rule
of combination, 62
of conditioning, 61
Dempster's rule of combination, 62
Dempster, Arthur P., 57
Dempster-Shafer theory, 57
descendant (hierarchy), 112
direct refinement, 112
directed acyclic graph, 43

elementary coverage factor, 105
elementary event, 26
elementary possibility assignment, 71
induced from random set, 75
elementary refinement operation, 112
empty-set outcome, 88
event, 26
elementary, 26
evidence, 59, 62
extended set universe, 90

factor potential, 44
focal set, 55
focal sets
consonant, 56
frame of discernment, 15–16, 19, 20, 26, 109, 110
decomposition, 35
refinement, 111
related, 112

graph, 36
directed, 36
undirected, 36
Graphical Model, 35–49

hierarchical attribute domain, 112
hierarchy, 111
definition of, 111
descendant, 112
parent, 112
root node, 112

imprecision, 23–24
independence
statistical
conditional, 39
independence map, 42
inference, 22
Inference Network, *see* Graphical Model
instantiation, 15–16
tuple-based formalization, 94, 95

knowledge, 12–14
graded, 24
knowledge representation, 15–32
ordinal, 24
probabilistic, 26
relational, 23
set-of-instances view, 19
state-of-the-world view, 20
symbolic, 22
Kolmogorov axioms, 26

Laplace correction, 28, 120, 121
Log-Likelihood, 135

marginal distributions, 102
Markov Network, 43
modus ponens, 22
multi-label instantiation
in hierarchy, 119

node separation
in directed graphs, 41
in undirected graphs, 41

object, 9–10
objective probability, 28
one-point coverage, 75, 89, 92

computation using relative coverage factors, 93
outcome, 26
overfitting, 132, 135

parent (hierarchy), 112
perfect map, 42
pignistic transformation, 66
Plato, 12
plausibility measure, 60, 70
polytree propagation, 45
Popper, Karl, 9
possibilistic aggregation problem, 82
possibility
aggregation problem, 78–83
distribution, 68
possibility distribution, 68
normalized, 68
possibility measure, 68
relation to measure theory, 68
possibility theory, 25, 67
premise, 22
probabilistic reasoning, 29
probability, 26–32
classical, 28
coherent conditional, 31
conditional, 31
estimation from data, 27–28
objective, 28, 29
subjective, 28, 29
probability distribution, 26
conditional, 29
marginal, 29
probability tree, 117
projection
of relations, 96
of tuples, 95
proportion, 27
proposition, 59
Protein-DNA interaction, 52

random experiment, 26
random set, 54, 91
consistent, 55
consonant focal sets, 56
one-point coverage, 75
properties of, 55
relation to possibility distributions, 75
reasoning, 22
refinement, 111
direct, 112
refinement operation
elementary, 112
relational database, 36
relations, 17
relative coverage factor, 92
restriction, 95
root
of a hierarchy, 112

Saccharomyces cerevisiae, 129
Saccharomyces Genome Database, 129
sample space, 26
separator set, 48
set reduction mapping
multivariate, 97
one-dimensional, 90
tuple-based, 97
SGD, 129
Shafer, Glenn, 58
simple graph, 36
singleton probability
error, 137
singletons
precise knowledge specification, 53
statistical independence
conditional, 39
sub-frame, 112
sub-frame expansion, 112
subjective probability, 28
subjective vs. objective probability, 29

transcription factor, 52

transferable belief model, 66
tree-structured domains, 109
tuple, 94
tuple-based formalization, 94
  attribute, 94
  instantiation, 94

u-separation, 41
uncertainty, 21, 26
  in hierarchies, 114–118
universe of discourse, *see* frame of discernment

variable elimination, 49
variance error, 132

weak law of large numbers, 27
weight of evidence problem, 66

yeast, 129