Incremental Inference on Higher-Order Probabilistic Graphical Models Applied to Constraint Satisfaction Problems

by Simon Frederik Streicher

Dissertation presented for the degree of Doctor of Philosophy in Electrical and Electronic Engineering in the Faculty of Engineering at Stellenbosch University

Promoter Prof. Johan du Preez

March 2022

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

This dissertation includes three original papers published in peer-reviewed journals or books. The development and writing of the papers (published and unpublished) were the principal responsibility of myself and, for each of the cases where this is not the case, a declaration is included in the dissertation indicating the nature and extent of the contributions of co-authors.

Date: March 2022

Abstract

Probabilistic graphical models (PGMs) are used extensively in the probabilistic reasoning domain. They are powerful tools for solving systems of complex relationships over a variety of probability distributions, such as medical and fault diagnosis, predictive modelling, object recognition, localisation and mapping, speech recognition, and language processing [5, 6, 7, 8, 9, 10, 11]. Furthermore, constraint satisfaction problems (CSPs) can be formulated as PGMs and solved with PGM inference techniques. However, the prevalent literature on PGMs shows that suboptimal PGM structures are primarily used in practice and a suboptimal formulation for constraint satisfaction PGMs.

This dissertation aimed to improve the PGM literature through accessible algorithms and tools for improved PGM structures and inference procedures, specifically focusing on constraint satisfaction. To this end, this dissertation presents three published contributions to the current literature:

- a comparative study to compare cluster graph topologies to the prevalent factor graphs [1],
- an application of cluster graphs in land cover classification in the field of cartography [2], and
- a comprehensive integration of various aspects required to formulate CSPs as PGMs and an algorithm to solve this formulation for problems too complex for traditional PGM tools [3].

First, we present a means of formulating and solving graph colouring problems with probabilistic graphical models. In contrast to the prevailing literature that mostly uses factor graph configurations, we approach it from a cluster graph perspective, using the general-purpose cluster graph construction algorithm, LTRIP. Our experiments indicate a significant advantage for preferring cluster graphs over factor graphs, both in terms of accuracy as well as computational efficiency.

Secondly, we use these tools to solve a practical problem: land cover classification. This process is complex due to measuring errors, inefficient algorithms, and low-quality data. We proposed a PGM approach to boost geospatial classifications from different sources and consider the effects of spatial distribution and inter-class dependencies (similarly to graph colouring). Our PGM tools were shown to be robust and were able to produce a diverse, feasible, and spatially-consistent land cover classification even in areas of incomplete and conflicting evidence.

Lastly, in our third publication, we investigated and improved the PGM structures used for constraint satisfaction. It is known that tree-structured PGMs always result in an exact solution [12, p355], but is usually impractical for interesting problems due to exponential blow-up. We, therefore, developed the "purge-and-merge" algorithm to incrementally approximate a tree-structured PGM. This algorithm iteratively nudges a malleable graph structure towards a tree structure by

selectively *merging* factors. The merging process is designed to avoid exponential blow-up through sparse data structures from which redundancy is *purged* as the algorithm progresses. This algorithm is tested on constraint satisfaction puzzles such as Sudoku, Fill-a-pix, and Kakuro and manages to outperform other PGM-based approaches reported in the literature [13, 14, 15]. Overall, the research reported in this dissertation contributed to developing a more optimised approach for higher-order probabilistic graphical models. Further studies should concentrate on applying purge-and-merge on problems closer to probabilistic reasoning than constraint satisfaction and report its effectiveness in that domain.

Uittreksel

Grafiese waarskynlikheidsmodelle (PGM) word wyd gebruik vir komplekse waarskynlikheidsprobleme. Dit is kragtige gereedskap om sisteme van komplekse verhoudings oor ‘n versameling waarskynlikheidsverspreidings op te los, soos die mediese en foutdiagnoses, voorspellingsmodelle, objekherkenning, lokalisering en kartering, spraakherkenning en taalprosessering [5, 6, 7, 8, 9, 10, 11]. Voorts kan beperkingvoldoeningsprobleme (CSP) as PGM’s geformuleer word en met PGM gevolgtrekkingtegnieke opgelos word. Die heersende literatuur oor PGM’s toon egter dat sub-optimale PGM-strukture hoofsaaklik in die praktyk gebruik word en ‘n sub-optimale PGM-formulering vir CSP’s.

Die doel met die verhandeling is om die PGM-literatuur deur toeganklike algoritmes en gereedskap vir verbeterde PGM-strukture en gevolgtrekking-prosedures te verbeter deur op CSP toepassings te fokus. Na aanleiding hiervan voeg die verhandeling drie gepubliseerde bydraes by die huidige literatuur:

- ‘n vergelykende studie om bundelgrafieke tot die heersende faktorgrafieke te vergelyk [1],
- ‘n praktiese toepassing vir die gebruik van bundelgrafieke in “land-cover”-klassifikasie in die kartografieveld [2] en
- ‘n omvattende integrasie van verskeie aspekte om CSP’s as PGM’s te formuleer en ‘n algoritme vir die formulering van probleme te kompleks vir tradisionele PGM-gereedskap [3].

Eerstens bied ons ‘n wyse van formulering en die oplos van grafiekkleurprobleme met PGM’s. In teenstelling met die huidige literatuur wat meestal faktorgrafieke gebruik, benader ons dit van ‘n bundelgrafiek-perspektief deur die gebruik van die automatiese bundelgrafiekkonstruksie-algoritme, LTRIP. Ons eksperimente toon ‘n beduidende voorkeur vir bundelgrafieke teenoor faktorgrafieke, wat akkuraatheid asook berekende doeltreffendheid betref.

Tweedens gebruik ons die gereedskap om ‘n praktiese probleem op te los: “land-cover”-klassifikasie. Die proses is kompleks weens metingsfoute, ondoeltreffende algoritmes en lae-gehalte data. Ons stel ‘n PGM-benadering voor om die georuimtelike klassifikasies van verskillende bronne te versterk, asook die uitwerking van ruimtelike verspreiding en interklas-afhanklikhede (soortgelyk aan grafiekkleurprobleme). Ons PGM-gereedskap is robuus en kon ‘n diverse, uitvoerbare en ruimtelik-konsekwente “land-cover”-klassifikasie selfs in gebiede van onvoltooide en konflikterende inligting bewys.

Ten slotte het ons in ons derde publikasie die PGM-strukture vir CSP’s ondersoek en verbeter. Dit is bekend dat boomstrukture altyd tot ‘n eksakte oplossing lei [12, p355], maar is weens eksponensiële uitbreiding gewoonlik onprakties vir interessante probleme. Ons het gevolglik die algoritme, purge-and-merge, ontwikkel om inkrementeel ‘n boomstruktuur na te doen.

Die algoritme hervorm ‘n bundelgrafiek stapsgewys in ‘n boomstruktuur deur faktore selektief te “merge”. Die saamsmeltproses is ontwerp om eksponensiële uitbreiding te vermy deur van yl datastrukture gebruik te maak waarvan die waarskeinlikheidsruimte ge-“purge” word namate die algoritme vorder. Die algoritme is getoets op CSP-speletjies soos Sudoku, Fill-a-pix en Kakuro en oortref ander PGM-gegronde benaderings waaroor in die literatuur verslag gedoen word [13, 14, 15]. In die geheel gesien, het die navorsing bygedra tot die ontwikkeling van ‘n meer geoptimaliseerde benadering vir hoër-orde PGM’s. Verdere studies behoort te fokus op die toepassing van purge-and-merge op probleme nader aan waarskynlikheidsredenasie-probleme as aan CSP’s en moet sy effektiwiteit in daardie domein rapporteer.

Acknowledgements

- First and foremost, I would like to thank my promoter, Prof. Johan du Preez, for your help, encouragement, and most of all, your friendship. Furthermore, thanks for using your beautiful photograph as the cover page of this dissertation.
- Secondly, I would like to thank my wife, Tina Streicher, for your support through the years of writing, especially for the last stretch. Thank you for your love and support; none of this would have been possible without your encouragement and perseverance.
- I would like to thank Lloyd Hughes and Ekaterina Chuprikova for collaborating with me.
- Also, thank you, Tarl Berry, Petro Wagner, Dirk Streicher, Jaco Briers, Elretha Britz, Jacques Smidt, and Francois Kamper for direct help and technical support in completing this work, Malan Kriel for patiently waiting for my full involvement in Auto Actuary, my parents for their continual encouragement, and God for overseeing it all.
- For funding a large portion of this research through bursaries, I would also like to thank Stone Three.

Declaration of publications[†]

With regard to the publications within this dissertation, the contributions of author S. Streicher were as follows:

[1] Chapter 3: S. Streicher and J. du Preez, "Graph Coloring: Comparing Cluster Graphs to Factor Graphs," in *Proceedings of the ACM Multimedia 2017 Workshop on South African Academic Participation*. SAWACMMM '17. New York, NY, USA: ACM, 2017, pp. 35–42.

- **90% S. Streicher:** Majority of the contributions. Paper writing and revision.
- **10% J. du Preez:** Idea and discovery of LTRIP algorithm. Critical revision of the paper.

[2] Chapter 4: L. H. Hughes, S. Streicher, E. Chuprikova, and J. du Preez. "A Cluster Graph Approach to Land Cover Classification Boosting." *Data*, 4(1), 2019. ISSN 2306-5729.

- **40% L. H. Hughes:** Idea, methodological, and experimental formulations. Data handling and execution of experiments. Paper writing and revision.
- **35% S. Streicher:** The formulation, design, and execution of the PGMs used in this work, specifically Section 4.3.1, Section 4.3.2, and the design of the potential functions in Section 4.3.4. Paper writing.
- **20% E. Chuprikova:** Idea, data preprocessing, and evaluation of results in comparison to existing approaches. Paper writing.
- **5% J. du Preez:** Critical revision of the paper.

[3] Chapter 5: S. Streicher and J. du Preez, "Strengthening Probabilistic Graphical Models: The Purge-and-merge Algorithm," *IEEE Access*, vol. 9, pp. 149 423–149 432, 2021.

- **90% S. Streicher:** Majority of the contributions. Paper writing and revision.
- **10% J. du Preez:** Research supervision and technical expertise in guiding the scope. Critical revision of the paper.

Declaration by co-authors[†]

The undersigned hereby confirm that (1) the declaration above accurately reflects the nature and extent of the contributions of the candidate and the co-authors, (2) no other authors contributed besides those specified above, and (3) potential conflicts of interest have been revealed to all interested parties and that the necessary arrangements have been made to use the material in this dissertation.

[†]Declaration with signatures is in possession of dissertation author and promoter.

To Tina

for your faith in my work when I had none left

Contents

Abbreviations

AC arithmetic circuit
ACE AC compilation and evaluation
AI artificial intelligence
ASTER Advanced Space-borne Thermal Emission and Reflection
ATKIS Amtliches Topographisch-Kartographisches Informations System
BP belief propagation
BU belief update
CART classification and regression tree
CLC2006 CORINE Land Cover 2006
CNF conjugate normal form
CNN convolutional neural network
CORINE Coordination of Information on the Environment
CSP constraint satisfaction problem
DEM digital elevation model
LSTM long short-term memory
LTRIP layered trees for the running intersection property
ML maximum likelihood
MODIS Moderate Resolution Imaging Spectroradiometer
OLI Operational Land Imager
OSM OpenStreetMap
PGM probabilistic graphical model
RF random-forest
RIP running intersection property
SDD sentential decision diagram
SER Sudoku explainer rating
SVM support vector machine
VGI volunteered geographic information

Glossary

arithmetic circuit An alternative representation of a Bayes network for rapid conditional and marginal queries – see ACE [16] for a software implementation of such a system.

ACE Software that compiles a Bayes network into an arithmetic circuit to answer conditional and marginal queries – available at ACE [16].

Bayes network A PGM structure representing a set of variables and their conditional dependencies via a directed acyclic graph – see Section 2.4 and Figure 2.2.

belief propagation A message-passing algorithm for performing inference on graphical models – well-defined on tree-structured graphs and results in exact inference. See Sections 1.1.1 and 2.8 and Algorithm 1.

belief update An equivalent algorithm to belief propagation that requires fewer calculations – introduced by Lauritzen and Spiegelhalter [17].

Calcudoku A similar puzzle to Killer Sudoku, but with the custom regions having both a number and an operator $+$, $-$, $\times$ or $\div$, such that applying a region's operator to its cells (e.g. $\square \times \square \times \square$) must yield the number attached to the region.

cardinality Used in this dissertation in the context of discrete random variables, where it refers to the size of the domain of the variable – e.g. a random variable representing a die roll has a cardinality of 6.

classifications boosting A method for combining multiple classifications into a stronger classification by compensating for the weaknesses of the individual classifications – see Chapter 4.

cluster graph A PGM structure that allows for multivariate message passing – see Sections 1.1.2, 3.3.2, 3.3.3, and 5.2.3.

constraint satisfaction problem The problem of assigning values to variables under a given set of constraints over those variables – see Sections 1.1.3 and 5.2.

determinism Used in this dissertation in the context of constraint satisfaction and potential tables, where it refers to zero-probability entries – i.e. states within a system that are deterministically impossible to be part of a solution.

domain Used in this dissertation in the context of discrete random variables, where it refers to the set of possible outcomes of a random variable – e.g. a random variable representing a die roll can be represented by the domain {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}.

error correction code Formulation for adding redundant information to data in order to detect possible errors later on and correct those errors – see Hamming (7,4) in Section 2.7 for an example.

exponential blow-up An informal term to indicate a non-linear, exponential growth of a solution space with regard to an increase in variables.

factor graph A simple PGM structure that supports univariate message passing – see Sections 1.1.2 and 3.3.1.

Fill-a-pix A paper-and-pencil adaptation of the classic computer game Minesweeper, where cells in a grid are pre-filled with digits to indicate how many neighbouring cells (including the cell with the digit) are to be painted in, in order to reveal an underlying pixel art image.

generalised arc consistency A type of consistency within a constraint satisfaction system, if a state shared by two variables are deterministic in one factor, it should be deterministic in all other factors – see Section 5.4.2 and Dechter et al. [18].

graph colouring The colouring (or labelling) of nodes in an undirected graph such that adjacent nodes do not have the same colour – see Section 3.2.1 and Figure 5.1.

junction tree A PGM structure for exact inference, similar to a cluster graph but with the condition that the graph is tree-structured – see Figure 2.14 for an example and Koller [12, p287] for variable elimination as a construction algorithm.

Kakuro A constraint puzzle similarly shaped to a crossword, where squares are to be filled in with digits 1 to 9 in order to sum up to an indicated number but without repeating a digit.

Killer Sudoku A similar puzzle to Sudoku but with additional custom regions and associated numbers, such that the sum of all cells within a region must sum to the associated number.

Kullbach-Leibler divergence A measure of the dissimilarity of one probability distribution compared to another distribution – see Section 2.9.

land cover classification Classifying a map according to land cover classes, such as “forest”, “grassland”, “water”, “artificial surface”, and many others – see Section 4.1.

LTRIP An algorithm to construct a cluster graph from set of factors – introduced in Section 3.3.3 and summarised in Section 5.2.3.

loopy graph A graph topology that allows for multiple paths between any two nodes

within the graph – i.e. a graph that contains cycles.

max marginalisation Similar to marginalisation over a probability distribution but instead of summing over grouped values, the maximum value in the group is taken. This operation can replace marginalisation during message passing if only the most likely assignment over the system is required.

maximal clique A subset of nodes from an undirected graph, where every node is adjacent to all other nodes and the subset is not extendable any further.

maximum spanning tree Remove edges from a loopy graph with weighted edges such that the result is a tree structure with maximum possible weight – see the Prim-Jarník algorithm [19] used by LTRIP in Section 3.3.3.

probabilistic graphical model Models that encode complex joint multivariate probability distributions using graphs – see Section 1.1.1 and Chapter 2.

purge-and-merge An algorithm to systematically simplify inference on constraint satisfaction problems – introduced in Section 5.4 and outlined in Algorithm 5.

random variable A variable with an uncertain value from a specific domain of possible values. See Section 2.3.

region graph A more general PGM structure than factor graphs and cluster graphs, see Sections 1.1.2.

running intersection property A necessary property for valid cluster graphs – for any two clusters sharing a variable, there must exist a unique path of clusters and sepsets between them containing the same variable. See Sections 3.3.2 and 5.2.3 and Koller [12, p347].

Shannon diversity index A metric based on Claude Shannon's formula for entropy for the variety of land uses in an area – see Dušek and Popelková [20] for a discussion on the usage and validity of this metric.

Sudoku A constraint puzzle with a partially filled 9×9 grid, where the digits 1 to 9 are to be assigned to each cell such that every digit appears only once in a region, with the regions as the nine rows, nine columns, and nine non-overlapping 3×3 sub-grids.

sum marginalisation The correct form of marginalisation over a probability distribution, where the sum is taken over grouped values.

tree-structured graph A graph topology where there is a unique path between any two nodes – i.e. there are no cycles in the graph.

volunteered geographical information The collection, analysis, and sharing of geographic information provided by individuals.

1

Introduction

1.1 | Background

1.1.1 | Probabilistic graphical models

Probabilistic graphical models (PGMs) originated with the discovery of belief propagation. Pearl [21] first proposed belief propagation as an exact inference procedure on trees. Then, after some investigation for its merit on loopy structures, he concluded that a loopy version of belief propagation can converge, but that "this asymptotic equilibrium is not coherent, in the sense that it does not represent the posterior probabilities of all nodes in the network"[22].

Independent discoveries in belief propagation were unknowingly made with turbo codes [23] as an iterative scheme for solving complex error codes. This came as a practical advancement in coding theory as it achieved near Shannon limit error-correcting coding and decoding but did not provide a general framework or a theoretical justification. McEliece et al. [24] later discovered that turbo codes are simply an application of loopy belief propagation on a loopy Bayes network that captures the error coding and channel noise.

By establishing the potential for PGMs in practical applications, a new interest in the field was developed for approximate inference via belief propagation schemes [25, 26, 27, 28]. This can be attributed to both the effectiveness of PGMs as well as the expressiveness. Intricate problems with multiple dependencies can easily be formulated into graphs and solved via PGM inference. This is useful for a variety of complex scenarios, where PGMs can be used to

- fill in a system's blind spot if a problem is sparsely defined,

- integrate an additional complexity level, such as measurement-noise parameters and other uncertainties, and
- solve problems with inverse relationships that are difficult to derive directly from calculus.

Thus, it is not surprising that PGMs are integral to various applications, such as medical and fault diagnosis, predictive modelling, object recognition, localisation and mapping, speech recognition, and language processing [5, 6, 7, 8, 9, 10, 11].

1.1.2 | Graph structures

Many advancements have been made in the representation and formulation of probabilistic reasoning problems; this is especially true in the choice of graph structures. For example, junction trees emerged from the variable elimination algorithm [22] and allow for belief propagation to determine exact marginal distributions. When allowing for cycles in the graph, junction trees can be extended to cluster graphs, and belief propagation can be extended to an approximate reasoning method called loopy belief propagation. Another structure is the factor graph, adapted from the physics literature [29]. Loopy belief propagation on this specific structure can be translated to minimising what is referred to in statistical physics as the Bethe free energy [30]. Yedidia et al. [31] developed this concept further and provided a generalisation of these structures and methods in the form of region graphs and generalised belief propagation (relevant to minimising what is referred to as Kikuchi free energy). To illustrate the different graph structures for further discussion, an example of each of these graph types is presented in Figure 1.1, with the order of generality as factor graph $\leq$ cluster graph $\leq$ region graph.

Graph design is one of the key elements in building an effective PGM. Factor graphs have been used in most practical applications where automatically generated graphs are required. Their deterministic structure allows for easy replication, and their relation to physics models and their simplicity allows for a more straightforward study of their behaviour. However, despite the importance of their energy-based physics relationship, they do not provide a guarantee on the accuracy obtained by loopy belief propagation [12, p529]. Also, due to univariate messages, pairwise correlations between variables are not as sufficiently propagated as in more general graph structures [12, p415]. Factor graphs are, therefore, less suited than cluster graphs and region graphs for loopy belief propagation.

A cluster graph is equivalent to a region graph of only two layers. A valid cluster graph (and factor graph) construction must adhere to the running intersection prop-

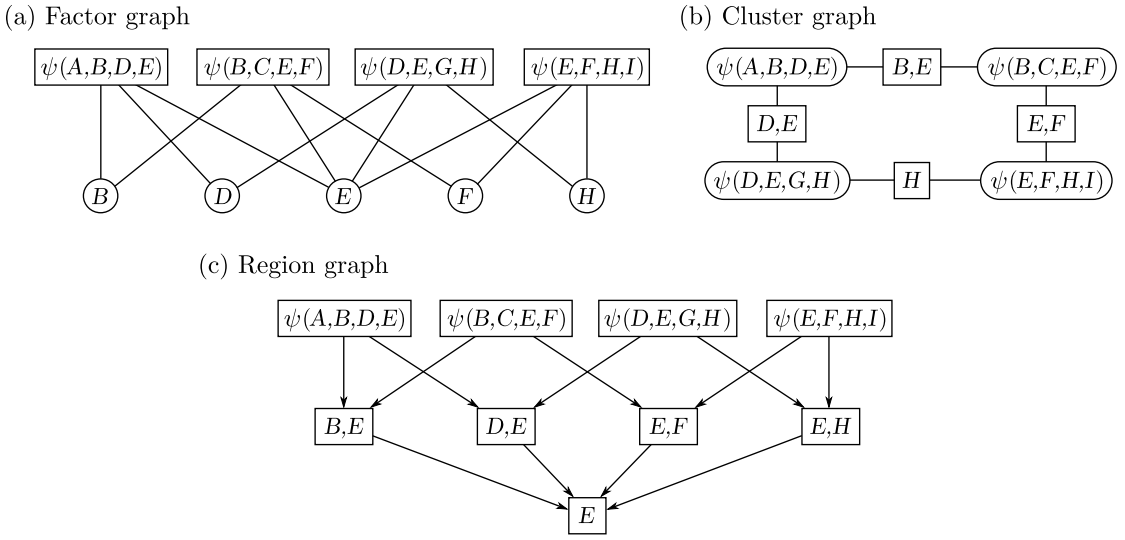


Figure 1.1: Example of three common types of PGM graphs configured from the same input factors $\psi(A, B, D, E)$, $\psi(B, C, E, F)$, $\psi(E, F, H, I)$, and $\psi(D, E, G, H)$. The region graph in (c) can be made equivalent to the cluster graph in (b) by replacing region $\boxed{E, H}$ with $\boxed{H}$ and removing region $\boxed{E}$ along with all links to it.

erty [12, p347]. While factor graphs trivially achieve this property, the extra freedom afforded to cluster graphs may result in multiple valid configurations for the same input factors. For constructing a cluster graph, Yedidia et al. [31] suggest fully creating all variable links between clusters and then, through some search heuristic, removing variables until the running intersection property is satisfied. Koller [12] elaborated on their experimental results along with the findings of Yedidia et al. [31] and Welling [32] and concluded that

- different graphs from the same input factors can lead to wildly different solutions [12, p404],
- the choice of cluster graph is a trade-off between computational cost and accuracy [12, p404], and
- that it is not obvious how to formulate a general-purpose cluster graph construction procedure [12, p429].

Region graphs have a far greater degree of freedom in their design than cluster graphs and factor graphs. Consequently, they adhere to a generalisation of the running intersection property based on nested containment and the calibration of so-called counting numbers. A counting number is associated with each region in order to weigh

the contribution of factors and variables to the system. Yedidia et al. [31] provide insight into efficient graph structures along with some methods to construct them. However, they clarify that "the problem of generating region graphs with highly accurate marginals is still an open research problem"[31]. Welling [32] proposed the "regional pursuit" algorithm as a sequential approach to designing region graphs from three operations they call "split", "merge", and "death".

It is unclear to what extent region graphs are superior to cluster graphs since an efficient algorithm for constructing cluster graphs is still mostly unexplored. Cluster graphs do have some advantages over deeply nested region graphs because of their structural simplicity. For example, they can accommodate the simplified formulation of belief propagation as defined on junction trees. Furthermore, a cluster graph will have fewer or equal regions than a region graph constructed from the same factors. This will result in fewer marginals to compute. Cluster graph construction already falls within a large space of optimisation. Therefore, it might be useful to explore cluster graph construction thoroughly before relying on region graph inference.

1.1.3 | Constraint satisfaction

A constraint satisfaction problem (CSP) involves a set of variables, a domain over each variable, and a set of constraints over these variables. A CSP solver is tasked with finding a solution (i.e. an assignment) over these variables. Dechter et al. [18] investigated the relationship between PGMs and constraint satisfaction. They proved that loopy belief propagation on cluster graphs with determinism in the network (i.e. zero-potentials) is reduced to an algorithm for generalised arc consistency. Any determinism propagated through the system cannot be reverted, and all determinism propagates within a finite number of iterations. This provides at least one formal justification for using belief propagation on loopy structures. Their investigation also shows that PGMs can be "flattened", such that sparse structures represent the factors in the system, thereby hiding zero-potentials and allowing determinism to reduce the potential space. Furthermore, replacing the marginalisation operations in belief propagation with max-marginalisation results in the same zero propagation but with less overall computation.

Although the constraint satisfaction properties of PGMs have been studied, the PGM literature does not provide a general-purpose CSP solver. One particular system for testing CSPs is Sudoku puzzles.

Moon and Gunther [33] investigated belief propagation on Sudoku puzzles using a factor graph to represent the constraint network. The system could, reportedly, not solve all Sudoku puzzles it encountered.

Goldberger [14] further explored the difference between belief propagation using max-marginalisation or sum-marginalisation. They report that a single round of belief propagation does not reliably solve Sudoku puzzles. For failed attempts, sum-marginalisation offers approximate marginals over the Sudoku factors that can be used to find the maximum likelihood. In contrast, the beliefs obtained through max-marginalisation will have an equal weighting for all non-zero potentials.

Khan et al. [15] combine belief propagation with Sinkhorn balancing. As with Goldberger [14], they aimed for a maximum likelihood estimation and not an exact solver. They reported an improvement but did not reliably solve Sudokus of moderate difficulty.

The PGM-based constraint satisfaction approaches listed here are all limited in one way or another. They are either inefficient in purging all the redundant potentials from the system or rely on heuristics to pick a maximum likelihood solution.

Other advances have been made to solve PGM-based constraint networks with tools from other domains. These approaches would typically convert a PGM into another domain and then solve it with domain-specific tools. Some examples include converting PGMs to Boolean satisfiability problems (such as conjugate normal form [34]), sentential decision diagrams [35], and arithmetic circuits [16]. For example, the arithmetic-circuit compilation and evaluation (ACE) software [16] can compile a factor graph (or a Bayes network) into an arithmetic circuit, which can be queried to infer a solution. These approaches are helpful, but they rely on tools outside the scope of traditional PGM structures and belief propagation algorithms.

1.2 | Objectives

This dissertation aims to explore some of the unresolved themes in the established PGM literature, specifically with regard to constraint satisfaction. We aim to develop inference techniques to express and solve constraint satisfaction problems more reliably. Our exact objectives are listed as

- exploring efficient cluster graph construction and comparing cluster graphs to the prevalent factor graph structure,
- investigating constraint satisfaction using PGMs and developing methods to solve high-order CSPs,
- applying our findings to a practical problem to verify our approach and the PGM approach in general.

1.3 | Contributions

This dissertation contributed to the following publications

[1] S. Streicher and J. du Preez, "Graph Coloring: Comparing Cluster Graphs to Factor Graphs," in *Proceedings of the ACM Multimedia 2017 Workshop on South African Academic Participation*. SAWACMMM '17. New York, NY, USA: ACM, 2017, pp. 35–42,

[2] L. H. Hughes, S. Streicher, E. Chuprikova, and J. du Preez. "A Cluster Graph Approach to Land Cover Classification Boosting." *Data*, 4(1), 2019. ISSN 2306-5729, and

[3] S. Streicher and J. du Preez, "Strengthening Probabilistic Graphical Models: The Purge-and-merge Algorithm," *IEEE Access*, vol. 9, pp. 149 423–149 432, 2021.

The direct contributions from author S. Streicher are

- a comparative study between cluster graphs and factor graphs, in which cluster graphs show great promise in comparison to factor graphs in [1],
- comprehensive integration of various aspects required to formulate graph colouring problems into PGMs in [1], further expanded to general constraint satisfaction in [3],
- a practical application of these principles in a land cover classification problem in the field of cartography in [2],
- the purge-and-merge algorithm, which makes it possible to simplify inference on complex CSPs systematically, developed as a combination of constraint satisfaction, LTRIP, belief propagation, and factor multiplication in [3], and
- an illustration of these tools in solving some very challenging puzzles, namely Sudoku, Fill-a-pix, Kakuro, and Calcudoku puzzles in [3].

Furthermore, a significant contribution published via this work is the general-purpose cluster graph construction algorithm, LTRIP in [1]; however, S. Streicher cannot be credited for discovering the basic algorithm. Nevertheless, the author can be credited for abstracting the "Connection-Weights" subroutine to allow for alternative minimisation functions.

2

Probabilistic graphical model basics

2.1 | Introduction

For the sake of completeness, this chapter describes basic underlying PGM techniques. These techniques will serve as building blocks for the later chapters of this dissertation, which covers incremental inference on higher-order PGMs. We, therefore, approach this chapter as a collection of relevant theory necessary to replicate the PGMs used in this work. First, we investigate the exponential blow-up found in the complexity of multivariate probabilities. Secondly, we introduce the concept of probability distributions and show how conditional independencies can be used to factorise large probabilistic spaces. Thirdly, we provide a computational model for representing factors and for computing factor operations. Finally, we introduce a technique for formulating a problem probabilistically using a cluster graph and solving the problem using belief propagation. Our example on the Hamming (7,4) code indicates that the system is an effective probabilistic inference tool that yields practical results.

2.2 | Why consider probabilistic graphical models

Graphical models emerged from the successes of the 1960s for the use of graph inference in Kalman filtering [36], error correction codes [24], and hidden Markov models [37]. Graphical models are a resourceful combination of graph theory and probabilistic inference techniques; they provide powerful tools for optimisation and probabilistic computation. They are used to encode dependencies among interacting variables to model a system's underlying probabilistic structure.

At the heart of PGMs lies the problem of exponential blow-up, where a problem's

slight growth can quickly lead to a computationally intractable large space. This can occur when finding a joint probability over even a modest number of variables.

A good example of this phenomenon is described in Shwe and Cooper [38] and in Theodoridis [39, p773]. A system for disease hypothesis inference on patients is explored by using a PGM for encoding diseases and symptoms. The model operates on

- n disease hypotheses $D_1, \ldots, D_n$,
- m symptom findings $F_1, \ldots, F_m$,
- a disease being either present or absent in a patient, and
- each finding being either observed or unobserved in a patient – an approximation for a symptom being present or not.

To understand the connections between these variables, see Figure 2.1, where the diseases and findings are represented as a Bayes network and an edge between disease D_i and finding F_j represents a corresponding link in the disease profile database. A simplification arising from this formulation is that the disease hypotheses are independent, and symptoms are conditionally independent given a disease hypothesis.

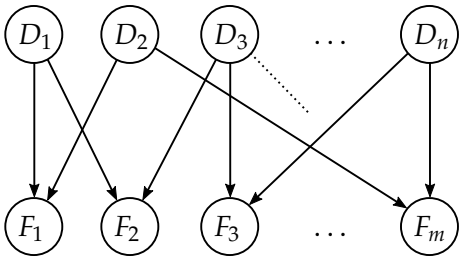


Figure 2.1: Bayes network showing example dependencies between diseases D_j and findings F_i in a medical network. Edges are only present where there is a direct link between a disease and a finding.

The goal of this model is to predict the presence of a number of diseases, given the presence of a set of findings. This can be achieved by calculating the joint probability distribution over all the variables $P(F_1, \ldots, F_m, D_1, \ldots, D_n)$ and then, for each patient, condition on the patient's findings to obtain the disease profile of that patient. The joint distribution can be expressed by (established later in Section 2.4)

$$P(F_1, \ldots, F_m, D_1, \ldots, D_n) = \prod_{i=1}^{n} P(F_i|\mathbf{D}_i) \prod_{j=1}^{m} P(D_j),$$

where $\mathbf{D}_i$ is the set of all associated diseases related to finding F_i. We like to point out here that the resulting joint distribution has a space of size 2^{n+m} and can easily grow to a computationally intractable scale.

When focusing on a single disease hypothesis, however, some reduction in the complexity can be obtained. We can then rephrase the problem as follows, given a set of findings $\mathbf{F}' = \mathbf{f}'$, what is the probability of a specific disease $D_j = d_j$? By defining $\mathbf{D}'$ as the set of all diseases related to $\mathbf{F}'$, we can use Equations 2.1 (established later in Section 2.4) to formulate the problem as

$$\begin{aligned} P(D_j|\mathbf{F}') &= \frac{P(\mathbf{F}'|D_j)P(D_j)}{P(\mathbf{F}')} \\ &= \frac{\sum_{D=d \forall D \in \mathbf{D}' \setminus \{D_j\}} \prod_{F_i \in \mathbf{F}'} P(F_i|\mathbf{D}_i) \prod_{D_k \in \mathbf{D}'} P(D_k)}{\sum_{D=d \forall D \in \mathbf{D}'} \prod_{F_i \in \mathbf{F}'} P(F_i|\mathbf{D}_i) \prod_{D_k \in \mathbf{D}'} P(D_k)}. \end{aligned}$$

If no additional factorisation is applied, the summation in the denominator will involve $2^{|\mathbf{D}'|}$ terms. This calculation is still computationally intractable for even a modest set of diseases $|\mathbf{D}'| = 500$. It is, therefore, clear that additional tools are required to reason about these problems computationally. The following sections will introduce efficient PGM structures and show how problems like these can be formulated and solved effectively using PGMs.

2.3 | Probability distributions

A random variable is a variable with an uncertain value. This value can be from a continuous domain (such as measuring the temperature in °C) or a discrete domain (such as rolling a die to determine a one-in-six outcome). In this work, we are primarily interested in discrete random variables and will mainly express probabilities in terms of a discrete probability mass function. The distribution over X, i.e. $P(X)$, is represented by a mass function $p_X(X{=}x)$, which we will mainly write in shorthand as either $p(X = x)$, or $p(x)$, whenever the meaning is clear from the surrounding context. Furthermore, we will liberally switch between expressing ideas in terms of probability distributions or mass functions.

A probability distribution represents the uncertainty inherent in one or more random variables, $P(X_1, \ldots, X_n)$, and can be expressed by a joint mass function $p(x_1, \ldots, x_n)$. The mass function is a mapping between the states over a group of variables and the probability of each state.

We can classify different types of distributions according to their context. The following structures are most important to our work: joint distributions, marginal distributions, conditional distributions, and potential functions [4].

- A joint distribution describes the combined probability of two or more random variables. For example the jointly distributed variables $X_1, X_2, \ldots, X_n$ have a joint distribution $P(X_1, X_2, \ldots, X_n)$ with mass function $p(x_1, \ldots, x_n)$.
- A marginal distribution is a distribution over a subset of the variables of a joint distribution, calculated by summing over unwanted variables, e.g. the marginal probability distribution $P(X_3, \ldots, X_n)$ with regard to $P(X_1, X_2, X_3, \ldots, X_n)$ is calculated as
$$P(X_3, \ldots, X_n) = \sum_{X_1=x_1, X_2=x_2} P(X_1, X_2, X_3, \ldots, X_n).$$
- A conditional distribution is a distribution with some of the variables observed (i.e. the values of those variables are certain), and the other random variables remain unobserved. The conditional distribution
$$P(X_3, \ldots, X_n | X_1{=}x_1, X_2{=}x_2)$$
is equivalent to setting $P(X_1, \ldots, X_n)|_{X_1=x_1 \text{ and } X_2=x_2}$ and normalising the result to have the sum of all its values equal to 1.
- A potential function is a function similar in structure to a probability mass function, often used as the intermediate result in a chain of computations. A potential function $\phi(X_1{=}x_1, \ldots, X_n{=}x_n)$, shorthand $\phi(x_1, \ldots, x_n)$, is a function over random variables X_1, ..., X_n with all its values greater than zero, i.e. $\phi(x_1, \ldots, x_n) > 0$. Therefore, a potential function $\phi(x_1, \ldots, x_n)$ is a more general case of a mass function $p(x_1, \ldots, x_n)$ without requiring that the sum of its values equals 1.

2.4 | Probability theory

We will start this discussion by listing the common theorems applicable to PGM literature (using $\perp\!\!\!\perp$ as the symbol for independence):

- Product rule: $P(X_1, X_2) = P(X_1|X_2)P(X_2)$
- Chain rule: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{i+1}{=}x_{i+1}, \ldots, X_n{=}x_n)$
- Bayes rule: $P(X_1|X_2) = \frac{P(X_2|X_1)P(X_1)}{P(X_2)}$
- Independence: $P(X_1, X_2) = P(X_1)P(X_2)$ iff $X_1 \perp\!\!\!\perp X_2$
- Conditional independence: $P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3)$ iff $X_1 \perp\!\!\!\perp X_2|X_3$
- Marginalisation: $\sum_{X_2=x_2} P(X_1, X_2) = P(X_1)$ (2.1)

By establishing these equations, we can now easily discuss the basic principles on which PGMs are built. We will show how some distributions can be factorised into a lower-dimensional representation and how some computational sequences can result in fewer calculations than others.

Given a joint distribution with n random variables $P(X_1, \ldots, X_n)$, we can use the chain rule from Equations 2.1 and factorise the distribution as

$$P(X_1, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2|X_3, \ldots, X_n)\ldots P(X_n). \tag{2.2}$$

Now consider two extreme cases about the underlying dependencies between the random variables: scenario 1, where no independencies exist between the variables involved and scenario 2, where all the variables are mutually independent – such as a series of coin flips.

For scenario 1, to marginalise with respect to one variable X_1 we sum over the other variables $X_2 \ldots X_n$ as

$$p(x_1) = \sum_{x_2}\sum_{x_3}\ldots\sum_{x_n} p(x_1, x_2, x_3 \ldots, x_n). \tag{2.3}$$

This is the general case where the structure cannot be exploited to reduce the computational cost. The number of summations is $\mathcal{O}(\text{card}(X_2) \cdot \text{card}(X_3)\ldots\text{card}(X_n))$, or in a simplified case $\mathcal{O}(k^{n-1})$, when the cardinality of all random variables is equal to k. Such a system can easily lead to an intractable computation, for even a modest variable count of $n = 100$ and cardinality of $k = 2$.

For scenario 2, we can employ the product rule to write the joint probability as

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2)\ldots P(X_n), \tag{2.4}$$

and the marginalisation over X_1 as

$$p(x_1) = P(x_1) \sum_{x_2} p(x_2) \sum_{x_3} p(x_3) \dots \sum_{x_n} P(x_n). \tag{2.5}$$

This reduces the complexity from $\mathcal{O}(k^{n-1})$ to $\mathcal{O}(k(n-1))$. For $n = 101$ and $k = 2$, such as in the case of 101 independent coin flips, the number of summations is reduced from 2^{100} down to 100. In reality, though, no summation is necessary for this example since all the $\sum_{x_i} P(x_i)$ terms are equal to 1.

Marginal posteriors

Since the two scenarios given above are extreme cases, most of the problems practically encountered will fall somewhere between these two extremes. For instance, given a distribution with the dependencies captured by the Bayes network in Figure 2.2, we will show how the marginal distributions can be calculated more directly than first calculating the joint distribution and then the marginals from that result.

In a Bayes network, the edges are drawn from a factor's right-hand side conditional variables to the left-hand side variables. For example, the network $(Z) \leftarrow (Y) \leftarrow (X)$ would represent the conditional distribution relationships $P(Y|X)$ and $P(Z|Y)$ and the full factorisation $P(X, Y, Z) = P(Z|Y)P(Y|X)P(X)$.

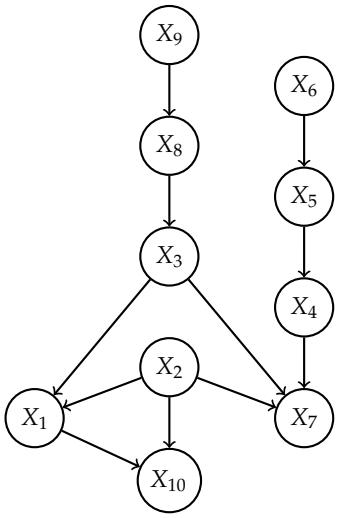


Figure 2.2: Example Bayes network depicting dependencies between random variables $X_1, \dots, X_{10}$. A factorisation of this graph can be found in Equation 2.6 in the main text.

The following is, therefore, a valid factorisation of the Bayes network in Figure 2.2

$$\begin{aligned} P(X_1, \ldots, X_{10}) = \\ & P(X_1|X_2, X_3) \cdot P(X_{10}|X_2, X_1) \cdot P(X_7|X_2, X_3, X_4) \cdot P(X_2) \cdot \\ & P(X_4|X_5) \cdot P(X_5|X_6) \cdot P(X_6) \cdot P(X_3|X_8) \cdot P(X_8|X_9) \cdot P(X_9). \end{aligned} \tag{2.6}$$

A calculation of the marginals $P(X_1, X_2, X_3)$ can be made more efficient than simply calculating the product of all the factors and then applying the marginalisation. Instead, the system's independencies can be exploited to allow for a piecewise marginalisation and multiplication sequence, such as

$$\begin{aligned} P(X_1, X_2, X_3) = \\ & P(X_1|X_2, X_3) \sum_{x_{10}} P(X_{10}|X_2, X_1) \sum_{x_4, x_7} P(X_7|X_2, X_3, X_4) P(X_2) \\ & \sum_{x_5} P(X_4|X_5) \sum_{x_6} P(X_5|X_6) P(X_6) \sum_{x_8} P(X_3|X_8) \sum_{x_9} P(X_8|X_9) P(X_9). \end{aligned} \tag{2.7}$$

Furthermore, the extraction of not only $P(X_1, X_2, X_3)$ but of all the marginal distribution can be achieved by using a single calculation flow. For this, we require an algorithm such as variable elimination or belief propagation on a junction tree [40]. A junction tree is a graph representation that acts as a map for the order of a marginalisation sequence. A junction tree avoids cycles by clustering distributions together that cannot be marginalised independently. For ill-structured problems, this can lead back to the original issue and result in a single joint distribution without an optimised marginalisation sequence.

In Section 2.8, we provide a formulation for belief propagation on tree-structured graphs. Section 2.9 introduces a heuristic for performing belief propagation on some ill-structured problems by allowing the PGM graph structure to contain cycles.

2.5 | Factors

A factor $\psi(X_1, \ldots, X_n)$ over random variables $X_1, \ldots, X_n$ describes the knowledge we have of those variables within a system – usually captured by a mass or potential function. In our work, the term factor is generally used in the context of factorising a system over groups of random variables. A mass function or potential function is generally used to represent (and work with) the underlying information. Therefore, in this work, the term factor is more linked to context than having a strictly distinct definition from the potential function it represents.

To use factors in a computational model, we need to represent their underlying data in terms of a data structure that can be stored and manipulated programmatically. For

example, since a potential function is a mapping between states and potentials, this mapping can be stored in any data structure that can capture this relationship, such as an n-dimensional array [41], a NamedArray [42], or a map or dictionary type [43, 44].

The choice of data structure can significantly affect the implementations of various factor operations, and great care needs to be taken to allow for applying labels to dimensions, re-ordering variables, aligning shared-scope between factors, and applying correct broadcasting rules. Figures 2.3 and 2.4 illustrate the same probability distributions but with two different data layouts.

$P(X, Y, Z)$

		y_1	y_2
		y_1	y_2
x_1	y_1	0.0162	0.0108
x_2	y_1	0.0228	0.0152
x_3	y_1	0.0210	0.0140
		z_1	

		y_1	y_2
		y_1	y_2
x_1	y_1	0.1485	0.0945
x_2	y_1	0.2090	0.1330
x_3	y_1	0.1925	0.1225
		z_2	

$P(X)$

x_1	0.27
x_2	0.38
x_3	0.35

$P(Y|Z=z)$

	z_1	z_2
y_1	0.600	0.611
y_2	0.400	0.389

Figure 2.3: Example joint distribution $P(X, Y, Z)$, univariate distribution $P(X)$, and conditional distribution $P(Y|Z)$. They are illustrated as 3D, 1D, and 2D tables respectively to highlight their dimensionality. Adapted from [4].

$P(X, Y, Z)$

x	y	z	$p(x, y, z)$
x_1	y_1	z_1	0.0162
x_1	y_1	z_2	0.1485
x_1	y_2	z_1	0.0108
x_1	y_2	z_2	0.0945
x_2	y_1	z_1	0.0228
x_2	y_1	z_2	0.2090
x_2	y_2	z_1	0.0152
x_2	y_2	z_2	0.1330
x_3	y_1	z_1	0.0210
x_3	y_1	z_2	0.1925
x_3	y_2	z_1	0.0140
x_3	y_2	z_2	0.1225

$P(X)$

x	$p(x)$
x_1	0.27
x_2	0.38
x_3	0.35

$P(Y|Z=z)$

y	z	$p(y|z)$
y_1	z_1	0.600
y_1	z_2	0.611
y_2	z_1	0.400
y_2	z_2	0.389

Figure 2.4: Example joint distribution $P(X, Y, Z)$, univariate distribution $P(X)$, and conditional distribution $P(Y|Z)$, illustrated as tables.

Note that some intricacies are involved in working with calculations on potential and mass functions, such as redundant dimensions. For example, if the distribution $P(X|Y)$ is represented as a table, it will require a two-dimensional structure. However, if $X \perp\!\!\!\perp Y$, then $P(X|Y) = P(X)$, which can be represented as a one-dimensional structure. Intricacies like these can be dealt with on a case-by-case basis or be programmatically baked into the data structures or factor operations.

2.6 | Factor operations

When encountering an equation such as $p(x_1, x_2, x_3) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2)\phi_3(x_1, x_3)$, it might not yet be clear how to perform the underlying mathematical operations. This section acts as an implementation guide for different factor operations used in the rest of our work. These operations are discussed by Koller [12] as factor multiplication, division, marginalisation, reduction, damping, and normalisation [12; Defs. 4.2, 10.7, 13.12, and 4.5; Eq. 11.14; and Ch. 4].

Using the factorisation $P(X, Y, Z) = P(X)P(Y|Z)P(Z)$ (with subsequent independencies $X \perp\!\!\!\perp Y, Z$) from Figure 2.3 as a baseline, we will show examples for each factor operation mentioned above.

2.6.1 | Multiplication

Our first example is the product $P(X, Y, Z) = P(X)P(Y, Z)$, and our second example is the product $P(Y, Z) = P(Y|Z)P(Z)$. The procedure is illustrated in Figures 2.5 and 2.6 as the product of each combination of state over the intersecting variables of the two distributions.

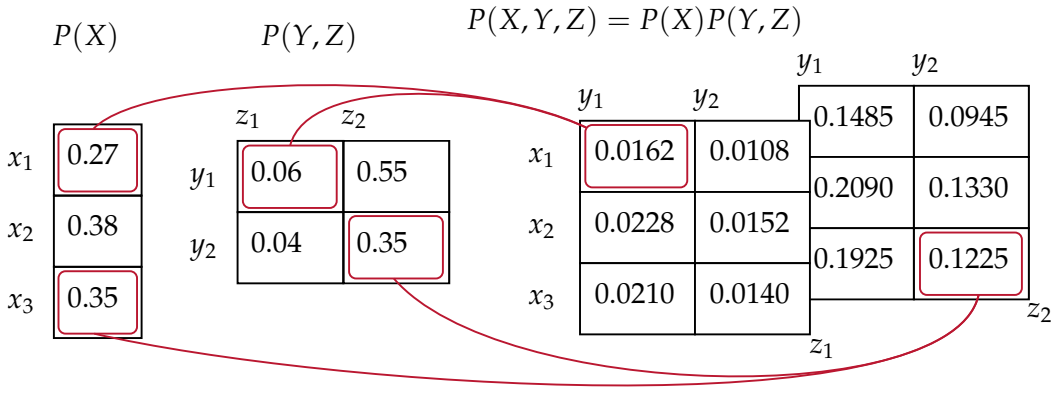


Figure 2.5: The product of two distributions. Note that the result has a structure of three dimensions, a total state space of 12 potentials, and $X \perp\!\!\!\perp Y, Z$. Adapted from [4].

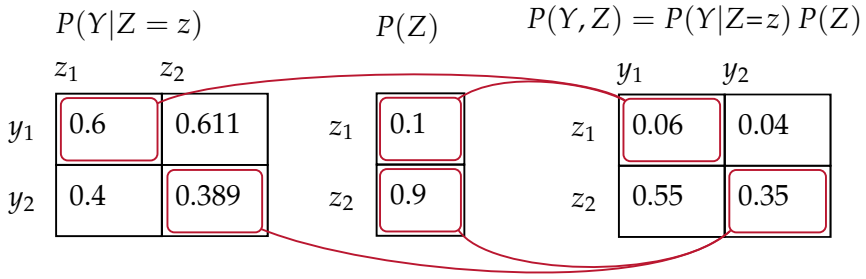


Figure 2.6: The product of two distributions with common scope Z. Note that the conditional distribution is conditioned on all possible outcomes of Z. Adapted from [4].

2.6.2 | Division

This procedure is the inverse of multiplication. It is, therefore, required that the variables in the denominator are a subset of the variables in the numerator. The calculation $P(Y|Z) = \frac{P(Y,Z)}{P(Z)}$ is displayed in Figure 2.7. It is accomplished by dividing matching variables in a component-wise fashion. Note that $\frac{0}{0}$ is defined as 0 in this context.

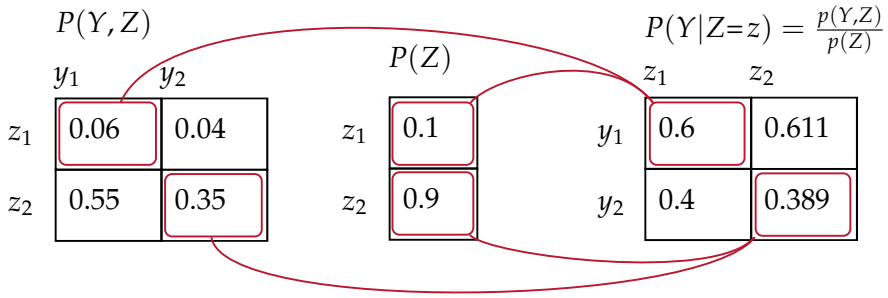


Figure 2.7: The division of two distributions. This is the inverse of the multiplication in Figure 2.6. Adapted from [4].

2.6.3 | Marginalisation

A marginal distribution is accomplished by summing over the joint distribution states not represented by the marginal. For example, the marginal distribution $P(Y,Z)$, calculated from the distribution $P(X,Y,Z)$, would require summing over X as $P(Y,Z) = \sum_{X=x} P(X,Y,Z)$. See Figure 2.8.

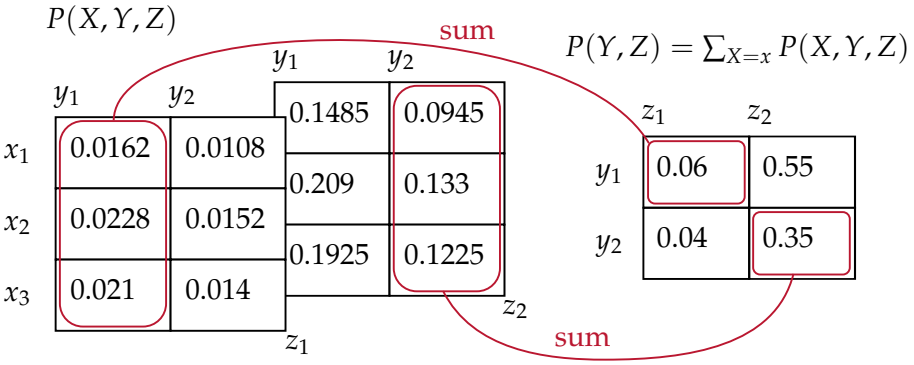


Figure 2.8: The marginalisation of a distribution. The variable X is removed from the table by summing over its domain. Adapted from [4].

2.6.4 | Conditioning (reduction)

When a random variable is reduced to a specific value, the resulting table is sliced (and normalised) according to that outcome. The distribution is thereby reduced to the dimensions of the variables with uncertain values. For example, see the process for extracting the distribution $P(X,Y|X{=}x_1)$ in Figure 2.9.

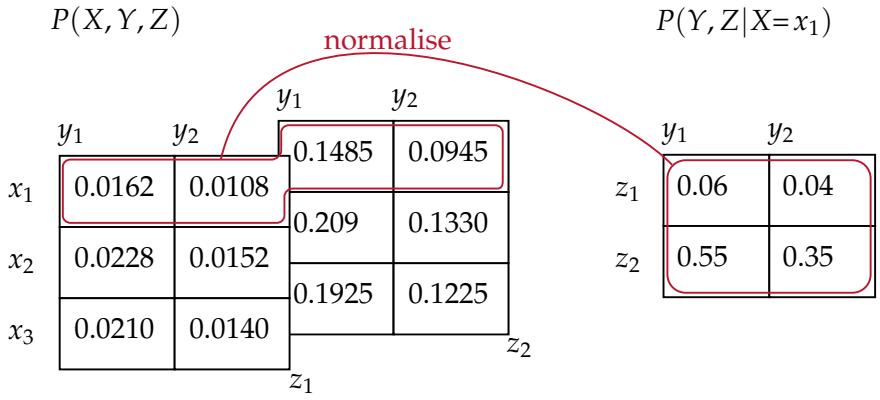


Figure 2.9: The reduction of a distribution, by observing $X{=}x_1$. Adapted from [4].

Furthermore, we can generalise the conditioning to any value, $X{=}x$, such that all possible conditionals over X are captured by a single table. This can be represented by the potential function $\phi(x,y,z) = P(Y{=}y, Z{=}z|X{=}x)$ as illustrated in Figure 2.10.

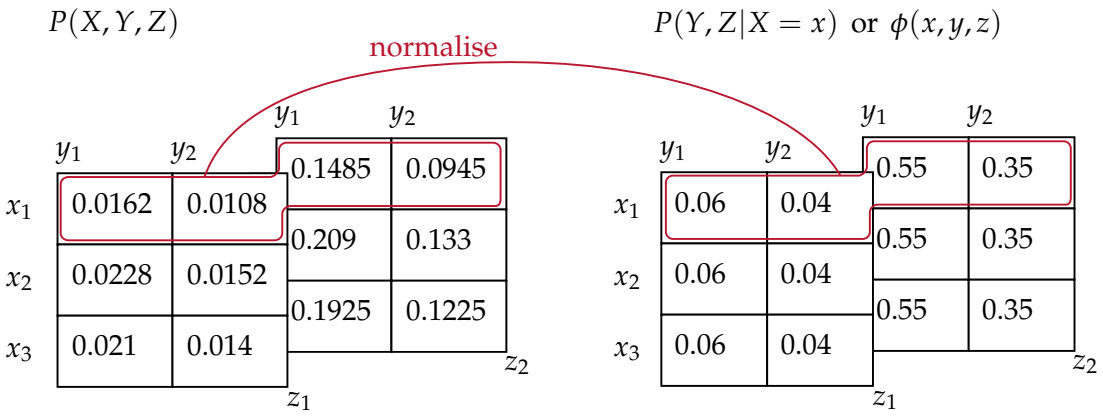


Figure 2.10: A conditional table capturing the random distributions for all conditioned states $X{=}x_1$, $X{=}x_2$, and $X{=}x_3$ as a single potential function. Adapted from [4].

2.6.5 | Damping (element-wise averaging)

Damping is not a fundamental factor operation but a specific biasing technique used during non-exact message passing as described in Section 2.9. However, for the sake of compiling these operations in one place, we present it here.

Given an optimisation problem with factor $\psi_{i-1}(y,z)$ as an intermediate result at step $i{-}1$ and $\psi_i(y,z)$ as an update at step i, it is common in some systems for such updates to overshoot. As a result, the system can be steered away from a good optimisation. To combat this, we can dampen the update by replacing it with an interpolation between itself and the previous result at step $i{-}1$. This can be achieved with a component-wise weighted average $\lambda\psi_i(y,z) + (1-\lambda)\psi_{i-1}(y,z)$, where $0 \leq \lambda \leq 1$. This process is illustrated in Figure 2.11. Note that for damping to be correctly weighted, the potentials need to be normalised first.

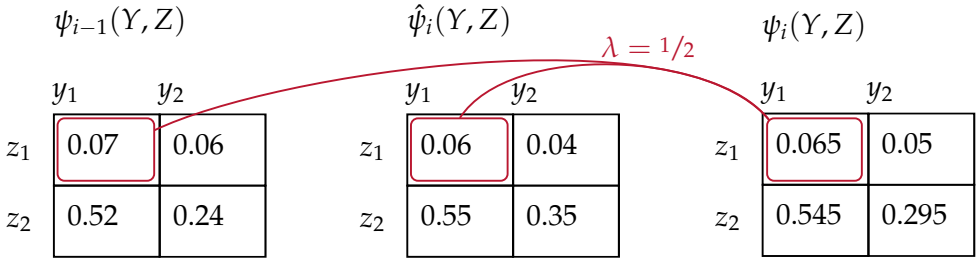


Figure 2.11: Damping as weighted element-wise averaging between the values of two factors, used in the context of message updates. Note that the initial update factor is indicated as $\hat{\psi}_i$, and the resulting dampened factor is indicated as ψ_i.

2.7 | Probabilistic formulation

We explain probabilistic formulation with the help of an example: the Hamming (7,4) code [4]. This idea is to encode a 4-bit message into a 7-bit sequence by introducing three extra parity bits. The goal is to include enough redundancy in the sequence to allow for message validation and error correction. Therefore, if the 7-bit sequence Hamming encoded sequence gets corrupted, the original 4-bit message can be recovered within a level of confidence.

The four message bits are b_1, b_2, b_3, b_4, the three parity bits are b_5, b_6, b_7, and their mathematical relationships are

$$\begin{aligned} b_5 &= b_1 \oplus b_2 \oplus b_3, \\ b_6 &= b_2 \oplus b_3 \oplus b_4, \\ b_7 &= b_1 \oplus b_3 \oplus b_4, \end{aligned} \tag{2.8}$$

with $\oplus$ as the XOR operator.

Given a message $b_1, \ldots, b_7$ is transmitted and received as $r_1, \ldots, r_7$, the receiver can extract the underlying transmitted sequence with a confidence related to the level of discrepancy within Equations 2.8. We can represent this problem probabilistically by assigning the random variables $B_1, \ldots, B_7$ and $R_1, \ldots, R_7$ to the underlying bits of the system. Then, we can draw the relationships between the variables as a Bayes network, see Figure 2.12, and factorise the joint distribution using the chain rule:

$$\begin{aligned} P(B_1, \ldots, B_7, R_1, \ldots, R_7) = \\ & P(R_1|B_1) \cdot P(R_2|B_2) \cdot P(R_3|B_3) \cdot P(R_4|B_4) \cdot P(R_5|B_5) \cdot P(R_6|B_6) \cdot P(R_7|B_7) \cdot \\ & P(B_5|B_1, B_2, B_3) \cdot P(B_6|B_2, B_3, B_4) \cdot P(B_7|B_1, B_3, B_4) \cdot \\ & P(B_1) \cdot P(B_2) \cdot P(B_3) \cdot P(B_4). \end{aligned} \tag{2.9}$$

To formulate this as a reasoning problem: we would like to estimate the values of the message bits $B_1, \ldots, B_7$, given the observations $r_1, \ldots, r_7$, i.e. determining the values $b_1, \ldots, b_7$ that maximises the distribution

$$P(B_1, \ldots, B_7 | R_1{=}r_1, \ldots, R_7{=}r_7). \tag{2.10}$$

This can naïvely be accomplished by multiplying all the factors together, applying the reductions $R_i{=}r_i$, and marginalising to the univariate distributions $P(B_i)$. Although this is viable for a problem with a small scope (such as this), it can easily become computationally intractable for problems with even a modest number of random variables. For instance, a 9×9 Sudoku grid has only 81 variables, each with a domain of 9 possibilities, resulting in a joint distribution of 9^{81} potentials.

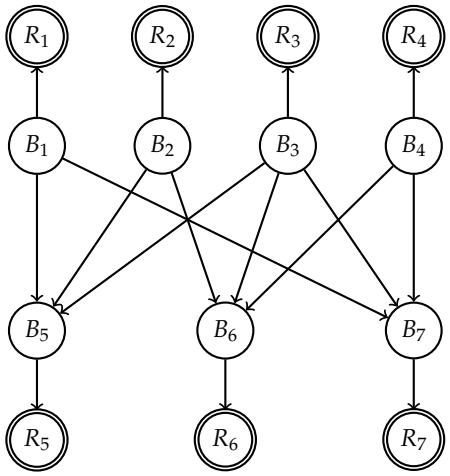


Figure 2.12: A Bayes network for the Hamming (7,4) probabilistic formulation, with message bits $B_1, \ldots, B_4$, parity bits B_5, B_6, B_7, and received bits $R_1, \ldots, R_7$. The double-lined circles indicate observed variables, e.g. $R_1{=}r_1, \ldots R_7{=}r_7$.

Factorisation

To formulate the problem into a PGM structure, we first need to find a suitable factorisation for the problem. Ideally, a system is simply factorised by applying the chain rule and using the resulting distributions as PGM factors. However, in practice, most factorisations are derived by using the available inputs, data, and logic. Thus, the problem is usually then worked backwards to a suitable factorisation.

For the Hamming (7,4) example, our choice of factors is the three mathematical relationships in Equation 2.8 and seven correlations between receive and send bits $P(R_i|B_i)$:

$$\begin{aligned} &\psi_1(B_5, B_1, B_2, B_3),\ \psi_2(B_6, B_2, B_3, B_4),\ \psi_3(B_7, B_1, B_3, B_4), \\ &\psi_4(R_1, B_1),\ \psi_5(R_2, B_2), \ldots,\ \psi_{10}(R_7, B_7). \end{aligned} \tag{2.11}$$

Note that the univariate factors $P(B_1), \ldots, P(B_4)$ are omitted since they carry no prior information in this case.

These factors can now be constructed into a suitable graph to assist with belief propagation. In Chapter 3, we investigate two graph configurations and find cluster graphs ideal for PGM tasks. For further discussion, see Section 3.3, where the benefits of cluster graphs are compared to factor graphs, and the graph construction algorithm LTRIP is provided (Section 3.3.3). As a result, a cluster graph is configured by using the factors as nodes and linking them up with sepsets, a connecting set of variables to facilitate message passing between nodes – see Figure 2.13.

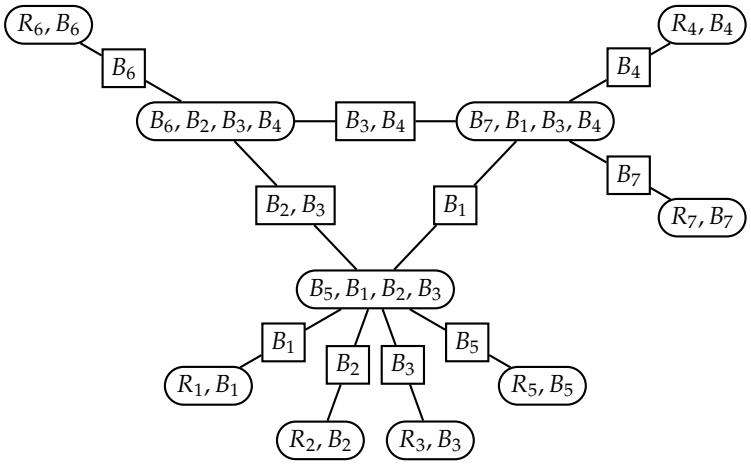


Figure 2.13: A cluster graph formulation for the recovery of a Hamming (7,4) encoded message. This graph is produced from the factors in Equation 2.11 using the LTRIP algorithm from Section 3.3. Note the sepset $\boxed{B_1}$ between $\overline{(B_5, B_1, B_2, B_3)}$ and $\overline{(B_7, B_1, B_3, B_4)}$ is not a full intersection since LTRIP enforces the running intersection property [12, p347].

2.8 | Belief propagation

Pearl [22] first discovered belief propagation as an algorithm for exact inference on tree-structured networks. We will, therefore, first demonstrate how to formulate this scheme on a tree structure and then return to the Hamming (7,4) example to show how to extend this scheme into an approximate reasoning algorithm for loopy structured PGMs. For now, we reuse the factorisation in Equation 2.6 to produce the structure in Figure 2.14 in order to assist our example.

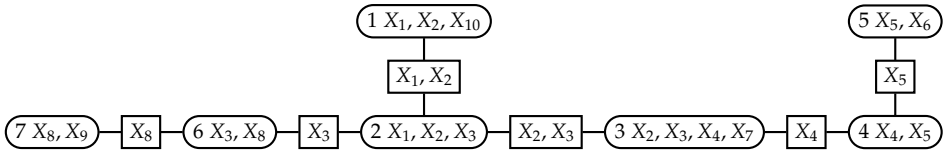


Figure 2.14: A tree-structured cluster graph produced from the factorisation of the Bayes network in Figure 2.2. See the main text for an example of applying belief propagation to this structure.

Belief propagation is built on two concepts: $\beta_i(\mathbf{C}_i)$ is the cluster belief for cluster i, and $\delta_{i\to j}(\mathbf{S}_{i,j})$ is the sepset belief for the sepset located between clusters i and j. We often denote these simply as β_i and $\delta_{i\to j}$ when it is clear from the context. β_i is designed to be the posterior distribution of factor ψ_i and $\delta_{i\to j}$ is designed to propagate information from cluster i to cluster j.

Performing belief propagation on a tree structure is similar to calculating the marginal distribution of each factor using an efficient factorisation and marginalisation sequence. It is essentially a generalisation of the example in Equation 2.7. The following two equations formulate these beliefs:

$$\beta_i = \psi_i \prod_{k \in \text{Adj}(i)} \delta_{k \to i}, \text{ and} \tag{2.12}$$

$$\delta_{i \to j} = \sum_{\mathbf{C}_i \setminus \mathbf{S}_{i,j}} \psi_i \prod_{k \in (\text{Adj}(i) \setminus \{j\})} \delta_{k \to i}, \tag{2.13}$$

with $k \in \text{Adj}(i)$ depicting neighbouring indices, and ψ_i the factor associated with $\mathbf{C}_i$, i.e. the prior. For example, this procedure is designed to steer a cluster such as $\{X_1, X_2, X_3\}$, with $\psi(X_1, X_2, X_3) = P(X_1|X_2, X_3)$, to the posterior belief $\beta(X_1, X_2, X_3) = P(X_1, X_2, X_3)$.

The message order for belief propagation on a tree structure is captured by Algorithm 1 (sum-product message passing [12, p348]). For example, applying this algorithm to the graph structure in Figure 2.14 results in the following message passing order [4]:

1. we first approach end node (5),

   $\delta_{5 \to 4} = \sum_{X_6} \psi_5$

   $\delta_{4 \to 3} = \sum_{X_5} \psi_4 \, \delta_{5 \to 4}$

   $\delta_{3 \to 2} = \sum_{X_4, X_7} \psi_3 \, \delta_{4 \to 3}$

2. then proceed to end node (7),

   $\delta_{7 \to 6} = \sum_{X_9} \psi_7$

   $\delta_{6 \to 2} = \sum_{X_8} \psi_6 \, \delta_{7 \to 6}$

3. then to end node (1),

   $\delta_{1 \to 2} = \sum_{X_{10}} \psi_1$

4. and finally, the order is reversed to propagate consensus back across the nodes,

   $\delta_{2 \to 1} = \sum_{X_3} \psi_2 \, \delta_{3 \to 2} \, \delta_{6 \to 2}$

   $\delta_{2 \to 6} = \sum_{X_1, X_2} \psi_2 \, \delta_{1 \to 2} \, \delta_{3 \to 2}$

   $\delta_{6 \to 7} = \sum_{X_3} \psi_6 \, \delta_{2 \to 6}$

   $\delta_{2 \to 3} = \sum_{X_4, X_7} \psi_2 \, \delta_{1 \to 2} \, \delta_{6 \to 2}$

   $\delta_{3 \to 4} = \sum_{X_5} \psi_3 \, \delta_{2 \to 3}$

   $\delta_{4 \to 5} = \sum_{X_6} \psi_4 \, \delta_{3 \to 4}$.

We can now calculate the cluster beliefs by applying Equation 2.12. For example the marginal distribution $P(X_1, X_2, X_3) = \beta_2(\mathbf{C}_2)$ is calculated as

$$\beta_2 = \psi_2 \, \delta_{1\to 2} \, \delta_{6\to 2} \, \delta_{3\to 2}.$$

Algorithm 1: Belief propagation. Adapted from [4].

Input: Cluster tree $\mathcal{T}$ (with clusters $\mathbf{C}_i$, edges $\mathcal{E}$, and sepsets $\mathcal{S}$).

1: // Calculate sepset beliefs
2: **while** any $\mathbf{C}_i$ is ready to pass a message to $\mathbf{C}_j$ **do**
3: $\quad \delta_{i\to j} := \sum_{\mathbf{C}_i \setminus \mathbf{S}_{i,j}} \psi_i \prod_{k\in(\mathrm{Adj}(i)\setminus\{j\})} \delta_{k\to i}$
4: **end while**
5: // Calculate cluster beliefs
6: **for** each cluster $\mathbf{C}_i$ **do**
7: $\quad \beta_i := \psi_i \prod_{k\in\mathrm{Adj}(i)} \delta_{k\to i}$
8: **end for**

Line 2 Cluster $\mathbf{C}_i$ is ready to pass a message to $\mathbf{C}_j$ when $\mathbf{C}_i$ has received all neighbouring messages except for $\delta_{j\to i}$.

There are slight variants on message passing and belief propagation. One such variation is called belief update message passing, also known as the Lauritzen-Spiegelhalter algorithm [17]. It is mathematically equivalent to belief propagation [12, p368] but is preferred in many cases, as it requires fewer calculations than belief propagation.

2.9 | Loopy belief propagation

If a cluster graph contains loops, a cyclic dependency between messages will result in a deadlock. However, we can heuristically solve this by passing messages iteratively until the beliefs converge. Note that it is proven that such a heuristic will not always converge and that the posterior beliefs can deviate significantly from the true marginals of the system [12, p407]. However, these schemes typically return practical results if implemented effectively [12, p407].

Loopy belief propagation introduces a new problem in the form of deriving a suitable message scheduling scheme. This is an active research problem [45] with many different implementations and trade-offs. However, a robust message passing scheme is essential to improving convergence and accuracy [12, p408].

Our focus is on a message passing scheme that prioritises the parts of the system that underwent significant updates. Our implementation is as follows: after a message $\delta_{i\to j}$ is propagated, all messages $\delta_{j\to k}\forall \text{Adj}(j) \setminus \{i\}$ are added to a priority queue using an error metric as weights. This allows us to prioritise the parts of the system that are furthest from converging. The adjustments necessary for belief propagation to accompany loopy graphs are shown in Algorithm 2 as loopy belief propagation [12, p397].

As an error metric, we use the Kullback-Leibler divergence as an approximation of the error between consecutive messages. However, note that the Kullback-Leibler divergence $D_{KL}(P||Q)$ measures the loss of information when a distribution P is approximated by a distribution Q, given as

$$D_{KL}(P||Q) = \sum_i P(i)\ \ln \frac{P(i)}{Q(i)}. \tag{2.14}$$

It is a nonsymmetric function, i.e. $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, and is, therefore, not a true distance or error metric. It does, however, report information about the similarity between the two distributions and is, in practice, still exceptionally efficient for these kinds of applications [46].

Since the algorithm is defined on potential functions and not on probability distributions directly, it is important to note the relationship between them, that is,

$$p(x_1, \ldots, x_n) = \frac{1}{Z}\phi(x_1, \ldots, x_n), \tag{2.15}$$

where Z is the normalisation constant $\sum_i \phi(i)$. Numerical stability is known to be a problem in long-running loopy belief propagation cycles [47]. It is, therefore, advisable to keep messages and beliefs normalised.

Algorithm 2: Loopy belief propagation. Adapted from [4].

Input: Cluster graph $\mathcal{T}$ (with clusters $\mathbf{C}_i$, edges $\mathcal{E}$, and sepsets $\mathcal{S}$).

1: $q := [\,]$
2: **for** each edge (i, j) in $\mathcal{E}$ **do**
3: // Initialise queue with miniscule priorities and a bias towards leaf nodes
4: priority $:= 10^{-10}(|\text{Adj}(i)| + |\text{Adj}(j)|)^{-1}$
5: q.push(priority $\rightarrow (i, j)$)
6: q.push(priority $\rightarrow (j, i)$)
7: // Initialise cluster messages as vacuous
8: $\delta_{i\rightarrow j} := \mathbf{1}$
9: $\delta_{j\rightarrow i} := \mathbf{1}$
10: **end for**
11: // Message passing
12: **while** q is not empty **do**
13: $(i, j) := q$.pop_highest()
14: $\delta_{\text{prev}} := \delta_{i\rightarrow j}$
15: $\delta_{i\rightarrow j} := \text{norm}\left(\sum_{\mathbf{C}_i \setminus \mathbf{S}_{i,j}} \psi_i \prod_{k\in(\text{Adj}(i)-\{j\})} \delta_{k\rightarrow i}\right)$ // Update message
16: // Propagate priorities
17: priority $:= \text{error}(\delta_{\text{prev}}, \delta_{i\rightarrow j})$
18: **for** $k \in \text{Adj}(j) - \{i\}$ **do**
19: **if** $(j, k) \in q$ **then**
20: remove (j, k) from q
21: **end if**
22: **if** priority $>$ chosen threshold **then**
23: q.push(priority$\rightarrow (j, k)$)
24: **end if**
25: **end for**
26: **end while**
27: // Calculate posterior beliefs
28: **for** each cluster $\mathbf{C}_i$ **do**
29: $\beta_i := \psi_i \prod_{k\in\text{Adj}(i)} \delta_{k\rightarrow i}$
30: **end for**

Line 8–9 An alternative initialisation is $\delta_{i\rightarrow j} := \text{norm}\left(\sum_{\mathbf{C}_i \setminus \mathbf{S}_{i,j}} \psi_i\right)$.

Line 15 Message damping can be applied by adding the line:

$$\delta_{i\rightarrow j} := (1 - \lambda)(\delta_{\text{prev}}) + (\lambda)(\delta_{i\rightarrow j}),$$

where λ is the damping factor.

Line 22 Local convergence is reached when all message updates are below a chosen threshold.

2.10 | Example

Hamming (7,4) model

We now have the tools to solve the Hamming (7,4) code as a PGM. We initialise the cluster graph factors of Equation 2.11 using the parity bit logic of Equation 2.8. Factor $\psi_1(B_5, B_1, B_2, B_3)$ represents the conditional distribution $P(B_5|B_1, B_2, B_3)$ via

$$\psi_1(B_5, B_1, B_2, B_3) = \begin{cases} 1, & \text{where } b_5 = b_1 \oplus b_2 \oplus b_3 \\ 0, & \text{where } b_5 \neq b_1 \oplus b_2 \oplus b_3 \end{cases}.$$

We similarly define the factors $\psi_2(B_6, B_2, B_3, B_4)$ and $\psi_3(B_7, B_1, B_3, B_4)$.

The factors $\psi_4(R_1, B_1), \ldots, \psi_{10}(R_7, B_7)$ are represented by $P(R_1|B_1), \ldots, P(R_7|B_7)$. If we assume a 10% chance for each bit to be erroneously flipped, we can set a probability of 90% for a transmitted bit being equal to a received bit and 10% otherwise:

$$\psi_4(R_1, B_1) = \begin{cases} 0.9, & \text{where } b_1 = r_1 \\ 0.1, & \text{where } b_1 \neq r_1 \end{cases}.$$

We assign the potential functions for $\psi_5, \ldots, \psi_{10}$ similarly.

Now that we have factor potentials, we can solve a Hamming (7,4) message by applying observations and running loopy belief propagation.

Example

Given that a message "1010" is to be sent over a noisy channel using a Hamming (7,4) code, the message is first encoded as "1010010" and then transmitted. After transmission, the message is recorded as "1110010", with an unknown bit-flip at b_2. We are now going to explore the use of belief propagation for detecting and recovering bit-flips.

First, we integrate the observations

$$R_1{=}1,\ R_2{=}1,\ R_3{=}1,\ R_4{=}0,\ R_5{=}0,\ R_6{=}1,\ R_7{=}0$$

by replacing the factors $\psi_4(R1, B1), \ldots, \psi_{10}(R7, B7)$ with reduced versions of themselves, $\psi\prime_4(B_1), \ldots, \psi\prime_{10}(B_7)$, with their potentials as

$$\psi\prime_4(1) = 0.9, \quad \psi\prime_5(1) = 0.9, \quad \psi\prime_6(1) = 0.9, \quad \psi\prime_7(1) = 0.1,$$
$$\psi\prime_8(1) = 0.1, \quad \psi\prime_9(1) = 0.9, \quad \psi\prime_{10}(1) = 0.1.$$

We can then run loopy belief propagation (Algorithm 2) on the cluster graph of Figure 2.13 and obtain posterior beliefs for each factor. By querying the posterior distributions, the system can extract the original message as "1010010" with very high confidence. To illustrate, posterior distributions from a resulting run of loopy belief propagation [4] are shown in Figure 2.15.

Beliefs $\psi\prime_4(B_5), \ldots, \psi\prime_{10}(B_7)$

	$x=0$	$x=1$
$P(B_1{=}x)$	01.01%	**98.99%**
$P(B_2{=}x)$	**98.01%**	01.99%
$P(B_3{=}x)$	00.03%	**99.97%**
$P(B_4{=}x)$	**98.99%**	01.01%
$P(B_5{=}x)$	**99.00%**	01.00%
$P(B_6{=}x)$	00.99%	**99.01%**
$P(B_7{=}x)$	**99.00%**	01.00%

Likely values $(B_1, \ldots, B_7) =$
$(1,0,1,0,0,1,0)$

Beliefs $\psi_1(B_5, B_1, B_2, B_3)$

B_5	B_1	B_2	B_3	$P(B_5, B_1, B_2, B_3)$
0	0	0	0	0.00000101
0	0	0	1	0.00000000
0	0	1	0	0.00000000
0	0	1	1	0.00991713
0	1	0	0	0.00000000
0	1	0	1	**0.98001150**
0	1	1	0	0.00009991
0	1	1	1	0.00000000
1	0	0	0	0.00000000
1	0	0	1	0.00000001
1	0	1	0	0.00010000
1	0	1	1	0.00000000
1	1	0	0	0.00000101
1	1	0	1	0.00000000
1	1	1	0	0.00000000
1	1	1	1	0.00986940

Figure 2.15: Probability distributions obtained from performing belief propagation on an erroneous Hamming (7,4) message of "1110010". Loopy belief propagation managed to identify and correct a bit-flip in b_2 with a 98% certainty, recovering the correct message of "1010010".

2.11 | Conclusion

The objective of this chapter was to present a basis for formulating and solving probabilistic reasoning problems using PGMs. We supplied all the theory and tools necessary to formulate, build, and solve the Hamming (7,4) code as a graphical model. The goal was to provide basic underlying PGM techniques as building blocks for the later chapters.

3

Graph colouring: comparing cluster graphs to factor graphs

Preface

This chapter presents *Graph Colouring: Comparing Cluster Graphs to Factor Graphs* [1][†], first presented at the ACM 25th International Conference on Multimedia (ACM MM 2017) at the Computer History Museum in Silicon Valley.

This publication is the first instalment of three to serve as the framework for our research on incremental inference on higher-order PGMs, applied to constraint satisfaction. The main contributions under author S. Streicher are

- a comparative study between cluster graphs and factor graphs, in which cluster graphs show great promise in comparison to factor graphs, and
- a comprehensive integration of various aspects required to formulate graph colouring problems into PGMs (further expanded to general constraint satisfaction in Chapter 5).

Furthermore, a significant contribution is the general-purpose cluster graph construction algorithm, LTRIP. Credit for the discovery of the algorithm goes to author J. du Preez, and credit for formulating the algorithm and abstracting the "Connection-Weights" subroutine goes to S. Streicher.

[†]Sections of this work have been published in:
S. Streicher and J. du Preez, "Graph Coloring: Comparing Cluster Graphs to Factor Graphs," in *Proceedings of the ACM Multimedia 2017 Workshop on South African Academic Participation*. SAWACMMM '17. New York, NY, USA: ACM, 2017, pp. 35–42.

By establishing the groundwork for graph colouring and cluster graph formulation, this chapter forms the basis for our larger project aimed to develop efficient PGM solutions to constraint satisfaction. Chapter 4 contributes to a cartography classification problem by using the region-based formulation established in Section 3.2.3 and applying the PGM-based inference found in Section 3.4.3. Chapter 5 introduces a higher-order PGM technique called purge-and-merge that combines LTRIP with factor multiplication to solve problems too difficult for other belief-propagation approaches, overcoming the limitations presented in Section 3.5.

Abstract

We present a formulation for solving graph colouring problems with probabilistic graphical models. In contrast to the prevailing literature that uses factor graphs for this purpose, we instead approach it from a cluster graph perspective. Noting the lack of algorithms to automatically construct valid cluster graphs, we provide such an algorithm (termed LTRIP). Our experiments indicate a significant advantage for preferring cluster graphs over factor graphs, both in terms of accuracy as well as computational efficiency.

3.1 | Introduction

Due to their learning, inference, and pattern-recognition abilities, machine learning techniques such as neural networks, probabilistic graphical models (PGMs), and other inference-based algorithms have become quite popular in artificial intelligence research. PGMs can easily express and solve intricate problems with many dependencies, making it a good match for problems such as graph colouring. The PGM process is similar to aspects of human reasoning, such as the process of expressing a problem by using logic and observation and applying inference to find a reasonable conclusion. With PGMs, it is often possible to express and solve a problem from easily formulated relationships and observations, without the need to derive complex inverse relationships. This can be an aid to problems with many inter-dependencies that cannot be separated into independent parts to be approached individually and sequentially.

Although the *cluster* graph topology is well established in the PGM literature [12, 18], the overwhelmingly dominant topology encountered in literature is the *factor* graph. We speculate that this is at least partially due to the absence of algorithms to *automatically* construct valid cluster graphs, whereas factor graphs are trivial to construct. To address this, we detail a general-purpose construction algorithm termed LTRIP (layered trees

for the running intersection property). We have been covertly experimenting with this algorithm for a number of years [48, 49, 4, 50].

The graph colouring problem originated from the literal colouring of planar maps. It started with the four-colour map theorem, first noted by Francis Guthrie in 1852. He conjectured that four colours are sufficient to colour neighbouring counties differently for any planar map. It was ultimately proven by Kenneth Appel and Wolfgang Haken in 1976 and is notable for being the first major mathematical theorem with a computer-assisted proof. In general, the graph colouring problem deals with the labelling of nodes in an undirected graph such that adjacent nodes do not have the same label. The problem is core to a number of real-world applications, such as scheduling timetables for university subjects or sporting events, assigning taxis to customers, and assigning computer programming variables to computer registers [51, 52, 53]. As graphical models became popular, message passing provided an exciting new approach to solving graph colouring and (the closely related) constraint satisfaction problems [33, 54]. For constraint satisfaction, the survey propagation message passing technique seems to be particularly effective [55, 56, 57, 58]. These techniques are primarily based on the factor graph PGM topology.

The work reported here forms part of a larger project aimed at developing an efficient alternative for the above message passing solutions to graph colouring. Cluster graphs and their efficient configuration are important in that work – hence our interest in those aspects here. Although we also provide basic formulations for modelling graph colouring problems with PGMs, this is not the primary focus of this chapter but instead only serves as a vehicle for comparing topologies.

The rest of this chapter is structured as follows. Section 3.2 shows how the constraints of a graph colouring problem can be represented as "factors". Furthermore, it is shown how these factors are linked up into graph structures on which inference can be applied. Section 3.3 discusses the factor graph and cluster graph topologies and algorithms for automatically configuring them. The former is trivial; for the latter, we provide the LTRIP algorithm in Section 3.3.3. Section 3.4 then integrates these ideas by expressing the well-known Sudoku puzzle (an instance of a graph colouring problem) as a PGM. The experiments in Section 3.5 show that the cluster graph approach is simultaneously faster and more accurate, especially for complex cases. The last two sections consider possible future exploration and final conclusions.

3.2 | Graph colouring with PGMs

This section provides a brief overview of graph colouring and PGMs, along with techniques for formulating a graph colouring problem as a PGM. We also explore the four-colour map theorem and illustrate how to solve these and similar problems through an example.

3.2.1 | A general description of graph colouring problems

Graph colouring problems are NP-complete – easily defined and verified but can be difficult to invert and solve. The problem is of significant importance as it is used in a variety of combinatorial and scheduling problems.

The general graph colouring problem deals with attaching labels (or "colours") to nodes in an undirected graph, such that (a) no two nodes connected by an edge may have the same label, and (b) the number of different labels that may be used is minimised. Our focus is mainly on the actual labelling of a graph.

A practical example of such a graph colouring is the classical four-colour map problem that gave birth to the whole field: a cartographer is to colour the regions of a planar map such that no two adjacent regions have the same colour. To present this problem as graph colouring, an undirected graph is constructed by representing each region in the map as a node and each boundary between two regions as an edge connecting those two corresponding nodes. Once the problem is represented in this form, a solution can be approached by any typical graph colouring algorithm. An example of this parametrisation can be seen in Figure 3.1 (a) and (b); we refer to (c) and (d) later on.

3.2.2 | PGMs to represent graph colouring problems

PGMs are used as a tool to reason about large-scale probabilistic systems in a computationally feasible manner. They are known for their powerful inference over problems with many interdependencies. It is often useful for problems that are difficult to approach algorithmically, with graph colouring being a specific example.

In essence, a PGM is a compact representation of a probabilistic space as the product of smaller, conditionally independent distributions called factors. Each factor defines a probabilistic relationship over the variables within its associated cluster – a cluster being a set of random variables. For discrete variables, this results in a discrete probability table over all possible outcomes of these variables. Instead of explicitly calculating the product of these factors (which typically is not computationally feasible), a PGM con-

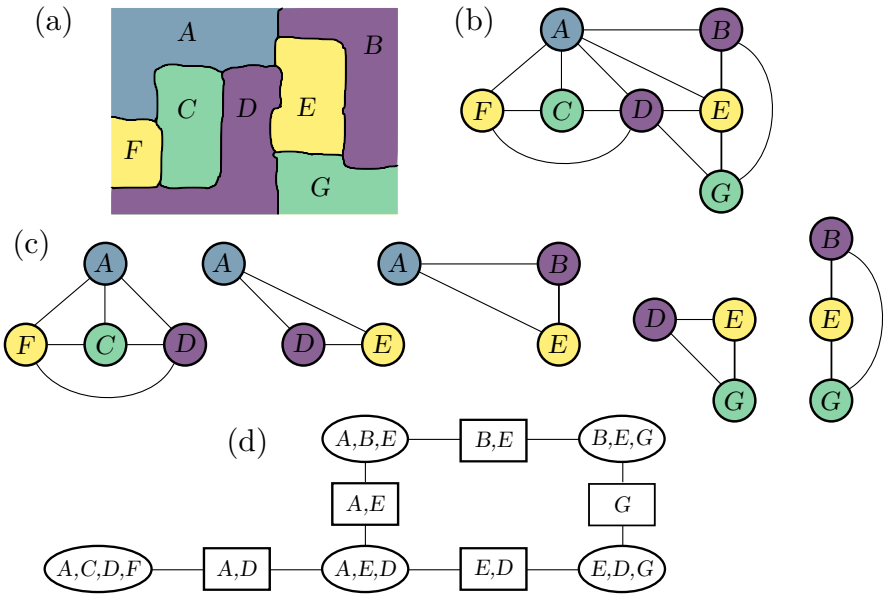


Figure 3.1: (a) A four-colour graph problem containing regions A to G, with (b) its graph colouring representation, (c) the maximal cliques within the graph, and (d) a cluster graph configuration for this problem. The ellipses represent the clusters and the boxes the sepsets – see the main text for more detail.

nects them into an appropriate graph structure. We apply inference by passing messages over the links in this structure until convergence is reached (see Section 2.9). In combination with the initial factor distributions, these converged messages can then be used to obtain the (approximate) posterior marginal distributions over subsets of variables.

To factorise a graph colouring problem, we first need to parametrise the problem probabilistically. This is achieved by allowing each node in the graph to be represented by a discrete random variable X_i, that can take on a number of states. For graph colouring, these states are the available labels for the node, e.g. four colours in the case of the four-colour map problem.

Now that we have the random variables of our system and the domains of these variables, we need to capture the relationship between them to represent it as factors in our PGM. For graph colouring, no two adjacent nodes may have the same colour; therefore, their associated random variables may not have the same state. One representation of this system would then be to capture this relationship using factors with a scope of two variables, each taken as an adjacent pair of nodes from the colouring graph. Although this fully represents the solution space, there is still a trade-off between accuracy and cluster size (cardinality) [59]. Fortunately, some configurations allow for larger clusters.

A clique is defined as a set of nodes that are all adjacent to each other within the

graph, and a maximal clique is defined as a clique that is not a subset of any other clique. In order to maximise factor scope, we prefer to define our factors directly on the maximal cliques of the graph. (We use the terms clique and cluster more or less interchangeably.) We can then set the discrete probability tables of these factors to only allow states where all the variables are assigned different labels. In the next section, we show an example of this.

After finalising the factors, we can complete the PGM by linking these factors in a graph structure. There are several valid structure variants to choose from – in this work, we specifically focus on factor graph and cluster graph structures. In the resulting graph structure, linked factors exchange information with each other about *some*, and not necessarily all, of the random variables they have in common. These variables are known as the separation set, or "sepset" for short, on the particular link of the graph. Whichever graph structure we choose must satisfy the so-called running intersection property (RIP) [12, p347]. This property stipulates that for all variables in the system, any occurrence of a particular variable in two distinct clusters should have a unique (i.e. precisely one) path linking them up via a sequence of sepsets that all contain that particular variable. Several examples of this are evident in Figure 3.1 (d). In particular, note the absence of the E variable on the sepset between the $\{B, E, G\}$ and $\{E, D, G\}$ clusters. If E were to be included, there would have been two distinct sepset paths containing E between those two clusters. This would be invalid, broadly, because it causes a type of positive feedback loop.

After establishing the factors and linking them in a graph structure, we can apply inference by using one of several belief propagation algorithms available (see Section 2.9).

3.2.3 | Example: The four-colour map problem

We illustrate the above by means of the four-colour map problem. The example in Figure 3.1 can be expressed by the seven random variables A to G, grouped into five maximal cliques as shown. There will be no clique with more than four variables (otherwise, four-colours would not be sufficient, resulting in a counter-example to the theorem). These maximal cliques are represented as factors with uniform distributions over their valid (i.e. non-conflicting) colourings. We do so by assigning either a possibility or an impossibility to each joint state over the factor's variables. More specifically, we represent the potential function $\phi(A, C, D, F)$ with a discrete table, assigning a "1" for outcomes where all variables have differing colours and a "0" for cases with duplicate colours.

For example, the factor belief $\phi(A, C, D, F)$ for the puzzle in Figure 3.1 is shown

in Table 3.1. These factors are connected into a graph structure – such as the cluster graph in Figure 3.1 (d). We can use belief propagation algorithms on this graph to find posterior beliefs.

Random variables →	A	C	D	F		
State →	1	2	3	4	1	$\leftarrow \phi(A, C, D, F)$
	1	2	4	3	1	
	1	3	2	4	1	
	1	3	4	2	1	
			⋮			
	4	3	2	1	1	
	elsewhere				0	

Table 3.1: A discrete table representing $\phi(A, C, D, F)$, where all possible combinations of outcomes for $\{A, C, D, F\}$ are captured.

We successfully tested this concept on various planar maps of size 100 up to 8000 regions. These were created by generating superpixels on arbitrary images using the SLIC algorithm [60] to serve as the initially uncoloured regions.

PGMs configured to utilise only binary probabilities always preserve all possible solutions [18]. The underlying reason is that a state considered as possible within a particular factor will always be retained unless a message from a neighbouring factor flags it as impossible. In that case, it is quite correct that it should be removed from the spectrum of possibilities. Dechter et al. [18] proved this by showing that applying belief propagation on this type of configuration results in a constraint satisfaction algorithm for generalised arc consistency.

However, in this four-colour map example, the space of solutions can, in principle, be prohibitively large. We force our PGM to instead find a particular unique solution, by firstly fixing the colours in the largest clique and, secondly, by very slightly biasing the other factor probabilities towards initial colour preferences. This makes it possible to pick a particular unique colouring as the most likely solution. An example of a graph of 250 regions can be seen in Figure 3.2.

3.3 | Factor vs cluster graph topologies

The graph structure of a PGM can make a big difference in the speed and accuracy of inference convergence. That said, factor graphs are the predominant structure in literature – surprisingly so, since we found them to be inferior to a properly structured cluster graph. Cluster graphs allow for passing multivariate messages between factors,

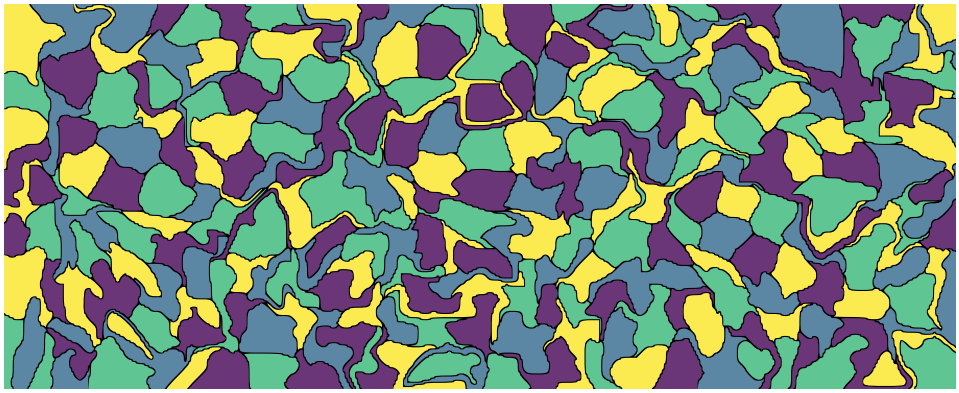


Figure 3.2: A generated planar map with its colouring results from a PGM labelling the 250 regions into one out of four colours.

thereby maintaining some of the inter-variable correlations already known to the factor. This is in contrast to factor graphs, where information is only passed through univariate messages, thereby implicitly destroying such correlations.

A search on scholar.google.com (conducted on June 28, 2017) for articles relating to the use of factor graphs versus cluster graphs in PGMs returned the following counts:

- 5590 results for *probabilistic graphical models "factor graph"*,
- 661 results for *probabilistic graphical models "cluster graph"*, and
- 49 results for *probabilistic graphical models "factor graph" "cluster graph"*.

Among the latter 49 publications (excluding four items authored at our university), no cluster graph constructions are found other than for Bethe / factor graphs, junction trees, and the clustering of Bayes networks. We speculate that this relative scarcity of cluster graphs points to the absence of an automatic and generic procedure for constructing good RIP satisfying cluster graphs.

3.3.1 | Factor graphs

A factor graph, built from clusters $\mathbf{C}_i$, can be expressed in cluster graph notation as a Bethe graph $\mathcal{F}$. For each available random variable X_j, $\mathcal{F}$ contains an additional cluster $\mathbf{C}_j = \{X_j\}$. Their associated factors are all uniform (or vacuous) distributions and, therefore, do not alter the original products' distributions. Each cluster containing X_j is linked to this vacuous cluster $\mathbf{C}_j$. This places $\mathbf{C}_j$ at the hub of a star-like topology, with

all the various X_j subsets radiating outwards from it. Due to this star-like topology, the RIP requirement is trivially satisfied.

The setup of a factor graph from this definition is straightforward, the structure is deterministic, and the placements of sepsets are well defined. Figure 3.3 provides the factor graph for the factors shown in Figure 3.1.

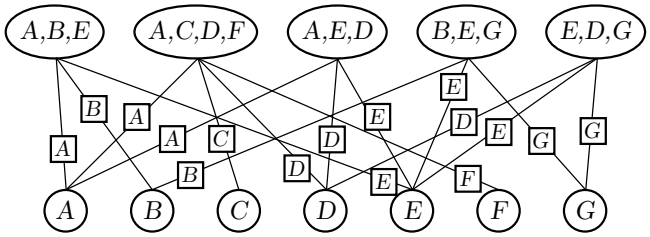


Figure 3.3: The Bethe factor graph topology applicable to Figure 3.1. Note the univariate sepsets are arranged in a star-like topology.

3.3.2 | Cluster graphs

A cluster graph $\mathcal{T}$, built from clusters $\mathbf{C}_i$, is a non-unique undirected graph, where

1. no cluster is a subset of another cluster, $\mathbf{C}_i \nsubseteq \mathbf{C}_j$ for all $i \neq j$,
2. the clusters are used as the nodes,
3. the nodes are connected by non-empty sepsets $\mathbf{S}_{i,j} \subseteq \mathbf{C}_i \cap \mathbf{C}_j$, and
4. the sepsets satisfy the running intersection property.

Point (1) is not strictly necessary (see the factor graph structure) but provides convenient computational savings. Moreover, it can always be realised by simply assimilating non-obliging clusters into a superset cluster via factor multiplication. Refer to Figure 3.1 (d) for an example of a typical cluster graph.

Although Koller [12, p404] provides extensive theory on cluster graphs, no general solution for its construction is provided. Indeed, they state that "the choice of cluster graph is generally far from obvious, and it can make a significant difference to the [belief propagation] algorithm." Furthermore, the need for such a construction algorithm is made clear from their experimental evidence, which indicates that faster convergence and an increase in accuracy can be obtained from better graph structuring. Therefore, since cluster graph theory is well established, an efficient and uncomplicated cluster graph construction algorithm will be useful. We provide the LTRIP algorithm for this purpose.

3.3.3 | Cluster graph construction via LTRIP

The LTRIP algorithm is designed to satisfy the running intersection property for a cluster graph $\mathcal{T}$ by layering the interconnections for each random variable separately into a tree structure and then superimposing these layers to create the combined sepsets. More precisely, for each random variable X_i available in $\mathcal{T}$, all the clusters containing X_i are interconnected into a tree structure – this is then the layer for X_i. After finalising all these layers, the sepset between cluster nodes $\mathbf{C}_i$ and $\mathbf{C}_j$ in $\mathcal{T}$ is the union of all the individual variable connections over all these layers.

While this procedure guarantees to satisfy the RIP requirement, there is still considerable freedom in exactly how the tree structure on each separate layer is connected. We were guided by the assumption that it is beneficial to prefer linking clusters with a high degree of mutual information. We, therefore, chose to create trees that maximise the size of the sepsets between clusters. The full algorithm is detailed in Algorithm 3. In addition, an example for constructing all the intermediate trees can be found in Figure 3.4, and an illustration of constructing a tree for a single variable in Figure 3.5. Furthermore, a reiteration and summary of the algorithm can be found in Section 5.2.3.

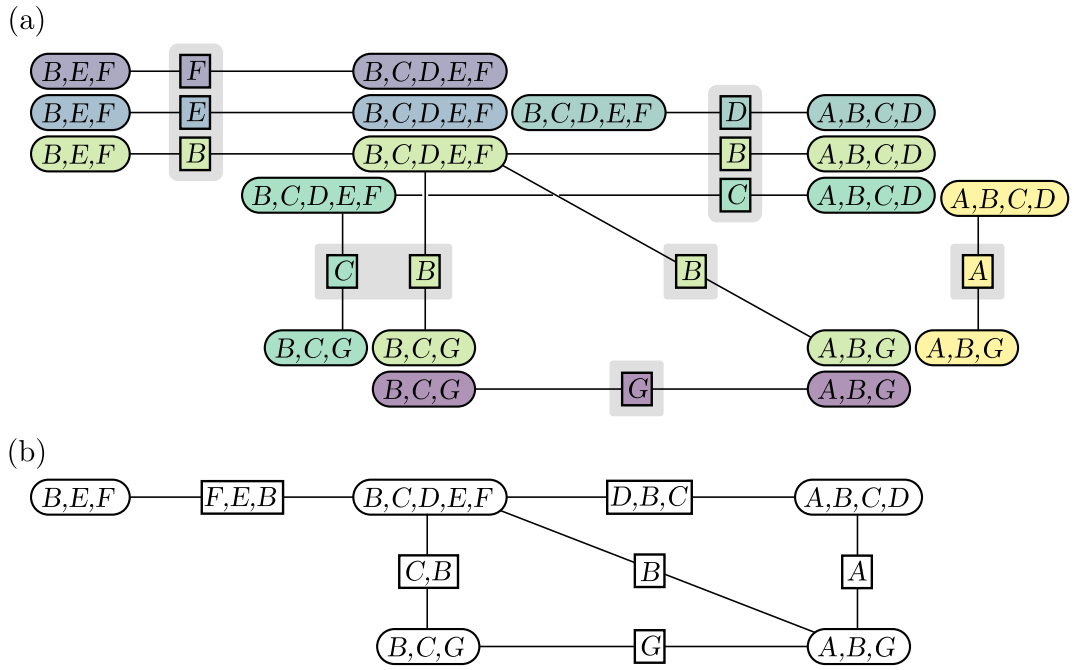


Figure 3.4: The intermediate tree structures resulting from the LTRIP procedure in Algorithm 3. The five clusters to be linked are $\{B, C, D, E, F\}$, $\{A, B, C, D\}$, $\{B, E, F\}$, $\{B, C, G\}$, and $\{A, B, G\}$. (a) The LTRIP procedure yielded seven intermediate tree structures, each a contribution from one of the seven variables A, B, C, D, E, F, and G. (b) The resulting cluster graph is simply a superposition of these tree structures.

Algorithm 3: LTRIP

Input: Set of clusters $\mathcal{V} = \{\mathbf{C}_1, \ldots, \mathbf{C}_N\}$, with subsets already assimilated into their supersets; i.e. $\mathbf{C}_i \nsubseteq \mathbf{C}_j \, \forall \, i \neq j$.

1: // Empty set of sepsets
2: $\mathcal{S} := \{\}$
3: **for** each random variable X found within $\mathcal{V}$ **do**
4: // This inner loop procedure is illustrated in Figure 3.5 (a)
5: $\mathcal{V}_X$:= set of clusters in $\mathcal{V}$ containing X
6: $\mathcal{W}_X := \text{Connection-Weights}(\mathcal{V}_X)$
7: // Add X to the appropriate sepsets
8: $\mathcal{P}_X$:= max spanning tree over $\mathcal{V}_X$ using weights $\mathcal{W}_X$
9: **for** each edge (i, j) in $\mathcal{P}_X$ **do**
10: **if** sepset $\mathbf{S}_{i,j}$ already exists in $\mathcal{S}$ **then**
11: $\mathbf{S}_{i,j}.\text{insert}(X)$
12: **else**
13: $\mathcal{S}.\text{insert}(\, \mathbf{S}_{i,j} = \{X\} \,)$
14: **end if**
15: **end for**
16: **end for**
17: $\mathcal{T}$:= cluster graph of $\mathcal{V}$ connected with sepsets $\mathcal{S}$

18: **function** Connection-Weights($\mathcal{V}_X$)
19: $\mathcal{W}_X := \{w_{i,j} = \left|\mathbf{C}_i \cap \mathbf{C}_j\right| \text{ for } \mathbf{C}_i, \mathbf{C}_j \in \mathcal{V}_X, \; i \neq j\}$
20: // Emphasise nodes strongly connected to multiple nodes
21: $m := \max(\mathcal{W}_X)$
22: **for** i **do**
23: // Number of maximal edges on this node
24: t_i := number of adjacent nodes j for which $w_{i,j} = m$
25: // Add to each edge touching this node
26: **for** j **do**
27: $w_{i,j}$ += t_i
28: **end for**
29: **end for**
30: **return** $\mathcal{W}_X$
31: **end function**

Line 8 For constructing a maximum spanning tree, see the Prim-Jarník algorithm [19].

Note that other (unexplored) alternatives are possible for the "Connection-Weights" function in the algorithm. In particular, it would be interesting to evaluate information-theoretic considerations as criteria.

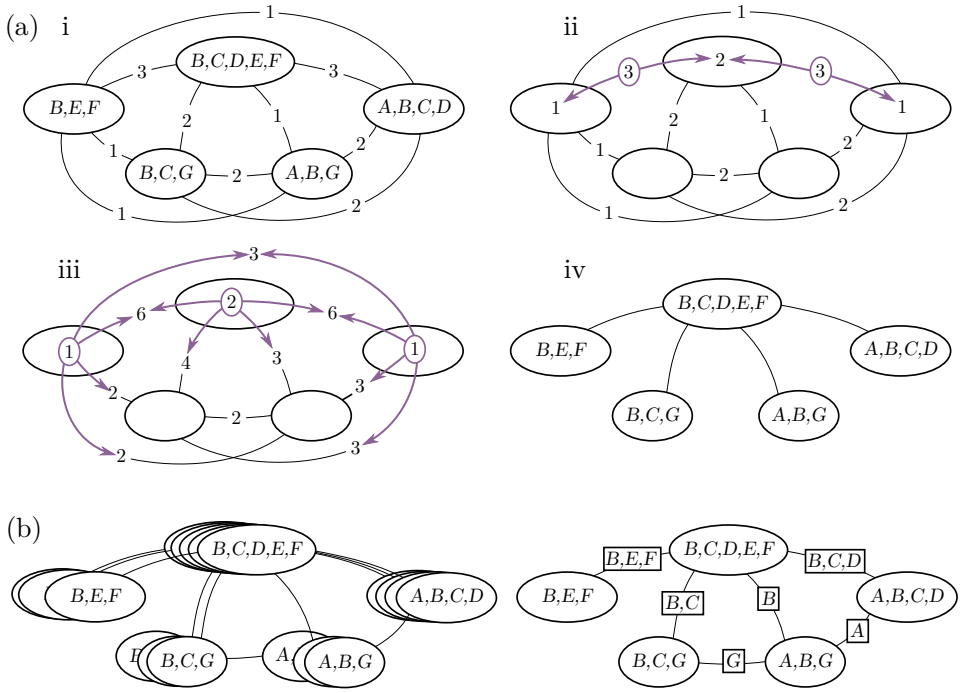


Figure 3.5: Illustration of constructing a cluster graph via the LTRIP procedure. The five clusters to be linked are $\{B, C, D, E, F\}$, $\{A, B, C, D\}$, $\{B, E, F\}$, $\{B, C, G\}$ and $\{A, B, G\}$.

Figure 3.5 (a): The procedure for joining up all clusters containing variable B into a tree. In sub-step (i), we set the initial connection weights as the number of variables shared by each cluster pair. In sub-step (ii), we identify the current maximal connection weight to be $m = 3$. In sub-step (iii), we note for each cluster how many of its links have maximal weight m. This number is added to all its connection weights. This emphasises clusters that are strongly connected to others. In sub-step (iv), we use these connection weights to form a maximal spanning tree, connecting all occurrences of variable B.

Figure 3.5 (b): Similarly constructed connection trees for all other variables are superimposed to yield the final cluster graph and sepsets.

3.4 | Modelling Sudoku via PGMs

The Sudoku puzzle is a well-known example of a graph colouring problem. A player is required to label a 9×9 grid using the integers "1" to "9", such that 27 selected regions have no repeated entries. These regions are the nine rows, nine columns, and nine non-overlapping 3×3 sub-grids of the puzzle. Each label is to appear exactly once in each region. Multiple solutions are possible if a Sudoku puzzle is under-constrained, i.e. too few of the values are known beforehand. A well-defined puzzle should have only one unique solution. We illustrate these constraints with a scaled-down 4×4 Sudoku (with 2×2 non-overlapping sub-grids) in Figure 3.6 (a).

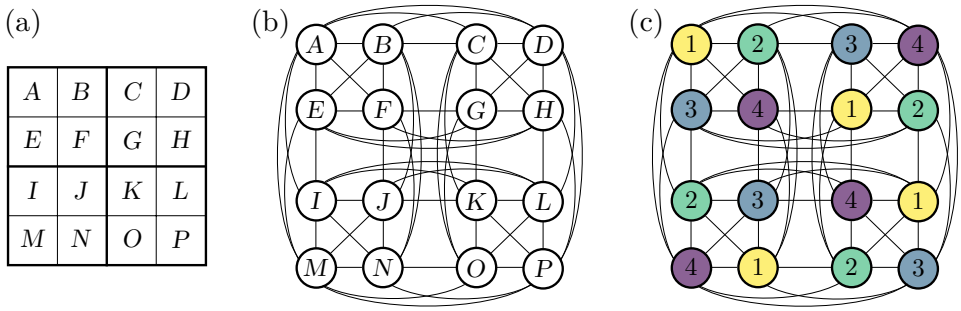


Figure 3.6: (a) An example of a 4×4 scaled Sudoku grid, with (b) its colouring graph, and (c) a non-unique colouring solution.

We use the Sudoku puzzle as a proxy for testing graph colouring via PGMs since this is a well-known puzzle with many freely available examples. However, it should be kept in mind that solving Sudoku puzzles per se is *not* a primary objective of this chapter, and in Chapter 5, we explore constraint satisfaction solvers in more detail. We now show how to construct a PGM for a Sudoku puzzle by following the same approach described for the four-colour map problem.

3.4.1 | Probabilistic representation

For the graph colouring and probabilistic representation of the Sudoku puzzle, each grid entry is taken as a node, and all nodes that are prohibited from sharing the same label are connected with edges, as seen in Figure 3.6 (b). It is apparent from the graph that each of the Sudoku's "no-repeat regions" is also a maximal clique within the colouring graph.

The probabilistic representation for the scaled-down 4×4 Sudoku is, therefore, 16 random variables A to P, each representing a cell within the puzzle. The factors of the system are set up according to the 12 cliques present in the colouring graph. Three examples of these factors, a row constraint, a column constraint, and a sub-grid constraint, are respectively $\{A, B, C, D\}$, $\{A, E, I, M\}$, and $\{A, B, E, F\}$. The entries for the discrete table of $\{A, B, C, D\}$ are precisely the same as those of Table 3.1. The proper 9×9-sized Sudoku puzzle used in our experiments is set up in exactly the same manner as the scaled-down version but using 27 cliques, each of size nine.

Also note, in the case of Sudoku puzzles, some variables are already assigned a value. To integrate this into the system, we formally "observe" those variables. There are various ways to deal with this, one of which is to purge all the discrete distribution states not in agreement with the observations. Following this, the variables can be purged from all factor scopes altogether.

3.4.2 | Graph structure for the PGM

We have shown how to parametrise the Sudoku puzzle as a colouring graph and how to parametrise this graph probabilistically. This allows capturing the relationships between the variables of the system via discrete probability distributions. The next step is to link the factors into a graph structure. We outlined factor graph construction in Section 3.3.1 and cluster graph construction via LTRIP in Section 3.3.3. Finally, we apply these two construction methods directly to the Sudoku clusters, thereby creating structures such as the cluster graph of Figure 3.7.

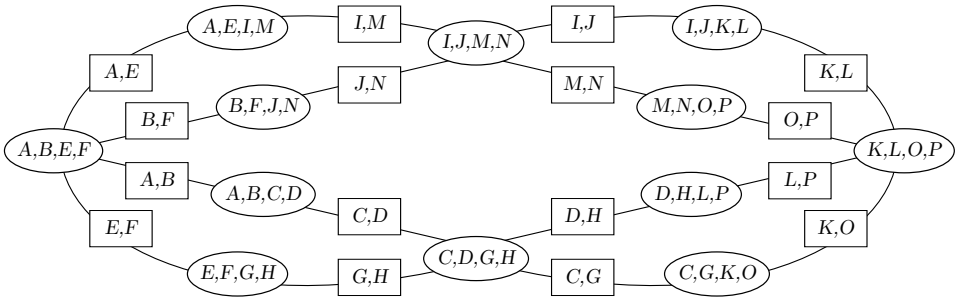


Figure 3.7: A cluster graph construction for the 4×4 Sudoku clusters.

3.4.3 | Message passing approach

For a full discussion on belief propagation, see Sections 2.8 and 2.9. The specific design choices for our PGM implementation are as follows:

- For the inference procedure, we used the belief *update* procedure, also known as the Lauritzen-Spiegelhalter algorithm [17].
- The convergence of the system and the message passing schedule are determined according to Kullback-Leibler divergence between the newest and immediately preceding sepset beliefs.
- Max-normalisation and max-marginalisation are used in order to find the maximum posterior solution over the system.
- All discrete distributions support sparse representations in order to make efficient use of memory and processing resources.

3.5 | Experimental investigation

As stated earlier, factor graphs are the dominant PGM graph structure encountered in the literature. However, this seems like a compromise since cluster graphs have traits that should enable superior performance. This section investigates the efficiency of cluster graphs compared to factor graphs by using Sudoku puzzles as test cases.

3.5.1 | Databases used

For our experiments, we constructed test examples from two sources, (a) 50 9×9 Sudoku puzzles ranging in difficulty taken from Project Euler [61], and (b) the "95 hardest Sudokus sorted by rating" taken from Sterten [62]. All these Sudoku problems are well-defined (with unique solutions), and their solutions are available for verification.

3.5.2 | Purpose of experiment

The goal of our experiments is to investigate both the accuracy and efficiency of cluster graphs compared to factor graphs. We hypothesise that properly connected cluster graphs, as constructed with the LTRIP algorithm, will perform better during loopy belief propagation than a factor graph constructed with the same factors.

Mateescu [59] shows that inference behaviour differs with factor complexities: a graph with large clusters is likely to be computationally more demanding than a graph with smaller clusters (when properly constructed from the same problem). However, the posterior distribution is likely to be more precise. We, therefore, want to also test the performance of cluster graphs compared to factor graphs over a range of cluster sizes.

3.5.3 | Design and configuration of the experiment

Our approach is to set up Sudoku tests with both factor graphs and cluster graphs using the same initial clusters. With regard to setting up the PGMs, we follow the construction methodology outlined in Section 3.4.

In order to generate graphs with smaller cluster sizes, we strike a balance between clusters of size two, using every adjacent pair of nodes within the colouring graph as described in Section 3.2.2 and using the maximal cliques within the graph, also described in that section. We do so by generating M-sized clusters from an N-sized clique (where $M \leq N$). We split the cliques by sampling all M-combination of variables from the N variable clique and keeping only a subset of the samples, such that every pair of adjacent nodes from the clique is represented at least once within one of the samples.

For experiments using the Project Euler database, we construct Sudoku PGMs with cluster sizes of three, five, seven, and nine variables in this manner. This results in graphs of 486, 189, 108, and 27 clusters, respectively. We compare the run-time efficiency and solution accuracy for both factor and cluster graphs constructed from the same set of clusters.

On the much harder Sterten database, PGMs based on clusters with less than nine variables were wildly inaccurate. We, therefore, limit those experiments to only clusters with nine variables.

3.5.4 | Results and interpretation

The results are reported in Figure 3.8. Cluster graphs showed superior accuracy for all the available test cases. Note that from our results, whenever a cluster graph failed to obtain a valid solution, the corresponding factor graph also failed. However, it happened regularly that a cluster graph succeeded where a factor graph failed, especially in more complex configurations.

In the case of small clusters, factor graphs are faster than cluster graphs. This is unexpected since cluster graphs built from small clusters are getting closer to factor graphs in terms of sepset sizes. We expected the execution speed to also get closer to each other in this case.

As the cluster sizes increase (especially when the problem domain becomes more difficult), the cluster graphs clearly outperform the factor graphs in terms of execution speed. Two explanations come to mind. Firstly, with the larger sepset sizes, the cluster graph needs to marginalise out fewer random variables when passing messages over that sepset. Since marginalisation is one of the expensive components in message passing, this should result in computational savings. Secondly, the larger sepset sizes allow factors to pass richer information to their neighbours. This speeds up the convergence rate, once again resulting in computational savings.

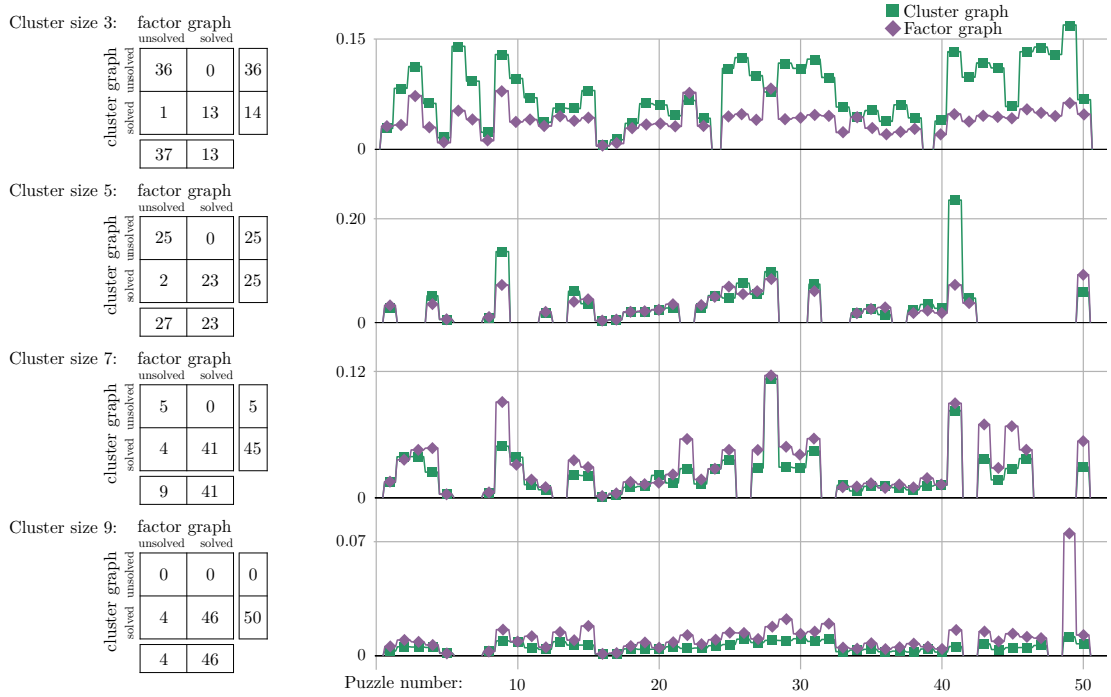
3.6 | Future work

The LTRIP algorithm is shown to produce well-constructed graphs. However, the criteria for building the maximal spanning trees in each layer can probably benefit from further refinement. In particular. we suspect that taking the mutual information between factors into account might prove useful.

Our graph colouring parametrisation managed to solve certain Sudoku puzzles successfully, as well as assigning colours to the four-colour map problem. This is a good

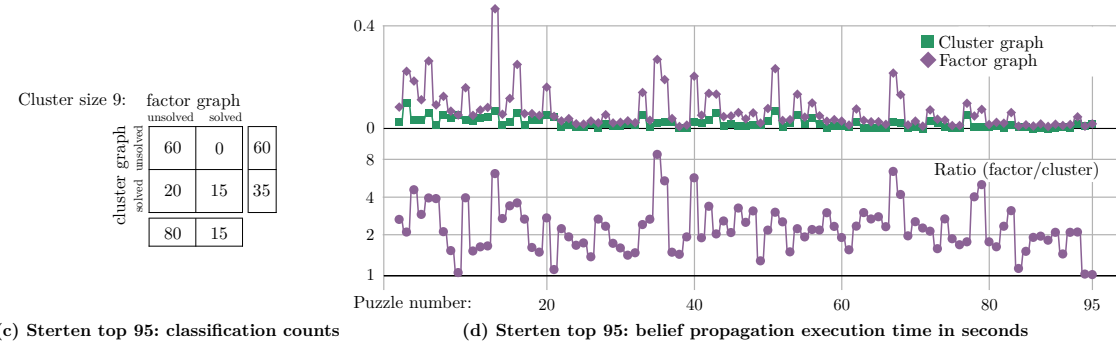
starting point for developing more advanced techniques for solving graph colouring problems.

In this chapter, we evaluated our cluster graph approach on a limited set of problems. We hope that the LTRIP algorithm will enhance the popularity of these problems as well as other related problems. This should provide evaluations from a richer set of conditions, contributing to a better understanding of the merits of this approach.



(a) Project Euler: classification counts

(b) Project Euler: belief propagation execution time in seconds



(c) Sterten top 95: classification counts

(d) Sterten top 95: belief propagation execution time in seconds

Figure 3.8: The results of our test cases. Note that whenever a cluster graph failed to obtain a valid solution, the corresponding factor graph also failed. Also, note that (b) includes only points where the factor graph and cluster graph posteriors are equivalent, whereas (d) includes all results (otherwise the plot would mostly be empty).

3.7 | Conclusion

The objective of this study was to illustrate how graph colouring problems can be formulated with PGMs, to provide a means for constructing proper cluster graphs, and to compare the performance of these graphs to the ones prevalent in the current literature.

The main contribution of this chapter is certainly LTRIP, our proposed cluster graph constructing algorithm. The cluster graphs produced by LTRIP show great promise compared to the standard factor graph approach, as demonstrated by our experimental results.

4

A cluster graph approach to land cover classification boosting

Preface

In the previous chapter, *Graph Colouring: Comparing Cluster Graphs to Factor Graphs*, we illustrated how to formulate graph colouring problems as PGMs. We also presented this formulation via examples such as Sudoku puzzles and the four-colour map problem (where a planar map's neighbouring regions are coloured differently using only four colours). Furthermore, we provide a general-purpose cluster graph construction algorithm named LTRIP. Our experimental results show that the cluster graphs produced by LTRIP have superior characteristics compared to factor graphs, both in terms of speed and accuracy. The publication presented here subsequently illustrates how these contributions are used to solve a region-based classification problem in cartography, similar to the four-colour map problem.

In this chapter, we present *A Cluster Graph Approach to Land Cover Classification Boosting* [2][†], a collaborative effort between the Electronic Engineering department of Stellenbosch University and the Department of Civil, Geo and Environmental Engineering of the Technical University of Munich. Author S. Streicher's contribution to this publication is the formulation, design, and execution of the PGMs used in this work; specifically (but not exclusively) Section 4.3.1, Section 4.3.2, and the design of the potential functions in Section 4.3.4.

This publication sets out to solve a practical problem in cartography and classification, where multiple (and contradictory) classifications for pieces of land are combined

[†]Sections of this work have been published in:
L. H. Hughes, S. Streicher, E. Chuprikova, and J. du Preez. "A Cluster Graph Approach to Land Cover Classification Boosting." *Data*, 4(1), 2019. ISSN 2306-5729.

and boosted to form a unified, more accurate classification.

The main contribution of this publication is in the form of a concrete application of the tools developed in Chapters 2 and 3; notably, approaching a region-based problem statement as a probabilistic reasoning problem and configuring the problem into a cluster graph using LTRIP. This approach produced a feasible, diverse, and spatially-consistent boosted land cover classification. In addition, a validation study reported an overall accuracy improvement compared to a reference dataset.

In conclusion, this paper serves as a practical demonstration of the power of cluster graphs. It establishes a substantial application of LTRIP for use in fields outside of Computer Science and Engineering. A solution to the land cover classification problem is made highly accessible through the use of LTRIP and the general PGM approach described in previous chapters.

Abstract

Land cover classification is complex due to algorithmic errors, the spatio-temporal heterogeneity of Earth observation data, variation in data availability, and reference data quality. This article proposes a probabilistic graphical model approach in the form of a cluster graph, to boost geospatial classifications and produce a more accurate and robust classification and uncertainty product. Cluster graphs can be used for reasoning about geospatial data such as land cover classifications by considering the effects of spatial distribution and inter-class dependencies in a computationally efficient manner. To assess the capabilities of our proposed cluster graph boosting approach, we apply it to the field of land cover classification. We make use of existing land cover products, GlobeLand30 and CORINE Land Cover, along with data from volunteered geographic information (VGI), namely OpenStreetMap (OSM), to generate a boosted land cover classification and the respective uncertainty estimates. Our approach combines qualitative and quantitative components by applying our probabilistic graphical model to data inputs as well as subjective expert judgments. Evaluating our approach on a test region in Garmisch-Partenkirchen, Germany, our approach boosted the overall land cover classification accuracy by 1.4% compared to an independent reference land cover dataset. Our approach was shown to be robust and was able to produce a diverse, feasible, and spatially-consistent land cover classification in areas of incomplete and conflicting evidence. On an independent validation scene, we demonstrated that our cluster graph boosting approach was generalisable even when initialised with poor prior assumptions.

4.1 | Introduction

The rapid development of remote sensing techniques and data processing algorithms has led to the availability of large amounts of Earth observation data at regular intervals. However, the usefulness of this data is limited to a small community of experts. Therefore, practical data analysis methods and applications based on remote sensing data have become an active area of research, specifically thematic mapping and land cover classification [63, 64, 65]. Land cover has been recognised as a key variable in environmental studies for deforestation, climate assessment, food and water security, and urban growth [66, 67].

Due to its high relevance, the demand for accurate and timely production of land cover classification has grown rapidly as governments push towards large-scale, and frequent monitoring of agricultural and urban environments [68]. While a magnitude of approaches exists for creating land cover classifications, the quality, resolution, and reclassification periods for these vary drastically [69, 70, 71, 72]. Additionally, many of these approaches make use of supervised learning where accurately labelled training data is required. This data is often manually annotated and thus suffers from human biases and varying accuracy. Furthermore, as the quality and quantity of this data directly affect the accuracy of the final classifier, the usability of these approaches is often limited to regions that depict similar features to the training dataset.

Overall, land cover classification is an inherently nuanced problem and has a large scope for error. These errors can largely be attributed to automated classification algorithm errors, temporal and spatial heterogeneity of Earth observation data, variation in the availability and quality of reference data, and the need for human intervention in labelling. Therefore, trade-offs are required to produce *good enough* land cover classifications. Thus, it becomes clear that no single land cover classification product can be produced to cover all use cases with a sufficient spatial resolution, coverage extent, accuracy, and granularity. Due to these reasons, there has been a growing interest in fusing different land cover classifications, including crowd-sourcing information, into a single more complete data product [73, 74, 75, 76]. These methods are based on various data boosting approaches that play off each dataset's strengths to gain a more complete classification and minimise errors. While these approaches have seen various successes, spatially-weighted land cover prediction and the assessment of its related uncertainties have received limited attention.

In this study, we address these challenges with four objectives:

1. to explore the potential of a probabilistic graphical model approach, using cluster

graphs, towards producing a more accurate land cover classification,

2. to exploit the potential of expert knowledge for probabilistic reasoning in land cover classification,

3. to perform uncertainty analysis on the outcome using the Shannon diversity index as a measure of uncertainty, and

4. to potentially contribute to OpenStreetMap data with missing land cover information.

To this end, we propose an efficient approach using cluster graphs for boosting spatial heterogeneous data. We then investigate the effect of priors (in the form of expert knowledge) on the accuracy of the proposed inference process. Finally, we analyse the accuracy and uncertainty of the boosted classification. We demonstrate the applicability of our method, using three major datasets, namely GlobeLand30 [77], Coordination of Information on the Environment (CORINE) [78], and volunteered geographic information (VGI) based on OpenStreetMap (OSM) [79].

The study proceeds as follows: First, in Section 4.2, we review the relevant literature on land cover boosting algorithms and accuracy assessment methods. Section 4.3 describes our proposed approach and architectural design. In Section 4.4, we report on results of the cluster graph approach for land cover boosting applied to data of Garmisch-Partenkirchen (Bavaria, Germany). Finally, we conclude with an overview of our findings.

4.2 | Related work

Over the years, many approaches have been proposed for land cover classification, of which the majority of these rely on satellite imagery and remote sensing techniques. Many of these approaches are based on simple linear methods and use hand-crafted features and simple classification schemes to create land cover maps. The most common of which is the maximum likelihood (ML) classifier. This approach has maintained steady popularity due to its availability and easy to use nature. However, as Pal and Mather [80] discuss, this approach does not provide as high-quality results as decision tree approaches, such as Gislason et al. [70].

Gislason et al. [70] proposed using a random-forest (RF) classifier in the form of a classification and regression tree (CART) to extract land cover classifications from multi-source remote sensing data. While CART remains a commonly applied method for land

cover classification (due to its low computational complexity and high interpretability), the produced classification tree does not generalise to vastly different environments. Based on this shortfall of RF approaches, Pal and Mather [71] proposed using support vector machines (SVMs) for remote sensing data classification. While they found that SVMs provided significantly better results, they also noted the sensitivity of the approach to datasets, parameter selection and class separability. These findings were reiterated in various studies into existing linear land cover classification approaches [81, 82].

More recently, there has been a departure from linear approaches and into the domain of non-linear classifiers, with a specific focus on deep learning techniques. Deep learning has gained considerable popularity due to the success of convolutional networks in image classification tasks. Most notably, the success of AlexNet in achieving state-of-the-art performance on the ImageNet classification task [83]. Based on the success of deep learning in conventional computer vision applications, remote sensing practitioners have turned their focus to exploiting these advancements for land cover classification [84, 85, 86]. Notably, Castelluccio et al. [86] proposed using convolutional neural networks (CNNs) for land cover classification. Following this CNN approach, Marmanis et al. [84] used existing CNNs, which were pretrained using ImageNet data, and applied these networks to remote sensing classification tasks. Following a different approach, Rußwurm and Körner [72] proposed using temporal data and a long short-term memory (LSTM) model for learning land cover classifications from underlying phenological features.

While existing techniques have all shown success in producing land cover classifications, they still suffer from a wide array of drawbacks that limit their overall usability and usefulness in large-scale and generalised land cover mapping applications [69]. For this reason, several studies have adopted data boosting approaches that combine the relative strengths of various classifiers to produce an improved land cover classification.

Chen et al. [76] suggested an approach to create a high-quality land cover classification using data fusion based on data from the Landsat-8 Operational Land Imager (OLI), Moderate Resolution Imaging Spectroradiometer (MODIS), China Environment 1A series (HJ-1A), and Advanced Space-borne Thermal Emission and Reflection (ASTER) digital elevation model (DEM). While this approach yields more accurate land cover classification results, it relies on the fusion of raw data and, thus, does not exploit well known and trusted land cover classification schemes or existing archives of land cover data.

Pérez-Hoyos et al. [87] approached this problem using existing land cover datasets,

such as CLC2006, CLC2000, MODIS, and GlobeCover, to create a synergistic land cover map of Europe. The thematic overlap was performed based on knowledge of the data quality and the formation of a common set of classes. Affinity scores between various classes in different datasets were then calculated, and a hybrid land cover classification was produced.

Following on from this approach, El-Deen Taha [88] proposed the use of a classifier ensemble to improve the accuracy of land cover classification approaches. RapidEye remote sensing imagery was classified using numerous well-known land cover classification approaches, such as SVM, CART, and artificial neural networks. The resulting classification maps were then combined using a bagging and boosting approach to generate a single boosted classification map.

The approaches towards boosting taken by Pérez-Hoyos et al. [87] and El-Deen Taha [88] resulted in boosted land cover classifications, showing a small overall accuracy improvement but consistent class-wise classification accuracy improvement over the initial classifications. However, neither approach inherently exploits expert knowledge or VGI data to support the boosting process. These two sources can provide critical prior information to resolve ambiguities and sensitivities when combining existing land cover classification datasets. Furthermore, the analysis of the uncertainties present in the final land cover classifications has largely been overlooked, despite its importance in understanding and utilising land cover classifications at a large scale.

Various measures have been proposed to describe data uncertainty quantitatively. These include scalar values like probability, error percentage, distance (e.g. from ground truth), standard deviation [89], and the Shannon diversity index, a quantitative estimator of complexity [90]. These measures have been used within the framework of various probabilistic approaches, such as Bayes networks, belief functions, interval sets, and fuzzy set theory [91]. However, few of these approaches have been applied to land cover classification boosting and, thus, this field remains predominantly unexplored. For this reason, we take inspiration from literature based within the realm of probabilistic graphical models (namely cluster graphs), graph colouring, and image processing. Our proposed approach follows the same general assumptions as problems where neighbouring relationships play a large role, such as the four-colour map problem (see Section 3.2.3), image segmentation [92], and image de-noising [93, 94].

4.3 | Land cover classification boosting with PGMs

This section introduces the concept of a cluster graph approach and describes how to formulate our land cover classification boosting problem as a graphical model. Furthermore, we describe how to apply inference on this model to perform classification boosting for our land cover problem.

4.3.1 | Cluster graphs

Probabilistic graphical models (PGMs) are a resourceful combination of graph theory and probabilistic inference techniques. A cluster graph is a type of PGM known for performing inference over problem spaces with many inter-dependencies and complex relationships. It can, therefore, be used to approach problems that are otherwise too difficult to define and solve algorithmically. In a general sense, cluster graphs can be seen as a method whereby a large system is broken down and clustered into smaller sections, such that these sections can be connected in a graph structure. The graph structure allows these smaller systems to communicate about their combined outcome and, thus, perform inference. Therefore, cluster graphs can be described as a tool to reason about large-scale probabilistic systems in a computationally efficient manner.

Cluster graphs provide a compact representation of a probabilistic space as the product of conditionally independent distributions referred to as factors. Each factor defines a probabilistic relationship over its associated cluster of variables. For discrete classification problems, such as our land cover classification problem, these factors will have potential functions related to prior beliefs or variable dependencies. In practice, these factors are built-up from any available knowledge and assumptions (i.e. educated guesses, or expert knowledge) about the variables and the relationships within the model.

As a means to inference, explicitly calculating the product of these factors is useful but typically not computationally feasible. A cluster graph rather connects factors into a graph structure with the factors as nodes and connections holding a set of variables, called a sepset. Information may be passed between factors through connecting sepsets in one of the many PGM inference techniques. Typically the factors of a cluster graph are initialised using the prior beliefs of the system. These beliefs are then updated by passing information about neighbouring sepset variables and factors until all the beliefs reach convergence. This produces approximations of the marginal distributions of the system and, therefore, a solution to the problem at hand.

A detailed discussion of cluster graph topologies is found in Section 3.3. The details of message passing in graphical models and determining convergence are discussed in

Sections 2.8 and 2.9.

4.3.2 | Proposed approach

Given the situation where multiple, independent, non-agreeing classifications exist, we can combine these classifications to obtain a new, more accurate classification in an approach referred to as classification boosting.

Many approaches of classification boosting exist and have varying degrees of success dependant on the problem. The most notable are naïve boosting, where the mode of the different classifications is selected as the output class, and ensemble methods, where an optimal combination of existing classifiers is learned to produce a new classification [95]. To this end, we propose the use of cluster graphs to solve the classification boosting problem.

Cluster graphs have numerous benefits over existing approaches since they make defining variables and relationships easy and exhibit powerful inference abilities. Furthermore, unlike the naïve and ensemble methods of classification boosting, cluster graphs can extrapolate expert knowledge to reclassify regions that were probabilistically unlikely in the original classification. Therefore, it can be argued that cluster graphs are positioned somewhere between learning a new classifier and pure classification boosting.

To model the land cover classification problem, we made the following assumptions about the nature of the land cover classification data we are using.

- The classifications are noisy observations that correlate to the true underlying class.
- The underlying land cover map can be sufficiently divided into small squares, similar to pixels, with the observations sub-sampled to correspond to these locations.
- Adapting Tobler's first law of geography [96], locations in close proximity have a higher likelihood to be of the same class than locations further away.

We simplified these assumptions into the following relationship model. First, we split our underlying map into $N \times M$ pixels, and assign a random variable $X_{i,j}$ to each pixel to represent the underlying class. Then, for data taken from K different approaches (in our case, K different land cover classification maps), we assign the variable $Y^k_{i,j}$ as an observation correlating to $X_{i,j}$ by sampling from the classification k at pixel (i, j). We also assign a relationship between each $Y^k_{i,j}$ and its associated state $X_{i,j}$. Finally, we

assign a relationship between all neighbouring pixels of $X_{i,j}$ to enforce Tobler's first law of geography. These variables and relationships are illustrated in Figure 4.1(a).

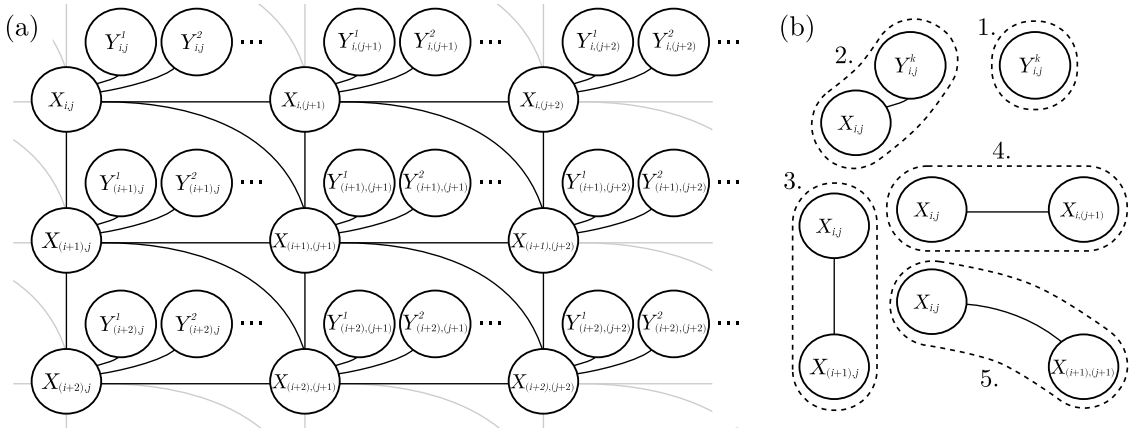


Figure 4.1: Graph (a) represents our relationship model for the land cover problem with nodes $X_{i,j}$ as the underlying classes and nodes $Y^k_{i,j}$ as observations taken from various classification approaches. The groupings in (b) represent our choice of factors as described in Table 4.1.

For a PGM formulation of this model, we choose the factors in a manner that would capture the relationships of the model and can easily be initialised from the land cover data. This choice is outlined in Figure 4.1(b) and Table 4.1, along with a description of the purpose of each factor in our model.

	Num. of factors	Factor variables	Purpose
1.	NMK	$\{Y^k_{i,j}\}$	Capture the different classifications for each variable
2.	NMK	$\{Y^k_{i,j}, X_{i,j}\}$	Relationship between observations and underlying classes
3.	$(N-1)M$	$\{X_{i,j}, X_{(i+1),j}\}$	Relationship between southward neighbours
4.	$N(M-1)$	$\{X_{i,j}, X_{i,(j+1)}\}$	Relationship between eastward neighbours
5.	$(N-1)(M-1)$	$\{X_{i,j}, X_{(i+1),(j+1)}\}$	Relationship between south-eastward neighbours

Table 4.1: Factor setup for the land cover PGM

Our factors are then initialised according to the land cover classifications and assumptions in the form of expert knowledge. The idea is simple: we assign a potential function to each factor according to an underlying relationship. More specifically, we set the observation variables $Y^k_{i,j}$ according to the land cover classification k: for hard classi-

fications, we have $P(Y^k_{i,j}{=}\text{classification}^k_{i,j}){=}1$ and $P(Y^k_{i,j}{\neq}\text{classification}^k_{i,j}){=}0$, and for soft classifications, we make use of joint probability tables initialised by expert knowledge about the likelihood of class co-occurrence in the defined region, and the class-specific classification quality, see Table 4.2. Furthermore, the variables in the relational factors are defined as more likely to have the same outcome, such as Table 4.1 factors 2 to 5.

	$X_{i,j} = a$	$X_{i,j} = b$	$X_{i,j} = c$
$X_{i,(j+1)} = a$	$p(a,a)$	$p(a,b)$	$p(a,c)$
$X_{i,(j+1)} = b$	$p(b,a)$	$p(b,b)$	$p(b,c)$
$X_{i,(j+1)} = c$	$p(c,a)$	$p(c,b)$	$p(c,c)$

Table 4.2: Prior probabilities table of $P(X_{i,j}, X_{i,(j+1)})$ for three classes a, b, and c, capturing expert knowledge. This is easily expandable to N classes, and the same table is usually applied to define relationship priors between eastward, southward and south-eastward neighbours

Using our defined and initialised factors, we (a) construct a cluster graph, (b) obtain a posterior distribution over this graph using PGM inference techniques and (c) extract the random variables $X_{i,j}$ from the posterior as the most likely land cover classifications. Thus completing the boosting process by creating the boosted land cover classification using the $X_{i,j}$ variables.

A summarised overview of the construction, inference, and settings used in this chapter are presented below.

1. Configuring a cluster graph is not a trivial task and requires some heuristics. For the construction of our cluster graphs, we use the LTRIP procedure described in Section 3.3.3 and further summarised in Section 5.2.3.

2. We used belief update, also known as the Lauritzen-Spiegelhalter algorithm [17], to perform inference over the graph.

3. The convergence of the system, as well as the scheduling of messages, is determined according to the Kullbach-Leibler divergence between the newest and immediately preceding sepset beliefs.

4. The distribution over a variable $X_{i,j}$ is found by locating a factor containing the variable and marginalising to that variable, i.e. $P(X_{i,j}) = \sum_{Y_{i,j}} P(X_{i,j}, Y_{i,j})$.

To better understand the proposed cluster graph architecture and inference process, see the reduced example presented in Figure 4.2. For illustrative purposes we only consider the spatial factors (lines 3-5 of Table 4.1), reduce the location grid to 4×4,

and use a constant probability to define all inter-class relationships. In practice, the prior probabilities are to be specified by expert knowledge, and all factors need to be included to boost the original land-cover datasets.

(a)

$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$
$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$
$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	$X_{4,4}$

(b)

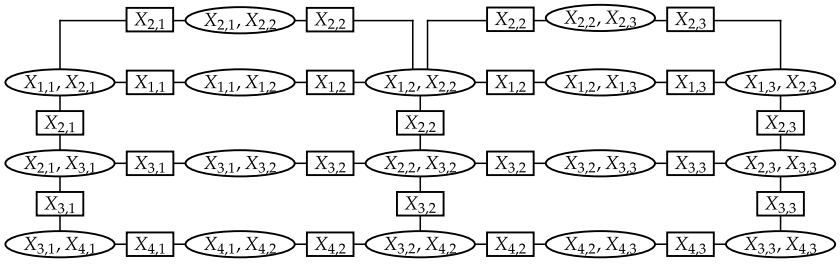
Factors: eastward neighbours

$\{X_{1,1}, X_{1,2}\}$	$\{X_{1,2}, X_{1,3}\}$	$\{X_{1,3}, X_{1,4}\}$
$\{X_{2,1}, X_{2,2}\}$	$\{X_{2,2}, X_{2,3}\}$	$\{X_{2,3}, X_{2,4}\}$
$\{X_{3,1}, X_{3,2}\}$	$\{X_{3,2}, X_{3,3}\}$	$\{X_{3,3}, X_{3,4}\}$
$\{X_{4,1}, X_{4,2}\}$	$\{X_{4,2}, X_{4,3}\}$	$\{X_{4,3}, X_{4,4}\}$

Factors: southward neighbours

$\{X_{1,1}, X_{2,1}\}$	$\{X_{1,2}, X_{2,2}\}$	$\{X_{1,3}, X_{2,3}\}$	$\{X_{1,4}, X_{2,4}\}$
$\{X_{2,1}, X_{3,1}\}$	$\{X_{2,2}, X_{3,2}\}$	$\{X_{2,3}, X_{3,3}\}$	$\{X_{2,4}, X_{3,4}\}$
$\{X_{3,1}, X_{4,1}\}$	$\{X_{3,2}, X_{4,2}\}$	$\{X_{3,3}, X_{4,3}\}$	$\{X_{3,4}, X_{4,4}\}$

(c)



(d)

	$X_{3,1}$	$X_{4,1}$	$P(X_{3,1},\ X_{4,1})^\dagger$
same	Cultivated	Cultivated	1
	Forest	Forest	1
	Grassland	Grassland	1
	⋮		⋮
diff.	Cultivated	Forest	0.8
	Cultivated	Grassland	0.8
	⋮		⋮

$\xrightarrow{X_{4,1}}$ $\xleftarrow{X_{4,1}}$

	$X_{4,1}$	$X_{4,2}$	$P(X_{4,1},\ X_{4,2})^\dagger$
same	Cultivated	Cultivated	1
	Forest	Forest	1
	Grassland	Grassland	1
	⋮		⋮
diff.	Cultivated	Forest	0.8
	Cultivated	Grassland	0.8
	⋮		⋮

† tables reflect non-normalised values and should be rescaled in order to sum to one

Figure 4.2: A simplified example of expressing our land cover classification problem as a cluster graph. In (a) we show a 4×4 location grid, in (b) we introduce factors describing neighbouring relationships to enforce Tobler's first law constraints, in (c) we show a cluster graph construction from these factors using LTRIP, and in (d) we highlight the use of discrete tables in message passing and the common variables through which information is passed. The joint probabilities defined here are merely for illustrative purposes, and should rather be specified by expert knowledge.

As a final note, since the number of factors grows by order NM, it is useful to split the problem into smaller sections that can be processed in parallel. We found it safe to assume that regions sufficiently far apart have near-zero influence on each other. Thus, we segmented the region into non-overlapping sub-regions with overlapping bound-

aries to enforce smoothness along the edges. We then ran the cluster graph process on each of these sub-regions in parallel. Finally, we stitched the posteriors together along the subregion boundaries while discarding the overlapping regions (which may contain conflicting results). For a more intuitive understanding of this process, please refer to Figure 4.3.

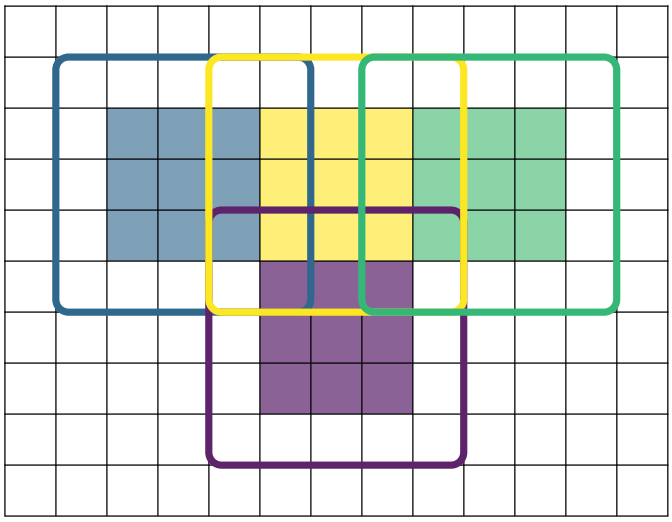


Figure 4.3: A simplified illustration of how a large region can be sub-divided and processed in a parallel manner while reducing the chance of artefacts along stitching boundaries.

4.3.3 | Datasets

We selected two commonly used land cover datasets to assess our proposed approach, namely GlobeLand30 [77] and CORINE Land Cover (CLC2006) for Germany [78]. Additionally, we use land cover data derived from volunteered geographic information (VGI), specifically OpenStreetMap (OSM) [79]. The test region of Garmisch-Partenkirchen, Germany (101 224 km^2) was selected due to the availability of high-quality datasets and sufficient diversity in the distribution of land classes. The land cover classifications are temporally independent of one another, each derived from imagery captured over different periods of time. Since large temporal change events can lead to inconsistencies, this can be a problem for boosting applications. However, the assumption was made that little temporal change has occurred over the dataset acquisition period due to the specific nature of the test location. Thus, despite the temporal data heterogeneity, we can still evaluate our cluster graph approach for boosting land cover classifications.

GlobeLand30 is a global-scale land cover product of 30m resolution for two baseline years (2000 and 2010). For this study, we make use of the 2010 version of the dataset. This dataset comprises 10 major land cover classes: cultivated areas, forests, grassland, shrubland, wetland, water bodies, tundra, artificial surfaces, bare land, and permanent snow and ice. However, only nine of these classes are present in our study region, as depicted in Figure 4.4, with examples of the visual appearance of each class.

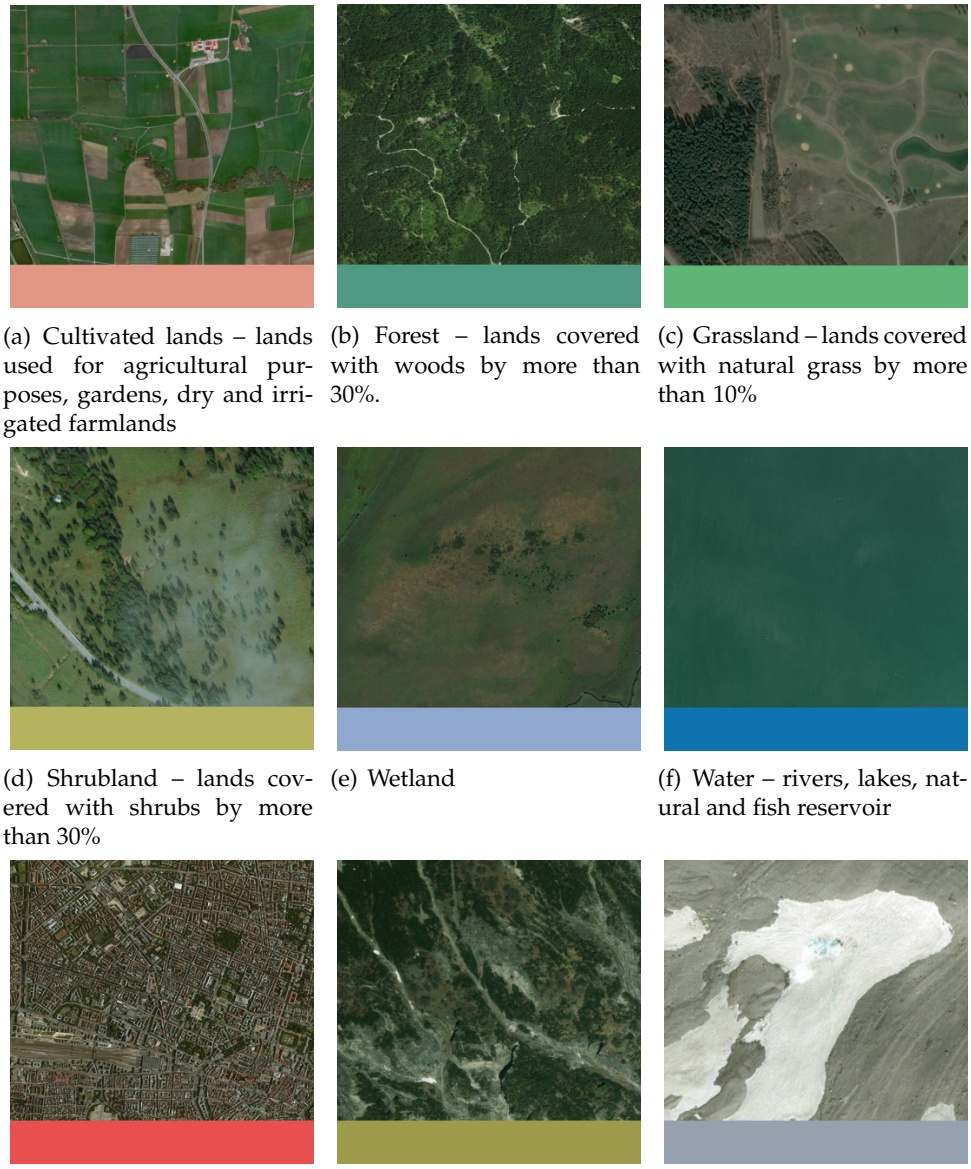
Raw remote sensing imagery was selected to coincide with the local vegetation growth season to reduce the effects of cloud cover on the creation of the GlobeLand30 dataset [64]. Thus, the land cover classification was created based on a mosaic of suitable images with minimal cloud occlusions. According to the data provider, the land cover classifications of our study area were generated from a mosaic of images acquired on 31st August 2009. Previous studies indicate that the overall classification accuracy of GlobeLand30 can range from 46% [97] and up to 80% [77, 64, 98]. Therefore, the true accuracy of the dataset is heterogeneous.

Land cover mapping that focuses on EU countries is available through the Coordination of Information on the Environment (CORINE) program. The CORINE Land Cover 2006 dataset (CLC2006) for Germany follows common European-wide CORINE nomenclature that consists of 44 classes, where 37 classes are relevant to Germany, and 29 classes are relevant to the study area. We scaled down the class complexity to provide a consistent classification for all the data sources. This was achieved by reassigning the 44 CORINE land cover classes to the 10 classes corresponding to the GlobeLand30 classification. The details of the reclassification processes can be seen in Table 4.3.

Table 4.3: Reclassification of CLC2006 land cover classes relevant for the study area based on the GlobeLand30 scheme.

Land cover classes	CLC2006, Pixel values	GlobeLand30, Pixel values
Cultivated	32 - 41	10
Forest	42 - 45	20
Grassland	46 - 47	30
Shrubland	48 – 49	40
Wetland	55 - 59	50
Water bodies	60 - 64	60
Tundra	-	70
Artificial surfaces	21 – 31	80
Bareland	50 – 53	90
Permanent ice and snow	54	100

Along with GlobeLand30 and CORINE (CLC2006), data from OSM plays an important role in this study as an auxiliary source. OSM is one of the most widespread and



(a) Cultivated lands – lands used for agricultural purposes, gardens, dry and irrigated farmlands

(b) Forest – lands covered with woods by more than 30%.

(c) Grassland – lands covered with natural grass by more than 10%

(d) Shrubland – lands covered with shrubs by more than 30%

(e) Wetland

(f) Water – rivers, lakes, natural and fish reservoir

(g) Artificial surface – lands modified by human activities

(h) Bareland – lands with less than 10% vegetation

(i) Permanent snow and ice

Figure 4.4: Classes available for our chosen study area.

well-recognised VGI projects. Although the OSM data is not specifically tailored to the needs of land cover mapping and the OSM data and user community is very diverse, the data has valuable input for the land cover classification. In our research, we implemented a method suggested by Chuprikova et al. [99] for deriving a land cover map from the OSM database, as shown in Figure 4.5. To preserve the entire content of the

database, we use a complete XML-encoded extract of the OSM database, representing our study area, instead of pre-processed Shapefiles distributed by OSM data providers. The data pre-processing has been done in a combination of automatic and manual OSM tag annotations to the GlobeLand30 classification scheme. For the derivation of the land cover map, a subset of the OSM tags, namely "amenity", "building", "historic", "land use", "leisure", "natural", "shop", "tourism", and "waterway" are considered. This mapping is only conducted for polygon features since point and line features do not provide immediate information about the coverage of an area. We define a mapping from the OSM attributes to the classes used in the GlobeLand30 classification scheme.

The data is then rasterised and resampled using a nearest neighbour approach to generate a land cover classification with the same resolution and spatial alignment as the GlobeLand30 dataset. The workflow for converting the OSM data into a land cover raster product is depicted in Figure 4.5.

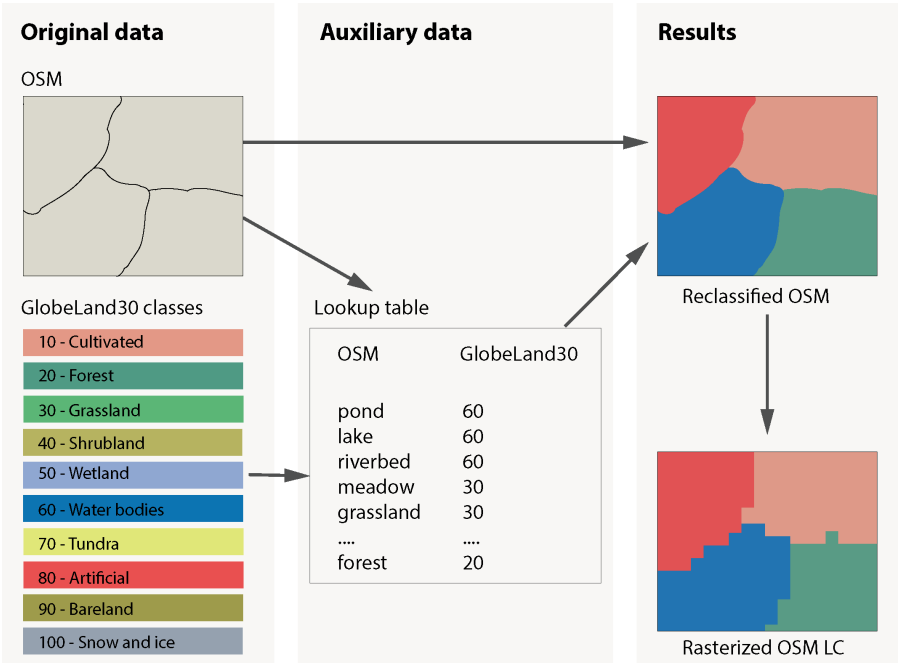


Figure 4.5: An overview of pre-processing steps for converting original OSM data into land cover. Source: [99].

To accurately access our classification boosting approach, a ground truth dataset is required. Although it is intractable to obtain a truly accurate land cover classification, it is still practical to use a *good enough* reference dataset. For our purpose, we chose a well maintained and frequently updated land cover classification dataset, the German national Amtliches Topographisch-Kartographisches Informations System (ATKIS). This

dataset represents a Digital Landscape Model of scale 1:10,000 and 1:25,000 (Basis-DLM), and was provided by an official national cartographic authority (Bundesamt für Kartographie und Geodäsie). Our selection was further motivated by the reported use of this dataset as a reference map by numerous other authors [98, 100].

Dataset preprocessing

Pre-processing and harmonisation of the classes among these three heterogeneous datasets were performed to simplify the construction of our cluster graphs. However, it should be noted that this step is not required and merely reduces the complexity in defining the inter-class relationships between the various datasets (by ensuring a common set of class labels exist for all the datasets).

The details of our pre-processing steps are outlined below:

1. All datasets are cropped to the municipal boundary of Garmisch-Partenkirchen.
2. The classes of OSM and CLC2006 are normalised to match the 10 classes specified by GlobeLand30.
3. The datasets were then rasterised with 30m pixel resolution and aligned to GlobeLand30 using the nearest neighbour resampling method as found in standard GIS software. (This ensured that each pixel was covering the same area of land. For OSM data, the process is more involved and is described in Figure 4.5.)
4. Individual rasters were then sub-divided into sub-regions as per the explanation in Section 4.3.2.

4.3.4 | Definition of priors and parameters

In addition to the cluster graph implementation and dataset, our approach requires inter-class relationships, a per map confidence factor, and the parameters of our sub-regions and boundary sizes.

The inter-class relationships are defined based on expert knowledge and fundamental laws of geography. Firstly, classes that are likely to occur next to each other are assigned to a high probability. In contrast, classes unlikely to neighbour each other – based on region, geography, and expert assumptions – are assigned a low value. Lastly, the self occurrence probability of each class $p(n, n)$ is assigned the highest value to add dependence on Tobler's first law of geography. Due to the nature of inference on cluster graphs via a form of consensus, the prior beliefs of the system do not need to be exact

probabilities but rather need to reflect the relative relationships between various classes. For instance, if $p(a,b) = p(a,c)$ and $p(a,b) >> p(a,d)$, it reflects that it is equally likely that class b or c could neighbour class a and that it is significantly less likely that class d would neighbour class a. The full potential function depicting the relationships used in our experiments is shown in Table 4.4.

Table 4.4: A potential function in the form of a discrete table for the inter-class relationships as defined by expert knowledge.

	cultivated	forest	grassland	shrubland	wetland	water	artificial	bareland	snow
cultivated	1	0.15	0.4	0.05	0.2	0.05	0.4	0.05	0.05
forest	0.15	1	0.05	0.4	0.2	0.05	0.15	0.05	0.05
grassland	0.4	0.05	1	0.2	0.15	0.15	0.05	0.05	0.05
shrubland	0.05	0.4	0.2	1	0.15	0.05	0.05	0.15	0.05
wetland	0.2	0.2	0.15	0.15	1	0.4	0.05	0.05	0.05
water	0.05	0.05	0.15	0.05	0.4	1	0.05	0.2	0.05
artificial	0.4	0.15	0.05	0.05	0.05	0.05	1	0.2	0.05
bareland	0.05	0.05	0.05	0.15	0.05	0.2	0.2	1	0.65
snow	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.65	1

In addition to introducing expert knowledge into the inference process through the inter-class table, we also include further expert knowledge in the form of a classification map confidence factor. This factor can weigh the confidence in each of the input datasets as a whole or in a class-wise manner. Also, the weighting factor can either be set by expert opinion or through more complex statistics.

To assess the effects of including VGI in the form of OSM data, we performed multiple experiments where the confidence in the OSM data was adjusted according to expert opinion. The confidence factors for the CLC2006 and GlobeLand30 datasets were kept constant to only assess the effect of VGI data which is often an incomplete and noisy source of land cover information.

The selected confidence factors are described in Table 4.5, and each of our experimental setups is detailed below:

1. **Scenario 1**: All land cover maps were assumed to be of equal quality, i.e. $P(Y_{i,j}^{k}\text{=Classification}) = 1$.
2. **Scenario 2**: OSM data was assumed to be less accurate overall, i.e. $P(Y_{i,j}^{\text{OSM}}\text{=Classification}) = 0.7$.
3. **Scenario 3**: OSM data was excluded completely from the boosting process.
4. **Scenario 4**: OSM data is assumed to be less accurate overall, except for grassland. The classes are weighted as follows: overall OSM weighting 0.75, cultivated 0.7, wetland 0.6 and grassland 1.0.

Table 4.5: Land cover map confidence scores for adding prior information about dataset confidence, this data is captured by the observation factors $(Y^k_{i,j}, X_{i,j})$ in Table 4.1. Four scenarios were evaluated to determine the effects of various expert assumptions. *Scenario 4: The OSM layer confidence was not uniformly weighted per class, with cultivated=0.7, grassland=1, wetland=0.6 and remaining classes=0.75

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
GlobeLand30	1	1	1	1
CLC2006	1	1	1	1
OSM	1	0.7	×	0.75*

Lastly, the sub-region and boundary sizes were defined such that each sub-region was square with a side length of $35px$ (1050m) and a boundary of $6px$ (180m). It was found that a total boundary width (left + right, or top + bottom), which is between 25% and 50% of the sub-region width or length, is appropriately large for the edge factors to have a negligible influence. In our case, this boundary was defined as $\frac{12px}{35px} \approx 35\%$.

4.4 | Results

In this section, we report on the results obtained using our cluster graph land cover boosting approach. We assess the accuracy of the land cover maps produced and evaluate the uncertainty (in the form of the Shannon diversity index) extracted during the classification boosting process. To perform our assessments, we use the dataset defined in Section 4.3.3 over the study area of Garmisch-Partenkirchen, Germany (see Figure 4.6). To asses thematic classification accuracy, we adopted the method of error matrix evaluation and provided accuracy metrics such as the overall accuracy, Kappa, and the class-wise balanced accuracy.

The overall accuracy represents the proportion of correctly classified pixels to the reference map. The Kappa coefficient $\kappa \in [-1, 1]$ indicates how well the classification performed compared to randomly assigned values. In other words, the Kappa values represent an agreement between two classifications, where $\kappa < 0$ show no agreement, $0 \leq \kappa \leq 0.4$ represent a small degree of agreement, $0.4 < \kappa \leq 0.6$ represents a moderate agreement, $0.6 < \kappa \leq 0.8$ indicate significant agreement, and $0.8 < \kappa \leq 1$ show strong agreement. Furthermore, we use the class-wise balanced accuracy [101] to evaluate the classification on a class-wise basis. The class-wise balanced accuracy represents correctly classified proportions for each class, which is essential as the classes are imbalanced in their distribution within the scene. This measure is favoured over the traditional consumer and producer accuracy as we are making a comparison to a reference

dataset rather than to true ground-truth data. The class-wise balanced accuracy, therefore, represents both the consumer and producer accuracy as a single accuracy measure weighted according to the class distribution in the reference dataset.

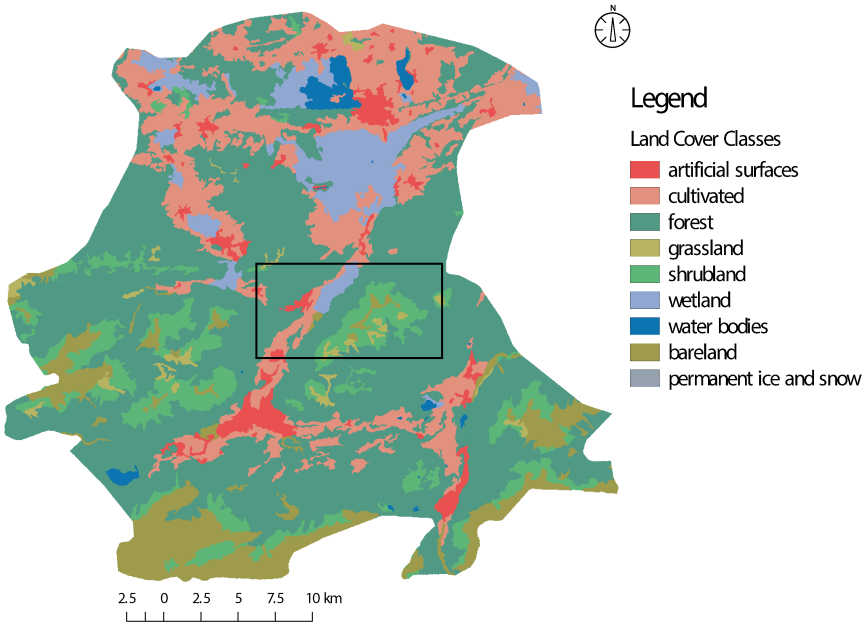


Figure 4.6: Overview map of the study area located in Garmisch-Partenkirchen, Germany. The land cover classifications include nine land cover classes based on the classification scheme adopted from GlobeLand30 (the tundra class is not present within the study area). The subset shows a selected area which is referred to during our more detailed discussions (i.e. Figure 4.8).

To quantitatively assess the accuracy of our boosted land cover maps, we compare our boosted land cover classifications against a reference land cover mapping. As a reference, we chose to use a national dataset called Amtliches Topographisch-Kartographisches Informations System (ATKIS), which represents a Digital Landscape Model with a scale of 1:250,000 (Basis-DLM) provided by an official cartographic authority (Bundesamt für Kartographie und Geodäsie). The remainder of the classification and uncertainty results are analysed qualitatively concerning the individual land cover classifications used as input to the factors of our proposed cluster graph formulation.

Effect of prior information

We tested four different scenarios to assess the effects of expert knowledge and the inclusion of volunteered geographic information (VGI). In all the scenarios, the initial inter-class priors were assigned according to Table 4.4. However, the overall confidence

factor for each dataset was altered on a global and class level (as described in Table 4.5 and Section 4.3.4).

Using these expert beliefs, we performed inference over the test region. Furthermore, we compared the final boosted classification to the ATKIS dataset to determine the effects of various sets of expert knowledge and VGI on the final boosted classification. The results of these comparisons can be seen in Figure 4.7 with a more detailed analysis of overall accuracy and Kappa in Table 4.6.

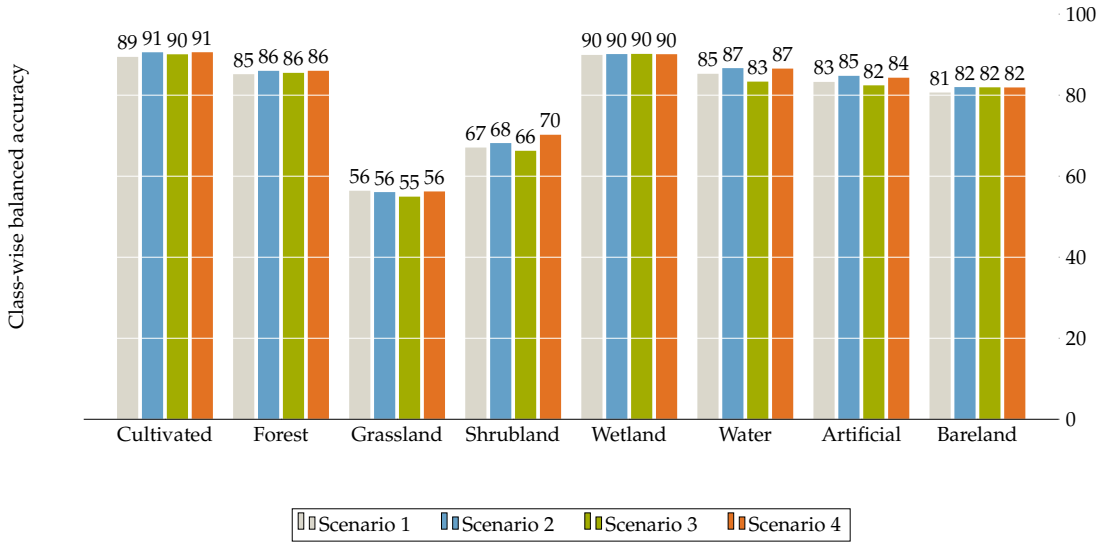


Figure 4.7: The class-wise balanced accuracy for each of our defined scenarios when compared to the ATKIS dataset.

Estimate	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Accuracy	0.7741	0.7798	0.7648	0.7828
Kappa	0.6581	0.6704	0.6506	0.6748

Table 4.6: Overall accuracy and Kappa results of various test scenarios with respect to ATKIS reference dataset

Qualitative assessment of scenarios

We further investigate the various scenarios in a qualitative manner by evaluating the boosted land cover maps. Figure 4.8 shows a subsection of our study area for each of

our four test scenarios, as well as for the input and reference data. Using this approach, we can gain a visual understanding and intuition for the performance of our approach.

Comparison to individual datasets

From Figure 4.7 and Table 4.6, it is clear that Scenario 4, with class-wise weighting, appears to provide the best results. For this reason, we will use Scenario 4 as the output from our approach to compare to the performance of the original land cover classifications of CLC2006, OSM, and GlobeLand30.

Comparing the classifications to the ATKIS dataset, we obtain class-wise accuracy as depicted in Figure 4.9 with an overall accuracy for the various land cover classifications as described in Table 4.7.

Estimate	CLC2006	OSM	GlobeLand30	Scenario 4
Accuracy	0.7452	0.7405	0.7697	0.7828
Kappa	0.6228	0.5013	0.6564	0.6748

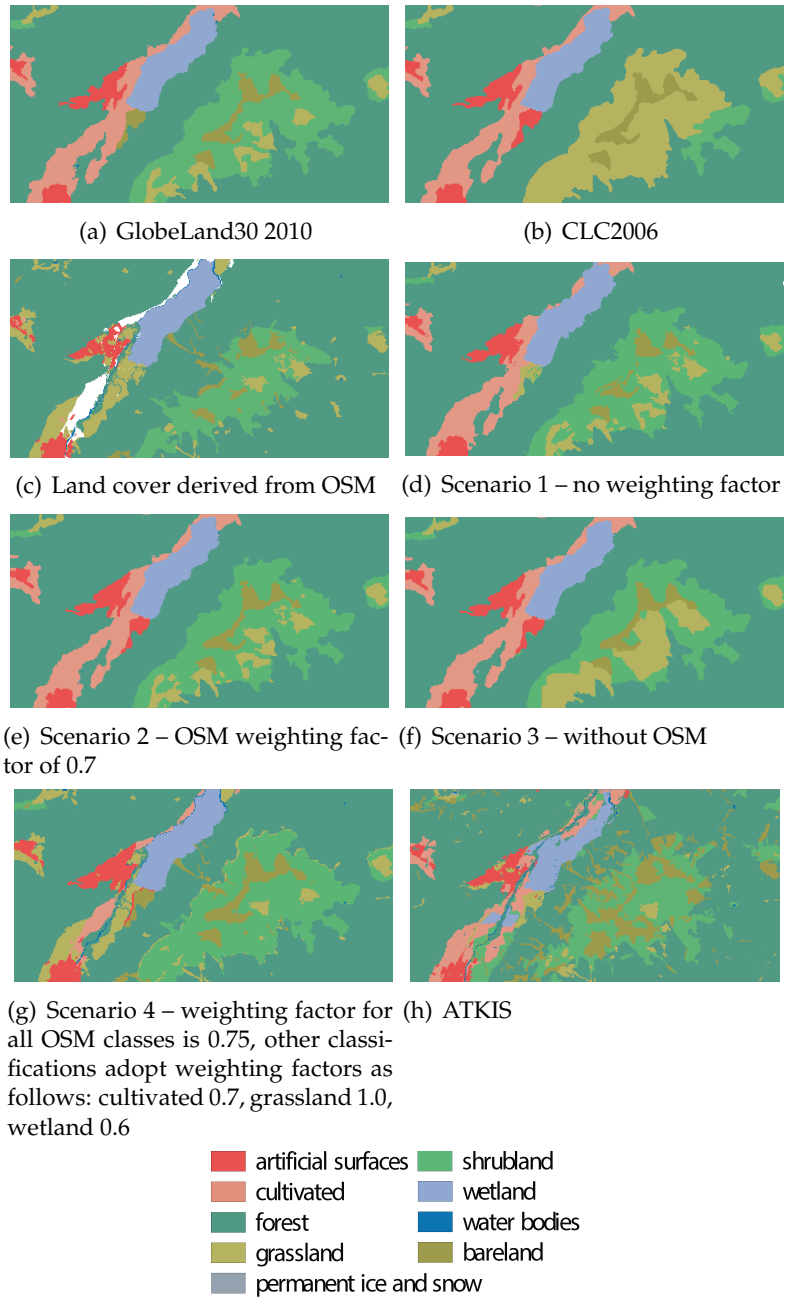
Table 4.7: Accuracy assessment and comparison of the original datasets to the proposed approach.

As mentioned, the overall accuracy represents the proportion of correctly classified pixels to the reference map. For this reason, the measure of accuracy is relative rather than absolute, as it depends on the quality of the reference data [102]. In our case, the ATKIS dataset provides a larger variety of classes and more detailed coverage than the datasets we are boosting. Thus, the overall accuracy is relative to the level of detail on the reference map. Nevertheless, the relative nature of the accuracy is an important factor to keep in mind when assessing the results with respect to existing work, which might employ another reference map or accuracy measure.

Additionally, we compare the prior land cover classifications and our best boosted classification to an aerial image and the ATKIS reference data in Figure 4.10. From this figure, the heterogeneity of the datasets and the incomplete nature of the OSM dataset is clearly visible.

Classification uncertainty

Due to the Bayesian framework in which cluster graphs are rooted, we obtain a measure of the probability for the likelihood of each class being present in each pixel of the boosted land cover map. Based on these probabilities, we can extract an uncertainty



(a) GlobeLand30 2010

(b) CLC2006

(c) Land cover derived from OSM

(d) Scenario 1 – no weighting factor

(e) Scenario 2 – OSM weighting factor of 0.7

(f) Scenario 3 – without OSM

(g) Scenario 4 – weighting factor for all OSM classes is 0.75, other classifications adopt weighting factors as follows: cultivated 0.7, grassland 1.0, wetland 0.6

(h) ATKIS

Figure 4.8: Overview of the land cover classification outputs based on the different scenarios. Scenario 1 – no weighting factor has been applied. Scenario 2 – OSM data is weighted by a factor of 0.7. Scenario 3 – no OSM data is included. Scenario 4 – the general weighting factor for OSM classes is 0.75, except for the following classes: cultivated 0.7, grassland 1.0, wetland 0.6.

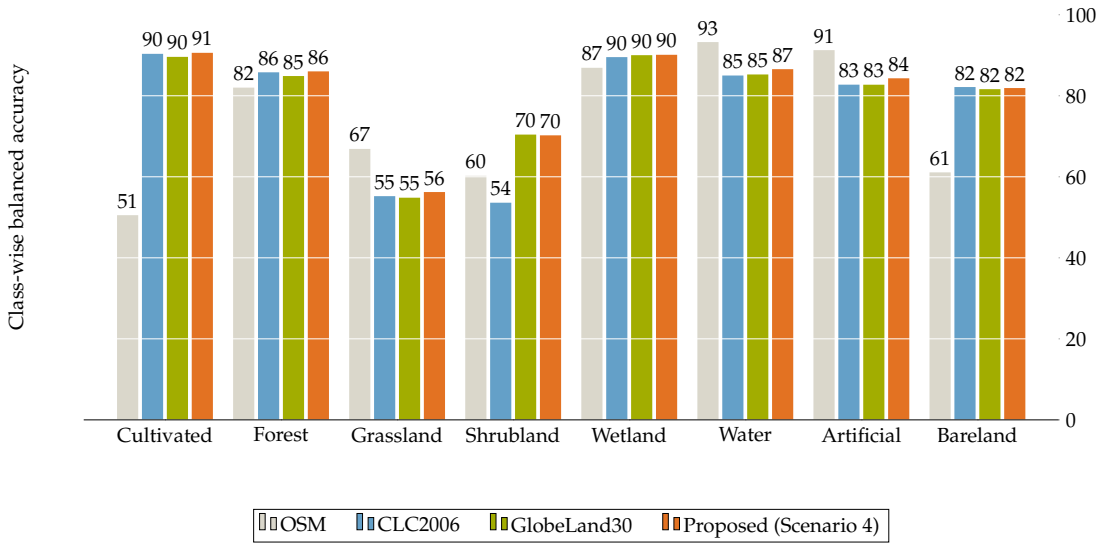


Figure 4.9: The balanced, class-wise accuracy of our approach and the original land cover maps when compared to the ATKIS reference dataset.

metric, Shannon diversity index, for each pixel in our boosted land cover classification map.

The uncertainty map for our approach (Scenario 4) and a zoomed-in section corresponding to the boosted region in Figure 4.10 are depicted in Figure 4.11.

Validation study

Unlike machine learning methods, Bayesian methods do not typically require an independent validation study as there is no differentiation between training and testing phases. However, we performed such a study to assess our general approach and choice of factors and priors. This study was conducted using the proposed cluster graph approach to an auxiliary study area of 84 850 km^2 within Germany (see Figure 4.12). To evaluate the generalisation of our approach with respect to the setting of priors and confidence factors, we kept these values as they were defined for the original test region (Scenario 4).

The validation study represents six land cover classes with the most extensive coverage of artificial and agricultural areas. By applying the same approach, we produced a boosted land cover classification with the overall accuracy given in Table 4.8. Figure 4.13 introduces a comparison of different land cover datasets over the validation area, and Figure 4.14 reports on the class-wise accuracy of our approach as well as the

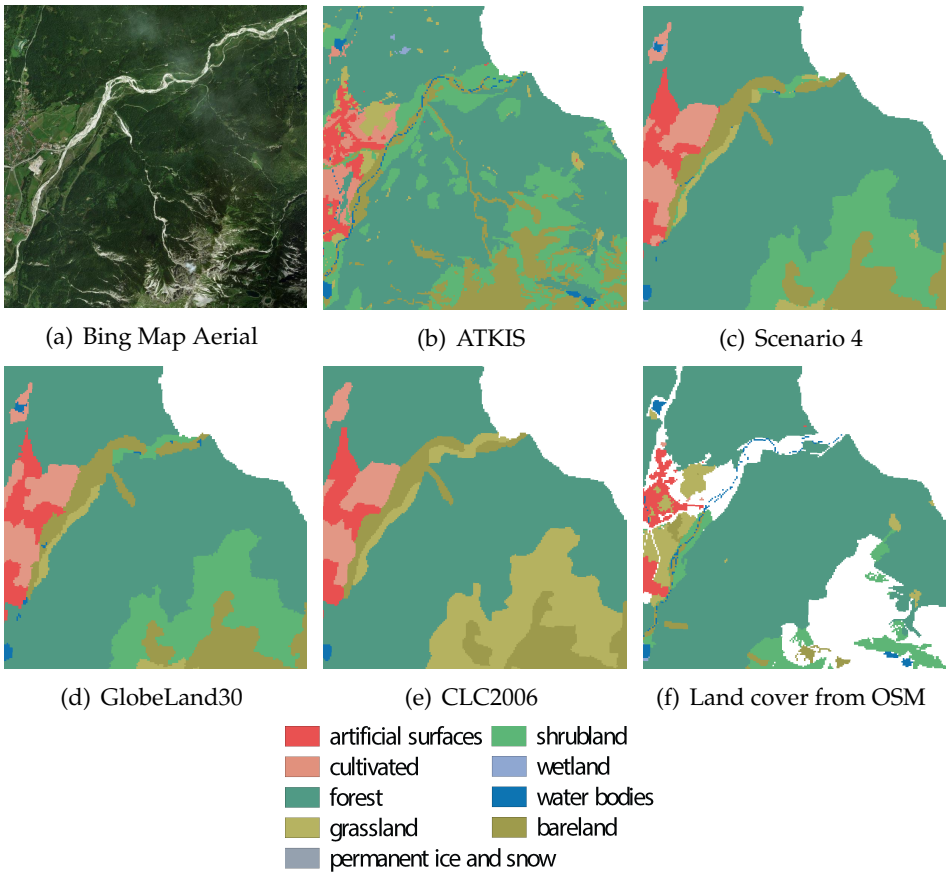


Figure 4.10: Comparison of the different land cover datasets over a selected region in our test scene. Due to inherent ambiguity in definition of classes such as shrubland and grassland, some areas with conflicting patterns are observed. Furthermore, the largely incomplete OSM data in the region can also be observed as the white (no data) pixels.

other datasets.

Estimate	CLC2006	OSM	GlobeLand30	Our Approach
Accuracy	0.8302	0.8009	0.8401	0.8434
Kappa	0.6952	0.6886	0.7054	0.7111

Table 4.8: Validation study: Accuracy assessment and comparison of the original datasets to the approach proposed.

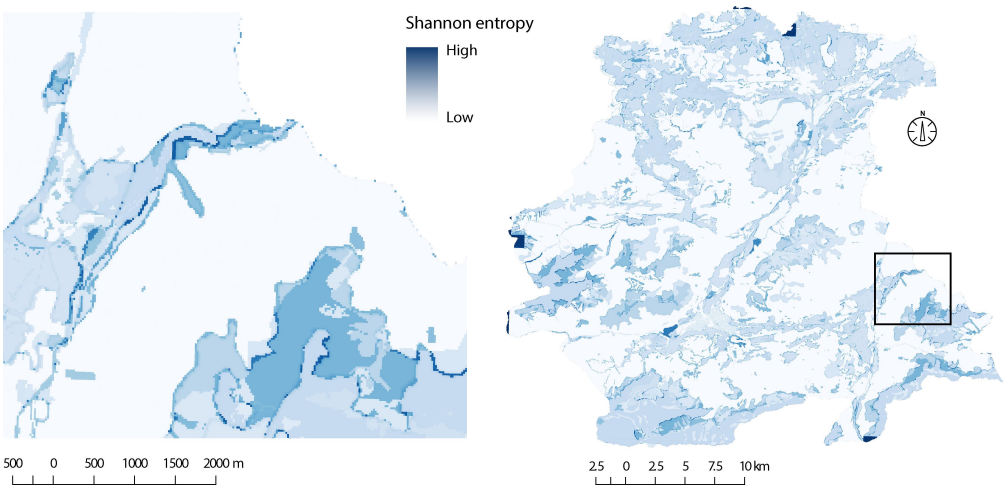


Figure 4.11: Land cover Shannon diversity index map (Scenario 4) depicting the uncertainty in the land cover classification. Low values represent patterns with the highest degree of thematic uncertainty when the land cover class was assigned.

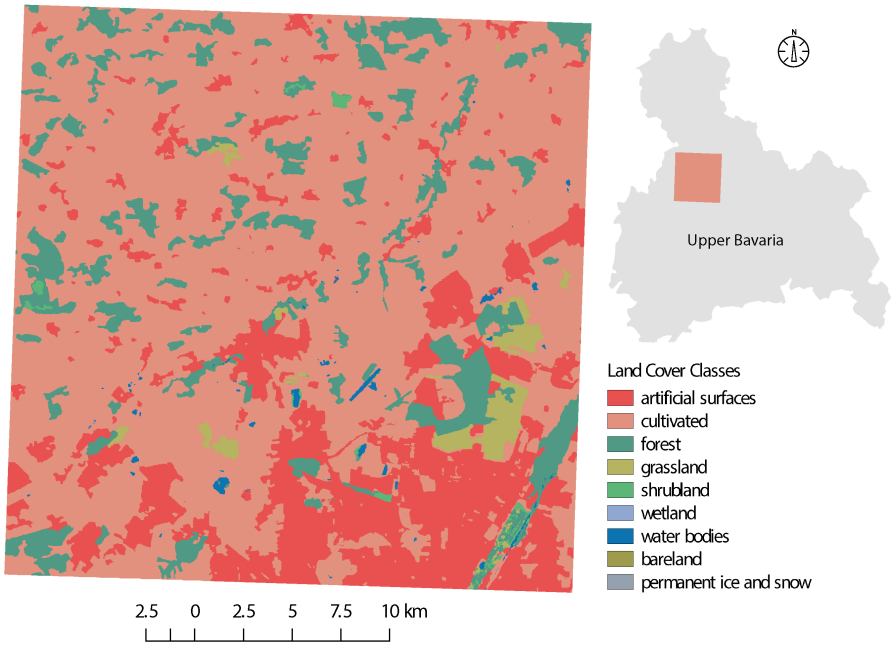
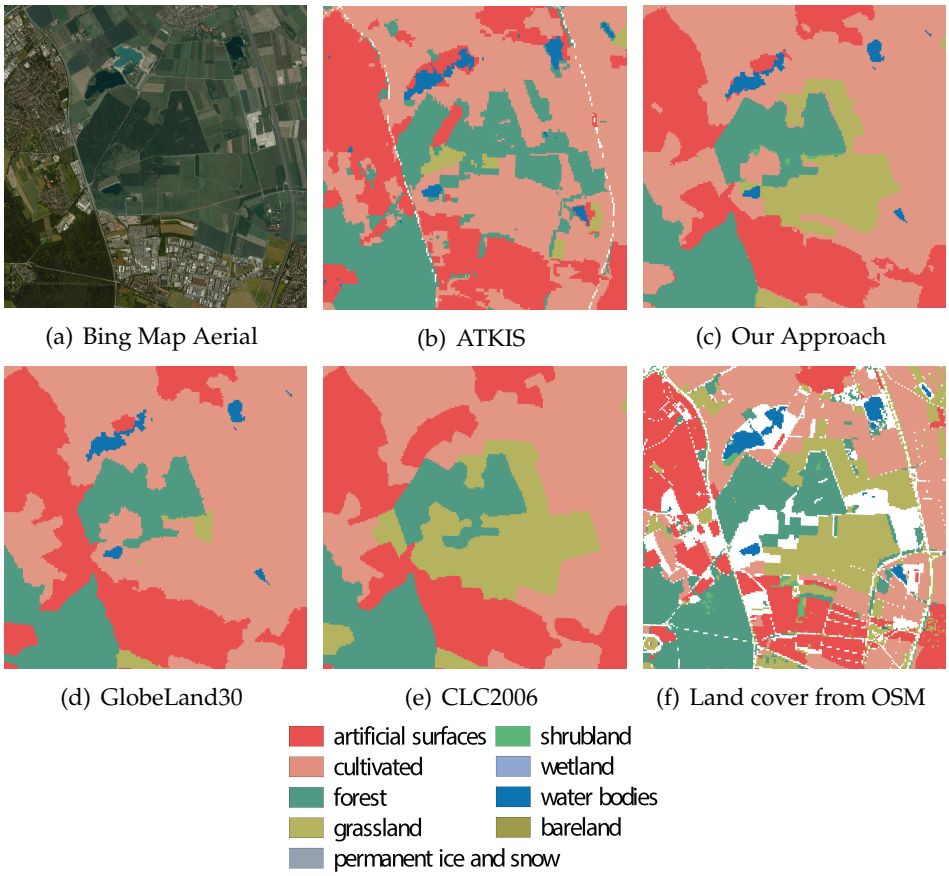


Figure 4.12: Overview map of the validation study area located in Upper Bavaria, Germany. The land cover classifications include six land cover classes based on the classification scheme adopted from GlobeLand30.

4.5 | Discussion

Generally, the results presented in Section 4.4 show that our boosting method can merge existing land cover classifications in a robust manner. Furthermore, it shows that we can



(a) Bing Map Aerial (b) ATKIS (c) Our Approach

(d) GlobeLand30 (e) CLC2006 (f) Land cover from OSM

(g) Legend

Figure 4.13: Validation study: Comparison of the different land cover datasets over a selected region in our validation scene. Due to inherent ambiguity in definition of classes such as grassland, some areas with conflicting patterns are observed in the original datasets (d) – (f). Our approach (c) exhibits spatial smoothness while still capturing smaller details even in regions with conflicting information.

generate an accurate uncertainty map of the boosted area, which is useful in the analysis of the boosted land cover map.

In this section, we further investigate these results and describe the capabilities and advantages of our method for land cover boosting, as well as describe its shortfalls.

Comments on the effect of priors

Since priors are set based on expert knowledge and are subjective by nature, it is difficult to fully analyse the effect of various decisions on accuracy. However, we can make a few key observations based on the results we presented. From Figure 4.7, we can see that all

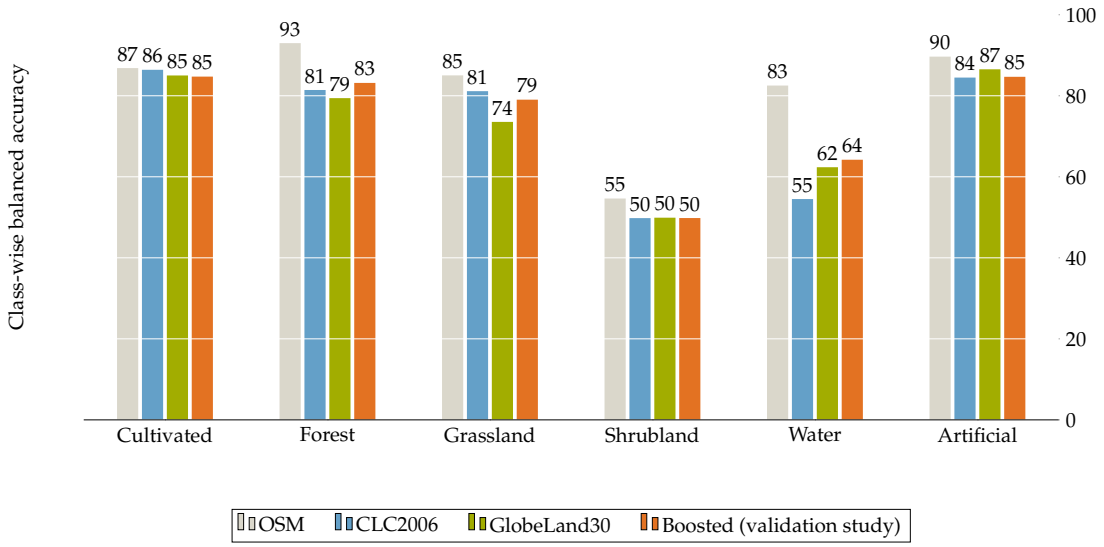


Figure 4.14: Validation study: The balanced, class-wise accuracy of our approach and the original land cover maps when compared to the ATKIS reference dataset.

scenarios which include all the datasets have similar accuracy. In contrast, scenario 3, which excludes OSM data altogether, has a slightly lower accuracy across many classes. However, the accuracy measure does not provide a complete picture of how this factor affects the overall boosted classification appearance. From Figure 4.8, we can see that the overall map confidence factor does not significantly change the final boosted classification. In contrast, class-wise confidence factors have a somewhat larger effect on the final booster classification – see Figures 4.8(d), 4.8(e), and 4.8(g). By adapting class-wise confidence factors of the VGI dataset, we can better capture fine details which are well represented in human labelled datasets (such as river and water features). These types of features are usually only present in land cover maps created with a relatively small minimum mapping unit and are, therefore, not often represented by existing large-scale land cover datasets such as GlobeLand30 and CLC2006.

Table 4.6 confirms that the inclusion of more detailed prior information leads to better overall accuracy, with a greater Kappa coefficient. However, the overall differences are relatively small and, therefore, can also be interpreted that the proposed cluster graph approach is robust to the setting of priors and can produce an overall accurate boosted classification for a range of prior configurations. This is likely due to the strong inference abilities of probabilistic graphical models, which allow this approach to determine the complex inter-dependencies between factors. This, coupled with our geographically centred design approach, appears to capture the most important relation-

ships and thus forgoes the need for strong expert priors to generate a reasonable land cover classification.

Apart from the effect of priors, Table 4.6 reveals that the inclusion of VGI has a positive effect on the overall accuracy of the final boosted classification. Thus, the inclusion of additional datasets, even if noisy and incomplete, might play a larger role in the overall accuracy than the configuration of priors.

Qualitative view of boosted classification accuracy

We compared the boosted classifications from our test scenarios over a diverse subsection of our study area using a purely qualitative approach. In Figure 4.8, we can see the classification differences between our four test scenarios, the original land cover classifications and the ATKIS reference datums.

Based on Figure 4.8 it is clear how the inclusion of various datum as well as the selection of priors affect the overall boosting process. Furthermore, we can observe how our approach preserves spatial consistency between neighbouring classes even when the input data sources are noisy, conflicting or incomplete. This observation is particularly clear in scenario 4, Figure 4.8(g), where the river feature from the OSM dataset is included in the boosted classification, but the area around the feature remains spatially consistent with CLC2006 and GlobeLand30.

Figure 4.10 depicts another sub-region of our test scene and the ability of our approach to perform reasonable inference in the presence of conflicting and missing data becomes quite apparent. The lower right area of the region has very sparse coverage in the OSM dataset and conflicting labels between GlobeLand30 and CLC2006. In this case, our approach can infer a suitable and accurate (with comparison to ATKIS and the aerial image) land cover mapping while still maintaining smaller features, such as lakes and grassland areas, as well as spatial consistency (smooth land cover mapping and reasonable neighbouring classes). Additionally, by referring to the produced land cover uncertainty map for this region in Figure 4.11, it is clear that the areas where land cover was inferred based on missing and conflicting evidence show higher uncertainty than areas where all three input data sources were in agreement. Thus our cluster graph approach can be said to be reasoning in a rational manner.

Furthermore, the results of our validation study, Figure 4.13, once again show how spatial consistency is preserved, and how even in the presence of noisy VGI and conflicting information, our cluster graph approach can boost the classifications to generate a detailed land cover classification. One particular area of interest is the area between the forest, artificial, and cultivated classes to the left of the centre of the image. The OSM

data is incomplete in this region, and the CLC2006 and GlobeLand30 classifications are conflicting. However, even in the presence of this, our boosting approach can infer a classification for this region in a spatially consistent manner. While the classification for the region in CLC2006 is grassland, and in GlobeLand30 is artificial, our boosted classification labels the area as cultivated. This labelling is likely the result of the strong preference for preserving Tobler's first law of geography, as well as strong priors for inter-class relationships. Upon inspection of the aerial image of this region, it can be seen that while our classification overlooks a small artificial region, the remainder of the classification is feasible.

While our approach does lose some granularity in smaller spatial areas, this could be considered a reasonable trade-off, given that some of the finer features are not present in more than one of the input data sources. Additionally, this smoothing effect could possibly be a consequence of the patch-based processing approach we employed, as detailed in Figure 4.3.

Quantitative view of boosted classification accuracy

Referring to the overall accuracy of our approach on the test and validation study areas in Table 4.7 and Table 4.8, our boosted land cover classification exhibits the highest overall accuracy compared to our ATKIS reference dataset. Furthermore, the Kappa coefficient of our land cover classification is significantly higher in both cases. In general, our approach far exceeds the accuracy of OSM and CLC2006 land cover and has a small improvement over GlobeLand30. As overall accuracy does not provide a complete picture of land cover classification accuracy, we further investigate and analyse the balanced class-wise accuracy of the prior data and our boosted classification.

Based on the respective balanced class-wise accuracy of our test scene, as depicted in Figure 4.9, it can be seen that our approach shows reasonable performance in all classes. While the accuracy of OSM does exceed our approach for some classes, note that our approach is never affected by poor class accuracy in any of the datasets. For instance, concerning bareland, shrubland, and cultivated classes, the accuracy of our approach is always better or on par with the other sets of input data. This property could be argued to be of more significance than being able to always achieve the best accuracy in each class. The reason for this is that our approach can perform at a consistent level, even in the presence of noisy input data and, therefore, can provide a higher quality land cover classification overall.

By examining the class-wise accuracy for our validation scene in Figure 4.14, the same observations can be made. While in the validation scene, OSM performs the best

overall, it should be noted that the confidence factor for the OSM dataset was not adjusted. Thus the evidence from the OSM dataset was down-weighted in the cluster graph. However, even with this low confidence in OSM, our approach was still able to extract value from the high accuracy OSM data to improve its accuracy over CLC2006 and GlobeLand30. This once again shows the robustness of a cluster graph approach in the presence of ill-defined priors. As the OSM dataset for this region is significantly less sparse than for our test region, the confidence factor should have been adjusted upwards. Furthermore, the region contains large areas of cultivated land, which is known to be a source of conflict among CLC2006 and GlobeLand30. Thus, prior to boosting the land cover classification, expert knowledge should have been used to adjust the confidence factors and priors for the region.

While class accuracy does not depict a large improvement over existing methods, the power of our proposed approach is in its ability to select and fuse the existing approaches in such a way as to have an overall better land cover classification than each of the individual land cover maps that were boosted. Furthermore, our approach can perform boosting in the presence of noisy, incomplete, and conflicting input data while preserving spatial consistency and producing an overall reasonable, detailed and still diverse land cover classification.

Comments on classification uncertainty

Perhaps one of the largest benefits of the proposed approach is that it provides the probability for each class occurring at each pixel. These class-wise probabilities can easily be exploited to generate uncertainty maps of the boosted land cover classification as well as for the original datasets. Due to the nature of the aerial imagery, which is often used to generate land cover classifications, inherent inter-class ambiguities exist due to the lack of height information. The generated uncertainty maps can help develop better algorithms for land cover classification by either forming part of the optimisation function or providing experts with clues as to which areas are often misclassified and why.

By comparing the uncertainty map, Figure 4.11, to the corresponding classification maps, Figure 4.10, it is clear that the uncertainty for each region tends to agree with intuition about the nature of certainty across the datasets. For example, regions that present conflicting information in two inputs and are missing information in the third are deemed to be more uncertain in the boosted map, while areas with agreement present a very low uncertainty. Perhaps one interesting observation is the low uncertainty in some regions where information is conflicting; this is likely due to the expert priors that enforce self-similarity based on Tobler's law of Geography.

Uncertainty maps can provide useful inputs into ecological and climactic research where uncertainty about land cover classifications can help improve models of land use dynamics and ecosystem stability. Furthermore, the class-wise probabilities could open the door for manual intervention where the top n classes which exhibit similar probabilities could be presented to practitioners for disambiguation and thus further improvement of the overall land cover map. This process could further be expanded to fine-tune the inter-class priors and thus improve the overall performance of the proposed approach.

4.6 | Conclusion

In this chapter, we presented a probabilistic graphical model approach to boosting of land cover classification maps. The formulation of the proposed solution took the form of a cluster graph that used observation and relational factors, along with expert knowledge to perform inference across multiple existing land cover classification data products. The study is applied to land cover classifications derived from remote sensing data, as they are among the crucial inputs to environmental analysis that supports research on topics such as climate change, deforestation, urban change, and population growth. Additionally, we made use of incomplete but accurate volunteered geographic information (VGI), namely OpenStreetMap (OSM), as an additional set of evidence for land cover classification boosting. Furthermore, we analysed how confidence factors could benefit from accurately labelled regions of data while reducing the effect of inaccurate and incomplete areas on boosting.

To improve the accuracy of land cover classification in the study region of Garmisch-Partenkirchen, Germany, our approach exploits existing expert knowledge and constraints such as Tobler's first law of geography. Taking expert knowledge into account enables a classification boosting process with more flexibility and robustness. Furthermore, this approach allows practitioners to customise the tool to their needs while still being robust enough to compensate for poor assumptions and initialisation.

Using the cluster graph approach, we produced a feasible, diverse, and spatially-consistent boosted land cover classification based on GlobeLand30, CLC2006, and OSM data. Our boosted classification exhibited an overall accuracy improvement of around 1.4% when compared against a reference land cover classification map of our test region. Furthermore, our approach was applied to a validation region without adjusting the priors and was shown to perform well even when initialised with sub-optimal priors.

In addition to producing accurate boosted land cover classifications, the proposed

approach can provide additional information on the uncertainty of the boosted classification and highlight commonly misclassified classes within our study region. These additional products are not available when using naïve boosting methods or learned ensemble methods and can provide important insights into better understanding land cover and land use dynamics.

5

Strengthening PGMs: the purge-and-merge algorithm

Preface

In the previous chapters, we established the groundwork for graph colouring, cluster graph formulation, and formulating a problem as a PGM. This chapter presents *Strengthening PGMs: The Purge-and-merge Algorithm* [3][†], which builds upon the previous work.

This publication expresses graph colouring from Chapter 3 as constraint satisfaction and shows how to formulate general CSPs as PGMs. This process is illustrated on a few logic puzzles such as Sudoku, Fill-a-pix, Calcudoku, and Kakuro.

The main contribution of the publication is a technique called purge-and-merge, which makes it possible to systematically simplify inference on complex CSPs. The algorithm is a combination of three established probabilistic techniques from Chapters 2 and 4. More specifically, purge-and-merge extends on

- the LTRIP algorithm for building cluster graphs,
- applying belief propagation on graph colouring (and CSP) factors, and
- merging factors into a joint space.

Furthermore, the technique successfully outperformed the PGM-based approaches mentioned in Chapter 3 and some other approaches from the probabilistic reasoning literature.

[†]Sections of this work have been published in:
S. Streicher and J. du Preez, "Strengthening Probabilistic Graphical Models: The Purge-and-merge Algorithm," *IEEE Access*, vol. 9, pp. 149 423–149 432, 2021.

Abstract

Probabilistic graphical models (PGMs) are powerful tools for solving systems of complex relationships over a variety of probability distributions. However, while tree-structured PGMs always result in efficient and exact solutions, inference on graph (or loopy) structured PGMs is not guaranteed to discover the optimal solutions [12, p391]. It is in principle possible to convert loopy PGMs to an equivalent tree structure, but this is usually impractical for interesting problems due to exponential blow-up [12, p336]. To address this, we developed the "purge-and-merge" algorithm. This algorithm iteratively nudges a malleable graph structure towards a tree structure by selectively *merging* factors. The merging process is designed to avoid exponential blow-up by way of sparse structures from which redundancy is *purged* as the algorithm progresses. We set up tasks to test the algorithm on constraint satisfaction puzzles such as Sudoku, Fill-a-pix, and Kakuro, and it outperformed other PGM-based approaches reported in the literature [13, 14, 15]. While the tasks we set focussed on the binary logic of CSP, we believe the purge-and-merge algorithm could be extended to general PGM inference.

5.1 | Introduction

We have successfully created flexible probabilistic graphical model (PGM) structures to solve constraint satisfaction problems (CSPs) that cannot be solved with existing PGM inference techniques. This entailed the creation of an exact CSP solver that preserves all solutions.

We did not set out to explore modern constraint satisfaction problem solving in general but rather to incorporate constraint satisfaction capabilities into PGMs. Central to this work is a PGM technique called purge-and-merge. It is the combination of three established probabilistic techniques: building cluster graphs, applying loopy belief propagation [18], and merging factors into a joint space. Together, these techniques enable purge-and-merge to allow the growth of factors via factor merging while also removing redundancies in the CSP problem space via loopy belief propagation. We can thus solve a range of CSPs that would be too intricate for either loopy belief propagation or factor merging. Our experimental study shows that purge-and-merge reliably solves problems too difficult for other belief-propagation approaches [1, 13, 14, 15].

Purge-and-merge provides higher-order reasoning for PGMs and constraint satisfaction. This technique would, therefore, be of benefit to any area that incorporates both these domains, such as

- classification and re-classification problems – e.g. image de-noising [12], scene classification [103], and classification boosting in Chapter 4;
- image segmentation – e.g. extracting superpixels using boundary constraints [104]; and
- hybrid reasoning – e.g. solving the game of Clue by combining logical and probabilistic reasoning [105] and solving a Sudoku visually by combining a handwriting input classifier with constraint satisfaction [106].

PGMs are tools that express intricate problems with multiple dependencies as graphs and PGM inference techniques such as message passing can be used to solve these graphs. As a result, PGMs are integral to a wide range of probabilistic problems [107] such as medical diagnosis and decision making [5], object recognition in computer vision [103], as well as speech recognition and natural-language processing [108].

Constraint satisfaction, in turn, is classically viewed as a graph search problem that falls under the umbrella of NP-complete problems. It originated in the artificial intelligence (AI) literature of the 1970s, with early examples in Mackworth [109] and Laurière [110]. Broadly, a CSP consists of a set of variables $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$, where each variable must be assigned a value such that a given set of constraints (clauses) $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$ are satisfied. Typical applications of CSPs include resource management and time scheduling [111], parity checking in error-correcting codes [112], and puzzle games such as Sudoku, Killer-Sudoku, Calcudoku, Kakuro, and Fill-a-pix [113].

Many advances have been made in solving highly constrained PGMs, i.e. PGMs with a large number of prohibited outcomes specified in their factors. This includes PGMs constructed from CSP clauses. A popular approach is to transform such a PGM to another domain and then to solve it with tools specific to that domain. This includes converting PGMs to Boolean satisfiability problems (such as conjugate normal form (CNF) [34]), sentential decision diagrams (SDD) [35], and arithmetic circuits (ACs) [16]. For example, the AC compilation and evaluation (ACE) system [16] compiles a factor graph or Bayes network into a separate AC, which can then be queried about the underlying variables. The drawback lies in the fact that an ACE query will yield a marginal for each queried variable, but it does not yield the joint distribution over all queried variables. ACE thus acts as a heuristic for selecting a single probable CSP outcome.

In contrast, our CSP solver can return a joint solution to a CSP problem (i.e. find its joint distribution) using PGM techniques.

There are trivial ways to formulate CSPs probabilistically and express them as PGMs [1, 33, 14, 15, 13, 114], and although most of them are aimed at specific CSPs, they share

the same basic approach. This amounts to (a) formulating the CSP clauses into PGM factors, (b) configuring the factors into a PGM graph structure, (c) applying belief propagation on this graph, and (d) using the most probable outcome as the solution to the CSP. Dechter et al. [18] provide a bridge between CSPs and PGMs by proving that zero-belief conclusions made by loopy belief propagation reduce to an algorithm for generalised arc consistency in CSPs.

There are limitations, however. Goldberger [14] highlights the difference between belief propagation (BP) with max-product and sum-product. They report that although max-product BP ensures the solution is preserved at all times, it is often hidden within a large spectrum of possibilities. This calls for additional search techniques. Meanwhile, sum-product BP acts as a heuristic to highlight a valid solution but can often highlight an incorrect one. Khan [15] tries to improve on the success rate of sum-product BP by combining it with Sinkhorn balancing. Although they report an improvement, the system could still not reliably solve high-difficulty Sudoku puzzles. In Chapter 3, we suggested a sparse representation for factors and promoted the use of a cluster graph over the ubiquitous factor graph. However, although the cluster graphs improved the accuracy and execution time of the system, this approach is not reliable as a Sudoku solver or CSP solver in general.

The above approaches are all limited in one way or another. They are either ineffective in purging redundant search space – due to their loopy PGM structure – or they rely on an unreliable heuristic to select a probable solution. In this work, we propose techniques to sidestep these limitations and iteratively nudge the graph towards a tree-structured PGM while preserving the CSP solution.

Our proposed technique employs the purge-and-merge algorithm. Purge-and-merge starts by constructing a CSP into a cluster graph PGM with sparse factors. It then *purges* redundancies from these factors by applying max-product belief propagation [18] and thereby propagating zero-belief conclusions. Next, it *merges* factors together to create cluster graphs that are closer to a tree structure. Finally, it constructs a new cluster graph from the factors. This process is repeated until a tree-structured cluster graph is produced. At this point, the exact solution to the CSP is found.

Purge-and-merge manages to reliably solve CSPs that are too difficult for the aforementioned approaches. We reason that a successful CSP approach, such as purge-and-merge, opens many new avenues for exploration in the field of PGMs. This may include hybrid models where rigid and soft constraints can be mixed. It may also be used in domains not previously suited for probabilistic approaches.

Our study is outlined as follows:

- In Section 5.2, we introduce CSP factors and show how they can be structured into a PGM. We also provide the design and techniques to build and solve a basic constraint satisfaction PGM.
- In Section 5.3, we investigate the limitations of PGMs as well as the trade-offs between the loopy-structured PGMs of small-factor scopes and the tree-structured PGMs of large-factor scopes.
- In Section 5.4, we provide a factor clustering and merging routine along with the purging methods necessary for our purge-and-merge technique.
- In Section 5.5, we evaluate purge-and-merge on a number of example CSPs such as Fill-a-pix and similar puzzles and compare them to the ACE system [16].

We found that with the purge-and-merge technique, PGMs can solve highly complex CSPs. We, therefore, conclude that our approach is successful as a CSP solver and suggest further investigation into integrating constraint satisfaction PGMs as subcomponents of more general PGMs.

5.2 | Constraint satisfaction using PGMs

In this section, we show how CSPs are related to PGMs. We express CSPs as factors, which can be linked in a PGM structure. We use graph colouring as an example and expand the idea to the broader class of CSPs.

Most constraint satisfaction problems are easily defined and verified, but they can be difficult to invert and solve. PGMs, by contrast, are probabilistic reasoning tools used to resolve large-scale problems in a computationally feasible manner. They are often useful for problems that are difficult to approach algorithmically – CSPs being one such example.

5.2.1 | A general description of CSPs

Constraint satisfaction problems are NP-complete. Nevertheless, they are of significant importance in operational research, and they are key to a variety of combinatorial, scheduling, and optimisation problems.

In general, constraint satisfaction deals with a set of variables $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ and a set of constraints $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$. Each variable needs to be assigned a value from the variable's finite domain $\mathrm{dom}(X_n)$, such that all constraints are satisfied. For

example, if X_n represents a die roll, then a suitable domain would be $\{1, 2, 3, 4, 5, 6\}$. Furthermore, if we define a CSP where two die rolls, X_1 and X_2, are constrained to sum to a value of 10, then the CSP solution (X_1, X_2) consists of the possible value assignments $(4, 6)$, $(5, 5)$, and $(6, 4)$.

The CSP constraints can be visualised through a factor graph. This is a bipartite graph where the CSP variables are represented by variable nodes (circles) and the CSP clauses by factor nodes (rectangles). The edges of the graph are drawn between factor nodes and variable nodes, such that each factor connects to all the variables in its scope. The scope of a factor is the set of all random variables related to that factor. To illustrate, we present the map colouring example in Figure 5.1.

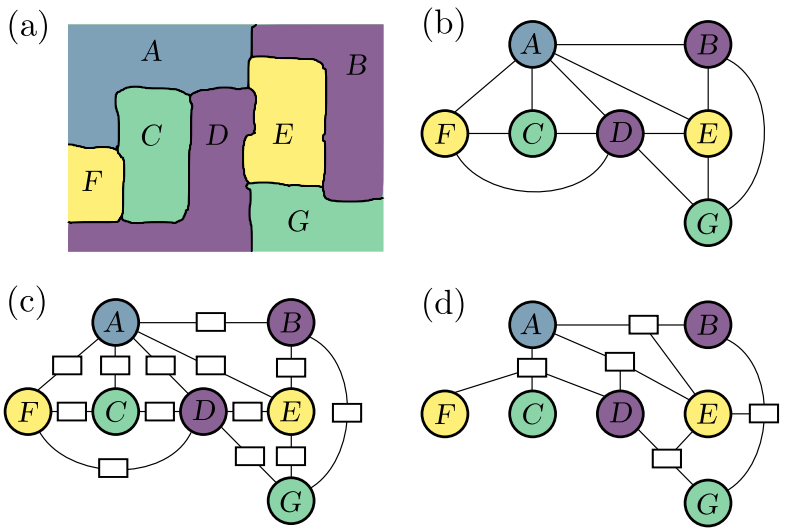


Figure 5.1: (a) A map colouring example, with (b) its graph colouring representation, and (c) and (d) two different factor graph representations using rectangles to represent CSP clauses.

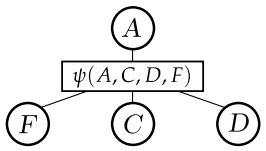
- Figure 5.1(a) shows a map with bordering regions. These regions are to be coloured using only four colours such that no two bordering regions may have the same colour.
- In Figure 5.1(b), we represent this map as a graph colouring problem, where the regions are represented by nodes and the borders by edges.
- Figure 5.1(c) shows a factor graph where the factors represent the CSP clauses. Note that each of these factors has a scope of two variables.
- In Figure 5.1(d), we show that the problem can also be expressed equivalently by combining factors differently. Here we have multiple constraints captured by a

single clause. As a result, we have fewer factors but larger factor scopes. (This example uses the maximal cliques in (b) as factors.)

5.2.2 | Factor representation

A factor graph representation will only show the clauses, the variables, and the relationship between the clauses and variables. The details of these relationships, however, are suppressed. To fully represent the underlying CSP, each factor must also express the relationships implied by the associated constraint. We do so by assigning a potential function to each clause in order to encode all valid local assignments concerning that clause. These assignments are captured by sparse probability tables. The tables list each local possibility as a potential solution and assign a value to that possibility. For CSPs specifically, we work with binary probabilities, ascribing "1" to any (valid) possibility and "0" to any impossibility enforced by the constraint. As an example, see Figure 5.2 for the sparse table representing the factor scope $\{A, C, D, F\}$ from Figure 5.1(d). (The use of sparse tables as PGM factors is also referred to as flattening [18].)

(a)



(b)

A	C	D	F	$\psi(A, C, D, F)$
1	2	3	4	1
1	2	4	3	1
1	3	2	4	1
1	3	4	2	1
⋮	⋮	⋮	⋮	⋮
4	3	2	1	1
elsewhere				0

Figure 5.2: (a) The map colouring clause $\{A, B, D, F\}$ from Figure 5.1, with factor $\psi(A, C, D, F)$, variables A, C, D, F, and variable domains $\text{dom}(A) = \text{dom}(B) = \text{dom}(C) = \text{dom}(D) = \text{dom}(F) = \{1, 2, 3, 4\}$. (b) A sparse table explicitly listing the non-zero entries in $\psi(A, C, D, F)$, and assigning all other entries to be zero.

It is worth noting that the factor graphs presented here are then not only a visually appealing representation for CSPs but are, in fact, PGMs. As such, PGM inference techniques such as loopy belief propagation and loopy belief update can be directly applied to these factor graphs.

In order to perform belief propagation using sparse tables, it is important to implement some basic factor operations; most importantly, see Section 2.6 on factor multiplication, division, marginalisation, conditioning, and damping.

5.2.3 | PGM construction

In essence, a PGM is a compact representation of a probabilistic space as the product of smaller, conditionally independent distributions. When we apply a PGM to a specific problem, we need to (a) obtain factors to represent these distributions, (b) construct a graph from them, and (c) use inference on this graph. In this section, we focus on graph construction.

A cluster graph is an undirected graph where:

- each node i is associated with cluster $\mathbf{C}_i \subseteq \mathcal{X}$,
- each edge (i, j) is associated with a separation set (sepset) $\mathbf{S}_{i,j} \subseteq \mathbf{C}_i \cap \mathbf{C}_j$, and
- the graph configuration satisfies the running intersection property (RIP) [12, p347].

The running intersection property requires that for all pairs of clusters containing a common variable, $X \in \mathbf{C}_i$ and $X \in \mathbf{C}_j$, there must be a unique path of edges, $(\hat{\imath}, \hat{\jmath})$, between $\mathbf{C}_i$ and $\mathbf{C}_j$ such that $X \in \mathbf{S}_{\hat{\imath},\hat{\jmath}} \, \forall \, (\hat{\imath}, \hat{\jmath})$.

Figure 5.3 provides two examples of a cluster graph configuration for the CSP clauses in Figure 5.1. In (a), we have a trivial connection called a Bethe graph. This is a cluster graph with univariate sepsets, an equivalent of the factor graph in Figure 5.1(d). In (b) we have a cluster graph with multivariate sepsets. This graph is generated from the same factors as the Bethe graph but using the LTRIP algorithm. The result is also referred to as a join graph [18].

In Section 3.5.4, we came to the conclusion that cluster graphs with multivariate sepsets have superior inference characteristics to factor graphs in terms of both speed and accuracy. The same is argued by Koller [12, p406], where it is shown that cluster graphs are a more general case of factor graphs without the limitation of passing messages only through univariate marginal distributions. With factor graphs, correlations between variables are lost during belief propagation, which can have a negative impact on the accuracy of the posterior distributions and on the number of messages required for convergence.

The LTRIP algorithm (Algorithm 3 from Section 3.3.3) is designed to configure factors into a valid cluster graph by following the RIP constraints. For each variable $X \in \mathcal{X}$, LTRIP builds a subgraph out of all clusters $\mathbf{C}_i$, where $X \in \mathbf{C}_i$. These subgraphs are then

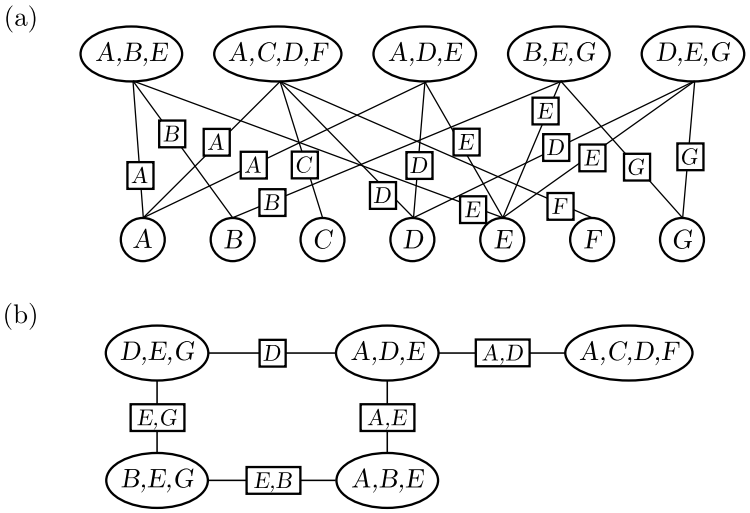


Figure 5.3: Two different cluster graph configurations for the map colouring example in Figure 5.1. (a) A Bethe graph configuration that satisfies RIP by connecting all CSP clauses via single-variable sepsets to singe-variable clusters. (b) A graph configuration generated by LTRIP with fewer edges and multivariate sepsets

superimposed in order to construct the sepsets of the final graph. In summary, the algorithm states that for each variable $X \in \mathcal{X}$, do the following:

- find all clusters $\mathbf{C}_i$ such that $X \in \mathbf{C}_i$,
- construct a complete graph over clusters $\mathbf{C}_i$,
- assign edge weights $w_{\hat{i},\hat{j}}$ to represent the similarity between neighbouring clusters,[1]
- connect the graph into a maximum spanning tree by using an algorithm such as the Prim-Jarník algorithm [19], and
- populate the tree's edges with intermediate sepset results $\mathbf{S}^X_{\hat{i},\hat{j}} = \{X\}$.

After the sepset results are populated for each variable, the sepsets $\mathbf{S}_{\hat{i},\hat{j}}$ of the final graph are taken as the union of the intermediate sepset results $\mathbf{S}_{\hat{i},\hat{j}} = \cup_{X \in \mathcal{X}} \mathbf{S}^X_{\hat{i},\hat{j}}$.

An example of the LTRIP algorithm for the graph in Figure 5.3(b) can be seen in Figure 5.4.

[1]The implementation from Section 3.3.3 uses cluster intersections as edge weights: $w_{\hat{i},\hat{j}} = |\mathbf{C}_{\hat{i}} \cup \mathbf{C}_{\hat{j}}|$. Other suggestions include mutual information or the entropy over the shared variables.

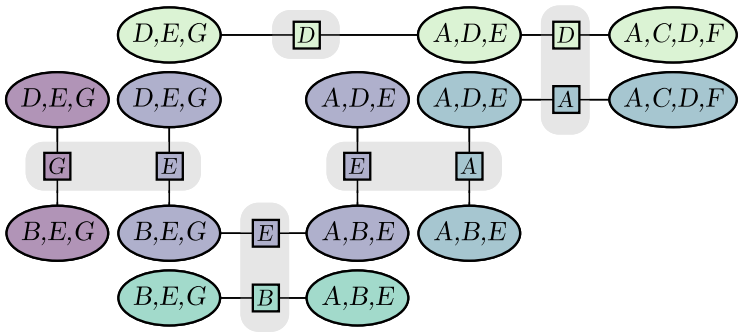


Figure 5.4: An example of applying the LTRIP algorithm in order to achieve the cluster graph construction from Figure 5.3(b). For each variable A, B, C, D, E, F, and G, a maximum spanning tree is constructed from its associated clusters and is populated with univariate sepsets. The resulting cluster graph is then created by taking the superposition of these intermediate trees.

5.2.4 | PGM inference

Our PGM approach extends the work in Chapters 2 and 3 and of flattened belief networks (i.e. sparse tables) [18]. The specific design choices for our PGM implementation are as follows:

1. The factors consist of sparse tables similar to those of Figure 5.2.
2. Graph construction is done using the LTRIP procedure.
3. We use inference via belief *update* (BU) message passing, also known as the Lauritzen-Spiegelhalter algorithm [17].
4. We use the Kullback-Leibler divergence as a comparative metric (and deviation error metric) between distributions.
5. Message passing schedules are set up according to residual belief propagation [115]. Messages are prioritised according to the deviation between a new message and the preceding message at the same location within the graph.
6. Convergence is reached when the largest message deviation falls below a chosen threshold.
7. Throughout the system, we use max-normalisation and max-marginalisation, as opposed to their summation equivalents.

Dechter et al. [18] proved that zero-belief conclusions made by loopy belief propagation are correct and equal to inducing arc consistency. This is true in the case of using

both sum or max operations [14, 18]. This means that the basic PGM approach does not guarantee a solution to the CSP, but it does guarantee that all possible solutions are preserved.

We found convergence to be faster with the max operations than with sum operations. Furthermore, the max operations maintain a unity potential for all non-zero table entries. This is in line with the constraint satisfaction perspective, where outcomes are either possible or impossible. Alternatively, if one is interested in a more dynamic distribution, the sum operations provide varying potentials that can be used as likelihood estimations [14].

Lastly, note that alternative message passing techniques such as warning propagation and survey propagation [55] are available. These two approaches attempt to elevate the solution from the problem space but cannot guarantee that the solution is retained. Our interest is in pursuing an approach where the full solution space is preserved.

5.3 | The limitations of PGMs

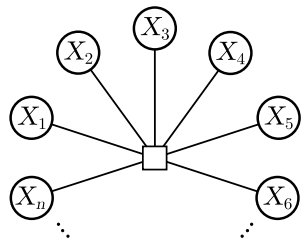
One of the main limitations of constructing CSP potential functions is the resources required to encode them. If the potential functions are encoded as probability tables, then at least all the non-zero potentials need to be listed. Such a list can grow exponentially with the number of factor variables. Therefore, not all CSPs are suitable to be expressed as sparse tables. A trivial example of an ill-suited problem would be a graph colouring problem with n fully connected nodes, as in Figure 5.5. The full space of the problem is n^n with $n!$ entries in the probability table.

Inference on loopy graphs is non-exact; it cannot guarantee a complete reduction to the solution space of a CSP [18]. In exchange, however, loopy graphs provide a great advantage: the ability to handle problems that would have required infeasibly large probability tables if constructed into tree-structured PGMs.

Consider the Sudoku puzzle. A player is presented with a 9×9 grid (with 81 variables) where each variable may be assigned a value of "1" to "9" and is constrained by the following 29 clauses: each row, each column, and each 3x3 non-overlapping subgrid may not contain any duplicates. Furthermore, a valid puzzle is partially filled with values such that only one solution exists. In Figure 5.6, we show the Sudoku-puzzle constraints as a graph colouring problem.

Since a valid Sudoku should only have a single solution, the posterior distribution, $P(X_1, X_2, \ldots, X_{81})$, can be expressed by a sparse table that covers the full variable scope and contains only a single entry. Each marginal distribution, $P(X_1, \ldots, X_9)$,

(a)



(b)

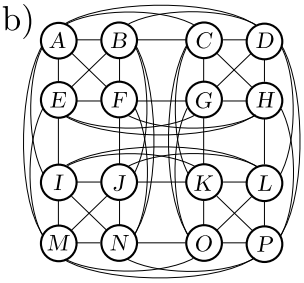
X_1	X_2	$\cdots$	X_{n-1}	X_n	$\psi(X_1, X_2, \cdots)$
1	2	$\cdots$	n-1	n	1
1	2	$\cdots$	n	n-1	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
n	n-1	$\cdots$	2	1	1
elsewhere					0

Figure 5.5: An example of an ill-suited problem with (a), its factor graph and (b), its sparse table containing $n!$ entries.

(a)

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

(b)



(c)

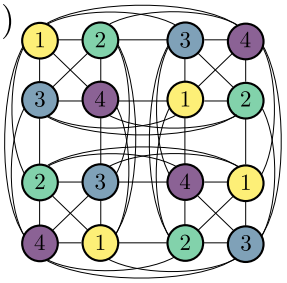


Figure 5.6: An example of a Sudoku puzzle as a graph colouring problem: (a) is a 4×4 version of a Sudoku puzzle, (b) connects the Sudoku variables in an undirected graph, and (c) shows one solution (of many) to this particular problem.

$P(X_{10}, \ldots, X_{18})$, $\ldots$, would then also hold only a single table entry. Yet, after the pre-filled values are observed, and each factor is set up according to its local constraint (the no-duplicated rule), the prior distributions can result in tables with as many as 9! entries. Therefore, a Sudoku solver would have to reduce these large initial probability tables into single-entry posteriors. Before we attempt such a solver, let us first consider two cases – (1) a loopy structure with small-factor scopes, and (2) a tree structure with large-factor scopes:

1. For model (1), we build a loopy-structured PGM directly from the prior distributions. Each factor, therefore, has the same variable scope as one of the clauses. For an inference attempt to be successful, factors should pass information around until

all sparse tables are reduced to single entries. In practice, however, the sparse tables are often reduced by very little. This is because inference on a loopy structure does not guarantee convergence to the final solution space [18].

2. For model (2), we use the most trivial tree structure: multiply all factors together to form a structure with a single node. The resulting factor will now contain one single table entry as the solution. This approach will often (or rather usually) fail in practice since we cannot escape exponential blow-up. In the process of multiplying factors together, the intermediate probability tables first grow exponentially large before the system settles on this single-entry solution.

Since we are confronted by the limitations of both small- and large-factor scopes, we propose a technique in the next section that mitigates these limitations.

5.4 | Purge-and-merge

In this section, we consider the various methods for purging factors and merging factors and combine these methods into a technique called purge-and-merge. It concludes with a detailed outline of the technique in Algorithm 5.

5.4.1 | Factor merging

Our aim in merging factors is to build tree-structured PGMs and to be able to perform exact inference. This can result in exponentially larger probability tables, so it is necessary to approach this problem carefully.

One approach is to cluster the factors into subsets that will merge to reasonably sized tables. To pre-calculate the table size of a factor product is, unfortunately, as memory-inefficient as performing the actual product operation. While we, therefore, cannot use exact table size as a clustering metric, we have investigated three alternative metrics. We propose (1) variable overlap, as in the number of overlapping variables between factors, (2) an upper-bound shared entropy metric, and (3) a gravity analogy that is built on entropy metrics. These methods are experimentally tested in Section 5.5 (Figure 5.7), with the gravity analogy showing the most potential.

Variable overlap

For variable overlap, we define the attraction between two factors, ψ_i and ψ_j, as

$$a_{i,j} = \left|\mathcal{X}_i \cap \mathcal{X}_j\right|,$$

with $\mathcal{X}_i$ and $\mathcal{X}_j$ the scopes of ψ_i and ψ_j respectively. Note the symmetrical relationship, in the sense that the attraction of ψ_j towards ψ_i can be defined as $a_{i,j} = a_{i \leftarrow j} = a_{j \leftarrow i}$.

Upper-bound shared entropy

Upper-bound shared entropy is proposed as an alternative metric to variable overlap. The definition for the entropy of a set of variables $\mathcal{X}$ is

$$\mathrm{H}(\mathcal{X}) = \sum_{x \in \mathrm{domain}(\mathcal{X})} -p(x) \log_2 p(x),$$

with a maximum upper bound achieved at the point where the distribution over $\mathcal{X}$ is uniform. This upper bound is calculated as

$$\hat{\mathrm{H}}(\mathcal{X}) = \log_2 |\mathrm{domain}(\mathcal{X})| .$$

We use this definition to define the attraction between clusters i and j as the upper-bound entropy of the variables they share:

$$a_{i,j} = \hat{\mathrm{H}}(\mathcal{X}_i \cap \mathcal{X}_j),$$

with symmetrical behaviour $a_{i,j} = a_{i \leftarrow j} = a_{j \leftarrow i}$. Note that maximal entropy is used as a computationally convenient proxy for entropy since calculating shared entropy directly is as expensive as applying factor product.

Gravity method

For the gravity method, we use gravitational pull as an analogy for attraction:

$$a_{i \leftarrow j} \propto m_i / r_{i,j}^2.$$

The idea is to relate mass m_i to how informed a factor is about its scope and distance $r_{i,j}$ to concepts regarding shared entropy.

Pseudo-mass: Mass equation $\mathrm{m}(\psi_i) = m_i$ is based on how informed factor ψ_i is about its scope $\mathcal{X}_i$. To parse this into a calculable metric, we use the Kullback–Leibler divergence of the distribution of ψ_i compared to a uniform distribution over $\mathcal{X}_i$.

Pseudo-distance: As a distance metric, we want to register two factors as *close together* if they have a large overlap and *far apart* if they have little overlap. We also do not want this metric to be influenced by a factor's size or mass.

We arrived at a metric using the entropy of the joint distribution, normalised by the entropy of the variables shared between the factors. By using upper-bound entropy in

our calculations, we arrive at distance

$$r_{i,j} = \mathrm{r}(\mathcal{X}_i, \mathcal{X}_j) = \log_2 \left(\frac{\hat{\mathrm{H}}(\mathcal{X}_i \cup \mathcal{X}_j)}{\hat{\mathrm{H}}(\mathcal{X}_i \cap \mathcal{X}_j)} \right).$$

Attraction: Finally, we define the attraction of ψ_j towards ψ_i as analogous to acceleration

$$a_{i \leftarrow j} = \frac{m_i}{r_{i,j}^2}.$$

Using the above metrics, we formulate a procedure for clustering our factors according to the mergeability between factors, as shown in Algorithm 4. Although the algorithm is specialised for the gravity method, it can easily be adjusted for different attraction metrics.

Via this procedure, we can cluster factors $\psi_1, \psi_2, \ldots, \psi_n$ into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_m$, where $m \leq n$. These clusters can then be incorporated into a PGM by calculating new PGM factors $\psi'_1, \ldots, \psi'_m$, by simply merging each cluster $\psi'_i = \prod_{\psi_j \in \mathcal{C}_i} \psi_j$.

5.4.2 | Factor purging

In this section, we show some methods for purging the probability tables of a constraint satisfaction PGM. We use the inference techniques from Section 5.2.4 along with some additional purging techniques:

Reducing variables: If for any factor ψ_j, a variable X is uniquely determined to be x_i, i.e. there are no non-zero potentials with $X \neq x_i$ in that factor, then observe $X{=}x_i$ throughout all factors and remove X from their scopes. This is a trivial case of node consistency [116].

Reducing domains: Likewise, if any domain entry $x_i \in \mathrm{dom}(X)$ has a zero probability to occur in a factor ψ_j, i.e. $P(X{=}x_i|\psi_j) = 0$ for any factor with X in its scope, remove x_i from $\mathrm{dom}(X)$, and remove all probability table entries from the system that allowed for $X{=}x_i$.

Propagating local redundancies: For any two factors, ψ_i and ψ_j, which have common variables, say $\{A, B, \ldots\}$, any zero outcome in ψ_i, i.e. $P(A{=}a, B{=}b, \ldots|\psi_i) = 0$, should also be zero for ψ_j, i.e. $P(A{=}a, B{=}b, \ldots|\psi_j) = 0$. The PGM inference from Section 5.2.4 is, in fact, an algorithm to enforce this relationship, as Dechter et al. [18] proved this to be an algorithm for generalised arc consistency.

We can now combine these techniques along with our merging techniques to build a PGM-based CSP solver.

Algorithm 4: Factor Clustering

Input: Factors $\psi_1, \ldots, \psi_n$ and threshold $\hat{H}_\tau$.

Output: Clustered sets of factors $\mathcal{C}_1, \ldots, \mathcal{C}_m$, with property $\hat{\mathrm{H}}(\text{vars in } \mathcal{C}_i) \leq \hat{H}_\tau \, \forall \, \mathcal{C}_i$.

// Initialise clusters and attractions
for each factor index i **do**
  $\mathcal{C}_i := \{\psi_i\}$
  $\mathcal{X}_i :=$ variables of ψ_i
  $m_i := \mathrm{m}(\psi_i)$
end for
for each i, j pair where $\left|\mathcal{X}_i \cup \mathcal{X}_j\right| > 0$ **do**
  $a_{i \leftarrow j} := m_i / \mathrm{r}(\mathcal{X}_i, \mathcal{X}_j)$
  $a_{j \leftarrow i} := m_j / \mathrm{r}(\mathcal{X}_j, \mathcal{X}_i)$
end for
// Dynamically merge clusters together
while any $a_{i \leftarrow j}$ are still available **do**
  $\hat{\imath}, \hat{\jmath} := \mathrm{argmin}_{i,j}(a_{i \leftarrow j})$
  if $\hat{\mathrm{H}}(\mathcal{X}_{\hat{\imath}} \cup \mathcal{X}_{\hat{\jmath}}) \leq \hat{H}_\tau$ **then**
    $\mathcal{C}_{\hat{\imath}} := \mathcal{C}_{\hat{\imath}} \cup \mathcal{C}_{\hat{\jmath}}$
    $\mathcal{X}_{\hat{\imath}} := \mathcal{X}_{\hat{\imath}} \cup \mathcal{X}_{\hat{\jmath}}$
    $m_{\hat{\imath}} := m_{\hat{\imath}} + m_{\hat{\jmath}}$
    for each $k \neq \hat{\jmath}$ where $|\mathcal{X}_{\hat{\imath}} \cup \mathcal{X}_k| > 0$ **do**
      $a_{\hat{\imath} \leftarrow k} := m_{\hat{\imath}} / \mathrm{r}(\mathcal{X}_{\hat{\imath}}, \mathcal{X}_k)$
      $a_{k \leftarrow \hat{\imath}} := m_k / \mathrm{r}(\mathcal{X}_k, \mathcal{X}_{\hat{\imath}})$
    end for
    remove $\mathcal{C}_{\hat{\jmath}}$, $\mathcal{X}_{\hat{\jmath}}$ and $m_{\hat{\jmath}}$
    remove $a_{\hat{\jmath} \leftarrow l}$ and $a_{l \leftarrow \hat{\jmath}}$, for any index l
  else then
    remove $a_{\hat{\imath} \leftarrow \hat{\jmath}}$
  end if
end while
return all remaining $\mathcal{C}_i, \mathcal{C}_j, \ldots$

5.4.3 | The purge-and-merge procedure

Having outlined all the building blocks needed for purge-and-merge, we can now describe the overall concept in more detail.

We start our model with factors of small-variable scopes by using the CSP clauses directly. We then incrementally transition towards a model with larger-factor scopes by clustering and merging factors. More specifically, we start with a PGM of low-factor scopes, purge redundancies from this model, progress to a model of larger-factor scopes,

and purge some more redundancies. We continue this process until our PGM is tree-structured and thus yields an exact solution to the CSP.

This incremental-factor growth procedure dampens the exponential blow-up of the probability tables and allows the model to incrementally reduce the problem space. The full procedure is outlined in Algorithm 5.

Algorithm 5: Purge-and-Merge

Input: Set of factors $\mathcal{F} = \{\psi_1, \ldots, \psi_n\}$.

Output: Solved variables $\mathcal{X}_s = \{x_i, \ldots\}$ and solved factors $\mathcal{F}' = \{\psi'_i, \ldots\}$.

$\mathcal{X}_s = \{\}$
while return conditions not met **do**
    $\hat{H}_\tau$:= an increasingly larger threshold
    // Factor clustering from Algorithm 4:
    $\mathbb{C}$:= Factor-Clustering($\mathcal{F}$, $\hat{H}_\tau$)
    $\mathcal{F}' := \{(\prod_{\psi_i \in \mathcal{C}} \psi_i)$ for each $\mathcal{C} \in \mathbb{C}\}$
    // LTRIP and LBU from Section 5.2.4:
    $\mathcal{G}$:= LTRIP($\mathcal{F}'$)
    $\mathcal{F}'$:= Loopy-Belief-Update($\mathcal{G}$)
    // Domain reduction from Section 5.4.2:
    Reduce-Domains($\mathcal{F}'$)
    // Variable reduction from Section 5.4.2:
    $\mathcal{X}$:= Reduce-Variables($\mathcal{F}'$)
    $\mathcal{X}_s := \mathcal{X}_s \cup \mathcal{X}$ // add solved variables
    if $\mathcal{G}$ is a tree structure **then**
        return $\mathcal{X}_s$, $\mathcal{F}'$
    end if
    $\mathcal{F} := \mathcal{F}'$
end while

5.4.4 | Algorithmic consistency

As a final reflection on purge-and-merge, we state that all steps taken by this algorithm are correct and result in a consistent CSP solver. Constraint satisfaction falls in the problem class of NP-complete [116], and any general CSP solvers such as purge-and-merge must, therefore, be at least NP-complete.

Dechter et al. [18] provide a prove that loopy belief propagation performed on cluster graphs with flattened tables and hard constraints reduces to generalised arc consistency. They also prove that zero-belief conclusions converge within $\mathcal{O}(n \cdot t)$ iterations of loopy belief propagation and that the algorithm results in a complexity of $\mathcal{O}(r^2 t^2 \log t)$ –

where n is the number of nodes in the cluster graph and t bounds the size of each sparse table. These metrics are, however, not particularly useful for purge-and-merge since the values n and t are expected to change as the algorithm progresses.

The merging of factors is exponential in time complexity, with some merge orders performing better than others [12, p287]. Finding an optimal merge order is, unfortunately, an NP-complete problem and is as difficult as the actual inference [40]. Some algorithms, such as variable elimination, can aid in this process [12, p287]. With variable elimination, the merging order is determined according to the marginalisation of each variable from the factors and the predicted effect it will have on the system as a whole. It should be noted that factor multiplication is commutative, and any merge order converges to the same solution.

Since purge-and-merge is a combination of belief propagation and factor merging, the full algorithm is bounded by factor merging and is thus also NP-complete. Purge-and-merge does, however, mitigate between loopy belief propagation and factor merging with the aim of preventing exponential blow-up in the factor-merging process.

In the next section, we investigate the performance of this procedure by applying it to a large number of CSP puzzles.

5.5 | Experimental study of purge-and-merge

In this section, we investigate the reliability of the purge-and-merge technique by solving a large number of constraint satisfaction puzzles. To compare results, we include Sudoku datasets used in other constraint satisfaction PGM reports [33, 14, 15, 13, 114], as well as the most difficult Sudoku puzzles currently available [117].

5.5.1 | Puzzle dataset

The Sudoku community has developed a large database of the hardest 9x9 puzzles [117] known to literature. The result is a unique collaborative research effort, spanning over a decade, using the widely accepted criterion of the Sudoku explainer rating (SER). This rating is applied by solving a puzzle using an ordered set of 96 rules ranked according to complexity. From the combination of rules required to solve the puzzle, the most difficult one of these rules is used as a hardness measure. The validity of SER as a difficulty rating is discussed by Berthier [113]. They found that SER highly correlates to external pure-logic ratings and can thus be used as a proxy for puzzle complexity. They do note that SER is not invariant to puzzle isomorphisms, i.e. two puzzles from the same validity-preserving group [118] can result in two different ratings.

In addition to the above set, we also compiled a database of constraint puzzles from various other sources to be used as tests. We verified each puzzle to have valid constraints using either PicoSat [119] or Google's OR-Tools [120]. All puzzles are available on GitHub [121], and were sourced as follows:

- 1000 **Sudoku** samples from the SER-rated set of the most difficult Sudoku puzzles, curated by Champagne [117],
- all 95 **Sudokus** from the Sterten set [62] used in Khan [15] and in Section 3.5,
- all 49151 **Sudokus** with 17-entries from the Royle's 2010 set [122] (an older subset of roughly 350000 puzzles was available to Goldberg [14] and Bauke [13]),
- 10000 **Killer Sudokus** from `www.krazydad.com` (labelled according to five difficulty levels).
- 4597 **Calcudokus** of size 9×9 from `www.menneske.no`,
- 6360 **Kakuro** puzzles from `www.grandgames.net`,
- 2340 **Fill-a-Pix** puzzles from `www.grandgames.net`, and
- a mixed set of fairly high difficulty, with one of each of the above puzzle types.

5.5.2 | Clustering metrics

Section 5.4.1 listed three metrics for the purge-and-merge procedure, namely (1) variable overlap, (2) upper-bound shared entropy, and (3) the gravity method. In order to select a well-adapted clustering method, we compared these three metrics on the Champagne dataset. Our approach was to allow purge-and-merge to run for 10*s* under the different clustering conditions and to then report on the largest table size for that run.

Under the naïve variable overlap metric, none of the puzzles came to convergence. When investigating further and re-running the first 10 puzzles without time restriction, they all ran out of physical memory. This is not surprising, as this metric does not account for the domain sizes of the variables, which can have a considerable impact on table size.

Of the remaining metrics, the gravity method had a 100% convergence rating within the 10*s* threshold, whereas upper-bound shared entropy had a 53% rating. Compared to upper-bound shared entropy, the gravity method also resulted in a smaller maximum table size in 74.7% of cases. A histogram representing the maximum table size for each run can be seen in Figure 5.7.

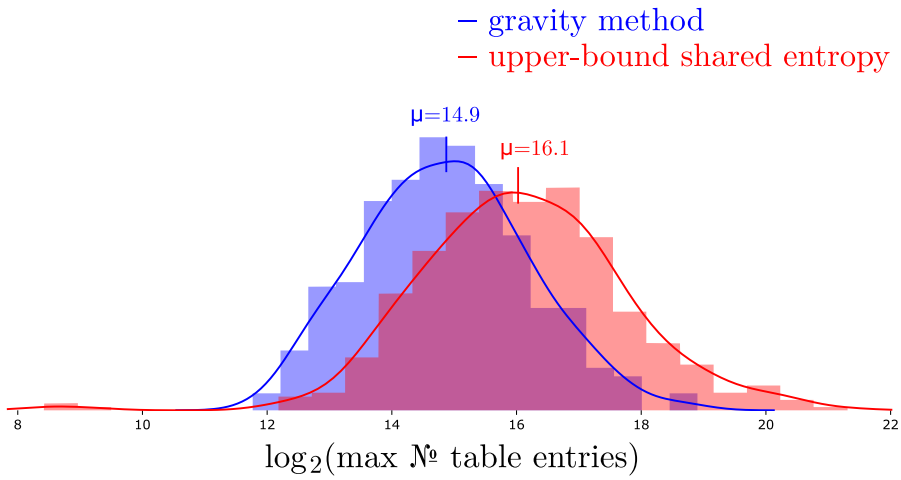


Figure 5.7: The maximum table produced in a purge-and-merge run using two different clustering methods: the upper-bound shared entropy method and the gravity method. All tests are run on the Champagne dataset. Only runs where both methods resulted in a convergence within 10*s* are displayed.

Since the gravity method performed better than the other metrics, we opted to use it in all further purge-and-merge processes.

5.5.3 | Purge-and-merge

All tests were executed single-threaded on an Intel® Core™ i7-3770K, with a rating of 3.50GHz and 4 cores / 8 threads in total. The purge-and-merge algorithm is available on GitHub [121] as a Linux binary with a command-line interface for running any of the puzzles used in our tests.

Purge-and-merge can solve all the Sudoku, Killer Sudoku, Kakuro, and Fill-a-Pix puzzles we have provided. In the case of Calcudoku, 1.4% of the puzzle instances reached the machine's physical RAM limitation of 32Gb. This indicates that purge-and-merge deals better with large numbers of small factors, as with Kakuro, rather than a small or medium number of large factors, as with Calcudoku and Killer Sudoku. Runtime metrics for the various puzzles can be seen in Figure 5.8.

Compared to the other available PGM approaches, purge-and-merge is the only method to achieve a 100% success rate with all the Sudoku puzzles it encountered. Moreover, purge-and-merge was tested on more cases than what is reported in any of the comparable literature.

If we compare purge-and-merge to the results in Section 3.5 for the Sterten [62] set, purge-and-merge is slower (see Figure 5.8). However, that approach is equivalent to a

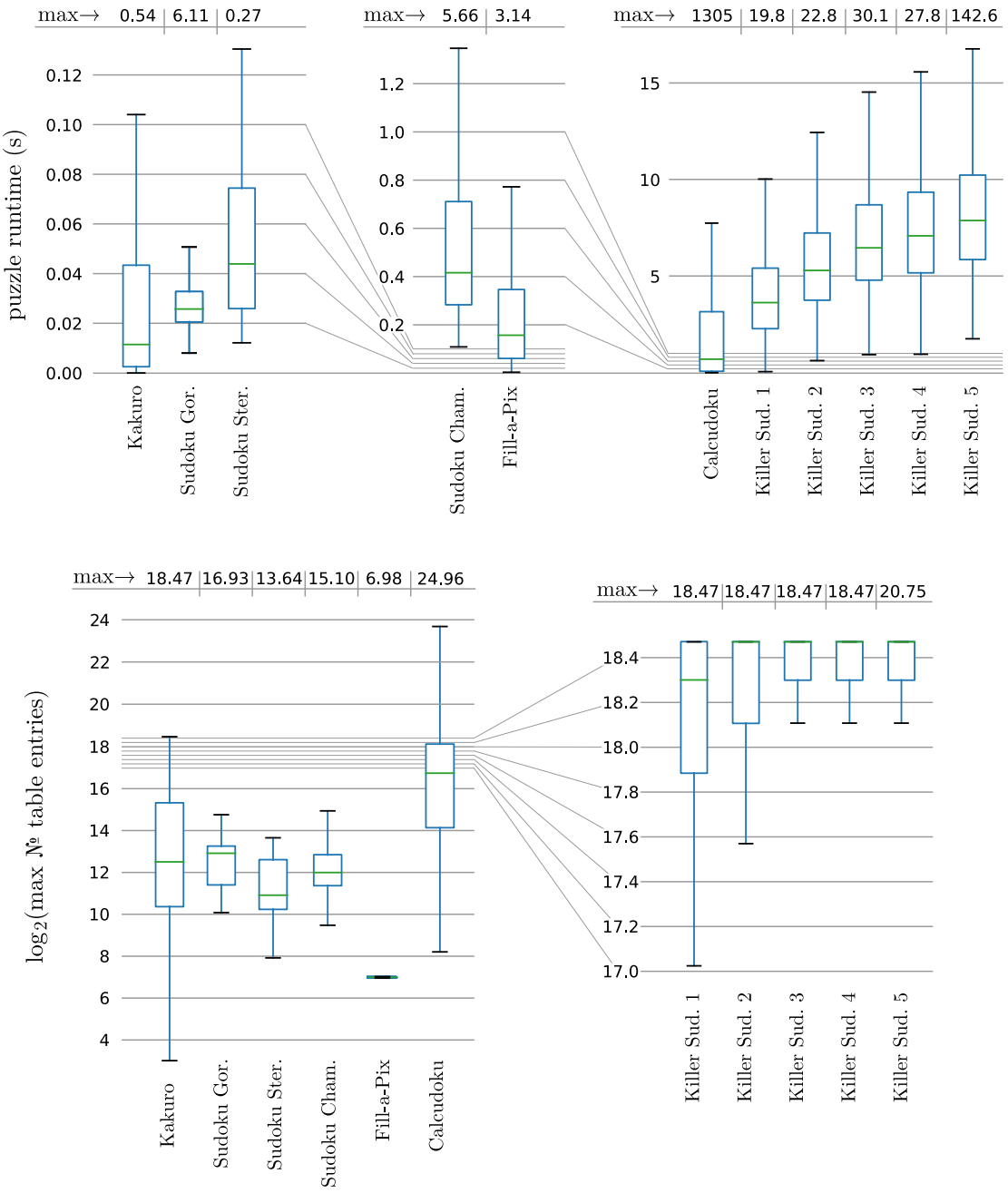


Figure 5.8: Runtime and size metrics for the purge-and-merge approach. The size metric indicates maximum entropy for any given factor during a purge-and-merge run, that is $\log_2(\text{maximum factor entries})$. The different Killer Sudoku sets are split according to reported difficulty, and the 1.4% of unsuccessful Calcudoku runs are not included in these plots.

single "purge" step in purge-and-merge. The full purge-and-merge method obtained a success rate of 100%, whereas the success rate in Section 3.5 is only 36.8%.

In comparison with the other Sudoku PGM literature, purge-and-merge and Chapter 3 are the only PGM approaches that ensure the full CSP solution space is preserved (i.e. no valid solutions are lost). Additionally, purge-and-merge allows the scope of the solver to increase up to the point where only the solution space is left.

The solution space is not preserved in the PGM approaches of Khan [15], Goldberger [14], and Bauke [13]; instead, they use sum-product BP to seek out a single likely solution from the problem space. Khan [15] provides us with a comparison between these three approaches, as shown in Table 5.1. From this table, it is clear that these reported PGM approaches are not well suited for Sudoku puzzles of medium and higher difficulty.

Research	Approach	Reported accuracy for 9×9 Sudokus
Bauke [13]	Sum-Product	53.2%
	Max-Product	70.6%
Goldberger [14]	Sum-Product	71.3%
	Max-Product	70.7% – 85.6%
	Combined Approach	76.8% – 89.5%
Khan [15]	Khan with 40 iterations	70%
	Khan with 200 iterations	95%
This paper	Purge-and-merge	100%

Table 5.1: The success rate of various PGM approaches on Sudoku puzzles, originally compiled by Bauke [13]. Note that this applies to puzzles far easier than our expansive set, which also includes the current most difficult Sudoku set [117].

5.6 | Comparison to the ACE system

The ACE system [16] is a related system for solving constraint satisfaction PGMs. ACE works by compiling Bayes networks and other factor graph representations into an arithmetic circuit, which can then be used to answer queries about the input variables.

ACE focuses on the marginals of the variables of the system and not on finding the joint distribution of the system. This distinction is important – purge-and-merge produces all the solutions to a CSP, whereas ACE will only report on the marginal of each variable.

To illustrate, if we take the first 10 puzzles from the Champagne dataset and arbitrarily remove a known entry, purge-and-merge finds 426, 380, 917, 799, 77, 476, 454,

1754, 777, and 796 answers for each puzzle, respectively. ACE, on the other hand, only reports on the domain of each unknown variable and is, therefore, unable to find any valid solution.

ACE approaches the problem in two stages. It first compiles a network along with its unknown variables into an arithmetic circuit. Then it uses the compiled network to answer multiple queries with respect to the unknown variables. Note that a single ACE circuit to represent all Sudoku puzzles is too large to fit in 32GB of memory due to the large number of possible solutions $\approx 6.67 \times 10^{21}$ [123].

To compare ACE with purge-and-merge, we parsed all the Champagne puzzles into ACE-compliant structures and then compiled each structure into an ACE circuit. In plotting the result in Figure 5.9, we discarded the loading and query times since all evidence was already incorporated into the structure.

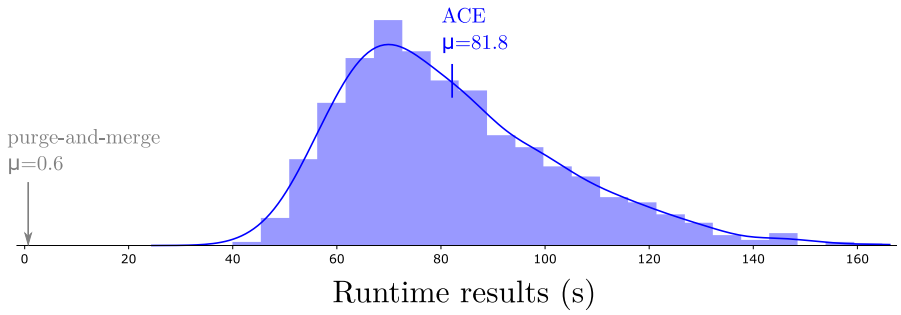


Figure 5.9: ACE runtime on the Champagne dataset. Only the network's compile times were recorded since the query times were negligible. For comparison, the average purge-and-merge runtime is indicated.

5.7 | Conclusion and future work

In general, the factors in a PGM can be linked up in different ways, resulting in different graph topologies. If such a graph is tree-structured, inference will be exact. However, more often than not, the graph structure will be loopy, which results in inexact inference [12, p381]. Transforming a loopy graph into a tree structure, unfortunately, is not always feasible – in all but the simplest cases, the resultant hyper-nodes will exponentially blow up to impractical sizes. Hence we are usually forced to work with message passing on a loopy graph structure.

The ubiquitous factor graph is the structure most frequently encountered in the literature – its popularity presumably stems from its simple construction. Previous work has shown that inference on factor graphs is often inferior to what can be obtained with

more advanced graph structures such as cluster graphs. Nodes in cluster graphs typically exchange information about multiple random variables, whereas a factor graph is limited to sending only messages concerned with single random variables [12, p406]. The LTRIP algorithm enables the automatic construction of valid cluster graphs. Despite their greater potency, however, they might still be too limited to cope with complicated relationships.

In our current work, we extend the power of cluster graphs by dynamically reshaping the graph structure as the inference procedure progresses. Semantic constraints discovered by the inference procedure reduce the entropy of some factors. Factors with high mutual attraction can then be merged without necessarily suffering an exponential growth in factor size. The LTRIP algorithm reconfigures a new structure that becomes progressively more sparse over time. When the graph structure morphs into a tree structure, the process stops with an exact solution. We refer to this whole process as purge-and-merge.

Purge-and-merge is especially useful in tasks that, despite an initially huge state space, ultimately have a small number of solutions. By hiding zero-belief conclusions from memory, purge-and-merge can perform calculations on subspaces within an exponentially large state space.

The purge-and-merge approach is not suited to tasks where the number of valid solutions would not fit into memory, as this would preclude a sufficient reduction in factor entropy. However, as the above results show, purge-and-merge enabled us to solve a wide range of problems that were previously beyond the scope of PGM-based approaches.

In comparison with ACE, we find purge-and-merge more suited to constraint satisfaction problems with multiple solutions, as well as puzzles with a problem space too large to be compiled into a single ACE network.

Our current approach relies on the increased sparsity of the resultant graphs to gradually nudge the system towards a tree structure. In future work, we intend to control that process more actively. This should result in further gains in efficiency, and it is our hope that it will conquer the couple of Calcudokus that still elude us.

6

Conclusion and future implications

The aim of this dissertation was to improve the PGM literature by providing accessible algorithms and tools for improved PGM inference. To this end, our main contributions are

- a comparative study between cluster graphs and factor graphs,
- boosted land cover classification as a practical application for using PGMs in the field of cartography, and
- purge-and-merge as a PGM formulation and technique for solving constraint satisfaction problems too complex for the traditional PGM approach.

This dissertation presented an overview of probabilistic graphical models (Chapter 2) and managed to build on that basis (Chapter 3) by configuring graph colouring factors into potential functions, providing a means for constructing cluster graphs from these factors, and comparing the performance of these graphs to the ones prevalent in the current literature. Through experimental results, we established that the cluster graphs produced by LTRIP have superior inference characteristics to the ones prevalent in the current literature.

Furthermore, these tools (established in Chapters 2 and 3) were shown to be effective in other domains. For example, we applied our PGM formulation to a problem in cartography, land cover classification boosting, in Chapter 4. In this, we illustrated how to approach relational problems as probabilistic reasoning problems and formulated them using PGMs. Our approach managed to reclassify a test region in Garmisch-Partenkirchen, Germany, and boosted the overall classification accuracy compared to an independent reference land cover dataset. This formulation can effectively be applied to a wide range of other probabilistic reasoning problems. We suggest using it for image

de-noising [12], extracting superpixels [104], and natural-language processing [108], for instance.

Lastly, the research reported in this dissertation contributed to the development of a more optimised approach for higher-order probabilistic graphical models applied to constraint satisfaction problems. This was presented in the form of an algorithm named purge-and-merge. Purge-and-merge extends the power of cluster graphs by dynamically reshaping a PGM graph structure as the inference procedure progresses. Inference is done via belief propagation and reduces the entropy of the factors. This allows factors to be merged without necessarily suffering an exponential growth in factor size. This routine is iteratively applied until a tree-structured graph is reached and exact inference is guaranteed. Purge-and-merge hides zero-belief conclusions from memory and can thereby perform calculations on subspaces within an exponentially large state space. It is especially useful for tasks that, despite an initial huge state space, ultimately have a small number of solutions.

Purge-and-merge was tested on a number of constraint satisfaction problems, such as Sudoku, Fill-a-pix, Kakuro, and Calcudoku puzzles, and managed to outperform other PGM-based approaches reported in the literature [13, 14, 15]. Finally, purge-and-merge was compared to another probabilistic reasoning approach, the ACE system [16], and was shown to be more robust when the problem domain had large factor sizes or multiple solutions.

Future work

Currently, purge-and-merge is aimed at constraint satisfaction problems. The purge routine is done by loopy belief propagation, after which the merge routine is done by clustering factors and merging them. This process is then repeated by reconfiguring everything into a new cluster graph structure. We restrict our CSP potential functions to only allow potentials of a binary nature, i.e. values of 0 or 1. If we were to allow the potentials to be continuous, i.e. $0 \leq \phi \leq 1$, then the posterior beliefs from belief propagation would not represent conditional distributions but rather that of marginal distributions. Since PGM inference is not designed for the use of marginal distributions as priors but rather for conditional distributions (corresponding to particular renditions of the chain rule in Equations 2.1), the following belief propagation step of purge-and-merge would effectively lead to overfitting of the beliefs established from the first iteration. We suggest a study into finding a scheme for calibrating the posterior beliefs of non-binary potential functions to be reused in subsequent rounds of belief propaga-

tion. This will allow purge-and-merge to be used in hybrid systems where probabilistic reasoning is mixed with constraint satisfaction.

Furthermore, it might also be possible to significantly reduce dimensionality for the purge-and-merge algorithm. Currently, the merge routine in purge-and-merge is performed via a factor product. This results in exponential blow-up in cases where the purge routine cannot reduce the valid system states effectively. We suggest exploring a more effective merging scheme, such as configuring factors into a junction tree (using variable elimination [12, p287], for example) rather than performing a factor product. These junction trees can then be used as hyper-nodes within a larger cluster graph. This scheme would also require a system to calculate messages from junction trees and pass messages from one junction tree to another. It would also require running tree-based belief propagation on the junction tree nodes and loopy belief propagation on the hyper structured cluster graph.

References

[1] S. Streicher and J. du Preez, "Graph coloring: Comparing cluster graphs to factor graphs," in *Proceedings of the ACM Multimedia 2017 Workshop on South African Academic Participation*, ser. SAWACMMM '17. New York, NY, USA: ACM, 2017, pp. 35–42. [Online]. Available: http://doi.acm.org/10.1145/3132711.3132717

[2] L. H. Hughes, S. Streicher, E. Chuprikova, and J. du Preez, "A cluster graph approach to land cover classification boosting," *Data*, vol. 4, no. 1, 2019. [Online]. Available: https://www.mdpi.com/2306-5729/4/1/10

[3] S. Streicher and J. du Preez, "Strengthening probabilistic graphical models: The purge-and-merge algorithm," *IEEE Access*, vol. 9, pp. 149 423–149 432, 2021.

[4] S. Streicher, "A probabilistic graphical model approach to solving the structure and motion problem," Master's thesis, Stellenbosch: Stellenbosch University, 2016.

[5] B. Wemmenhove, J. M. Mooij, W. Wiegerinck, M. Leisink, H. J. Kappen, and J. P. Neijt, "Inference in the Promedas medical expert system," in *Artificial Intelligence in Medicine*, R. Bellazzi, A. Abu-Hanna, and J. Hunter, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 456–460.

[6] I. Rish, "Distributed systems diagnosis using belief propagation," in *Proc. of the 43rd Annual Allerton Conf. on Communication, Control and Computing*. Citeseer, 2005, pp. 1727–1736.

[7] M. Zhang, D. D. Morris, and R. Fu, "Ground segmentation based on loopy belief propagation for sparse 3D point clouds," in *2015 International Conference on 3D Vision*, 2015, pp. 615–622.

[8] F. Dellaert, "Factor graphs: Exploiting structure in robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 141–166, 2021.

[9] T. Glarner, M. M. Momenzadeh, L. Drude, and R. Haeb-Umbach, "Factor graph decoding for speech presence probability estimation," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5.

[10] O. Bernier, P. Cheung-Mon-Chan, and A. Bouguet, "Fast nonparametric belief propagation for real-time stereo articulated body tracking," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 29–47, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731420800101X

[11] R. M. C. Taylor and J. A. du Preez, "ALBU: An approximate loopy belief message passing algorithm for LDA to improve performance on small data sets," 2021. [Online]. Available: https://arxiv.org/abs/2110.00635

[12] D. Koller and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*, 1st ed. MIT Press, 2009.

[13] H. Bauke, "Passing messages to lonely numbers," *Computing in Science and Engineering*, vol. 10, no. 2, pp. 32–40, 2008. [Online]. Available: http://dx.doi.org/10.1109/MCSE.2008.60

[14] J. Goldberger, "Solving Sudoku using combined message passing algorithms," 2007, bar Ilan's Faculty of Electrical and Computer Engineering. [Online]. Available: https://web.archive.org/web/20151127060738/eng.biu.ac.il/~goldbej/papers/Sudoku_BP.pdf

[15] S. Khan, S. Jabbari, S. Jabbari, and M. Ghanbarinejad, "Solving Sudoku using probabilistic graphical models," 2008, accessed: 2016-03-04. [Online]. Available: https://web.archive.org/web/20160304193943/ece.ualberta.ca/~madjid/Files/Publications/TR081231.pdf

[16] A. Darwiche and M. Chavira, "ACE, an arithmetic circuit compiler," 2007, accessed: 2021-09-20. [Online]. Available: http://reasoning.cs.ucla.edu/ace

[17] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 157–224, 1988.

[18] R. Dechter, B. Bidyuk, R. Mateescu, E. Rollon, H. Geffner, and J. Halpern, "On the power of belief propagation: A constraint propagation perspective," *Heuristics, Probabilities and Causality: A Tribute to Judea Pearl*, 2010.

[19] R. C. Prim, "Shortest connection networks and some generalizations," *The Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[20] R. Dušek and R. Popelková, "Theoretical view of the shannon index in the evaluation of landscape diversity," *AUC Geographica*, vol. 47, no. 2, pp. 5–13, 2017.

[21] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artificial intelligence*, vol. 29, no. 3, pp. 241–288, 1986.

[22] ——, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[23] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1," in *Proceedings of ICC'93-IEEE International Conference on Communications*, vol. 2. IEEE, 1993, pp. 1064–1070.

[24] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's "belief propagation" algorithm," *IEEE Journal on selected areas in communications*, vol. 16, no. 2, pp. 140–152, 1998.

[25] I. Rish, K. Kask, and R. Dechter, "Empirical evaluation of approximation algorithms for probabilistic decoding," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, p. 455–463.

[26] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, p. 467–475.

[27] J. S. Yedidia, W. T. Freeman, Y. Weiss *et al.*, "Generalized belief propagation," in *NIPS*, vol. 13, 2000, pp. 689–695.

[28] Y. Weiss and W. Freeman, "On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 736–744, 2001.

[29] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[30] T. Heskes *et al.*, "Stable fixed points of loopy belief propagation are minima of the Bethe free energy," *Advances in neural information processing systems*, vol. 15, pp. 359–366, 2003.

[31] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on information theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[32] M. Welling, "On the choice of regions for generalized belief propagation," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, ser. UAI '04. Arlington, Virginia, USA: AUAI Press, 2004, p. 585–592.

[33] T. K. Moon and J. H. Gunther, "Multiple constraint satisfaction by belief propagation: An example using Sudoku," in *2006 IEEE Mountain Workshop on Adaptive and Learning Systems*, July 2006, pp. 122–126.

[34] J. E. Whitesitt, *Boolean Algebra and Its Applications*. Courier Corporation, 2012.

[35] A. Choi, D. Kisa, and A. Darwiche, "Compiling probabilistic graphical models using sentential decision diagrams," in *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, 2013, pp. 121–132.

[36] T. Huang, D. Koller, J. Malik, G. Ogasawara, B. S. Rao, S. J. Russell, and J. Weber, "Automatic symbolic traffic scene analysis using belief networks," in *AAAI*, vol. 94, 1994, pp. 966–972.

[37] O. Ronen, J. Rohlicek, and M. Ostendorf, "Parameter estimation of dependence tree models using the EM algorithm," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 157–159, 1995.

[38] M. Shwe and G. Cooper, "An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network," *Comput Biomed Res*, vol. 24, no. 5, pp. 453–475, Oct 1991.

[39] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 1st ed. USA: Academic Press, Inc., 2015.

[40] F. G. Cozman *et al.*, "Generalizing variable elimination in Bayesian networks," in *Workshop on probabilistic reasoning in artificial intelligence*. Citeseer, 2000, pp. 27–32.

[41] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[42] D. van Leeuwen, "Julia type that implements a drop-in wrapper for AbstractArray type, providing named indices and dimensions," 2021. [Online]. Available: https://github.com/davidavdav/NamedArrays.jl

[43] B. Stroustrup, A. Koenig, and B. E. Moo, "C++," *Encyclopedia of Software Engineering*, 2002.

[44] G. van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[45] C. Verburgh, "Parallelising inference on cluster graphs," Master's thesis, Stellenbosch: Stellenbosch University, 2020.

[46] A. T. Ihler, J. W. F. III, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *Journal of Machine Learning Research*, vol. 6, no. 31, pp. 905–936, 2005. [Online]. Available: http://jmlr.org/papers/v6/ihler05a.html

[47] J. Mooij and H. Kappen, "Sufficient conditions for convergence of loopy belief propagation," *arXiv preprint arXiv:1207.1405*, 2012.

[48] S. Streicher, W. Brink, and J. du Preez, "A probabilistic graphical model approach to the structure-and-motion problem," in *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, Nov 2016, pp. 1–6.

[49] C. F. de Villiers, "Bayesian signal processing of Doppler radar data," Master's thesis, Stellenbosch: Stellenbosch University, 2016.

[50] D. Brink, "Using probabilistic graphical models to detect dynamic objects for mobile robots," Ph.D. dissertation, Stellenbosch: Stellenbosch University, 2016.

[51] R. M. Lewis, *A guide to graph colouring: algorithms and applications*. Springer, 2015.

[52] E. K. Burke, D. G. Elliman, and R. Weare, "A university timetabling system based on graph colouring and constraint manipulation," *Journal of Research on Computing in Education*, vol. 27, no. 1, pp. 1–18, 1994.

[53] P. Briggs, "Register allocation via graph coloring," Ph.D. dissertation, Rice University, 1992.

[54] L. Kroc, A. Sabharwal, and B. Selman, "Counting solution clusters in graph coloring problems using belief propagation," in *Advances in Neural Information Processing Systems*, 2009, pp. 873–880.

[55] A. Braunstein, M. Mézard, and R. Zecchina, "Survey propagation: An algorithm for satisfiability," *Random Structures & Algorithms*, vol. 27, no. 2, pp. 201–226, 2005.

[56] E. N. Maneva, E. Mossel, and M. J. Wainwright, "A new look at survey propagation and its generalizations," *CoRR*, vol. cs.CC/0409012, 2004. [Online]. Available: http://arxiv.org/abs/cs.CC/0409012

[57] L. Kroc, A. Sabharwal, and B. Selman, "Survey propagation revisited," *CoRR*, vol. abs/1206.5273, 2012. [Online]. Available: http://arxiv.org/abs/1206.5273

[58] D. E. Knuth, *The Art of Computer Programming, Volume 4, Fascicle 6: Satisfiability*, 1st ed. Addison-Wesley Professional, 2015.

[59] R. Mateescu, K. Kask, V. Gogate, and R. Dechter, "Join-graph propagation algorithms," *Journal of Artificial Intelligence Research*, vol. 37, pp. 279–328, 2010.

[60] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[61] C. Hughes, "Project Euler," 2012, accessed: 2017-07-03. [Online]. Available: https://projecteuler.net

[62] Sterten, "Sudoku dataset," accessed: 2020-06-29. [Online]. Available: https://web.archive.org/web/20200629023120/magictour.free.fr/top95

[63] O. Arino, J. J. Ramos Perez, V. Kalogirou, S. Bontemps, P. Defourny, and E. Van Bogaert, "Global land cover map for 2009 (GlobCover 2009)," l' European Space Agency (ESA) & Université catholique de Louvain (UCL), 2012. [Online]. Available: https://doi.org/10.1594/PANGAEA.787668

[64] J. Chen, J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu, W. Zhang, X. Tong, and J. Mills, "Global land cover mapping at 30m resolution: A POK-based operational approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, pp. 7–27, 2015.

[65] C. Giri, B. Pengra, J. Long, and T. R. LoveLand, "Next generation of global land cover characterization, mapping, and monitoring," *International Journal of Applied Earth Observation and Geoinformation*, vol. 25, no. 1, pp. 30–37, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jag.2013.03.005

[66] Y. Ban, P. Gong, and C. Giri, "Global land cover mapping using Earth observation satellite data: Recent progresses and challenges," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, no. February, pp. 1–6, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0924271615000131

[67] J. Chen, I. Dowman, S. Li, Z. Li, M. Madden, J. Mills, N. Paparoditis, F. Rottensteiner, M. Sester, C. Toth, J. Trinder, and C. Heipke, "Information from imagery: ISPRS scientific vision and research agenda," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S092427161500218X

[68] L. Martino and M. Fritz, "New insight into land cover and land use in Europe," *Statistics in Focus*, vol. 3, 2008.

[69] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 185–201, 2002.

[70] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[71] M. Pal and P. Mather, "Support vector machines for classification in remote sensing," *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007–1011, 2005.

[72] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, p. 129, 2018.

[73] S. Fritz, I. McCallum, C. Schill, C. Perger, R. Grillmayer, F. Achard, F. Kraxner, and M. Obersteiner, "Geo-wiki.org: The use of crowdsourcing to improve global land cover," *Remote Sensing*, vol. 1, no. 3, pp. 345–354, 2009. [Online]. Available: http://www.mdpi.com/2072-4292/1/3/345

[74] A. Pérez-Hoyos, F. García-Haro, and J. San-Miguel-Ayanz, "A methodology to generate a synergetic land-cover map by fusion of different land-cover products," *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, pp. 72–87, oct 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243412000906

[75] S. Gengler and P. Bogaert, "Integrating crowdsourced data with a land cover product: A bayesian data fusion approach," *Remote Sensing*, vol. 8, no. 7, p. 545, 2016. [Online]. Available: http://www.mdpi.com/2072-4292/8/7/545

[76] B. Chen, B. Huang, and B. Xu, "Multi-source remotely sensed data fusion for improving land cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 124, pp. 27–39, feb 2017.

[77] M. Brovelli, M. Molinari, E. Hussein, J. Chen, and R. Li, "The first comprehensive accuracy assessment of GlobeLand30 at a national level: Methodology and results," *Remote Sensing*, vol. 7, no. 4, pp. 4191–4212, 2015.

[78] L. Aune-Lundberg and G.-H. Strand, *CORINE Land Cover 2006. The Norwegian CLC2006 project*. Norsk institutt for skog og landskap, 2010.

[79] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *Pervasive Computing*, vol. 7, no. 4, pp. 12–18, Oct. 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4653466&tag=1

[80] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sensing of Environment*, vol. 86, no. 4, pp. 554–565, 2003.

[81] Y. Shao and R. S. Lunetta, "Comparison of support vector machine, neural network, and cart algorithms for the land-cover classification using limited training data points," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 70, pp. 78–87, 2012.

[82] C. Huang, L. Davis, and J. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[84] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2016.

[85] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[86] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv preprint arXiv:1508.00092*, 2015.

[87] A. Pérez-Hoyos, F. García-Haro, and J. San-Miguel-Ayanz, "A methodology to generate a synergetic land-cover map by fusion of different land-cover products," *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, pp. 72 – 87, 2012.

[88] L. G. El-Deen Taha, "Classifier ensemble for improving land cover classification," *International Journal of Circuits, Systems and Signal Processing*, vol. 10, 2016.

[89] H. Griethe and H. Schumann, "The visualization of uncertain data: Methods and problems," *Proceedings of SimVis '06*, vol. vi, no. August, pp. 143–156, 2006.

[90] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006. [Online]. Available: http://dx.doi.org/10.1111/j.2006.0030-1299.14714.x

[91] C. D. Correa, Y. H. Chan, and M. Kwan-Liu, "A framework for uncertainty-aware visual analytics," *VAST 09 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings*, pp. 51–58, 2009.

[92] L. Zhang and Q. Ji, "Image segmentation with a unified graphical model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1406–1425, Aug 2010.

[93] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[94] B. Schiele, "Lecture notes in probabilistic graphical models and their applications," November 2017. [Online]. Available: https://www.mpi-inf.mpg.de/fileadmin/inf/d2/GM/2017/gm-2017-1120-imageprocessing.pdf

[95] B. Ghimire, J. Rogan, V. R. Galiano, P. Panday, and N. Neeti, "An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA," *GIScience & Remote Sensing*, vol. 49, no. 5, pp. 623–643, 2012.

[96] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970. [Online]. Available: http://www.jstor.org/stable/143141

[97] B. Sun, X. Chen, and Q. Zhou, "Uncertainty assessment of GlobeLand30 land cover data set over Central Asia," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B8, pp. 1313–1317, 2016. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLI-B8/1313/2016/

[98] J. J. Arsanjani, L. See, and A. Tayyebi, "Assessing the suitability of GlobeLand30 for mapping land cover in Germany," *International Journal of Digital Earth*, vol. 9, no. 9, pp. 873–891, 2016.

[99] E. Chuprikova, L. Liebel, and L. Meng, "Towards seamless validation of land cover data," *Proceedings. ICC 2017 - International Cartography Conference*, pp. 1–10, 2017.

[100] H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on OpenStreetMap," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 700–719, 2014.

[101] L. Mosley, "A balanced approach to the multi-class imbalance problem," Ph.D. dissertation, Iowa State University, 2013, accessed: 2019-05-21. [Online]. Available: https://web.archive.org/web/20190502134555/lib.dr.iastate.edu/cgi/viewcontent.cgi?article=4544&context=etd

[102] H. Veregin, "Defining data quality," *Geographical Information Systems*, vol. 1, pp. 177–189, 1999.

[103] M. R. Boutell, J. Luo, and C. M. Brown, "Factor graphs for region-based whole-scene classification," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, pp. 104–104.

[104] L. Zhao, Z. Li, C. Men, and Y. Liu, "Superpixels extracted via region fusion with boundary constraint," *Journal of Visual Communication and Image Representation*, vol. 66, p. 102743, 2020.

[105] T. W. Neller and Z. Luo, "Mixed logical and probabilistic reasoning in the game of Clue," *ICGA Journal*, vol. 40, pp. 406–416, 2018, 4. [Online]. Available: https://doi.org/10.3233/ICG-180063

[106] M. Mulamba, J. Mandi, R. Canoy, and T. Guns, "Hybrid classification and reasoning for image-based constraint solving," in *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*. Springer, 2020, pp. 364–380.

[107] L. E. Sucar, "Probabilistic graphical models," *Advances in Computer Vision and Pattern Recognition.*, vol. 10, pp. 978–1, 2015.

[108] J. Zeng, "A topic modeling toolbox using belief propagation," *Journal of Machine Learning Research*, vol. 13, no. Jul, pp. 2233–2236, 2012.

[109] A. Mackworth, "Consistency in networks of relations," *Artificial Intelligence*, vol. 8, no. 1, pp. 99–118, 1977.

[110] J. Laurière, *Éléments de programmation dynamique*, ser. Recherche opérationnelle appliquée 3. Gauthier-Villars New York, 1979.

[111] A. Cesta, S. Fratini, and A. Oddi, "Planning with concurrency, time and resources: A CSP-based approach," in *Intelligent Techniques for Planning*. IGI Global, 2005, pp. 259–295.

[112] R. G. Gallager, *Information Theory and Reliable Communication*. Springer, 1968, vol. 2.

[113] D. Berthier, *Pattern-Based Constraint Satisfaction and Logic Puzzles*. Lulu Press, 2013, arXiv preprint arXiv:1304.1628.

[114] L. Ananthagopal, "Application of message passing and sinkhorn balancing algorithms for probabilistic graphical models," MSc dissertation, San Jose State University, 2014.

[115] G. Elidan, I. McGraw, and D. Koller, "Residual belief propagation: Informed scheduling for asynchronous message passing," *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*, 2006.

[116] C. Lecoutre, "Constraint networks: Techniques and algorithms," 2009.

[117] Champagne, "The hardest Sudokus," 2019. [Online]. Available: http://forum.enjoysudoku.com/the-hardest-sudokus-new-thread-t6539.html

[118] E. Russell and F. Jarvis, "Mathematics of Sudoku II," *Mathematical Spectrum*, vol. 39, no. 2, pp. 54–58, 2006.

[119] A. Biere, "Picosat essentials," *Journal on Satisfiability, Boolean Modeling and Computation (JSAT)*, 2008.

[120] "Google OR-Tools," 2019, accessed: 2021-09-20. [Online]. Available: http://developers.google.com/optimization

[121] Streicher, "Purge-and-merge GitHub," 2019, accessed: 2021-09-20. [Online]. Available: http://github.com/heetbeet/purge-and-merge

[122] G. Royle, "Minimum Sudoku," accessed: 2091-08-21. [Online]. Available: https://web.archive.org/web/20190821092327/staffhome.ecm.uwa.edu.au/~00013890/sudokumin.php

[123] B. Felgenhauer and F. Jarvis, "Mathematics of Sudoku I," *Mathematical Spectrum*, vol. 39, 01 2006.