

Imperial College London
Department of Computing

Tracking in Information Space

Helen Elizabeth Greenhough

Supervised by Professor Chris Hankin

27 February 2019

Submitted in part fulfillment of the requirements
for the degree of Doctor of Philosophy

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

I declare that this thesis represents my own original work.

Where I draw upon the work of others, I reference appropriately.

Acknowledgements

I have had the pleasure of receiving the support of a number of brilliant people throughout the course of my research. I would like to express my sincerest thanks to: my supervisor Professor Chris Hankin for taking this journey with me. I could not have asked for a better supervisor. My second supervisor Professor Michael Huth. A brilliant Goose to Chris' Maverick. Thank you for your early involvement and introductions. My sponsor for making this PhD possible and all your insight and pointers over the last three years. My family and friends with a special thank you to my parents for your invaluable checking of my Punctuation, Grammar and Spelling (PuGS). I promise I won't ask you to proof read anything¹ again.

To esteemed faculty members for their practical help in my research (in alphabetical order by surname), Professors: John Hassard (Imperial College London), Averil MacDonald (University of Reading), Ron Matthews (Defence Academy of the United Kingdom), Bill Proud (Imperial College London) and Francesca Toni (Imperial College London).

To Petras Saduikis for your valuable and timely coding support.

To my Imperial colleagues and friends: the team at the Institute for Security Science and Technology, in particular Denise McGurk for all of your assistance navigating Imperial systems; Angela Goldfinch (Department of Computing librarian) for your fast and positive responses to my queries; Luís Piña for the many hours you spent assisting me with the current landscape of virtual machines, terminal and coding and your constant un-sticking of my stucked-ness; Tom

¹quite as long^a

^a(maybe)

McGrath for helping me navigate Tensor Decomposition and for your work with the Imperial Maths Help Desk; Andreas Mattavelli for your assistance un-breaking Latex and pseudo humour; Nouredin Sadawi for helping me to get up and running with Anaconda and Amazon web services; Bruce Garvey for introducing me to the world of Morphological Analysis and Fibonacci; my fellow PhD students Jonathan Gudgeon and Shireen Seakhoa-King for your emotional support and sharing of wine and PhD specific *life-hacks*; Oana Cocarascu for assistance with my Machine Learning code; Dr Paul Beaumont for help with notation and to my *twinned* PhD students Martina Patone and Paul Gilbert at Southampton University I wish you all the best in your endeavours.

During the course of my PhD I was pleasantly surprised by how ready complete strangers were to help: Arthur Weiss for providing your materials regarding open sources; Bob Schrag for giving me access to the argumentation software FUSION; Adam Funk (University of Sheffield) for providing the files for the Semantic Pattern Recognition and Annotation Tool (SPRAT); Jacob Schreiber for your quick assistance with Pomegranate; Aaron Schein for your help with the Bayesian Poisson Tucker Decomposition code; Adrian Groza and Oana Popa (Technical University of Cluj-Napoca) for your insight regarding the General Architecture for Text Engineering (GATE) tool for Information Extraction; an individual (who does not wish to be publicly attributed) for your timely and valuable proof reading support. I reserve a special mention for Sutanay Chaudary (of Pacific Northwest National Laboratory) and Kalev Leetaru (of the Global Data on Events, Location and Tone (GDELT) Project) for becoming so involved through the course of my research and providing so much guidance on the realities of Natural Language Processing.

A final thank you to all the people who have taken the time to post to github, stackoverflow, youtube, coursera and udemy and the many developers adding to coding libraries and backing them up with supporting resources.

Abstract

A long-term ambition of researchers is to develop readily available *expert systems* to support decision makers. There is great interest in combining web-based information and Machine Learning techniques, which are enjoying a renaissance attributed to the ready availability of computing resources, to answer real-world strategic questions. Given the contribution of industry to overall national economic health, one such question is what industry mix is necessary to support national strategy? The proposed conceptual graph model and accompanying framework, applied the strengths of machines and computational techniques to inform national industrial policy in a data-driven paradigm. This thesis makes three contributions: the development of a novel conceptual graph model and associated framework, both based on a multi-disciplinary literature review, and their subsequent demonstration using both simulated binary and real-world statistical data. The graph model, drawing from experience in the government defence sector, linked national intent to industry via intermediate nodes such as capability. The wider framework combined supervised Bayesian Belief Network Machine Learning with web-scraping via the graph model to infer national intent. Proof of concept was achieved through the use of both simulated and real-world data, confirming the ability of the chosen inference approach to learn the required relations. Ultimately the framework was able to identify industry variables which best explain observed capability for two real-world case studies although it failed to draw inferences across industries using real data. Further avenues of investigation were addressed, including the use of unstructured text as a data source and alternative Machine Learning methods to account for time-series data. The potential of the web as a data source is truly staggering. This thesis goes some way to bridging the gap between this vast store of information and its incorporation in an evidence-driven pipeline in support of strategic decision making.

Contents

Acknowledgements	3
Abstract	4
List of Figures	11
List of Tables	15
Notation	15
1 Introduction	19
1.1 Research Aim and Novelty	19
1.2 Motivation and Impact	19
1.3 Contribution	20
1.4 Thesis Structure	21
2 State of the Art	25
2.1 Introduction	25
2.2 Exploiting Web Information for Decision Support	25
2.3 National Intent and Industry	32
2.4 Conclusion	34
3 Conceptual Graph Model	35
3.1 Introduction	35
3.2 Core Ideas	35
3.2.1 Strategic Industry and Industrial Policy	35

3.2.2	Capability	45
3.2.3	Intent	46
3.3	Conceptual Graph Model	50
3.3.1	Graph Model	50
3.3.2	Representation of Time	54
3.3.3	Graph Adjacency Matrix	54
3.4	Conclusion	55
4	Research Framework	57
4.1	Introduction	57
4.2	Core Ideas	57
4.2.1	Web Content	57
4.2.2	Information Exploitation	60
4.2.3	Pomegranate: BBN ML Algorithm	67
4.3	Research Framework	70
4.4	Conclusion	72
5	Framework Demonstration 1 & 2	73
5.1	Introduction	73
5.2	Demonstration 1: Simulated Data	75
5.2.1	Method	76
5.2.2	Results	83
5.2.3	Discussion and Conclusion	91
5.3	Demonstration 2: Web Scraping	93
5.3.1	Method and Results: OECD Input Data	93
5.3.2	Method and Results: OECD Target Data	96
5.3.3	Method and Results: IEA Target Data	97
5.3.4	Discussion and Conclusion	98
5.4	Demonstration 2: Case Study 1	100
5.4.1	Method	100

5.4.2	Results	104
5.4.3	Discussion and Conclusion	107
5.5	Demonstration 2: Case Study 2	110
5.5.1	Method	110
5.5.2	Results	111
5.5.3	Discussion and Conclusion	115
6	Extensions	117
6.1	Introduction	117
6.2	Extension 1: Treatment of Time	117
6.2.1	DBN	117
6.2.2	LSTM Networks	118
6.3	Extension 2: Text as a Data Source	120
6.3.1	The IE Task on Natural Language Text	120
6.3.2	Ontologies	122
6.3.3	Application to Research Framework	124
7	Discussion and Conclusion	127
7.1	Summary of Achievements	127
7.2	Research Aims and Key Contributions	128
7.3	Future Work	133
	Bibliography	134
	Appendices	153
A	Pseudocode	155
B	Copyrights	163

List of Figures

3.1	Examples of modeling a capability in accordance with MoDAF. Figures are adapted from public sector information licensed under the Open Government Licence v3.0. [MoD, 2010; HMG, 2018]	47
3.2	The conceptual overview for DM2 diagrams, illustrating the relationship between capability and material via performing resources as defined by the DM2 ontology. No copyright restrictions. [McDaniel, 2011]	48
3.3	Illustration of the novel conceptual graph model, a key contribution of this thesis. Solid lines indicate an observable relationship. Dashed lines indicate an inferred relationship.	51
3.4	Two representations of a capability mapped to constituent products. On the left, intermediate nodes of components and sub-components are shown. On the right, the direct mapping of capability to products is shown.	52
4.1	Trivial example of a BBN showing grass (G), weather (W) and sprinkler (S) nodes.	63
4.2	Illustration of ML basic pipeline, showing: different types of data, transformation into required format, split into training and test sets for ingest by learning algorithm, producing a trained model which is then evaluated using the test data. Contains images under the creative commons licence [Nielsen, 2017; Wikispaces, 2018; Karpathy, 2018; Garg, 2015].	66
4.3	Pomegranate's BayesianNetwork.from_samples function, showing input parameters.	67
4.4	Pomegranate's BayesianNetwork.predict function, showing input parameters.	69
4.5	Top level Research Framework architecture as a pipeline diagram.	70

4.6	Top level ML architecture for model development. Showing use of training and test datasets to train the model and evaluate its performance.	71
5.1	Scope of first demonstration in relation to complete Research Framework.	74
5.2	Scope of second demonstration in relation to complete Research Framework.	75
5.3	Simulated data high-level method and detail of Process.	77
5.4	Constraint graph for example where $ Ind = Cap = 3$. A BBN created using this constraint would only have edges $\langle Ind_i, Cap_c \rangle$	77
5.5	Detail of transform data process. A dotted line indicates an optional process.	79
5.6	Detail of evaluate process, illustrating the method for comparing the model's prediction for missing data with the correct answers for four performance metrics.	81
5.7	Adapted example output of Pomegranate's BayesianNetwork.from_samples algorithm, a BBN with its corresponding performance. An edge represents a 1 in the corresponding adjacency matrix. The performance metrics of precision, recall and F_1 by binary category for the prediction of the Cap_3 node is shown. 1.00 is the maximum value for these metrics. Note the high performance due to the low $ Ind $	82
5.8	Plot of average ratio split vs average percentage correctness of returned model. Colour indicates $ Nat $, size indicates $ Cap $, and shape indicates $ Ind $. Data is filtered, showing the lower and higher values of $ Ind $. Each point is the average value of 30 runs.	85
5.9	Plot of $ Ind $, $ Cap $ vs average percentage correctness of returned model. Colour indicates $ Nat $ in the line graph. Data is filtered, showing the lower values of $ Cap $ as the pattern continues for higher values. The distribution of results shows there is no significant difference as $ Cap $ changes.	86
5.10	Plot of performance metrics average percentage correct and average F_1 vs $ Ind $ which is in turn related to ratio split. The F_1 metric tends to 1 as ratio split tends to 0, and percentage correct tends to 0, making F_1 a poor choice to evaluate performance for the simulated data case.	87

5.11 Heat map showing average correctness of returned model for a given percentage sparsity and noise.	88
5.12 Graph showing average correctness of returned model for a given percentage sparsity and noise.	89
5.13 Graph showing standard deviation of average correctness for three values of percent sparsity.	90
5.14 OECD data scraping high level method.	94
5.15 Stages of retrieving names of available datasets.	95
5.16 Result of extracting a database via an API query.	96
5.17 Result of parsing extracted database.	97
5.18 Filtered IEA data downloaded as a csv file from UK Data Service website. . . .	98
5.19 IEA data after pre-processing.	98
5.20 Demonstration 2 Method.	101
5.21 Case Study 1 example model and performance output for an OECD dataset . .	106
5.22 Understanding correlation between industry variables and ratio.	109
5.23 Case Study 2 example model and performance output for an OECD dataset. . .	113
6.1 A static and dynamic n -order form of a BBN.	118
6.2 LSTM and memory cell.	119
6.3 Example of extending triple to n -ary relations via use of intermediate class node with multiple slots.	124
6.4 Conceptual Graph Model rendered as an ontology.	125
A.1 Top level code investigating effect of variables on performance.	156
A.2 Algorithm for creating constraint graph.	157
A.3 Algorithm for creating simulated data.	157
A.4 Algorithm for adding noise and sparsity to data.	158
A.5 Algorithm for splitting data.	158
A.6 Algorithm for calculating percentage correctness.	159
A.7 Parsing response r to extract English natural language and codified names. . . .	160

A.8	Iterating through available databases, filtering databases based on keywords and extracting using API query.	161
B.1	Permission for use of DM2 image, via email, from D. McDaniel c/o B. McDaniel.	164
B.2	Request for permission for use of neural net image, via email, to A. Karpathy. .	165
B.3	Permission for use of spider chart image, via email, from L. Garg.	166
B.4	Permission for use of LSTM cell image, via email, from C. Olah.	167
B.5	Request for permission for use of n-ary relation, via email, to W3C.	168

List of Tables

1.1	Relation between thesis contribution listed at Section 1.3 and Chapters 3 to 5. .	21
4.1	Weather state probability table	63
4.2	Sprinkler state probability table	64
4.3	Grass state probability table	64
4.4	Probability calculations	64
5.1	Sections describing data extraction methods.	93
5.2	Use of data sources for Demonstration 2 Case Studies.	93
5.3	Case Study 1 model performance.	105
5.4	Case Study 2 model performance. Note that the performance is for a single model, with the exception of Parents which is the average performance and standard deviation of 102 models.	112
B.1	Copyright summary for images	163

Notation

Sets and Logic

Unless otherwise stated, a set is indicated by an upper case letter, A , or as a three letter camel case code, Abc .

Unless otherwise stated, members of sets are enclosed by curly parenthesis and given by lower case letters: $A = \{a_1, a_2, \dots, a_n\}$ and $Nat = \{nat_1, nat_2, \dots, nat_3\}$.

Set membership: $a_1 \in A$. Element a_1 is a member of set A .

Subset: $A \subseteq B \Leftrightarrow \forall x(x \in A \Rightarrow x \in B)$. Set A is a subset of set B if and only if members x of A are also members of B .

Proper subset: $A \subset B \Leftrightarrow A \subseteq B \wedge A \neq B$. Set A is a proper subset of B if and only if A is a subset of B and A does not equal B .

The number of members in a set (cardinality) is given by $|A|$.

Negate: $\neg a$.

Statistics

Probability of observing event A given event B has occurred: $P(A \mid B)$.

Expected value of A: $E(A)$

Graphs, Vectors and Matrices

The name for an edge is given by lower case italicised letters, sometimes with an underscore:

has, *is_a*.

The graph adjacency matrix for two nodes with states from sets A and B is given by:

$$\mathbf{AB} = \begin{matrix} & \begin{matrix} b_1 & b_2 & \dots & b_b \end{matrix} \\ \begin{matrix} a_1 \\ a_2 \\ \dots \\ a_a \end{matrix} & \begin{pmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_b \\ a_2b_1 & a_2b_2 & \dots & a_2b_b \\ \dots & \dots & \dots & \dots \\ a_ab_1 & a_ab_2 & \dots & a_ab_b \end{pmatrix} \end{matrix}$$

A directed graph can be described by a set of triples which defines which nodes are connected by an edge. The convention is the direction of the edge goes from left to right: $\langle node_1, node_2 \rangle$.

A vector is given by a bold lower case letter: $\mathbf{x}$

A matrix is given by a bold upper case letter: $\mathbf{X}$

The size of a matrix is indicated by subscript, the first value giving the number of rows, r , and the second value giving the number of columns, c : $\mathbf{X}_{r \times c}$.

General

Product function: $\prod_{i=1}^n a_i = a_1 * a_2 * \dots * a_n$

Sum function: $\sum_{i=1}^n a_i = a_1 + a_2 + \dots + a_n$

Chapter 1

Introduction

1.1 Research Aim and Novelty

This research develops and demonstrates a flexible framework to infer and track the strategic intention of nations from web-sourced observations of their industrial bases. The framework encompasses a novel conceptual graph model relating industry to national intent via intermediate nodes. The graph model and wider framework draw upon experience in the government defence sector, in particular defence capability development and intelligence analysis.

As an example of a system exploiting web information in support of decision making, it is a modest contribution when compared to the large programmes detailed in Chapter 2, but it does stand out as a rare example of applying web scraping and reasoning in support of a strategic level decision. Further it is potentially a unique application of modern computational techniques to the task of systematically evaluating national intent from industry observations in a data-driven paradigm.

1.2 Motivation and Impact

Decision making is ubiquitous. Decisions can be small in terms of having limited impact while others are large. When impact is significant, decision makers attempt to invest resources in order to make informed decisions. However in a complex system, such as our world, compromises

are made and the resulting decisions are often suboptimal. A long-term ambition of researchers is to develop readily available *expert systems*¹ to support decision makers. In recent years the ready availability of computer storage and processing has facilitated both the explosion of web-based information and advances in applications of Machine Learning (ML) techniques. Given this, there is great interest in combining these two resources to answer real-world strategic questions.

Given the contribution of industry to overall national economic health, one such question is what industry mix is necessary to support national strategy? More specifically, can web-based information, combined with supervised ML techniques, exploit the collective global (historical) experience to robustly inform national industrial policy?

The UK government recognises its role in supporting economic growth through industrial strategy with the stated commitment to ‘position the UK at the forefront of the industries of the future’ and to make the UK ‘the world’s most innovative economy’ [HMG, 2017, p. 23, 9]. This research has the potential to inform UK industrial strategy that aims to grow the national economy, which was worth approximately £2 trillion in 2017. Further, it can be employed by other governments and trans-national economic forums such as the G8, G12 and the Office for Economic Cooperation and Development (OECD) which together were responsible for 49%, 53% and 63% of the \$75 trillion global economic output in 2016. [ONS, 2018; WB, 2018]

1.3 Contribution

Taking a graph-based approach relating industry to national intent, Bayesian Belief Network (BBN) ML provides a basis for developing a predictive model. This model is trained on historical industrial data scraped from the web and its performance evaluated. As such, this thesis represents a practical application of modern computation methods to exploit web-based

¹Computer programs able to emulate a human expert to make inferences based on stored knowledge of a given topic [Kendal and Creen, 2007].

information in support of a real-world strategic problem. Its three significant contributions are:

1. The development of a novel conceptual graph model abstracting the problem space, based on a multi-disciplinary literature review.
2. The development of an accompanying flexible framework (graph model + web scraping + ML), hereafter referred to as the Research Framework or Framework.
3. The demonstration of the Framework, specifically:
 - BBN ML approach applied to simulated data.
 - BBN ML approach applied to two real-world case studies using statistical data scraped from the web.
 - Case Study 1: relation between industry and renewable electricity generation as a proportion of total electricity generation.
 - Case Study 2: relation between industry and receipt of international aid.

In addition this thesis expands upon two potential routes for further research: ML techniques that account for time-series data and the use of text as a data source combined with ontological techniques.

1.4 Thesis Structure

This thesis is based on three substantive chapters each addressing a contribution as shown at Table 1.1.

Table 1.1: Relation between thesis contribution listed at Section 1.3 and Chapters 3 to 5.

Thesis contribution	Chapter
1. Novel conceptual graph model	3
2. Research Framework	4
3. Framework demonstration	5

Chapter 1 Describes the research aim and novelty, its motivations, anticipated impact, thesis contributions and, as fulfilled here, the structure of the thesis.

Chapter 2 Contextualises this work with respect to literature addressing the exploitation of web information in support of decision makers and inference of national intent from industrial bases. The concept of *strategic* industry, the idea that nations value some industries over others and act favourably towards them, is thoroughly explored in this Chapter as it forms a central theme of this thesis.

Chapter 3 Describes the abstraction of the research problem as a novel conceptual graph model based on a multi-disciplinary literature review addressing key concepts such as strategic industry, capability and intent, which is also documented.

Chapter 4 Describes the developed Research Framework based on an investigation into the exploitation of web-based information.

Chapter 5 Documents the elements of the Research Framework that are demonstrated, the methods used, the results obtained and brief discussion of the individual demonstration results. The chapter is structured as follows:

- Demonstrated Framework: Summary of the elements of the Research Framework demonstrated in this thesis.
- Demonstration 1: Simulated data. Here the method, results and discussion of the BBN ML technique applied to simulated data is presented.
- Demonstration 2: Web scraping statistical data. Here the method, results and discussion for extracting web-based OECD and International Energy Agency (IEA) data is presented.
- Demonstration 2: Case study 1. Here the method and results obtained by a BBN ML technique on the relation between national energy industry and renewable electricity generation, as a ratio of total electricity generation is presented.
- Demonstration 2: Case study 2. As above but examining the relation between national industry and receipt of international aid as a ratio of Gross Domestic Product (GDP) per Capita.

Chapter 6 Two routes for further research are discussed. First, the examination of the use of alternative ML techniques that account for time-series data. Second, the use of text data as an input to the Research Framework combined with an ontological approach within the Framework.

Chapter 7 Brings together contributions from each chapter and the overarching thesis discussion and conclusions are presented.

Chapter 2

State of the Art

2.1 Introduction

This thesis applies computational techniques to infer the relationship between national industry and intent by means of a conceptual graph model. No examples were found in the literature addressing this particular challenge. As a consequence, in order to contextualise this thesis and draw comparisons with other work, it is necessary to broaden the scope of this review. The overarching theme of this thesis is the exploitation of web information in support of decision makers. This is further broken down for the purpose of this review into web mining, reasoning and decision support. The second theme examined is the inference of national intent from industrial bases.

2.2 Exploiting Web Information for Decision Support

This is an active area of research and encompasses multiple challenges. Of most relevance to this thesis is work that addresses the tasks of web mining, reasoning (including use of ML) and decision support. To this end three combinations of four tasks are examined: end-to-end frameworks, reasoning for decision support, ML for decision support and web (text) mining. Where possible, the review addresses work with a defence and security application as these were felt most analogous to this thesis' proposed Framework.

End-to-End Frameworks The following are examples of pipelines that use web-based information and reasoning techniques in support of decision making in an end-to-end system and in that sense describe the work that is most similar to this thesis, although the applications and aims of each system are distinct.

DEFENDER *Syndromic surveillance* aims to provide early warning to administrative bodies in order to better respond to disease outbreaks. The use of web-sourced data for these systems has received recent interest. A representative example of such a system is DEFENDER, a symptom tracking tool, that takes as primary input live stream of geolocated Twitter data and a daily Rich Site Summary (RSS) download of news data, both publicly available data sources. The Twitter data is used in two ways. The first is the geo and temporal tracking of tweets matched to symptoms. The second is to form a network where the nodes represent clusters of users by geographic location, and the weighted edges represent the movement of Twitter users between two nodes within a 24 hour period. The result is a graph database where nodes are characterised by symptom counts and edges by daily movement. The two main algorithms in DEFENDER for decision support are the Early Aberration Reporting System syndromic surveillance algorithm for predicting a disease outbreak, and Autoregressive Integrated Moving Average (ARIMA) regression for nowcasting numbers of presented cases for four diseases and forecasting the numbers of symptom related tweets by geographic node [Thapen et al., 2016].

ICEWS Another example of an early warning system is the Department of Defense (DoD) Integrated Crisis Early Warning System (ICEWS) which focuses on crisis prediction to allow early response by administrative bodies. The ICEWS seeks to predict crisis at the subnational through to international scales to support resource allocation decisions. The aim of the system is to forecast five types of events over a three month period, where events include domestic political crisis and insurgency. It consists of five separate forecasting models which are then integrated using a sixth model, a BBN, that is understood to have been developed using supervised ML although this is not immediately clear from the literature. Several data sources are used, including extensive Subject Matter Expert (SME) elicitation. The data source of interest here is a news corpus collected from the surface web which was parsed automatically into an events

database in accordance with a standard political science coding system [O’Brien, 2010].

OSI Work sponsored under the auspices of the United States (US) Office of the Director of National Intelligence’s Intelligence Advances Research Projects Activity (IARPA) Open Source Indicators (OSI) program uses public data such as news feeds, blogs and web search queries to forecast events like civil unrest, outbreaks of disease, political and humanitarian crises, mass migration, economic instability and resource shortages. Results include systems to predict civil unrest using microblogging site Tumblr data, model online support for terror groups to predict large-scale violent events, and predict election outcomes using Twitter. [Gaurav, Srivastava, Kumar, and Miller, 2013; Xu, Lu, Compton, and Allen, 2014; Johnson, Zheng, Vorobyeva, Gabriel, Qi, Velasquez, Manrique, Johnson, Restrepo, Song, and Wuchty, 2016]

Memex The US Defense Advanced Research Projects Agency (DARPA) Memex programme goal is to counter human trafficking which typically uses the deep web to advertise services and recruit victims. Memex uses, amongst other tools, Stanford University’s DeepDive system (described in more detail later in this section) to extract relevant content from unstructured web data, specifically online advertisements for factory and sex workers. [Shen, 2016a,b; Stanford, 2016b]

Watson IBM’s Watson system performs a diverse range of tasks including chatbot building and converting audio to text. The application of interest here is decision support using an automatically constructed ontology. Watson has been applied to a number of decision domains but the area garnering much interest is its application to healthcare and government. Examples include cancer treatment recommendations, drug discovery and crime prediction. Watson uses Natural Language Processing (NLP) and ML to interpret unstructured data, such as web data, via an ontology and uses several hundred different question-answer algorithms to develop the most likely solution to a user’s question. Watson relies on Knowledge Graph (KG) YAGO, developed at the Max Planck Institute for Informatics which itself is an ontology developed from NLP of web data. [Barbosa, Wang, and Yu, 2013; Bielza and Larranaga, 2014; IBM, 2016; Stanford, 2016b; Somashekhar, Seplveda, Puglielli, Norden, Shortliffe, Kumar, Rauthan,

Kumar, Patil, Rhee, and Ramya, 2018; IBM, 2017, 2018].

Comparison The described frameworks were undertaken as part of large-scale (multi-year, multi-team) efforts, demonstrating the significant interest in exploiting web information in support of decision making. The primary difference between this body of research and this thesis is the nature of the decision being supported. The above examples principally address tactical decisions, that is decisions that need to be taken and/or acted on in a short time-frame, potentially over a small geographical area. The aim of the Research Framework aims to support a strategic decision, one that can be taken/acted upon over a longer time-frame at (inter)national scales. Another difference is that typically the work found focused on particular types of web sources such as social media or news feeds. This is thought to be due to two reasons. The first being the desire to make contributions to the state-of-the-art in parsing web sources with specific challenges, i.e. noisy tweet content. The second is it by default de-scopes the challenge of quality. While the demonstration of this Framework is limited to two web sources, the Framework is intended to be source agnostic. Addressing quality in this paradigm remains an outstanding challenge.

Reasoning for Decision Making The following are examples of work applying mathematical reasoning methods in support of decision making¹.

HMM Using a Hidden Markov Model (HMM), Vasanthan Raghavan, Aram Galstyan and Alexander Tartakovsky, predict the state of two terror groups over time to support decisions on counter-terror measures. To this end the work sets out to predict peaks and troughs of terror group activity. The data came from open source databases on terror activity. The researchers assume the observed terror attacks can be explained by unobserved internal group dynamics and thus apply a two-state HMM, representing the simplest non-trivial case, where the group states correspond to active and inactive. They found that their model performed on par with a baseline model but predicted the time of the next attack more accurately. [Raghavan, Galstyan, and Tartakovsky, 2013]

¹Quite by accident, the majority of examples are terror related.

BBN Using a BBN, Mohammed Olama, Glenn Allgood, Kristen Davenport and Jack Schryver look at the same challenge, again with a focus on developing a predictive system. Rather than the status of a terror group, the model is more akin to a risk assessment and outputs the ‘relative threat of an attack against a particular asset by a particular threat’ [Olama, Allgood, Davenport, and Schryver, 2010, p.3]. The structure and parameters of the BBN are based on SME solicitation. The group then apply the model to simulated input data. The group find the strength of the approach is the BBN’s ability to represent uncertainty, a typical feature of data. [Olama, Allgood, Davenport, and Schryver, 2010]

Information Fusion Rather than focusing on a particular scenario, others address the more general challenge of fusing heterogeneous information as part of an assessment framework. Using argumentation, Alice Toniolo et al. apply an adaptation of Praken’s argumentation scheme in order to compare hypotheses by structuring evidence. This work is set within a wider framework that enables collaboration between multiple users to form hypotheses and gather evidence, evaluate its credibility and capture its meta-data. FUSION is another framework, developed by Robert Schrag et al, that also addresses the fusion of heterogeneous information but this time combining argumentation with BBN in order to represent the uncertainty associated with evidence credibility and the strength of the supporting/refuting evidence. In both cases the focus of the work is the development of a Graphical User Interface (GUI) for Intelligence Analysts. [Toniolo, Norman, Etuk, Cerutti, Ouyang, Srivastava, Oren, Dropps, Allen, and Sullivan, 2015; Schrag, McIntyre, Richey, Laskey, Wright, Kerr, Johnson, Ware, and Hoffman, 2017]

Comparison The above work is similar with respect to this thesis in that they each apply a reasoning framework in support of decision making. The starkest similarity is between this thesis and the work of Raghavan et al, who develops a data-driven model and indeed both use time-series data. However the work of Raghavan et al is narrowly scoped to the application of HMMs to a particular problem, and, like the other examples, puts little emphasis on adaptability. In the case of Olama et al, their system requires extensive SME input and is therefore not readily adaptable. It is unfair to criticise the generic fusion systems on these

grounds as their focus is the GUI and facilitating the manual process of structuring evidence. These examples demonstrate the need for, and perhaps confirms the difficulty of, developing frameworks without a need for significant SME input.

ML for Decision Making The application of ML to decision support has received a lot of attention and numerous examples can be found. Domains where ML is being extensively applied include medical, legal and cyber security, due in part because of the data richness of these domains and necessity. Certainly in the case of cyber security, human operators just cannot react fast enough and adversaries are themselves rapidly adapting. Examples include: a comparison of performance of several ML techniques for predicting instances of sepsis and acute kidney injury in postoperative patients; a discussion on the potential of ML to support self-representing (in particular vulnerable/disadvantaged) litigants; a survey on cyber attack detection, including the application of ML. [Thottakkara, Ozrazgat-Baslanti, Hupf, Rashidi, Pardalos, Momcilovic, Bihorac, and Xie, 2016; John, 2017; Babiker, Karaarslan, and Hoscan, 2018].

A direct comparison of this work with the proposed Framework is not warranted as this thesis does not make a contribution to the field of ML. However it is relevant that ML is being used in the decision support domain and again appears limited to tactical decisions. Further re-enforcing the belief that this Framework is at the very least a rare example of an ML application to strategic decision support.

Web (text) Scraping The proposed Framework is demonstrated in this thesis using the relatively easy case of structured web data. But by design it is ambivalent and further work recommends addressing the more challenging case of unstructured text data. The following are examples of work focused on the Information Extraction (IE) task from text data. Automatic KG/ontology construction addresses the challenge of extracting triples (or n-ary facts) from unstructured text and transforming them into a graph-like database. This text can be web-sourced or from a given corpus. A general overview of KG building is given in Krzywicki, Wobcke, Bain, Calvo Martinez, and Compton [2016] and a comparison of systems in Nickel,

Murphy, Tresp, and Gabrilovich [2016]. The motivation for building KGs and ontologies is to enable down-stream analysis of the newly-structured information such as within a reasoning and decision making framework. A comparison is not relevant as this thesis does not make a contribution to the domain of IE but is included as such systems are relevant to future research directions of the Framework.

DeepDive A seminal effort in this area is by researchers at Stanford University on their DeepDive system. The culmination of their work, described in Zhang, Ré, Cafarella, De Sa, Ratner, Shin, Wang, and Wu [2017], is an automatic end-to-end system to build domain-specific KGs. It takes two inputs. The first input is unstructured and structured information, the later in the form of an existing KG. The second is the user-defined schema on which to perform extractions. DeepDive uses distant supervision to perform additional extractions and returns triples with an associated confidence score. The team at Stanford has applied DeepDive to a number of diverse domains to demonstrate its flexibility including paleobiology, genomics and law enforcement (see the description of Memex earlier in this section).

NELL Another notable example of automatic KG construction is the Never Ending Language Learner (NELL) developed at Carnegie Mellon University (CMU). The input and approach is similar to that of DeepDive but NELL, in addition to the distantly supervised approach, incorporates an online feedback loop, allowing the public to rank new extractions. The intention for the NELL system is unique in that the researchers set out to develop an architecture that continuously learns (it has been in continuous operation since January 2010) and like many other KG builders, the resultant KG is generic rather than tied to a domain. [Carlson, Betteridge, Kisiel, Settles, Hruschka Jr, and Mitchell, 2010]

NOUS A proposal for the automated construction of domain specific KGs is made by Sutanay Choudhury et al. of Pacific Northwest National Laboratory. The system takes existing KGs and combines them with bespoke extractions in order to generate a domain specific KG. Of interest to the researchers is using the system to track dynamic topics such as emerging technologies. [Choudhury, Agarwal, Purohit, Zhang, Pirrung, Smith, and Thomas, 2016]

GDEL GDEL is an ambitious project with multiple facets. It is of interest to this research as an example of web-scale multilingual IE. The project takes as input principally publicly available online multilingual news feeds but has recently expanded its capability to include social media and image feeds. Extracts include events, people, places, organisations and sentiment. [Leetaru, 2018]

Unsupervised Examples The previous examples incorporate some form of supervision, be that distance or human re-enforcement. At the forefront of research addressing the domain ontology building problem are unsupervised systems. An example of a comprehensive system is by a group at CMU: Dana Movshovitz-Attias and William Cohen, dubbed Knowledge Base-Latent Dirichlet Allocation (KB-LDA). This system takes as input a domain corpus and identifies, via an unsupervised technique, the latent schema and associated triples [Movshovitz-Attias and Cohen, 2015]. Similar work is also being conducted at the Indian Institute of Science by Madhav Nimishakavi and colleagues, called Schema Induction Using Coupled Tensor Factorisation (SICTF) which has also been extended to address n -ary extraction. [Nimishakavi, Saini, and Talukdar, 2016; Nimishakavi, Gupta, and Talukdar, 2018].

Comparison As previously noted, a direct comparison of this work with the proposed Framework is not desirable except to note that these systems have already been integrated into wider reasoning and decision making frameworks and thus should also prove to be a promising research direction for this thesis.

2.3 National Intent and Industry

The second important theme of this thesis is the link between national industry and strategic national intent. A few nations explicitly define a relationship between their national industrial bases and intention with respect to military procurement and it has been independently observed by researchers examining the concept of National Innovation System (NIS). The United Kingdom (UK), United States of America (USA) and Japan have all published material to this effect and this is explored in depth at Section 3.2.1.3. It is also known that some nations

evaluate the intention of other nations through examination of their industrial bases but few examples of methodology were found in the literature let alone evidence of applying modern computational methods to the task. This offers confidence that this Framework is unique in this respect, at least in the open literature.

NIS and Defence Industry NIS encompasses the idea that long-term economic growth cannot purely be explained by the raw inputs of materials and labour, but also by the production, dissemination and use of ‘economically useful knowledge’ [Lundvall, 1992, p.2]. The work in this area largely focuses on comparing national development with industrialisation paths. However authors such as Richard Nelson, show that a number of nations² have a NIS shaped specifically by security concerns. These nations exhibit significant spending on defence related Research and Development (R&D) and procurement from R&D orientated industries. Further that for nations with large military procurement programmes, the corresponding defence industry has high levels of government involvement. [Nelson, 1993]

Threat Assessment A common model for conducting threat assessment is based on the work of David Singer who equates perceived threat to the product of estimated capability and estimated intent [Vandepeer, 2011; Singer, 1958]. In the case of the USA, analysts use a combination of direct and indirect (proxy) measures to quantify capability, the latter includes a measurement of a nation’s defence and energy and power industries³ [Vandepeer, 2011]. An example of a threat assessment related to the Singer model is provided by Scott Barclay et al, in work conducted by the predecessor of DARPA. The aim is to determine the intention of a nation to develop ‘independent nuclear weapons production capability within the next five years’ [Barclay, Brown, Kelly, Peterson, Phillips, and Selvidge, 1977, p.118]. The model takes the form of a BBN split into indicators, associated activities and evidence. Industry activities are included in this model, in particular the status of centrifugal enrichment plants. [Barclay, Brown, Kelly, Peterson, Phillips, and Selvidge, 1977]

²Such as Japan, Korea, Taiwan, Israel, Brazil and Argentina.

³See Section 3.2.3 for discussion on the definition of intent for threat assessment.

2.4 Conclusion

In contrasting the work of this thesis with that of others, it is necessary to examine its broader scope, namely computational approaches to exploiting web information for decision support and the systematic assessment of national intent via industry. In the former case, this thesis is a modest contribution compared to large programmes such as IBM’s Watson and Stanford University’s DeepDive. But from this examination, this thesis does stand out as a rare example of applying source agnostic web scraping and reasoning to adaptable strategic level decision support. In the latter case, this thesis is potentially unique in the application of the strengths of computational techniques to the task of systematically evaluating national intent from industry observations in a data-driven paradigm.

Chapter 3

Conceptual Graph Model

3.1 Introduction

The purpose of this chapter is to describe the developed novel conceptual graph model and associated literature. The multi-disciplinary literature review is described at Section 3.2, addressing the core ideas of strategic industry, capability and intent. These ideas are incorporated into the conceptual graph model, which is described in detail at Section 3.3 and represents the first contribution of this thesis.

3.2 Core Ideas

3.2.1 Strategic Industry and Industrial Policy

3.2.1.1 Definition and Origin

Several definitions for *Strategic Industry* are offered by the literature. While these definitions differ in detail, fundamentally they agree that it addresses whether or not a given industry deserves special interventions from government in order to gain or preserve some national-level benefit.

The term itself can be traced to the economist Albert O’Hirschman in his 1958 work *Strategy of Economic Development* [Hirschman, 1958]. However the idea is documented as early as

the 18th Century with the emergence of modern economic theory. Yet governments, or their historic equivalents, have been intervening in chosen industries since at least medieval times. For example the economic significance of glass manufacture and ship building to Venice led the republic to execute emigrating Murano glass workers and Venetian shipwrights, lest they take their skills to other kingdoms. [Smith and McCulloch, 1995; Landes, 1998]

3.2.1.2 Early Economic Thought

Addressing the question of whether governments should act favourably towards certain industries, Adam Smith, credited by many as the father of modern political economics, makes the case against such behaviour. Smith's 1776 seminal work on political economy, *An Inquiry in the Nature and Causes of the Wealth of Nations*, was written at a time of great industrial development and when some industries had secured protection from foreign competition through government actions such as imposing high duties or bans on certain imports. [Smith and McCulloch, 1995]

Smith believed that the prevailing economic wisdoms of the day¹, to protect domestic industry from foreign competition and to impose export bans or heavy duties on gold and silver, were logically flawed. Nations believed that when trade is balanced between two nations the net accumulation of gold and silver (the medium of exchange) by each nation is zero. But when trade is unbalanced one nation accumulates these metals. Since nations wished to accrue these commodities in order to fund wars, penalties were imposed for trading with nations if it resulted in a flow of gold and silver out of the country.

Arguing strongly against this logic, Smith makes the case that both nations gain if there were *free market* conditions – under which any exchange would result in an increase in revenue for one nation and an increase of goods or services for the other, this latter part being ignored

¹which he calls the *mercantile system* to reflect how it evolved to favour merchants and manufacturers rather than society as a whole.

by the previous logic. Smith supports his argument with evidence of *free ports* being enriched rather than ruined as the prevailing doctrine of the day would have predicted.

Later in the 19th Century, and taking the opposing stance, Georg List, Sampson Lloyd and Joseph Nicholson make the case for government intervention in industry in certain situations. In their work *The National System of Political Economy* they refute Smith directly, arguing that he addresses a global economy; treats wealth simply as a product of labour and raw materials; and fails to account for national interest and *productive power*, which addresses what they refer to as *mental capital* which includes education, justice and defence [List, Lloyd, and Nicholson, 1885]. In their view government intervention in industry is justified for two reasons. First, on the basis that in doing so it allows the industry to become the cheaper producer in time over foreign imports through greater efficiencies in production derived from a developed indigenous manufacturing base. Secondly for ‘industrial independence in the case of war’ or political differences which would cause interruption of supply chains [List et al., 1885, p.145].

Since these early works, the arguments for and against government intervention in industry have been revisited numerous times by subsequent economists and remains a controversial topic. Regardless of the economic arguments for free markets or protectionism, governments continue to intervene in industries for a variety of reasons, a selection of which are explored next.

3.2.1.3 Motivations for Intervention

While the concept itself arguably has roots in economic theory with concessions to national security, context is all important when determining what makes an industry strategic to a nation – not just its economic impact. PESTLE analysis² offers a way to categorise these motivations, here two such motivations: Social (Defence) and Environment are briefly expanded upon.

²An analytical tool frequently used in business, listing categories of the macro-environment. It stands for: Political, Economic, Sociological, Technological, Legal and Environmental [Cadle, Paul, and Turner, 2014].

Defence As described in the previous section (3.2.1.2), the provenance of the strategic industries concept touched on the national interest in times of conflict. Specifically when it served the nation to not only be economically well-off in order to fund conflict abroad, but to also be technologically well-off, to conduct successful military operations and to be self-reliant in terms of access to supplies (both military and domestic) in the event of conflict [Smith and McCulloch, 1995]. In more recent years this perspective has proven itself to have remained relevant and shared by several nations, some manifestations of which are described below. These also represent examples of nations explicitly defining a relationship between national industrial bases and intention with respect to military procurement as discussed in Section 2.3.

USA From the 1950s the USA enshrined the importance of its industrial base to national security through law in the form of the US Defence Production Act and Title 10 of the US Code. The government’s role includes evaluating the ability of the domestic industry to meet defence production requirements, correcting shortfalls in industrial capacity, and designating critical and strategic materials. [USGov, 1956; US Congress, 2009]

OECD In the 1990s members of the OECD were concerned that the globalisation of supply chains that had begun in the 1950s had introduced risks regarding an individual nation’s ability to conduct military operations due to reliance on other nations³. Investigating this concern, the OECD confirmed that some national militaries were largely dependent on foreign suppliers, for example the USA’s reliance on Japan for semiconductors (this reliance is not unique to the USA nor the military) [OECD, 1991].

UK In 2002 the UK Ministry of Defence (MoD), recognising the emerged globalised structure of industry, stated that there is a need to retain certain industries in the UK in order to ensure security of supply for the military, the ability to deliver large scale defence programmes and research key technologies [MoD, 2002]. A few years later in 2005 the MoD listed the specific aspects of military capabilities that require a commensurate UK-based industry. For

³This risk is exemplified by the case of Russia’s procurement of helicopter carriers. These carriers were being built in France. In 2014 France delayed delivery of the carriers in response to the developing situation in the Crimea. Later both carriers intended for Russia were sold to Egypt. [Thomas, 2014; Staff, 2016]

example, the UK is willing to go to the international market for military helicopters but wishes to retain expertise within the UK in rotor-blades, vibration management and electronic architecture. At this point the MoD introduces the term ‘appropriate sovereignty’ for the purpose of ensuring ‘operational independence’ [MoD, 2005, p.6-7]. Subsequently, in 2012, the MoD defined the characteristics of military capability that are considered strategic such as capabilities that are dependent on highly classified information like the nuclear deterrent, weapons and propulsion systems [MoD, 2012b]. This trend has continued and has been re-enforced: the 2015 Strategic Defence and Security Review encourages specific sectors, such as cyber security, and defines exceptions to open competition regarding defence procurement. The 2017 National Shipbuilding Strategy states the ‘UK will need to retain the ability to design, build and integrate warships’ [MoD, 2017b, p. 33]; the 2017 update of the Defence Industrial Policy identifies developing the synergy between the UK Defence Industrial base and economic prosperity. [HMG, 2015; MoD, 2017a,b]

Japan At the end of the Second World War, Japan’s industrial base was essentially destroyed. The country initially relied on the USA for the lease of military equipment and logistics [Dosi, Freeman, Nelson, Silverberg, and Soete, 1988]. Since then Japan has actively managed its industry, including its defence industry, by using foreign expertise and developing indigenous abilities systematically through the Ministry of International Trade and Industry (MITI). This strategy has paid off for Japan, becoming the most recent entrant on the international defence market in 2014. While domestic production is favoured, Japan notes that defence procurement will not be solely indigenous but will draw appropriately from abroad through international arrangements in co-development and co-production, licensing⁴, and imports. The domestic and international approaches are seen as complementary rather than a compromise. Having a strong high technology defence industry puts Japan in a strong position to negotiate in co-production, co-development and imports. Further, these international procurement mechanisms offer the opportunity for Japan to gain access to technology and, if desired, a route to eventual domestic capability. Japan identifies which defence equipment will be procured domestically and which will depend on international arrangements. For example, its ability to produce tanks will be

⁴An agreement for domestic production of foreign technology.

maintained indigenously, and its jet aircraft requirements will continue to be met through international joint procurement while seeking a role for Japanese companies in production and maintenance. [JpnMoD, 2014, 2015]

Environment Government intervention in industry in order to gain an environmental benefit has roots in both economics and property law which determines how environmental assets are owned and used. Government intervention tends to occur when the specifics of an environmental asset's characteristics, ownership, and agents' rights and obligations result in an inefficient use of an asset. In other words, a market failure. The *tragedy of the commons* and the *free rider problem* are examples of market failures which may cause governments to intervene. [Grafton, Adamowicz, Dupont, Nelson, Hill, and Renzetti, 2008; Tietenberg, 2012]

When the characteristics of an environmental asset are equivalent to a product on the market in that it is highly exclusive, transferable and enforceable, free markets work well for the efficient use of the asset. The free market model however does not lead to efficient use of assets that no one owns nor has control over its use, so-called *open access* assets. Here *externalities* operate that lead to market failure and the environmental asset becomes a candidate for government intervention. An externality exists where the welfare of one agent is affected by the activities of another agent and these activities are out of the first agent's control, such as a person dumping waste in a river thus affecting the enjoyment of the river by others. The majority of environmental assets are considered open access but differ in their exclusivity, how one agent's use of an asset precludes others, and rivalry, how easy it is to prevent other agents from using an asset. [Grafton, Adamowicz, Dupont, Nelson, Hill, and Renzetti, 2008; Tietenberg, 2012]

Tragedy of the Commons *Common Pool* assets are open access with either low or high exclusivity and high rivalry. Examples include wild fish and oil. Agents are incentivised to take as much as they can, leading to overexploitation, which is a market failure known as the tragedy of the commons. Continuing with the example of wild fish, a case of a replenishable but depletable common pool resource, historically the majority of marine protein (unlike land-sourced protein) was taken directly from the wild rather than farmed. The unrestricted ability

of agents to take fish from the wild, twinned with demand, led to overexploitation. This in turn caused population crashes that resulted in a higher cost to harvest the same amount and damaged the ability of agents to fish from the same stock in future years. [Grafton, Adamowicz, Dupont, Nelson, Hill, and Renzetti, 2008; Tietenberg, 2012]

Observed interventions by governments are generally made on the basis of sustainability of fishing stocks and protection of species or habitat. Public policy interventions have included changing property rights of commonly-held waters to private ownership, making the asset highly exclusive, transferable and enforceable, thus allowing the benefits of the free market to take over. This particular intervention had startling results by incentivising the development of farmed fish which saw a rise in global consumption from 4% in 1970 to 46% in 2008. Other examples of government intervention by public policy include: taxing the fishing effort, applying quotas, buy-back of fishing fleet capacity and introduction of protected areas. These interventions are not without their issues. The private property rights applied to bodies of water has facilitated fish farming but has also created pollution, encouraged disease, and caused inefficient harvesting of smaller fish to feed the farmed fish. Other measures such as restricting fishing times has also had counter-productive results. An extreme case being the 1982 Herring Season in Prince William Sound. The season lasted just four hours, but the ensuing intensive fishing effort resulted in a catch that exceeded the quota. [Tietenberg, 2012]

To compound the issue, the market failure regarding wild fishing is not only a national issue but a trans-national one, as illustrated by the *cod wars* between the UK and Iceland. To address this, international law permits nations to declare Exclusive Economic Zones for up to 200 Nautical Miles from its low-tide mark, within which that nation, if capable, can regulate fishing. Outside these zones international law has developed such as applying limitations on whaling. [Tietenberg, 2012]

Free Rider Problem *Public Good* assets are neither exclusive nor rivalrous. Examples include air, water, landscape and biodiversity. In this case there is no incentive for agents to care for these assets. This leads to the free rider problem. A key example examined here is

pollution. [Grafton, Adamowicz, Dupont, Nelson, Hill, and Renzetti, 2008; Tietenberg, 2012]

Pollution can be defined by three dimensions. First, it can be defined by its ability to be absorbed by the environment. Stock pollution, such as plastic, accumulates, whereas fund pollutants, such as carbon dioxide, are able to be absorbed back into the environment provided the flow of waste does not exceed the capacity of the environment to either transform it into a non-harmful form or dilute it so it is no longer at harmful concentrations. The second is its sphere of influence from source in vertical and lateral distance: local, regional, global. The third dimension is the nature of the source. A point source is easily identifiable as the source of the pollution whereas a non-point source is not and, as a result, is harder to measure.

Market failure occurs because neither the emitters of pollution nor the consumers of the resulting product directly bear the cost of the pollution. The cost is distributed so it is not in the interest of the polluter nor consumer to do anything to reduce their pollution. While this was not an issue when quantities of waste being produced were small, it has become a growing concern as urbanisation, industrialisation and consumption have increased – increasing the rate at which waste, and hence pollution, is produced. [Beck, 1959; JICA, 2005]

Different industries differ in the type and quantity of pollution they produce. Raw material processing and energy production are industrial sectors frequently termed as *dirty industries* in that, while they make up a small proportion of total industry, they are responsible for the majority of the waste. [Jänicke, Binder, and Mönch, 1997]

While some authors note that the most polluting industries have been the principal targets of anti-pollution legislation in at least two industrialised nations, many of the landmark policies are not targeted at specific industries but at specific pollutants – possibly because the interventions need to reflect the properties of the targeted pollutant. Although it could be argued that these policies have the biggest impact on the industries which are the biggest producers of those pollutants. The literature also suggests that there is an economic angle to these interventions,

with cost-benefit analysis and associated methodologies being major drivers of government intervention (or not). Examples of interventions include imposing limits on pollution rate and a charge (or tax) on the emissions themselves or on the commodity that produces the emission [Tietenberg, 2012]. Intervention of course is not without issue. The *pollution haven effect* theorises that polluting industries will relocate to countries with less stringent pollution regulations, although this is disputed. [Jänicke, Binder, and Mönch, 1997; Kellenberg, 2009]

3.2.1.4 Modes of Intervention

Many modes of intervention by governments are documented in the literature, resulting in a plethora of categorisation schemes for government instruments. Commonly, government policy is categorised as fiscal and trade. The former addressing tax-rates and government spending. The latter addressing cross-border trading of goods and services [Rhodes, 2014]. Others seek to quantify the scale of intervention by governments, such as Charles Anderson who defines a scale of intervention, ranging from free market to regulatory [Anderson, 1977].

The categorisation of primary interest with respect to this research, in line with the idea of strategic industries, is vertical industrial policy. The purpose of industrial policy is to address production inefficiencies caused by externalities. Within this set of policies are vertical and horizontal policies. The former target specific industries and/or sectors and the latter does not discriminate between industry/sector. Examples of policy are diverse. Traditional *hard measures* encompass subsidies and public procurement and are exemplified by supporting a specific company against foreign competition, such as in the case of Airbus [Gribben, 2006]. More recently emerged *soft measures* include government agencies facilitating networks, as the US government has done through various agencies, and R&D funding [Wade, 2012]. Policy aims are equally diverse, from encouraging emerging industries that are considered drivers of future economic growth, to reducing shocks and encouraging rationalisation of industries in decline, or to encourage entry of firms into certain geographic areas to achieve economies such as lower transport costs and encouraging technical spill-overs, where an activity in one sector has benefits in another [Begg, 2008].

It is worth noting here that from the 1970s the prevailing stance of many Western governments and (Western-led) international economic bodies was promotion of free-market conditions. This is reflected in international policies. The European Union (EU), from conception, outlaws state aid to specific industries and/or sectors as does the World Trade Organisation (WTO). The World Bank lending and grant policy to low-income countries penalises nations exhibiting industry favoritism, for example by quantifying the extent of free trade [Wade, 2012].

While the concept of vertical industrial policy might be anathema to many, as Partizio Bianchi, Sandrine Labory and Robert Wade note, governments continue to implement it in spirit if not in name. In the 1990s the term *competitiveness policy* was used by European institutions [Bianchi and Labory, 2006]. The 2008 global financial crisis precipitated state intervention. For example in the US, sector specific measures were implemented to shore up bank lending, automotive, energy, medical, pharmaceuticals and information technology sectors [Wade, 2012]. Most recently the imposition of tariffs on steel and aluminum imports by the current US administration which, while officially citing security concerns, are widely considered as a protectionist move [Trump, 2018a,b; Donnan, 2018]. There are however signs that the prevailing stance is shifting. Bianchi, Labory and Wade comment that the World Bank and EU are reflecting upon the hard anti-vertical policy stance and are considering circumstances under which these policies may have benefit [Bianchi and Labory, 2006; Wade, 2012].

International legislation combined with the surreptitious approach to vertical industrial policy has contributed to the challenge for researchers wishing to empirically measure the extent (and effects) of such policies as there is little official reporting on the measures taken [Bianchi and Labory, 2006]. The lack of official reporting on the matter and obscuration of industry/sector specific assistance, suggests a new approach is needed that can take advantage of unofficial reporting and/or glean insight from what little data there is.

3.2.2 Capability

For the purpose of this research, the definition of capability is drawn from the government defence sector due to the maturity of the concept and its flexibility in application across sectors. Three methods of capability definition are documented here: Architecture Frameworks, Defence Lines of Development and Ontological Frameworks.

Architecture Frameworks An Architecture Framework, or Enterprise Architecture Framework, standardised under the International Standards Organisation (ISO), is defined as a conceptual model description of entities and the relationships between them [ISO, 2011]. There are numerous frameworks, many of which are aimed at the delivery of Information Technology systems but this system has also been applied to managing defence capability, of which four national architecture frameworks: Australian, Canadian, American and British were found. In general these defence architectures use Unified Modeling Language (UML) and its extensions to highlight different aspects of the capability being managed, specifically entities and relationships between them. UML was developed to represent software engineering architectures, but also has several extensions, such as Systems Modeling Language, designed for other applications.

In the case of the UK, the MoD Architecture Framework (MoDAF) is used for defence capability planning and change management. Of interest to this Research Framework is the representation of capabilities as: [MoD, 2012a; HMG, 2010b,c; MoD, 2010].

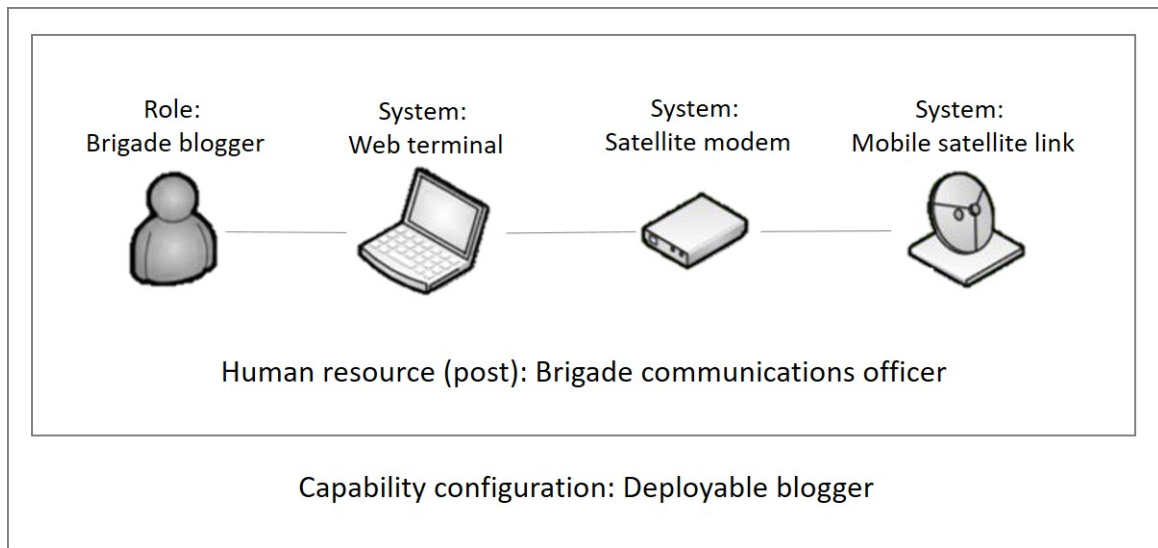
- Combinations of resources. Resources can be physical, software or human as shown in Figure 3.1a.
- A taxonomy. In this way a given capability can be represented as: a special case of another capability; a necessary (dependent) part of another capability; part of a cluster(s) of capabilities; and aggregated capabilities. A dependency example is shown in Figure 3.1b.
- Nodes. With each node having a relationship to another node, such as material flow.

Defence Lines of Development Defence Lines of Development (DLoD) identifies all the elements that must be considered in delivering a capability. In this thinking, a capability consists of: Training, Equipment, Personnel, Information, Concepts and Doctrine, Organisation, Infrastructure and Logistics (TEPIDOIL) as illustrated in Figure 3.1c [HMG, 2010a]. The US equivalent of TEPIDOIL is: Doctrine, Organisations, Training, Material, Leader development, Personnel, Facilities and Infrastructure, accompanied by the mnemonic DOTMLPFI. Capability can be linked to industry via the Equipment DLoD and the US equivalent, Material. This element of a capability is physical and at some point must be manufactured and assembled. For example an equipment element of a defence capability might be a ship. But this ship is also made up of numerous components, and these can be further broken down into sub-components until fundamental elements are reached which ultimately have to be extracted in some way from the environment. Industry is involved at all steps of this supply pipeline.

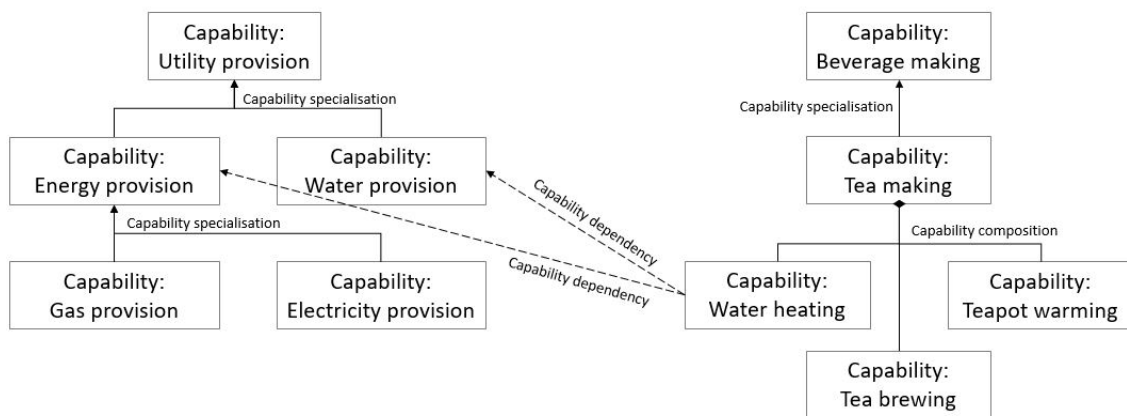
Ontological Framework Closely related to the Architecture Framework, the Ontological Framework sets out to scope what a capability is (and is not) and provides a system of classification. The US, via the Department of Defence Architecture Framework 2.0 Meta Model (DM2), the UK, via the MoDAF Ontological Data Exchange Mechanism, and an international version, International Defence Enterprise Architecture Specification (IDEAS) for exchange, provide examples of ontological frameworks for the defence case. The conceptual overview for DM2 is shown in Figure 3.2. An ontology provides a standard specification to define a domain's concepts and the relations between them. The purpose of an ontology is to facilitate information exchange via an international standard, typically Resource Description Framework (RDF) [W3C, 2014]. No examples were found of a capability being described via an ontology but is of interest to this Research Framework for its potential, once populated, for capturing capabilities and their constituent parts.

3.2.3 Intent

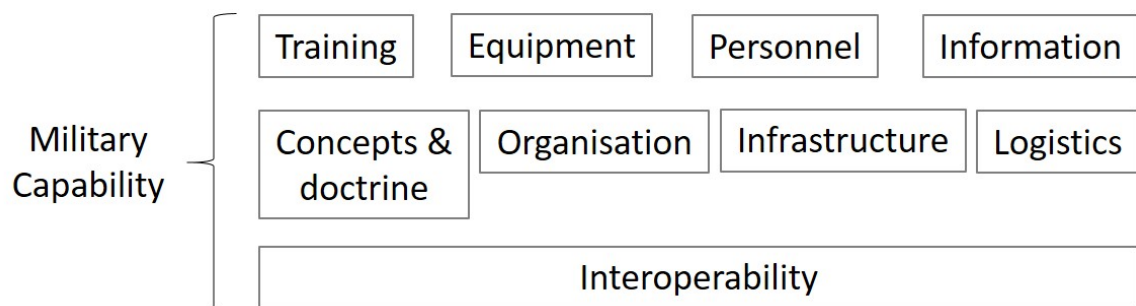
Here definitions of intent are documented from three disciplines, showing the scope of possibilities for intent definition within this research. Of principal interest is the field of cognitive



(a) Example illustrating a capability as a combination resource types.



(b) Example illustrating the dependency relation between capability entities.



(c) Example illustrating the components of a capability with respect to TEPIDOIL.

Figure 3.1: Examples of modeling a capability in accordance with MoDAF. Figures are adapted from public sector information licensed under the Open Government Licence v3.0. [MoD, 2010; HMG, 2018]

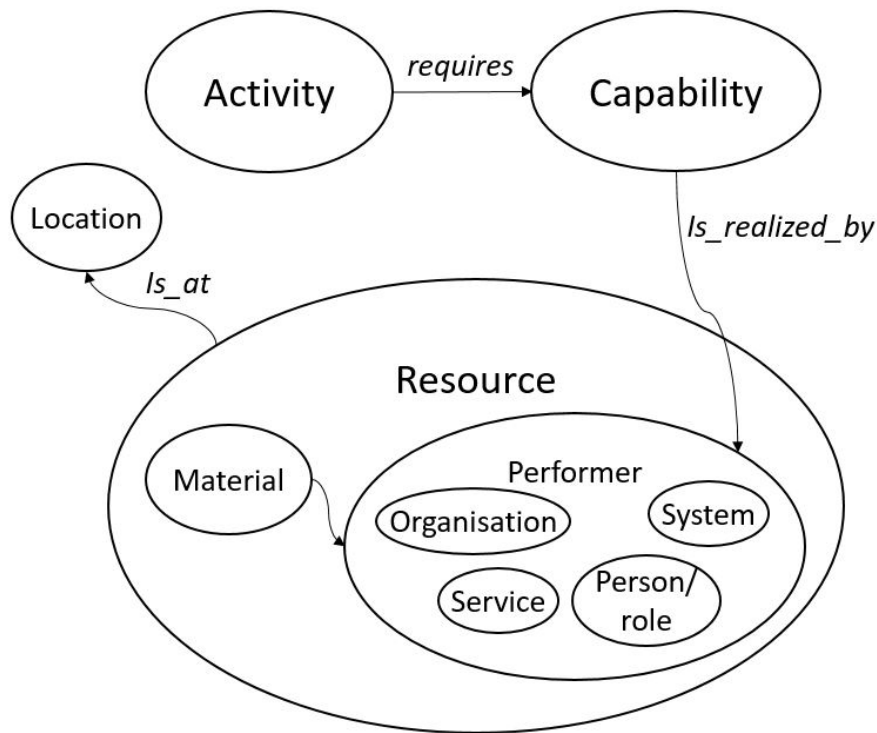


Figure 3.2: The conceptual overview for DM2 diagrams, illustrating the relationship between capability and material via performing resources as defined by the DM2 ontology. No copyright restrictions. [McDaniel, 2011]

science which is a multi-disciplinary area that has applied computational techniques, Artificial Intelligence (AI), to modelling human faculties and is inspired by theories of human cognition from psychology.

Threat Assessment Threat assessment, a type of risk assessment, is conducted by government defence departments/ministries⁵. A review of Western states' approaches to threat assessment is provided by Charles Vandepeer who notes the extensive use of models similar to that described by David Singer in his 1958 paper *Threat Perception and Armament Tension*. The Singer model equates perceived threat to the product of estimated capability and estimated intent. Regarding intent estimation, there are some, such as Richard Best, who believe that due to its nature it is impossible to measure. This has not prevented others from undertaking this task. In his thesis, Vandepeer observes that to measure intent analysts use indicators, or observables, to infer and/or estimate current and future intent. In his review Vandepeer found that the following three indicators are frequently used in intent assessment: military capability

⁵See Section 2.3 for a discussion on the relation between industry and intent.

assessment, state ideology and assessment of state rhetoric. The latter two often limited to those of a state’s political leader. [Vandepeer, 2011; Singer, 1958; Best, 2006]

Theories of Human Cognition The Theory of Mind considers research into understanding a person’s ‘ability to explain and predict one’s own and other’s behavior’ [Harbers, Meyer, and Van Den Bosch, 2012, p.333]. The term was coined by David Premack and Guy Woodruff in their seminal paper *Does the chimpanzee have a theory of mind?* and offers its definition as when an ‘individual imputes mental states to himself and to others’ [Premack and Woodruff, 1978, p.515], where mental states include: intention, knowledge, belief, thinking, doubt, guessing, pretending and liking. The authors define intent as synonymous to purpose [Premack and Woodruff, 1978]. There are two competing theories regarding how the Theory of Mind arises: Theory –Theory and Simulation Theory. Theory –Theory explains a person’s ability to explain another’s actions by using the concepts of belief, goal and intent. Simulation Theory involves an individual using knowledge to reason from the perspective of another individual. [Harbers, Meyer, and Van Den Bosch, 2012]

Artificial Intelligence In the field of AI, planning is an established area of research and has been refined as Plan, Activity and Intent Recognition (PAIR). These algorithms enjoy diverse applications to computer agents including computer games and human-robot interaction. Broadly, planning algorithms consist of a plan, π , which is a sequence of actions, $a_1, a_2, \dots, a_n \in A$, which move an agent from an initial state, I , via intermediate states, $\in S$, to a goal, G . The goal being the states that satisfy an agent’s objective and are akin to intent [Freedman and Zilberstein, 2018]. This discipline gives definitions of intent as an agent’s goal, interpreted as an agent’s state that satisfies an objective, or as the agent’s next action in a sequence.

PAIR PAIR algorithms take as input a sequence of observations, $o_1, o_2, \dots, o_n \in O$ and returns a member from a library, L . What constitutes an observation and the membership of the library depends on what is being recognised – a plan, an action or an intent. For plan recognition, the observations are the agent’s actions, $O = A$, and the plan is returned from a

library of potential plans, $L = \pi_1, \pi_2, \dots$. Intent recognition also takes the agent's actions as observations but returns either the goal, G , from a library of potential goals, $L = G_1, G_2, \dots$, or a predicted action, $L = A$. Activity recognition takes observations of lower level actions, $O = \sum$, and returns the higher level action, $L = A$. [Freedman and Zilberstein, 2018]

BDI Belief, Desire, Intent (BDI) algorithms are an extension of PAIR algorithms. As before, an agent has goals, and it selects suitable plans from a library and takes actions to fulfill those goals. However plan selection is constrained by beliefs and it can take into account its understanding of the beliefs and goals of other agents and their inferred plans. [Harbers, Meyer, and Van Den Bosch, 2012]

3.3 Conceptual Graph Model

3.3.1 Graph Model

Based on the Literature Review documented in Section 3.2, the derived conceptual graph model identifying the relation between national industrial bases and national intention is shown in Figure 3.3 as a directed graph. The basic premise for this graph is that national intentions can be inferred from national capability which in turn is a function of industry, as capabilities consist, in part, of products which are manufactured by companies which are members of industries. Further, governments act on their industries through policies, which here are broadly interpreted due to the diversity of ways governments target support to industries. The definitions of each node and edge is described next, noting that in demonstrating this graph as part of the wider Framework in Chapter 5, some node definitions are adapted.

While a directed graph was chosen to represent the conceptual model, alternatives exist, such as a hypergraph or the use of UML's richer notation. A directed graph was used for two reasons. First, it offers the simplest encoding of the required information in a readily understandable way. Second, the ease of translating the graph model into requirements for down-stream web-scraping and interpretation as a BBN (a directed acyclic graph) – the chosen ML algorithm.

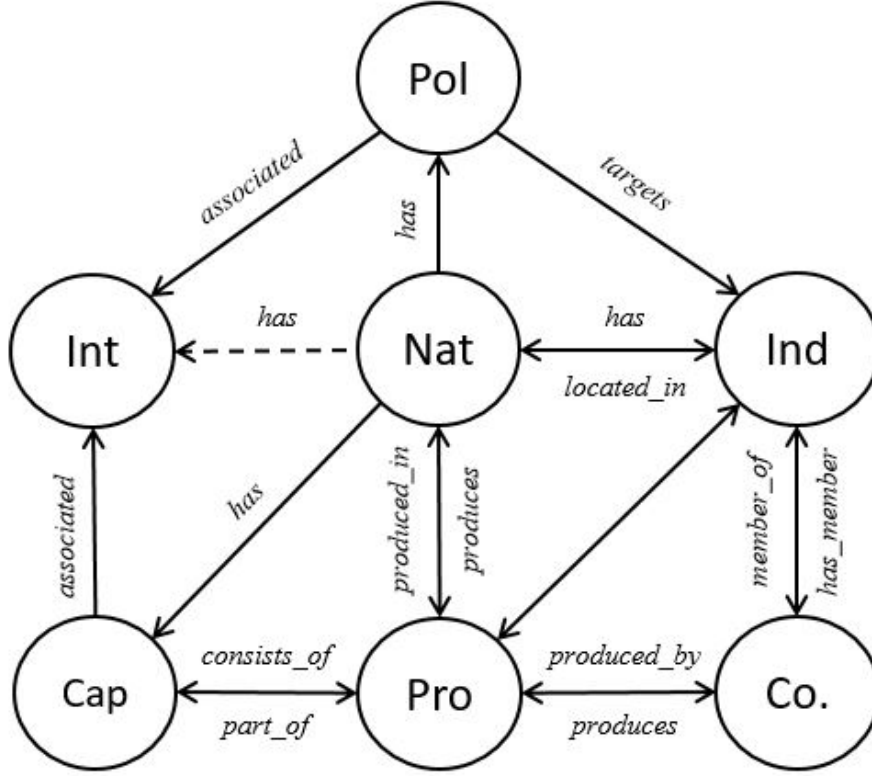


Figure 3.3: Illustration of the novel conceptual graph model, a key contribution of this thesis. Solid lines indicate an observable relationship. Dashed lines indicate an inferred relationship.

3.3.1.1 Node Definition

Nation The nation node represents individual demarcated nations⁶ and can take states given by a finite categorical set Nat ⁷. For example $Nat \supset \{Albania, Belgium, Bulgaria\}$ or generically $Nat = \{nat_1, nat_2, nat_3, \dots, nat_n\}$.

Industry The industry node represents categorised industries and can have states given by a taxonomic hierarchical set Ind where each member of Ind is also a set. An example of such a set is the United Nations (UN) International Standard Industrial Classification (ISIC) system which has a tree structure with four levels. The top level set is referred to as *Section* and is broken down by *Division*, *Group*, and *Classes*. For example, a Section-level member is *A. Agriculture forestry and fishing*, which consists of Division-level members: *01 Crop and animal*

⁶Notwithstanding current debate on sovereignty of some nations.

⁷Typically sets are represented by a single capital letter e.g. N but a three letter camel case notation here has been adopted to aid the reader in relating notation to its meaning, driving a multi-letter approach as some named sets share the same initial letter.

production, hunting and related service activities, 02 Forestry and logging, and 03 Fishing and aquaculture. [UNSD, 2017]

Capability The capability node represents a compositional containment hierarchy where a capability is broken down into its constituent parts to an arbitrary level of abstraction given by *Cap*. This framework only considers material aspects of a capability which are mapped to products. In this way the set can be simplified, removing intermediate levels and leaving a one-to-many mapping of a capability to products. This is illustrated in Figure 3.4. No extant public classification system was found for capabilities therefore a capability for the purpose of this framework remains user-defined.

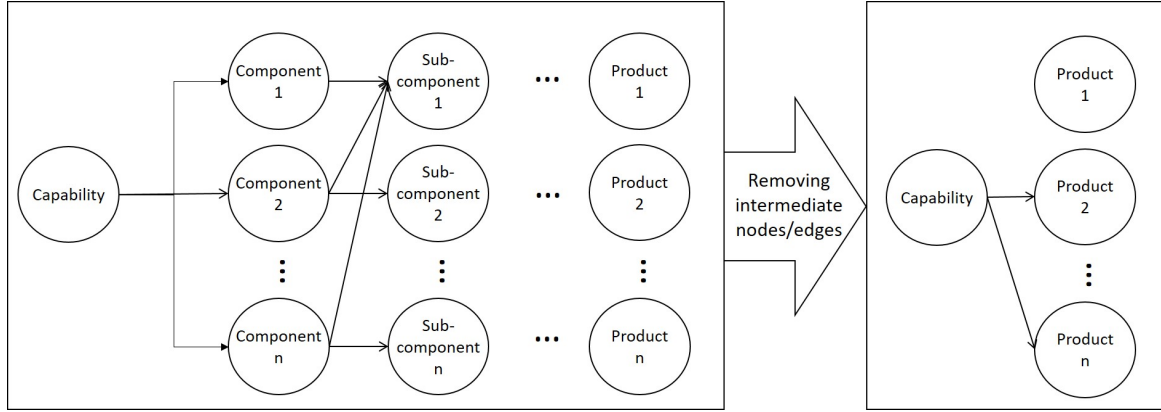


Figure 3.4: Two representations of a capability mapped to constituent products. On the left, intermediate nodes of components and sub-components are shown. On the right, the direct mapping of capability to products is shown.

Product The product node could be represented by a compositional hierarchy, but a simpler approach would be to represent it as a finite categorical set, *Pro*. There are several schema for categorising products, including the UN Standard Products and Services Code (UNSPC). [UN, 2018]

Company The company node is represented as a taxonomic hierarchical set *Co*.⁸

Intent As seen in Section 3.2.3, intent is a nebulous concept. Intent was considered in two ways for the purpose of this Framework. In the first instance, intent was considered

⁸Rather than following the three letter abbreviation established here for the company node, *Co*. was adopted as the conventional abbreviation for company as this was thought to bring more clarity over *Com*.

to be linked to capabilities such that the observation of a given set of capabilities indicated a given intention. In this way intent was considered an unlimited set of non-mutually exclusive options. In the second instance intent was considered with respect to the change in a capability over time, specifically a finite set of mutually exclusive completely exhaustive states where $Int = \{increasing, decreasing, maintaining\}$.

Policy Governments, which are assumed here to be akin to nations, utilise levers that affect industry, as described in Section 3.2.1.4. While many such levers exist, this framework addressed industrial policies that act in time and can be considered as favourable, unfavourable or neutral towards a given industry, giving $Pol = \{+ve, -ve, neutral\}$.

3.3.1.2 Edge Definition

Here the edges represented in Figure 3.3 are defined as relationships between the nodes.

- $\langle Nat, Ind \rangle, \langle Ind, Nat \rangle$. In accordance with Section 3.2.1, the edge between the nation and industry is shown as bi-directional to represent the reciprocal relationships: nation *has* industry and industry *located_in* nation. For one edge to exist, the other must also exist.
- $\langle Nat, Pol \rangle, \langle Nat, Int \rangle, \langle Pol, Ind \rangle$. As demonstrated in Section 3.2.1 and 3.2.3, a national government *has* policies that *target* its industries for given motivations. In this way, national intention can be inferred.
- $\langle Co., Pro \rangle, \langle Pro, Co. \rangle, \langle Co., Ind \rangle, \langle Ind, Co. \rangle, \langle Pro, Nat \rangle, \langle Nat, Pro \rangle$. Companies *produce* one or more products. By virtue of which product(s) is(are) produced, a company is a *member_of* one or more categories of industry. Products are *produced_in* a nation. This picture is complicated by diversification where companies produce multiple products across multiple nations and belong to multiple industries. Further, not all products produced by a company are produced in all nations that the company operates in. Globalisation also adds complexity. For example a nation may import products and thus have a capability without the commensurate industrial presence.

- $\langle Pro, Cap \rangle, \langle Cap, Pro \rangle, \langle Nat, Cap \rangle$. As already shown in Section 3.2.2, products are considered *part_of* a capability. Similarly a nation *has* capabilities. What may or may not constitute a capability is highly context dependent and for this Research Framework capability is considered to be defined by the user.
- $\langle Cap, Int \rangle, \langle Pol, Int \rangle$. A central hypothesis of this thesis is that national intention can be inferred from observations of capability which in turn can be inferred from industry and industrial policy.

3.3.2 Representation of Time

The relationships between nodes are dynamic. For instance the set of industries a nation has varies in time, as does the products that a company produces. Similarly the nodes can be considered dynamic, for example companies are established and disbanded over time, new industries emerge and others become obsolete. There is also lag in the system with respect to the time it takes for a policy to affect industry and thus have an effect on capability. At the most basic level, a capability may not be established until some time after the required industry has been established, which in turn may take some time to develop after suitable policies are put in place. The treatment of time is discussed in more detail in Chapter 6.

3.3.3 Graph Adjacency Matrix

Graphs may also be represented by an adjacency matrix. Each node in Figure 3.3, can take multiple values which become the rows of a vector, for instance $Nat = \{nat_1, nat_2, nat_3, \dots, nat_n\}$. The relationship between two high level nodes, such as Nat and Ind is given by a two dimensional matrix and is interpreted as national industry. When time is considered, it becomes a three dimensional matrix. Taking a slice of this three dimensional matrix at time t gives Equation (3.1), shown as a binary matrix, where values of one and zero indicate the presence and absence of an industry within a nation respectively.

$$NatInd(t) = \begin{matrix} & Ind_1 & Ind_2 & \dots & Ind_i \\ \begin{matrix} Nat_1 \\ Nat_2 \\ \dots \\ Nat_n \end{matrix} & \begin{pmatrix} 1 & 1 & \dots & 0 \\ 0 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & 0 \end{pmatrix} \end{matrix} \quad (3.1)$$

3.4 Conclusion

This chapter documents the novel conceptual graph model and its supporting literature, the first contribution of this thesis. In the next chapter, the graph model is incorporated into the Research Framework that is developed based on a multi-disciplinary literature review, and forms the second contribution of this thesis.

Chapter 4

Research Framework

4.1 Introduction

The purpose of this chapter is to describe the developed Research Framework and associated literature, thus representing the second contribution of this thesis. The multi-disciplinary literature review is described in Section 4.2. The Research Framework is described in detail in Section 4.3.

4.2 Core Ideas

4.2.1 Web Content

Web content is a large and growing, dynamic, heterogeneous (text, video, images, audio), multi-lingual, largely redundant (same info in multiple places), noisy (irrelevant content), undemocratic, observational, non-vetted dataset of structured, semi-structured and unstructured information. [Liu and Chang, 2004; Sprague, Grippink, Manyika, Moodley, Chappuis, Pattabiraman, and Bughin, 2014]

Not all web information is accessible. The *surface web* refers to the web that is indexable by search engines (those without a bot exclusion protocol). *Deep web* refers to web content that can be accessed if the address is known or a search is performed within a website. *Dark web*

is a sub-set of the deep web, and is only accessible via special web browsers such as The Onion Router (TOR). The remainder is not publicly available, for instance protected by passwords or some other form of access control. [Bergman, 2001; BrightPlanet, 2014; Liu and Chang, 2004; Sprague, Grijpink, Manyika, Moodley, Chappuis, Pattabiraman, and Bughin, 2014]

Size The growth of the web is undisputed, both in terms of content (estimated by number of indexable web pages, websites, web servers or amount of content in bytes) and use (estimated by number of users, number of devices, internet traffic and content generation). It is one of the largest datasets known to man [Baeza-Yates and Ribeiro-Neto, 2011]. Its size and growth rate are promising with respect to its use as a source to fulfil information requirements although it should not be assumed that all requirements can be fulfilled in this way.

Unrepresentative Web content is unevenly distributed in terms of geography and language [Graham, Hogan, Straumann, and Medhat, 2014; InternetWorldStats, 2016]. A number of social, technological, economic, political, and legal factors have contributed to this situation and these will be described next. In terms of content, Sub-Saharan Africa, the Middle East, North Africa and many nations in Latin America and Central Asia are under-represented with five countries in particular having several barriers to participation in effect: Bangladesh, Ethiopia, Nigeria, Pakistan and Tanzania. This uneven participation in the web, combined with relevance, indicates that information extractions will also be unrepresentative along these dimensions. This is of particular importance to the Research Framework as it may produce better results for some nations than for others. [Wu et al., 2015; Sprague et al., 2014]

Social A number of social reasons have been identified indicating why individuals/groups may be excluded from participation in the web: remoteness, illiteracy, gender, age, and lack of readiness (having the relevant skills, awareness and acceptance). It is estimated in 2015 that: one billion people are illiterate; there is no relevant language content for two billion people; females are less likely than males to use the internet due to a variety of issues, and approximately two thirds of unconnected people do not know what the internet is. Geographic differences in participation have also been observed, with the most connected populations located in North

America and the least connected in South Asia. [Wu et al., 2015; Sprague et al., 2014]

Technology The infrastructure required to gain access to the web is not uniformly distributed. Power provision, internet connections, availability of devices and usability of interfaces all contribute to this uneven geographic distribution. Regarding internet connection, on average 96% of the global population has access to a 2G mobile network. About 1.6 billion people in 2015 lived outside broadband coverage, mainly in remote areas of developing countries [Wu et al., 2015].

Economic Affordability of internet access and devices is another barrier to participation, in 2015 over 3 billion people are estimated to be unable to afford mobile broadband packages – the most common way to access the internet. It costs more to install infrastructure to cover remote areas, and this, combined with the lower number of users makes it an unattractive prospect to providers to cover remote regions [Wu et al., 2015].

Legal and Political In addition to legislation that defines illegal content, such as child pornography, protection of freedom of speech and incitement of hatred, some nations specifically seek to restrict and censor access to web content. Work by initiatives such as Freedom House, WebIndex, OpenData and Reporters without Borders have investigated web censorship. Motivations for web censorship include political and religious sensitivities with nations giving a number of reasons for censorship such as crime prevention, counter-terrorism, and public morality [Warf, 2011]. Nations involved in extensive web censorship, according to Freedom House 2015 index, include China, Syria, Iran, Ethiopia and Cuba [Statistica, 2015].

No Quality Control A major feature of web content is the lack of quality control, resulting in a great variance in content quality from peer reviewed to maliciously false e.g. fake news. Further complicating the quality issue, but not unique to web data, is that quality itself is perceived relative to the user. A given piece of information may be considered fit-for-purpose by one user but not another. A number of frameworks exist addressing information quality, some of which have been summarised by Shirlee-ann Knight and Janice Burn as they propose their own framework for assessing quality of web-based information. Examples of quality features

that appear in a number of different frameworks include: accuracy, source, completeness and currency. Regarding web content, other factors are also considered such as properties of the web-page itself including: last time document was modified; ratio of broken to working links on the page and number of links to the website. Knight and Burn propose a quality assessment that is conducted by a web crawler, meaning that the web content is only indexed if it meets quality standards [Knight and Burn, 2005].

This highlights an issue for the Research Framework that filters may have been applied at various stages of the information retrieval process, both by a web crawler and at the retrieve and rank process which are both expanded upon in the following section. Also the question of web content quality has at least four dimensions covering the properties of: information in isolation, the website, the author and the user. The author is important from the perspective of quality because they generate web content for a variety of reasons including compliance, marketing or even by mistake [Weiss, 2014].

4.2.2 Information Exploitation

The exploitation of information in its widest sense is an ubiquitous activity. Some disciplines have formalised methodologies or best practices for information exploitation and as a result many terms are used to describe this activity such as *Intelligence Analysis* or *Decision Making under Uncertainty*. These terms in turn have multiple definitions, but for the purpose of this research the salient features of information exploitation are that it requires: data and/or information as input; quality assessment of input; filtering of relevant input; integration of selected input; interpretation (taking into account uncertainties) and application (use of output). [MoD, 2011; NATO, 2013]

This research is focused on two aspects of the information exploitation challenge: sourcing of input from the web and integration of input, in particular inference or reasoning with the in-

put. To address this we look at the disciplines of web mining, mathematical inference/reasoning and ML.

Web Mining Web mining, or data mining of the web, refers to a set of automated techniques used to extract structured data of interest from potentially unstructured data/information and, in some cases pattern discovery within the extracted data. For the purpose of this thesis the term web scraping will be used to refer to the extraction task. Web mining can be divided into three classes: web content mining, web structure mining, and web usage mining. Here we are interested in the first of these, web content mining, which is itself also quite broad. Due to the wealth of different types of data available through the web it is a necessarily extensive discipline in computer science including natural language processing. The Framework demonstration is conducted using structured statistical data. This data source is examined next. However, the use of unstructured data is also of interest to this Framework and this is further expanded upon in Chapter 6.

Standardised Statistical Data As described by Tim Berners-Lee and fellow authors, in the seminal paper *The Semantic Web*, the Semantic Web refers to an international effort led by the World Wide Web Consortium (W3C), an international collaboration to establish standards for the web, and to make web-based information more readily interpretable by machines. It consists of several parts: markup languages described below, ontologies (a taxonomy and inference rules), and digital signatures. [Cailliau, c.1995; Berners-Lee, 2000; Berners-Lee, Hendler, and Lassila, 2001]

- Hypertext Markup Language (HTML) used to author web pages and web applications.
- Extensible Markup Language (XML) is used to encode documents in a manner that is both human and machine readable.
- RDF a method to represent knowledge as (graph) triples (node, edge, node).
- Uniform Resource Identifier (URI) a way of uniquely identifying a web resource.

The standard of interest here is the Statistical Data and Metadata eXchange (SDMX), which resulted from an international effort to standardise the way statistical data, and its metadata,

is exchanged [SDMX.org, 2018]. The approach employed by this research to scrape statistical data is to query an SDMX encoded database via an Application Program Interface (API). The result of which is then parsed for the data of interest.

Mathematical Inference Inference is the process of creating new data/information from existing data. Inference Methods include BBN, HMM, Dempster-Shafer and Argumentation. Here we focus on BBN.

Bayesian Belief Networks BBNs are based on the well-known Bayes Theorem which states that the probability of observing A given B has occurred is given by Equation (4.1). The observables A and B are variables which are represented by nodes in a BBN. The Joint Probability Distribution for a generalised BBN with nodes A_1 to A_n is shown by Equation (4.2). This can be simplified by taking into account only connected nodes in the BBN, as shown by Equation (4.3).

$$P(A|B) = \frac{P(B | A)P(A)}{P(B)} \quad (4.1)$$

$$P(A_1, A_2, \dots, A_n) = \prod_{i=1}^n P(A_i | A_{i+1}, \dots, A_n) \quad (4.2)$$

$$P(A_1, A_2, \dots, A_n) = \prod_{i=1}^n P(A_i | \text{Parents}(A_i)) \quad (4.3)$$

A well-known trivial example is the observation that the grass is wet and wanting to know the probability that this is due to rain or the garden sprinkler [Russell and Norvig, 1995]. This example is illustrated in Figure 4.1. In this model, the state of the grass (G) is *wet* or not wet ($\neg wet$) and is determined by the state of the weather (W) and the state of the sprinkler (S) which can be *rain* or not rain ($\neg rain$) and *on* or *off* respectively. To calculate the probability that the observation that the grass is wet due to rain is given by substituting the variables of interest into Equation (4.1) to give Equation (4.4a). From there substitutions for the numerator and denominator can be found using Equation (4.3), the sum rule and rearranging to give Equations (4.4b) and (4.4c). Substituting these into Equation (4.4a) gives Equation (4.4d).

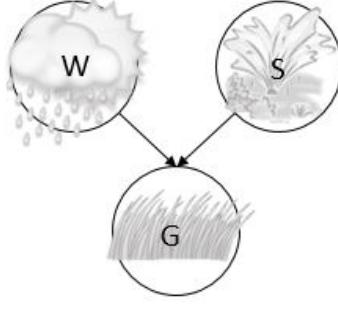


Figure 4.1: Trivial example of a BBN showing grass (G), weather (W) and sprinkler (S) nodes.

$$P(W = \text{rain} \mid G = \text{wet}) = \frac{P(G = \text{wet} \mid W = \text{rain})P(W = \text{rain})}{P(G = \text{wet})} \quad (4.4a)$$

$$P(G = \text{wet} \mid W = \text{rain}) = \frac{P(W = \text{rain}, G = \text{wet})}{P(W = \text{rain})} \quad (4.4b)$$

$$P(G = \text{wet}) = P(G = \text{wet} \mid W, S) \quad (4.4c)$$

$$P(W = \text{rain} \mid G = \text{wet}) = \frac{P(W = \text{rain}, G = \text{wet})}{P(G = \text{wet} \mid W, S)} \quad (4.4d)$$

$$P(W = \text{rain} \mid G = \text{wet}) = \frac{0.6 \times 1}{0.006 + 0.004 + 0.594} = \frac{0.6}{0.604} = 0.9934 \quad (4.4e)$$

This method requires that something is known about the probabilities of the state of the grass, weather and sprinkler, as well as the state of the grass given states of the weather and sprinkler, the prior probabilities. These are shown in Tables 4.1 to 4.4. Substituting these values into Equation (4.4d) gives Equation (4.4e).

Table 4.1: Weather state probability table

<i>rain</i>	$\neg \text{rain}$
0.6	0.4

Table 4.2: Sprinkler state probability table

<i>on</i>	<i>off</i>
0.01	0.99

Table 4.3: Grass state probability table

	Weather	<i>rain</i>		$\neg rain$	
	Sprinkler	<i>on</i>	<i>off</i>	<i>on</i>	<i>off</i>
Grass	<i>wet</i>	1.0	1.0	1.0	0
	$\neg wet$	0	0	0	1.0

Table 4.4: Probability calculations

<i>W</i>	<i>S</i>	$P(G = wet)$	$P(W) \cdot P(S)$
<i>rain</i>	<i>on</i>	1	$0.6 \cdot 0.01 = 0.006$
$\neg rain$	<i>on</i>	1	$0.4 \cdot 0.01 = 0.004$
<i>rain</i>	<i>off</i>	1	$0.6 \cdot 0.99 = 0.594$

Machine Learning ML falls within the field of AI and is based on algorithms that develop models based on data rather than the model being explicitly programmed. There are many different types of ML algorithms and types of models that can be learnt. At the simplest level of abstraction a ML algorithm develops, aka *learns*, the best function, f , that maps the input variable, x , to the output variable y , see Equation (4.5). In the case of multiple input and output variables, these are represented by the vectors $\mathbf{x}$ and $\mathbf{y}$ as shown by Equation (4.6).

$$y = f(x) \quad (4.5)$$

$$\mathbf{y} = f(\mathbf{x}) \quad (4.6)$$

Types of Machine Learning ML algorithms can be categorised along a number of dimensions. One such dimension is the way in which the algorithm learns the model, and is considered as using either a supervised, unsupervised, semi-supervised or reinforcement learning method. Supervised learning refers to algorithms that use a set of data where the output y is known for the input x , also known as a labelled dataset. Unsupervised learning algorithms use data without a known output, an unlabelled dataset. These algorithms, rather than predict

the outcome y , find structures within the data x . Semi-structured, as the name suggests, is a mixture of the two, namely the data has some examples with known output and others without. Reinforcement learning occurs when the algorithm receives feedback from the system on how well it is doing at performing a task. A second way to characterise machine learning algorithms is by the type of output they produce. Classification algorithms assign inputs to two or more classes, i.e. a discrete result. In contrast, regression algorithms produce results which are continuous. Clustering algorithms are similar to classification algorithms except the classes are not known a priori; the algorithm has to deduce them. A third characterisation, and the final one described here, is by families of machine learning techniques which work in similar ways such as Bayesian, Decision Trees, Support Vector Machines, and Instance Based.

Basic Pipeline A basic supervised machine learning task, illustrated in Figure 4.2, is approached as follows. First the data is transformed into the required structure for the chosen ML algorithm and is divided into training and test datasets. The chosen ML algorithm ingests the training set and learns the parameters that are most likely to produce the correct outcome y for a given x . This is the trained model. Next the test set is used to evaluate the performance of the model by comparing the model’s prediction of the output against the known ones. There are different methods of splitting the dataset and training the model so that it does not become very good at predicting the test set, known as over fitting, but instead can generalise well to unseen data. This is not further expanded on.

Performance Typically, performance of a ML categorisation model is measured by precision, recall and a combined metric known as F , see Equations (4.7), (4.8), and (4.9). Precision and recall are both measures of how good the model is at predicting a given category. Precision compares the number of times the category is correctly predicted as a proportion of the total number of times the category is predicted. Recall compares the number of times the category is correctly predicted as a proportion of the total number of opportunities for a correct prediction. The F metric is a combination of precision and recall. There are several versions of the F metric, weighting precision and recall as befits the situation and follows the notation F_β where β represents the relative weight of recall to precision. For the purpose of this research there is

no particular emphasis to preference the precision or recall metric, the F_1 metric (unweighted) is therefore used. In the case of a multi-categorical model, precision, recall and F are calculated for each category in addition to the average over all categories. This average can be weighted to account for any imbalance (skew) in the number of examples in each category.

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (4.7)$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (4.8)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (4.9)$$

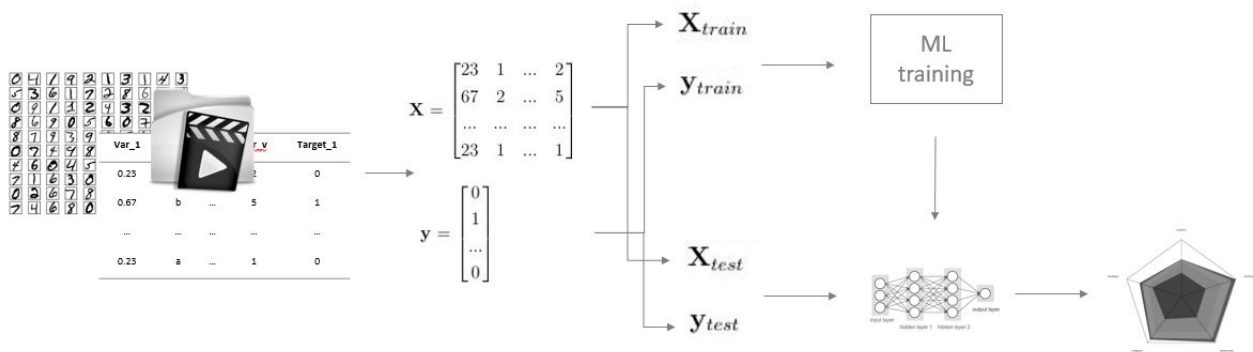


Figure 4.2: Illustration of ML basic pipeline, showing: different types of data, transformation into required format, split into training and test sets for ingest by learning algorithm, producing a trained model which is then evaluated using the test data. Contains images under the creative commons licence [Nielsen, 2017; Wikispaces, 2018; Karpathy, 2018; Garg, 2015].

4.2.3 Pomegranate: BBN ML Algorithm

This thesis employs the ML library *Pomegranate*, a Python library developed by Jacob Schreiber [Schreiber, 2018b]. Two functions were extensively used in this research and are explained here: *BayesianNetwork.from_samples()* and *BayesianNetwork.predict()*. Note that this algorithm takes the input and output variables as one matrix which will be referred to as $\mathbf{X}$.

Structure and Parameter Learning from Training Data The *from_samples()* function takes data and outputs the learnt structure (graph) of a BBN and its prior probability tables. The parameters used are illustrated in Figure 4.3. The first parameter, $\mathbf{X}_{\text{train}}$, is a data matrix where the columns represent the variables of interest and the rows represent each sample. The variables form the nodes of the resultant BBN. The second parameter specifies the structure learning algorithm. The two algorithms used in this research, *Exact* and *Chow-Liu* are described below. The final parameter is the constraint graph which restricts the search space of potential BBN structures by specifying which nodes are allowed to be linked by directed edges. In the illustration at Figure 4.3, variable var_1 can only be a parent of variables var_2 and/or var_3 . All other edges are excluded. The *from_samples* algorithm investigates the potential structures of Directed Acyclic Graphs (DAG) for the given nodes and constraint graph and applies a scoring function, the Minimum Description Length (MDL), which balances model complexity with performance. MDL is explained in greater detail below.

BayesianNetwork.from_samples($\mathbf{X}_{\text{train}}$, structure learning algorithm, constraint graph)

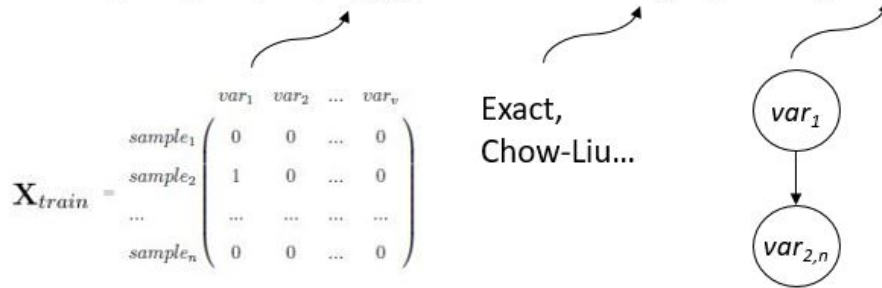


Figure 4.3: Pomegranate’s *BayesianNetwork.from_samples* function, showing input parameters.

The *Exact* Algorithm An exact inference algorithm sets out to find an optimal structure for a BBN by examining all possible DAGs and has been proven to be Nondeterministic

Polynomial (NP) hard for graphs other than those that are singly connected [Yuan, Malone, and Wu, n.d.; Guo and Hsu, 2002; Schreiber, 2018a]. There are a number of exact algorithms; Haipeng Guo and William Hsu give a number of examples in Guo and Hsu [2002]. The exact algorithm implemented by Pomegranate combines dynamic programming and A* algorithms. The dynamic programming applied has computational complexity to $O(n2^n)$ where n is the number of nodes in a BBN [Yuan, Malone, and Wu, n.d.]. The A* algorithm reduces the search space to improve computation time. The result is an *optimal* BBN. The A* algorithm is described by Changhe Yuan, Brandon Malone and Xiaojian Wu in Yuan, Malone, and Wu [n.d.]. [Schreiber, 2018a].

The Chow-Liu Algorithm Given that exact algorithms are NP-hard and therefore do not scale to graphs with large numbers of nodes, a number of researchers have examined approximate approaches [Guo and Hsu, 2002]. One such method is the Chow-Liu tree building algorithm which is $O(n^2)$, overcoming the NP issue by relying on a singly connected graph, namely a tree [Chow and Liu, 1968; Schreiber, 2018a]. The method is explained in detail at [Chow and Liu, 1968] and briefly here. First, the sample joint frequency of all pairs of variables are calculated, this is simply a ratio of the number of occurrences of a pair of variables for given values to the total number of occurrences of the pair for all values. From these the *sample mutual information*, $\hat{I}$, is calculated, again for each pair of variables. The measure of *mutual information* between two variables is explained by Chow and Liu as ‘a measure of closeness in approximating’ one variable by another [Chow and Liu, 1968, p. 463]. For example, if two variables are drawn from the same probability distribution, the measure of *mutual information* would equate to one. The tree is then formed by treating $\hat{I}$ as a weight and ordering the branches by decreasing value.

Minimum Description Length The MDL sets out to find the *best* graph for predicting the data and is based on Jorma Rissanen’s principle of data compression. The goal is to minimise the sum of two parameters, one describing the size of the BBN in terms of the number of nodes and edges and the size of the BBN’s model of the training data, both measured by the number of bits needed to describe them. The latter term tends to favour completely connected

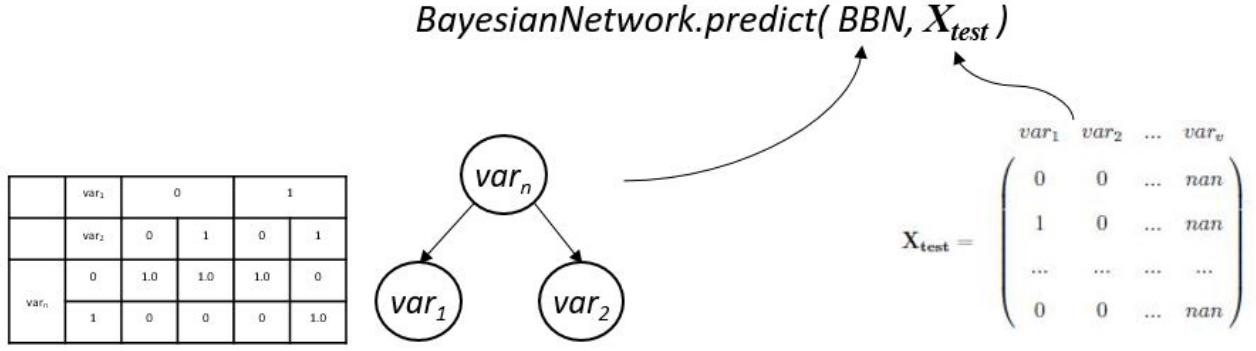


Figure 4.4: Pomegranate’s BayesianNetwork.predict function, showing input parameters.

graphs, while the former penalizes them. The graph that satisfies the minimal description length is one that balances predictive accuracy with graph size. [Friedman, Geiger, and Goldszmidt, 1997; Tian, 2013]

Constraint Graphs Constraint graphs have two functions. The first is to give the user the ability to define which nodes can be connected by directed edges, thus representing a priori information about the graph in addition to the sample data. The absence of an edge in the constraint graph means that edge cannot appear in the learned structure of the BBN. As a consequence, the second function of the constraint graph is to reduce the potential search space of the structure learning algorithms by excluding a number of potential DAGs and hastening the computation.

Prediction from Test Data The *predict()* function takes the BBN generated by the previous function, which consists of a graph and prior probability table, and a data matrix $\mathbf{X}_{test}$ with missing values as shown at Figure 4.4¹. The outputs are the most likely value for those missing numbers. The algorithm uses the Maximum Likelihood Estimate and loopy belief propagation algorithms. The Maximum Likelihood Estimate simply returns the most likely value. The loopy belief propagation is an approximate algorithm², that calculates the probability of a node being in some state given the states of the other nodes. [Schreiber, 2018b]

¹The missing variables are equivalent to the output $\mathbf{y}_{test}$ variables referred to in Section 4.2.2 and shown in Figure 4.2. Although in this case the missing values can appear anywhere in the matrix, not just within a column (vector).

²It is exact on some graphs, such as trees.

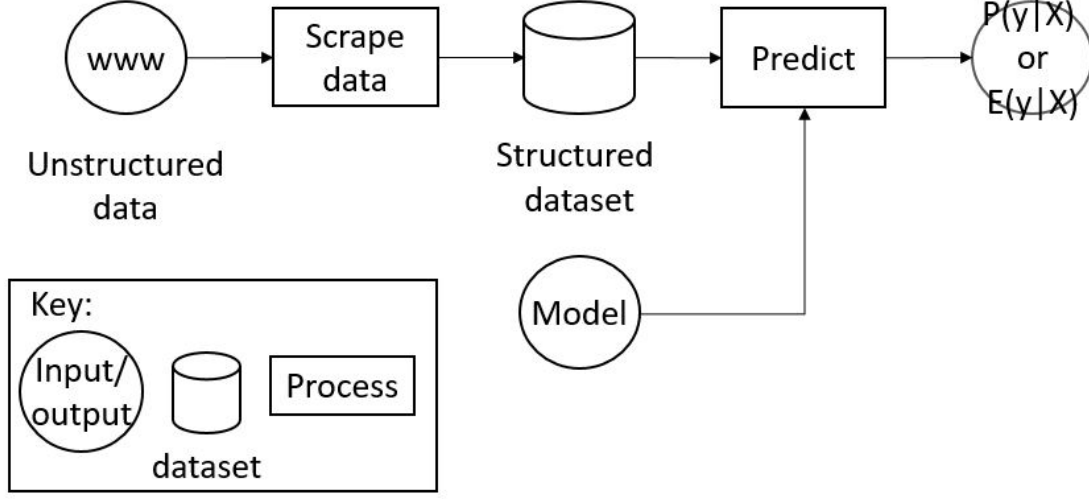


Figure 4.5: Top level Research Framework architecture as a pipeline diagram.

Loopy Belief Propagation Belief propagation makes inferences on graphs, such as a BBN. The nodes with observations inform the *marginal distribution* for each node without an observation. The algorithm attempts to converge the marginal distributions for each unobserved node. The Loopy variation works on many types of graphs and although convergence is not guaranteed, it has been shown to be a good approximation. [Ihler, Fisher, and Willsky, 2005]

4.3 Research Framework

Bringing together web scraping, the conceptual graph model and ML, the Research Framework is illustrated at Figure 4.5. The Framework takes data requirements from the graph model, the required data is scraped from the web and transformed into a structured dataset, which combined with the ML model, also defined by the graph model, makes predictions of the target variable.

Given the conceptual graph, the aspiration for the ML model is to predict the expected value of a Nation's intent at a given time, $E(NatInt(T))$ given values of national industrial policy and national industrial capability and the associated company-product mixes at previous time periods. This is represented mathematically as:

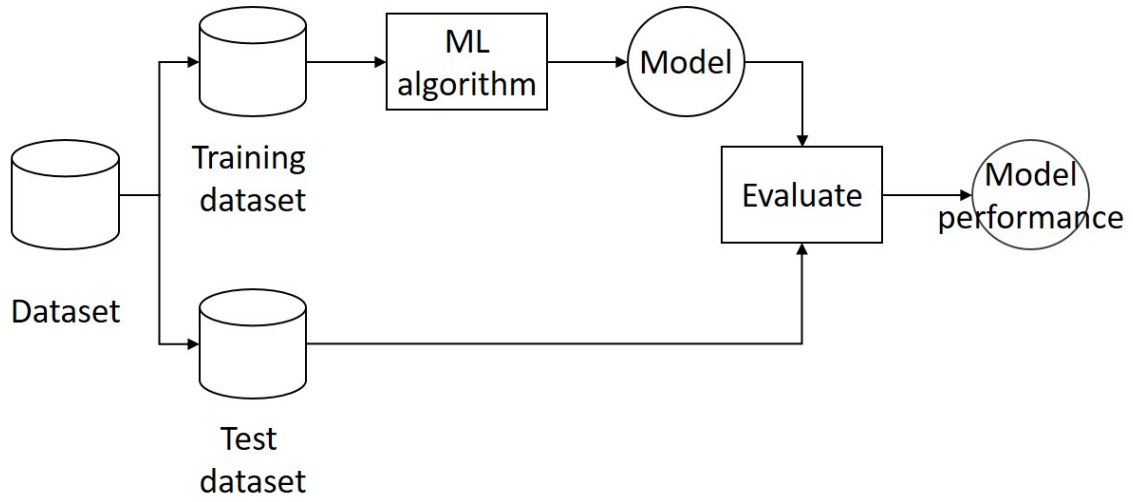


Figure 4.6: Top level ML architecture for model development. Showing use of training and test datasets to train the model and evaluate its performance.

$$E(\text{NatInt}(T) | \text{NatIndPol}(t < T), \text{NatIndCapProCo}, (t < T)) \quad (4.10)$$

There are many options for rendering the problem as a ML problem. For instance to treat it as a continuous regression problem or a categorical prediction. Another decision is when a prediction is made. Options include one prediction per time-step or, taking a sequence prediction approach, one or more predictions per n time-steps. Without precluding any of these options, the process of model development is shown at Figure 4.6 for a simple case of supervised ML. The scraped data is split into training and test datasets. The training set is used to train the model which is subsequently evaluated using the test set. The output being a model with known performance. This model can then be used repeatedly either at the behest of the user or when new data is available.

The framework illustrated is generic, it is agnostic to the type of data provided, the way data is subsequently structured and the type of ML algorithm used. For the purpose of this PhD several instances of this Research Framework are implemented and are described in the next chapter.

4.4 Conclusion

This chapter has documented the Research Framework and its supporting literature, the second contribution of this thesis. In the next chapter, the methodology and results for demonstrating aspects of the Research Framework is documented and conclusions for each demonstration shown.

Chapter 5

Framework Demonstration 1 & 2

5.1 Introduction

Chapter Aim and Structure The purpose of this chapter is to document the two demonstrations of the Framework, in each case describing the methodology employed, the results found and associated discussion. It fulfils the third contribution of this thesis, Framework demonstration. The structure of this chapter is as follows:

- Section 5.2 Demonstration 1 Simulated data (2 experiments).
- Section 5.3 Demonstration 2 Web scraping statistical data.
- Section 5.4 Demonstration 2 Case Study 1: Renewable electricity generation.
- Section 5.5 Demonstration 2 Case Study 2: State aid.

Demonstrated Framework The framework demonstrations aim to prove the concept of linking the graph conceptual model with web scraping and ML. The first demonstration focuses on the graph model and the selected ML algorithm: BBN ML, thus excluding the web scraping step through the use of simulated data. The second demonstration incorporated the web scraping step, using the example of statistical data encoded in a standard format. This necessitated the adaptation of interpretation of the sets *Ind* and *Cap*. In both cases the input and target variables considered are categorical. The scope of each demonstration is described

below, with greater details covered in the respective methodology sections. Of note is that only a portion of the conceptual graph model is demonstrated and that despite the focus of national intent as a key aspect of the conceptual graph model, this node is not incorporated in the demonstrations, but is identified as an area for future research. Also, the edges between the nodes that are incorporated in the demonstrations are considered as directed for the use in conjunction with a BBN.

Demonstration 1: Simulated data The first demonstration uses simulated binary data of *Nat*, *Ind* and *Cap* nodes, although for the purpose of this demonstration $|Nat|$ was equivalent to the number of examples from a ML perspective. A BBN ML algorithm was applied in a supervised manner to create a model to predict $E(Cap | Ind)$. The use of simulated data allowed: confirmation that the BBN could extract relations between national industry sets and national capability; investigation into the effect of the size of *Ind*, *Cap* and *Nat* sets on performance; and investigation of two significant characteristics of web data. As noted in Section 4.2.1, web data has no quality control. It was therefore of interest how sensitive the performance of the trained model was to noise¹ in the data. Second there is uncertainty in how complete any dataset will be, so the sensitivity in performance of the trained model to sparse² data was also of interest. An illustration of what was demonstrated in relation to the Research Framework is in Figure 5.1.

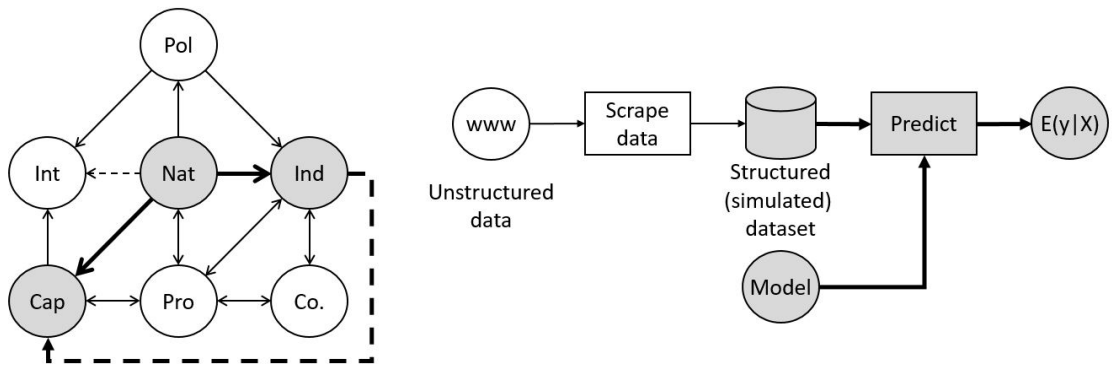


Figure 5.1: Scope of first demonstration in relation to complete Research Framework.

¹Here *noise* refers to incorrect data.

²Here *sparsity* refers to how complete the data is, i.e. is not a reference to matrices with a majority of elements that are equal to zero

Demonstration 2: Scraped statistical data Here statistical data in SDMX format is scraped from the web and categorised. To facilitate the use of statistical data, the definitions of the industry and capability nodes were adapted to have states of industrial statistical variables and a measurement of capability respectively. As before, BBN ML was applied. This demonstration framework was applied to two real-life case studies: renewable energy and state aid funding. The focus of this demonstration was the use of real-world data. An illustration of what was demonstrated in relation to the Research Framework is in Figure 5.2.

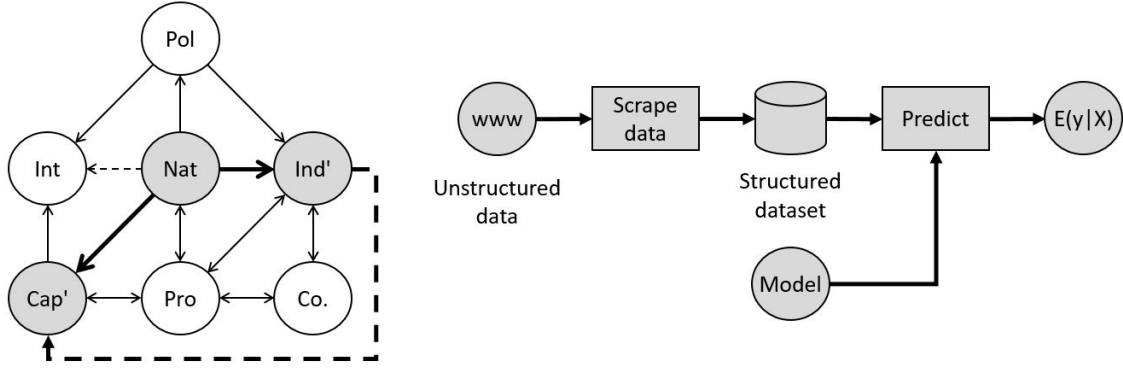


Figure 5.2: Scope of second demonstration in relation to complete Research Framework.

5.2 Demonstration 1: Simulated Data

As stated in Section 5.1, the purpose of Demonstration 1 is to confirm that BBN ML can learn relations between national industry and national capability, and determining the effect on model performance of the parameters: size of Ind , Cap , and Nat sets, the effect of noise in and sparsity of the input data. Two experiments are conducted using simulated data: the first varies the size of the three sets given by $|Nat|$, $|Ind|$ and $|Cap|$, the second varies percentage noise and sparsity of the input data. These experiments address elements of the conceptual graph model and Research Framework as illustrated in Figure 5.1. For the purpose of this demonstration $|Nat|$ is equivalent to the number of samples in an ML algorithm.

5.2.1 Method

The overall method for Demonstration 1 is shown in Figure 5.3. In Figure 5.3a parameters are initialised and a process is looped through until each combination of parameters has been evaluated multiple times. The process begins by creating simulated data and a constraint graph. Then the simulated data is transformed, resulting in training data and two versions of the test data. The training data and constraint graph are used to train a model using BBN ML as shown in Figure 5.3b. The resulting model is evaluated and the performance captured. In order to evaluate performance, the typical metrics of precision, recall and F_1 are used in addition to *percentage correctness* which is explained at Section 5.2.1.5.

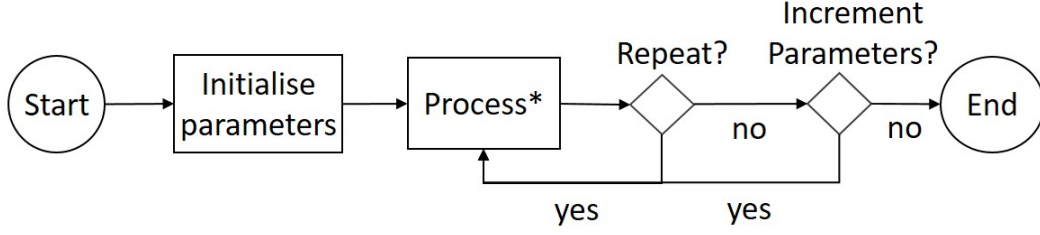
An example output of a model and its performance, is shown in Figure 5.7. The inner loop repeats the process to get multiple performance metrics for each combination of parameters. The outer loop increments the experiment parameters. Figure A.1 shows the pseudocode for these steps. Below each step is explained in greater detail and illustrated as necessary using an example where $|Nat| = |Ind| = |Cap| = 3$.

5.2.1.1 Create Constraint Graph

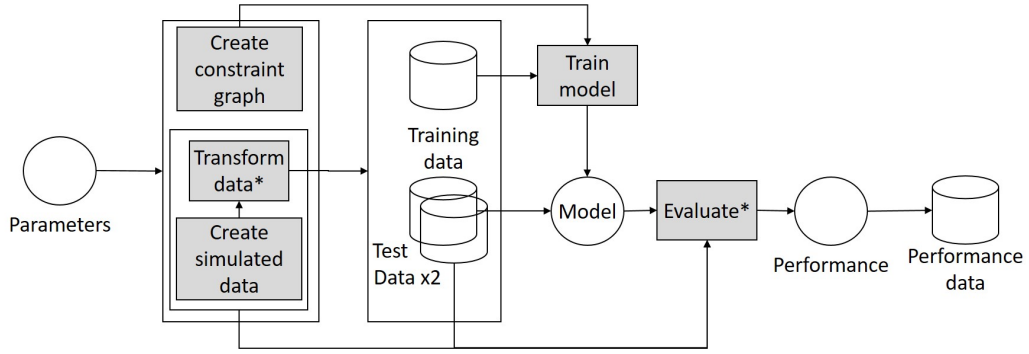
The pseudocode for this algorithm is shown in Figure A.2. The input of this step is the settings of parameters $|Ind|$ and $|Cap|$. The output is a graph described by a set of triples defining which nodes in the model may be linked, i.e. how the model is to be *constrained*. Here only the industry nodes can be linked by directed edges to the capability nodes, described by a triple $\langle Ind_i, Cap_c \rangle$ where $i = 1$ to $|Ind|$ and $c = 1$ to $|Cap|$. A zero-based indexing convention is adopted for labelling the model nodes: industry nodes are labeled 0 to $(|Ind| - 1)$, capability nodes are labeled $|Ind|$ to $(|Ind| + |Cap| - 1)$. A constraint graph for an example where $|Ind| = |Cap| = 3$ is shown in Figure 5.4.

5.2.1.2 Create Simulated Data

The input of this step is the settings of $|Nat|$, $|Ind|$ and $|Cap|$. The output is three binary matrices: $NatInd$ $|Nat| \times |Ind|$, $IndCap$ $|Ind| \times |Cap|$ and $NatCap$ $|Nat| \times |Cap|$. These



(a) Simulated data high level method illustrating two loops to repeat the process, * shown in more detail at Figure 5.3b, multiple times for each combination of parameter setting.



(b) Process from Figure 5.3a shown in detail. * Transform data and evaluate processes shown in more detail at Figures 5.5 and 5.6 respectively. Shaded processes explained in more detail in Sections 5.2.1.1 through 5.2.1.5.

Figure 5.3: Simulated data high-level method and detail of Process.

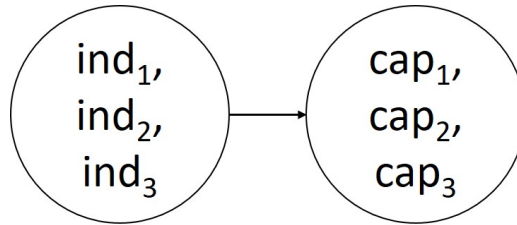


Figure 5.4: Constraint graph for example where $|Ind| = |Cap| = 3$. A BBN created using this constraint would only have edges $\langle Ind_i, Cap_c \rangle$.

represent the *ground truth* data in the form of adjacency matrices. First the *NatInd* and *IndCap* matrices are created and populated with binary data using a uniform distribution. Second *NatCap* is created in order to be consistent with *NatInd* and *IndCap*. An example for $|Nat| = |Ind| = |Cap| = 3$ is shown at Equations: (5.1), (5.2) and (5.3). In accordance with Equation (5.1), *Nat*₁ has industries *Ind*₁ and *Ind*₂ but not *Ind*₃. It can be inferred from Equation (5.2) that *Nat*₁ does not have the requisite industries to support any of the capabilities as shown by Equation (5.3). For demonstration purposes it is arbitrary which two of the three matrices is populated using random binary data and which is calculated. Similarly, the uniform distribution was also arbitrarily chosen. The pseudocode for creating simulated data is shown in Figure A.3

$$NatInd = \begin{matrix} & \begin{matrix} Ind_1 & Ind_2 & Ind_3 \end{matrix} \\ \begin{matrix} Nat_1 \\ Nat_2 \\ Nat_3 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{matrix} \quad (5.1)$$

$$IndCap = \begin{matrix} & \begin{matrix} Cap_1 & Cap_2 & Cap_3 \end{matrix} \\ \begin{matrix} Ind_1 \\ Ind_2 \\ Ind_3 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{matrix} \quad (5.2)$$

$$NatCap = \begin{matrix} & \begin{matrix} Cap_1 & Cap_2 & Cap_3 \end{matrix} \\ \begin{matrix} Nat_1 \\ Nat_2 \\ Nat_3 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (5.3)$$

5.2.1.3 Transform Data

The input of this step is the simulated data in the form of three matrices and the settings of percentage sparsity and noise. The output is the training dataset and two versions of the test dataset. Three transformations are applied: addition of noise and sparsity, concatenation and split as shown in Figure 5.5. Each transform is described in more detail next.

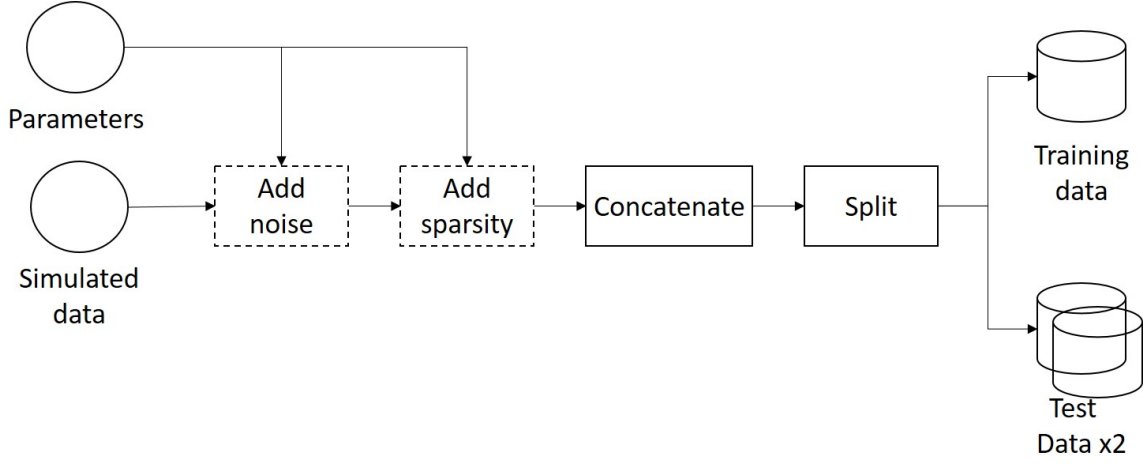


Figure 5.5: Detail of transform data process. A dotted line indicates an optional process.

Add Noise and Sparsity This is an optional step when investigating parameters of percentage noise and sparsity. Two functions were created to add random noise to a matrix and remove data as needed, their pseudocode is shown in Figure A.4. The noise function takes as input a matrix $\mathbf{A}$ and the setting for percentage noise. Using a mask, the required percentage of random elements of the matrix, a_{ij} , are replaced by a random binary number and the matrix returned. The sparsity function works in a similar way to the noise function. It takes as input a matrix $\mathbf{A}$ and the setting for percentage sparsity. Using a mask, the required percentage of random elements of the matrix a_{ij} , are replaced by a non value such as Not a Number (NaN) and the matrix returned. As an example, applying these transforms to Equation (5.1) gives Equation (5.4) where element Nat_1Ind_2 has been removed and element Nat_2Ind_2 has been changed to an incorrect value. Note that for the $NatCap$ matrix, all but the final column is transformed. This is because this column, relating to Cap_c will be used to make predictions and calculate performance metrics.

$$NatInd' = \begin{matrix} & Ind_1 & Ind_2 & Ind_3 \\ \begin{matrix} Nat_1 \\ Nat_2 \\ Nat_3 \end{matrix} & \begin{pmatrix} 1 & NaN & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \end{matrix} \quad (5.4)$$

Concatenate Here the $NatInd$ and $NatCap$ matrices are concatenated to give a $|Nat| \times (|Ind| + |Cap|)$ shaped matrix. Applying this transform to the matrices by Equations (5.1) and (5.3) gives the matrix shown by Equation (5.5).

$$NatIndCap = \begin{matrix} & Ind_1 & Ind_2 & Ind_3 & Cap_1 & Cap_2 & Cap_3 \\ \begin{matrix} Nat_1 \\ Nat_2 \\ Nat_3 \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \quad (5.5)$$

Splitting The final transformation in this step is splitting the $NatIndCap$ matrix into two matrices, or datasets, by row. In this application the matrix is split randomly, with 70% forming the training dataset and 30% forming the test dataset. Two versions of the test dataset are created. The first version is the ground truth. The second has the final column, which corresponds to Cap_c , replaced with no values and is used in the evaluate step. The pseudocode is shown in Figure A.5.

5.2.1.4 Train Model

The input of this step is the constraint graph and training dataset. The output is a BBN model with structure and parameters given by a graph and prior probability tables. The structure represents the learnt $IndCap$ adjacency matrix. As previously mentioned, the Pomegranate Python library is used in this demonstration, specifically the *BayesianNetwork.from_samples* function with the *exact* algorithm. An example model structure output is shown in Figure 5.7.

5.2.1.5 Evaluate

Two types of performance are calculated in the evaluate step, as shown in Figure 5.6. The first is percentage correctness, a metric developed for this demonstration. The second is standard ML metrics of precision, recall and F_1 .

Percentage Correctness The input of this step is the simulated $IndCap$ matrix representing the ground truth and the model graph. The output is a metric *percent correctness*. To

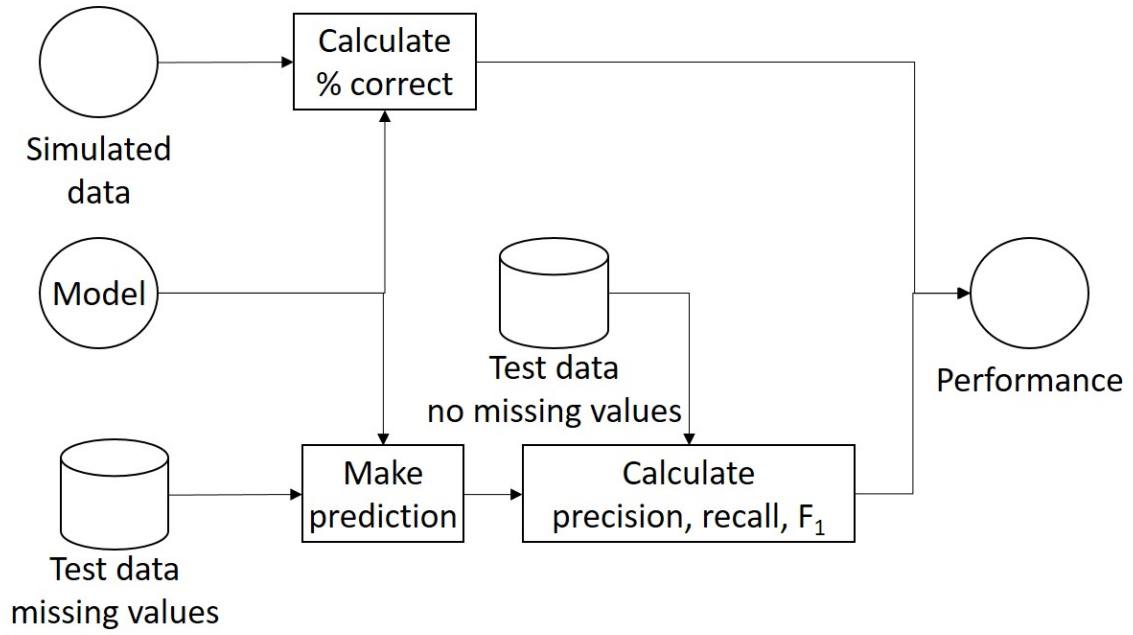


Figure 5.6: Detail of evaluate process, illustrating the method for comparing the model's prediction for missing data with the correct answers for four performance metrics.

calculate this metric, the model graph is transformed into its corresponding adjacency matrix, $IndCap_{model}$. Each element of $IndCap$ is compared with $IndCap_{model}$ and the percentage match of elements calculated. An exact match across all elements corresponds to a 100% correctness. No match corresponds to 0% correctness. It must be noted that due to the binary nature of the matrices, as noise tends to 100%, percent correctness will tend to 50%. The pseudocode for this algorithm is shown in Figure A.6.

ML Metrics The input of this step is the two versions of the test datasets and the model. First, the model is used with the test data with missing values to calculate predictions for those missing values. These predictions are then compared with the complete version of the dataset using the scikit learn Python library to calculate precision, recall and F_1 metrics. Example performance metrics by binary category for a trained model are shown in Figure 5.7.

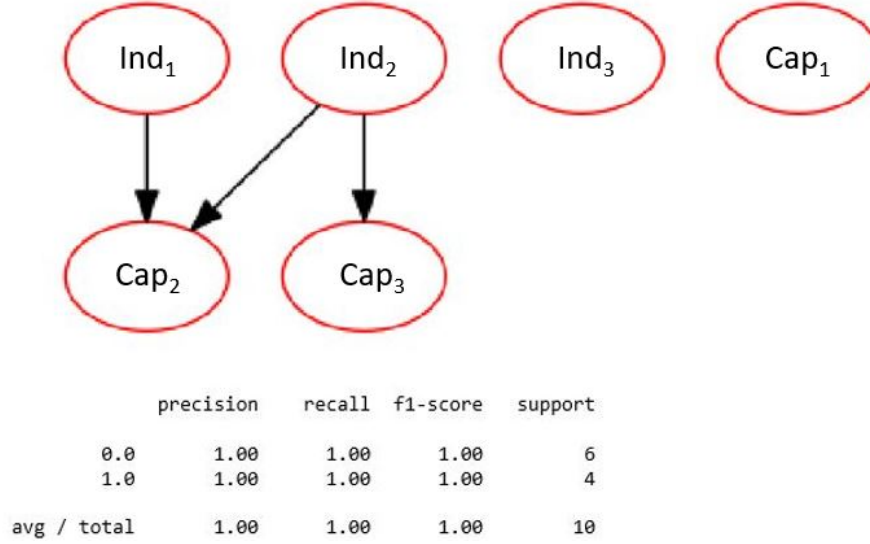


Figure 5.7: Adapted example output of Pomegranate’s BayesianNetwork.from_samples algorithm, a BBN with its corresponding performance. An edge represents a 1 in the corresponding adjacency matrix. The performance metrics of precision, recall and F_1 by binary category for the prediction of the Cap_3 node is shown. 1.00 is the maximum value for these metrics. Note the high performance due to the low $|Ind|$.

5.2.1.6 Experiments

As already noted, two experiments were conducted:

Experiment 1 The parameters $|Nat|$, $|Ind|$ and $|Cap|$ were varied as follows, with 30 repeats of data creation and model training and evaluation being conducted for each combination of parameter settings in order to calculate an average performance:

- $|Nat|$ took values: 100, 200, 300, 500, 1000
- $|Ind|$ and $|Cap|$ took values in range 3 to 57 in steps of 6
- Percentage noise = Percentage sparsity = 0
- Repeats = 30

Experiment 2 The parameters percentage noise and sparsity were varied as follows, with 50 repeats of data and model training and evaluation being conducted for each combination of variable setting in order to calculate an average performance. The baseline for $|Nat|$,

$|Ind|$ and $|Cap|$ were set at 200, 3 and 3 respectively. These values were chosen based on the results of experiment 1 because the correctness of the returned model $\rightarrow 100\%$, thus any change in performance can be attributed to percentage noise and sparsity.

- $|Nat| = 200$
- $|Ind| = |Cap| = 3$
- Percentage noise took values in range 0 to 100% in steps of 1%
- Percentage sparsity took values in range 0 and 100% in steps of 1%

5.2.2 Results

5.2.2.1 Experiment 1: Changing $|Nat|$, $|Ind|$, $|Cap|$

Result 1 As $|Nat|$ increases, the higher the percentage correctness of the returned model compared to the ground truth for a given $|Ind|$ as illustrated in Figures 5.8 and 5.9. In Figure 5.8 this can be observed by looking at the spread of points in the y dimension within each shape grouping where the shape indicates a given value of $|Ind|$ and the colours indicate $|Nat|$. For higher values of $|Nat|$, the greater the average percent correct of the resultant model measured along the y -axis. Noting that this effect dissipates as percentage correctness tends to the extremes (0.5 and 1.0). In Figure 5.9 this effect can also be observed. The minor x -axis represents increasing $|Cap|$, the major x -axis represents increasing $|Ind|$ and colour represents $|Nat|$ as before. For a given $|Cap|$, $|Ind|$, i.e. reading within a block, the spread in percent correct is determined by $|Nat|$, with higher values resulting in higher performance but with the effect diminishing at the extremes of performance (0.5 and 1.0). The lower bound of 0.5 being equivalent to a random model in this case and is a property of the binary data used.

These results are as expected as $|Nat|$ is akin to the number of samples for this demonstration and in supervised ML, the greater the number of samples, the more accurate the returned model. The reason percent correct tends to 0.5 for all $|Nat|$ when $|Ind|$ increases is due to the corresponding $NatCap$ matrix containing only a few positive training examples.

Conversely, for low values of $|Ind|$, there are sufficient training examples for all values of $|Nat|$ investigated to maximise performance. If $|Nat|$ were sufficiently increased/decreased respectively, a difference in performance is expected.

Result 2 As $|Ind|$ increases, the percentage correctness of the returned model decreases as illustrated in Figures 5.8 and 5.9.

This is due to the corresponding *NatCap* matrix having an increasing proportion of zeros and thus providing very few positive training examples for the ML algorithm.

Result 3 $|Cap|$ did not have a significant effect on the percentage correctness of the returned model as indicated in Figure 5.8 and more clearly in Figure 5.9.

Changing $|Cap|$ has no effect on the ratio of zeros to ones in the *NatCap* matrix therefore does not hinder the ML training.

Result 4 A high correlation was noted between the percentage correctness of the returned model and the ratio in which examples (or support) for each category were present in the data, referred to here at the *ratio split*³. The more even the split (ratio split = 0.5), the more correct the returned model and conversely tending to 1 and 0.5 respectively as shown in Figure 5.8. Once again, this observation is expected as it reflects the suitability of the data for training the ML algorithm.

Result 5 The use of percentage correct of model is justified in the case of simulated data where F_1 is shown to tend to 1 as $|Ind|$ increases and ratio split tends to 0 as shown in Figure 5.10. This is due to the $P(true\ positive)$ tends to 1 and $P(false\ positive)$ and $P(false\ negative)$ tend to 0 for a matrix with a high proportion of one category. Substituting this into Equation 4.9 gives 1.

³Here ratio split refers to the ratio of category 1 to category 0, in other words, the relative representation of one class over another or skew.

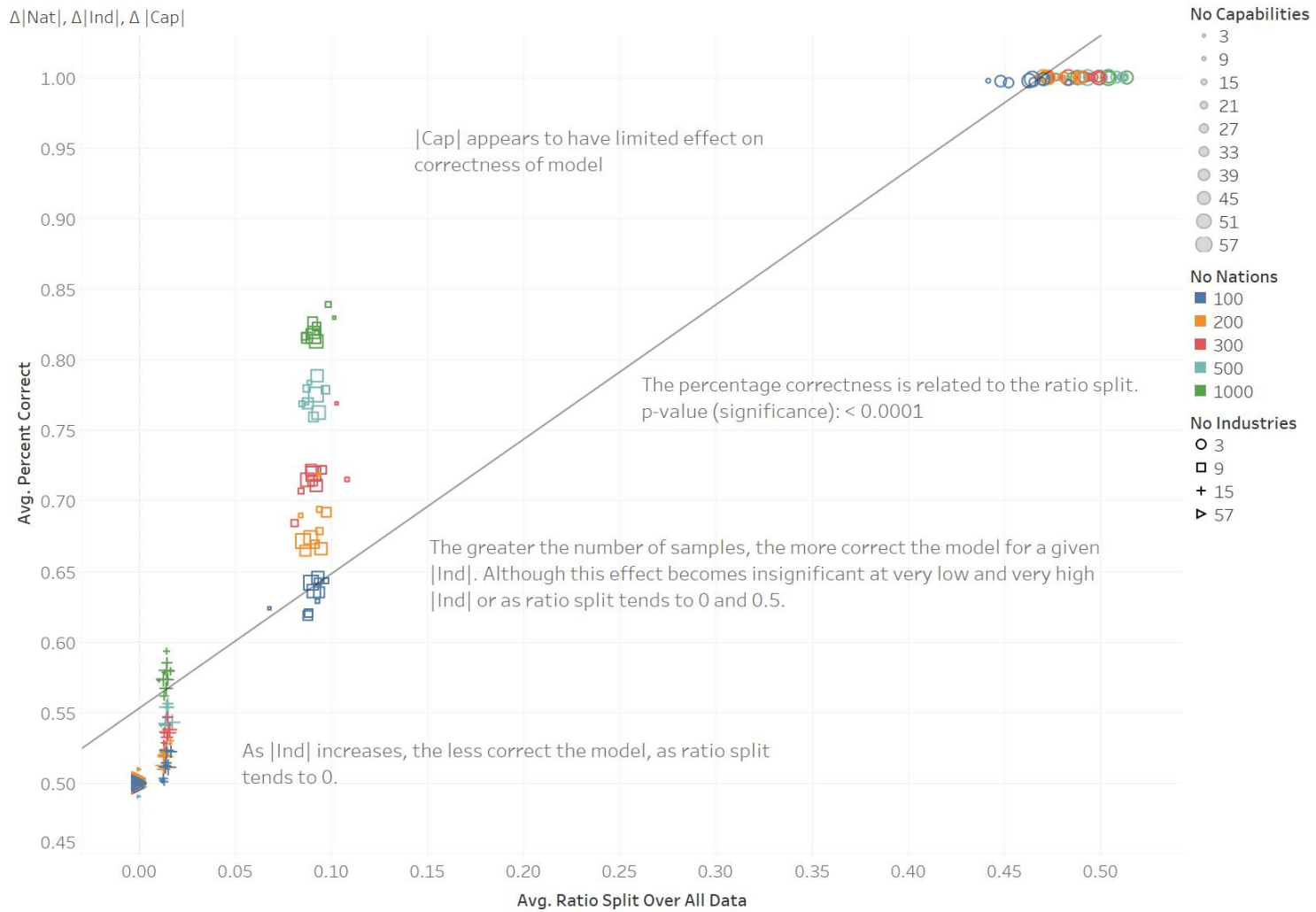


Figure 5.8: Plot of average ratio split vs average percentage correctness of returned model. Colour indicates $|Nat|$, size indicates $|Cap|$, and shape indicates $|Ind|$. Data is filtered, showing the lower and higher values of $|Ind|$. Each point is the average value of 30 runs.

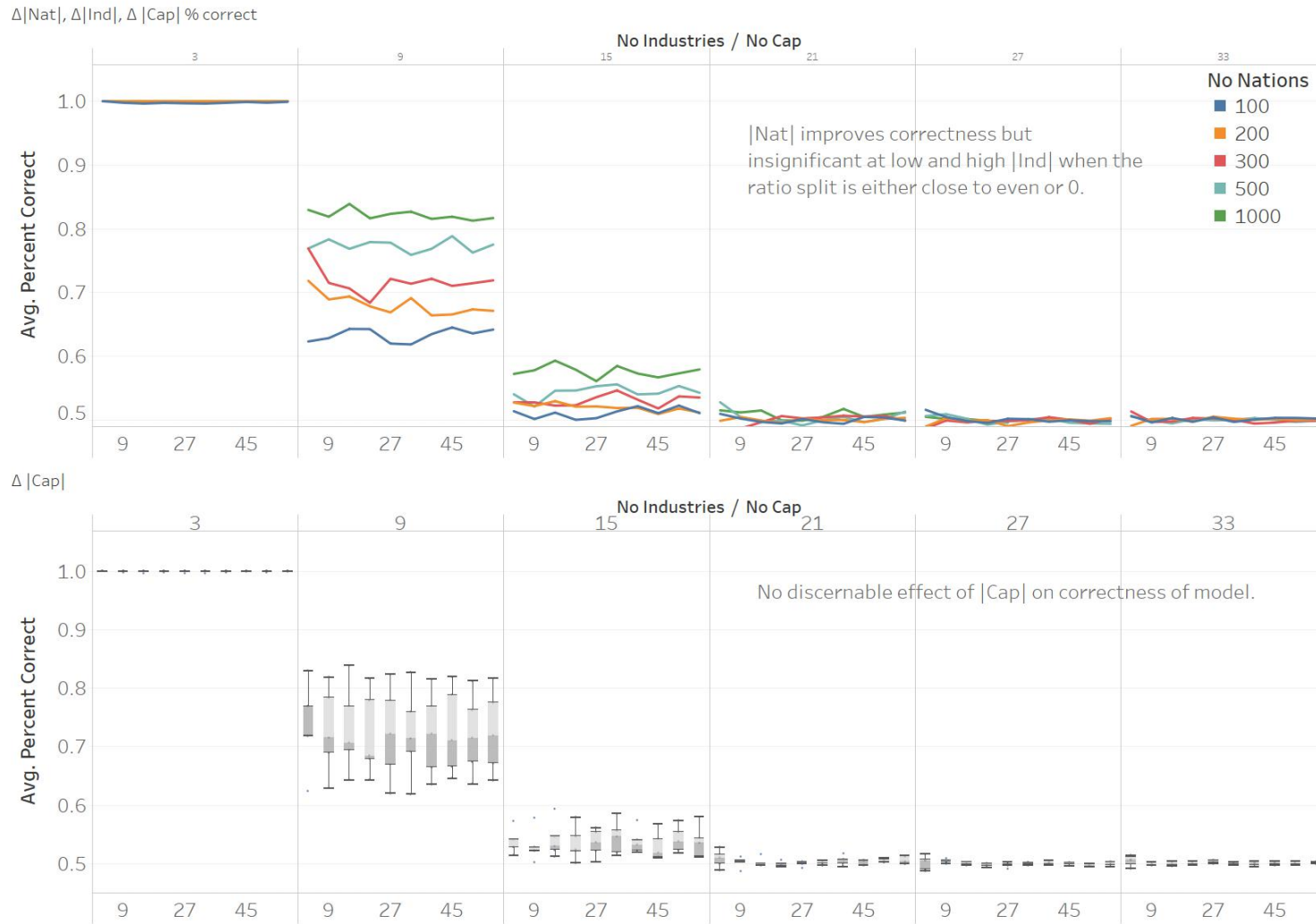


Figure 5.9: Plot of $|Ind|$, $|Cap|$ vs average percentage correctness of returned model. Colour indicates $|Nat|$ in the line graph. Data is filtered, showing the lower values of $|Cap|$ as the pattern continues for higher values. The distribution of results shows there is no significant difference as $|Cap|$ changes.

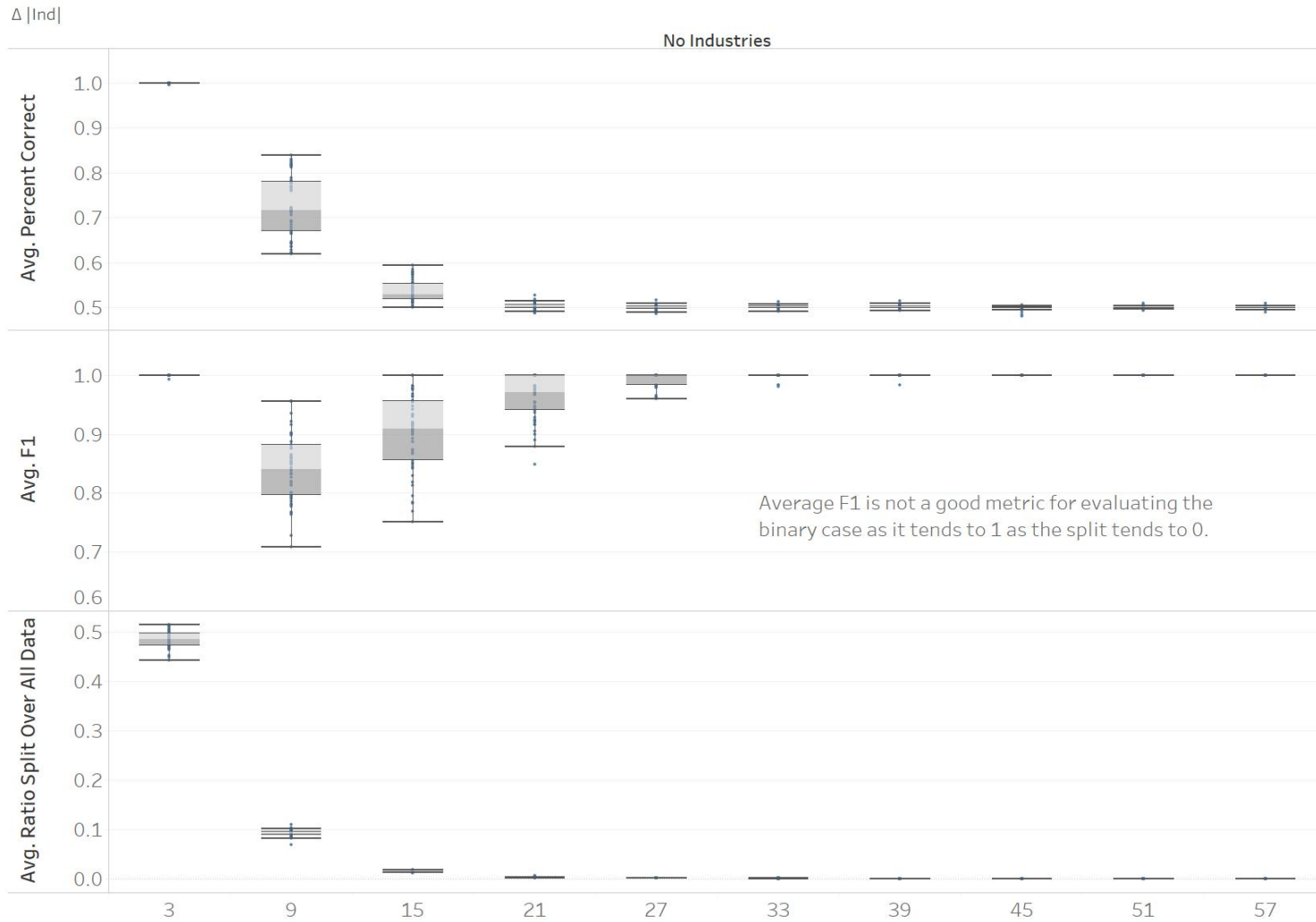


Figure 5.10: Plot of performance metrics average percentage correct and average F_1 vs $|Ind|$ which is in turn related to ratio split. The F_1 metric tends to 1 as ratio split tends to 0, and percentage correct tends to 0, making F_1 a poor choice to evaluate performance for the simulated data case.

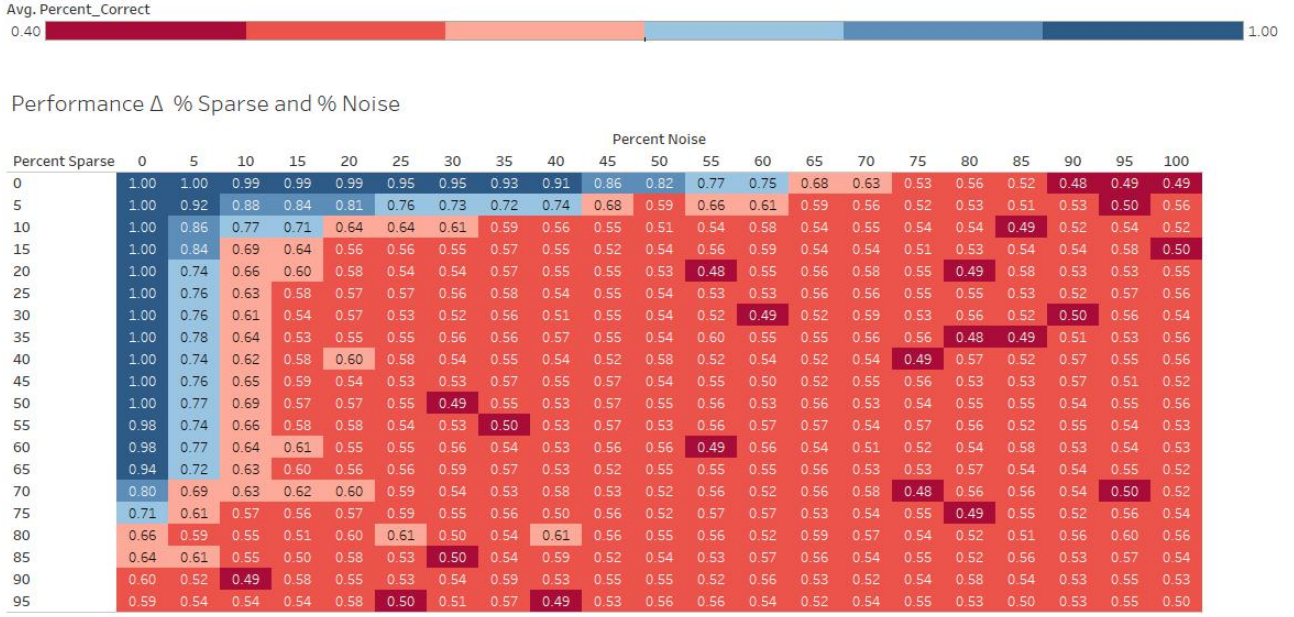


Figure 5.11: Heat map showing average correctness of returned model for a given percentage sparsity and noise.

5.2.2.2 Experiment 2: Changing percentage noise and sparsity

Result 1 The performance of the BBN ML model is affected by noise and sparsity, with the correctness of the returned model tending to 0.5 (equivalent of a random model), as noise and sparsity tend to 100% as shown in Figures 5.11 and 5.12. Example standard deviation of percentage correctness for three values of percent sparsity shown in Figure 5.13. This was as anticipated.

Result 2 The BBN ML used is less sensitive to sparsity than noise, this is shown in Figure 5.12 as a plateau in percentage correctness for percent noise between 0 – 10% over the range 15 – 55% for sparsity. It is also demonstrated when comparing the gradients of the curves for percent sparsity and percent noise = 0 (the blue line) in the two charts, changing sparsity requires a much higher percent noise before a change in average correctness is seen than for the reverse.

This was partly as anticipated as the implemented BBN algorithm was chosen for its ability to handle missing data.

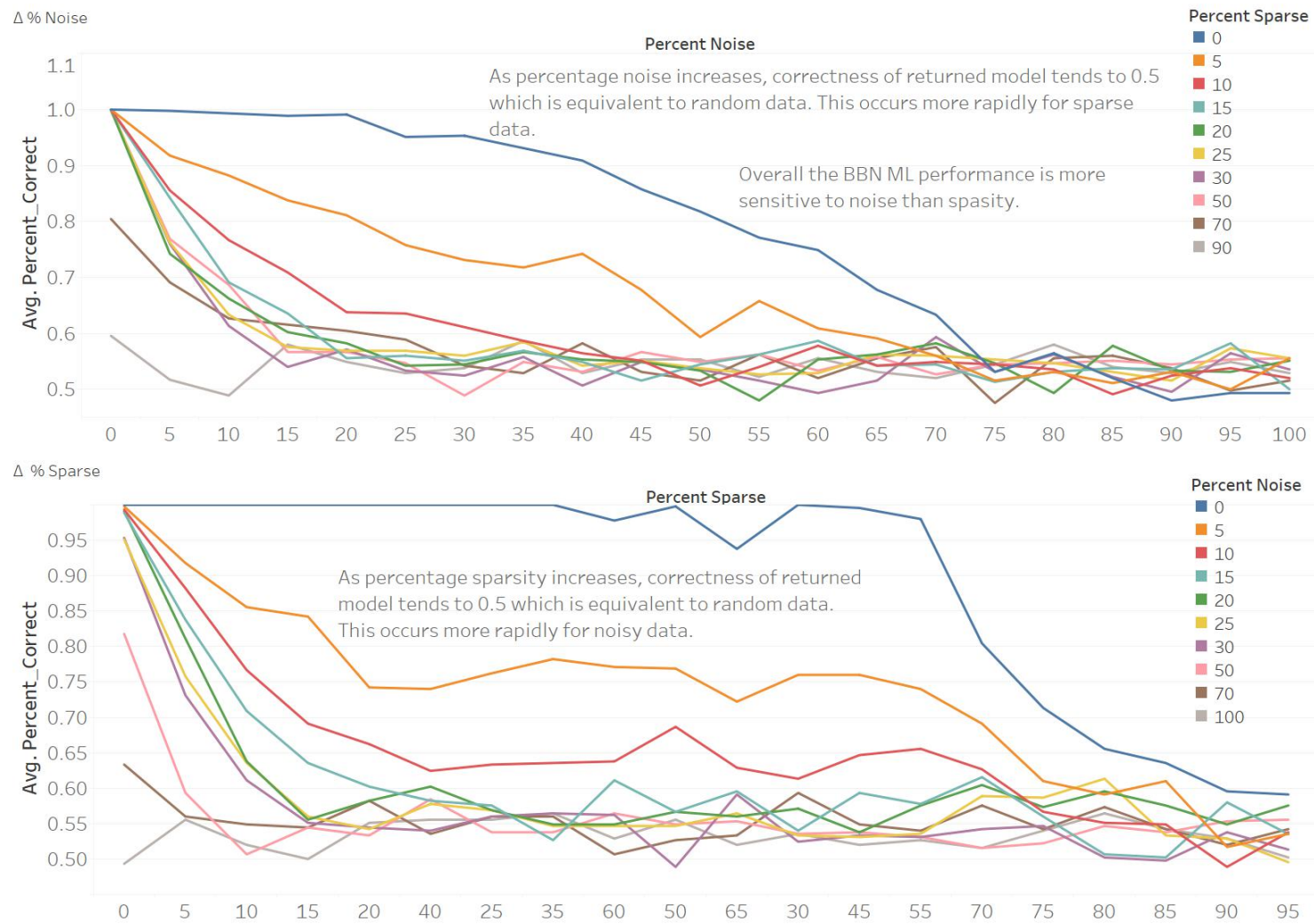


Figure 5.12: Graph showing average correctness of returned model for a given percentage sparsity and noise.

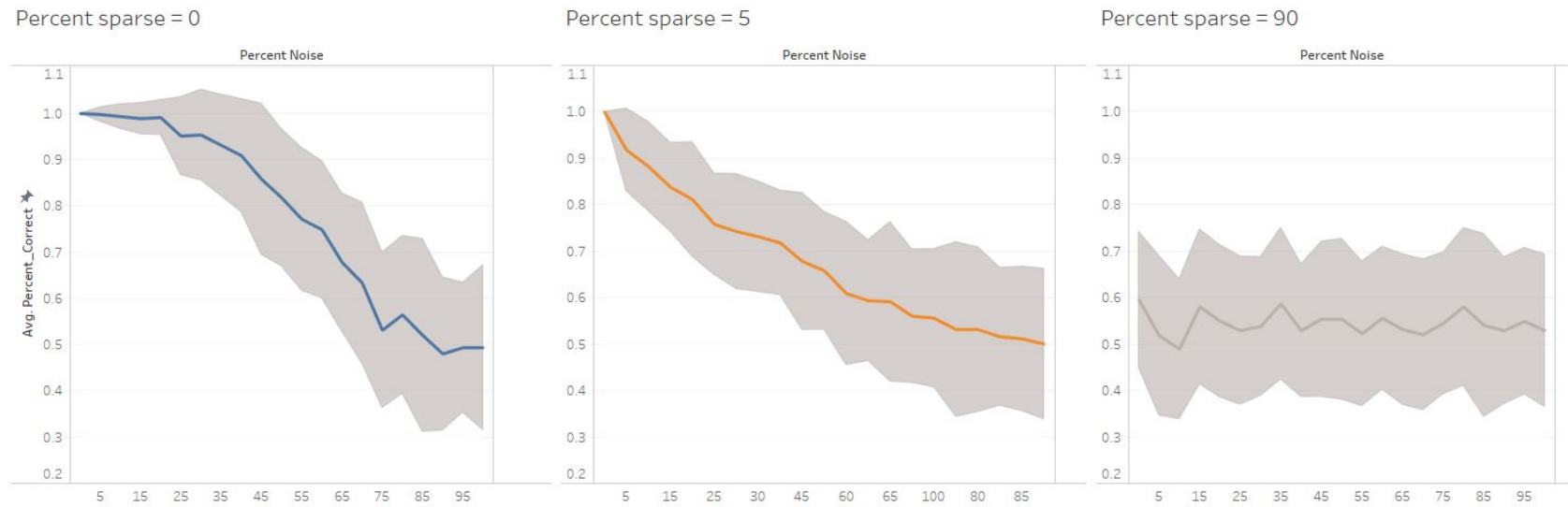


Figure 5.13: Graph showing standard deviation of average correctness for three values of percent sparsity.

5.2.3 Discussion and Conclusion

Using simulated binary data, the purpose of Demonstration 1 was to confirm that BBN ML can learn relations between national industry and national capability sets and to understand the effect on model performance of: $|Nat|$, $|Ind|$ and $|Cap|$ and percentage noise and percentage sparsity of input data. The results show that the BBN ML can learn $IndCap$ from $NatInd$ and $NatCap$ for simulated binary data and the effect on model performance of changing the size of Nat , Ind , and Cap and percentage sparsity and percentage noise was quantified. The results were largely as anticipated, specifically that model performance improved as $|Nat|$ increased. This is because $|Nat|$ in this case is equivalent to number of samples and the performance of ML algorithms is known to improve as the number of samples increases. It was however surprising to find that while smaller values of $|Ind|$ result in higher performance, $|Cap|$ had no significant effect on performance. This is thought to be due to the $NatCap$ matrix having a high proportion of zeros.

Regarding percentage noise and sparsity of input data, the results were as anticipated, with performance decreasing as sparsity of these parameters increases. However the results did refine the understanding of the relative sensitivity of the BBN ML algorithm to the two parameters, with performance being more sensitive to noisy versus sparse input data, although performance does drop off rapidly for data that has both missing values and noise. Of note is the plateau in performance for percent noise between 0% and 10% and sparsity over the range 15 – 55%. In addition to the anticipated results, two additional results were found. First that F_1 did not prove to be a good performance metric for this case, necessitating the development of another metric, percentage correctness, to compare the BBN ML algorithm’s generated $IndCap$ with the ground truth matrix. Second that surprisingly low ratio splits can be tolerated before performance is adversely affected. The reason F_1 is not a good metric in this case is because out of the multiple runs for each parameter, when the ratio split was low, the number of models produced that only had samples and hence predictions in one category increased. These models therefore had precision, recall and F_1 scores of 1. Because there were many more of these, when averaging across the multiple runs, the score average tends to 1.

The three consequences for application of this BBN ML to this Research Framework are: In scraping data the strategy should prioritise correctness over completeness, ideally taking advantage of the identified plateau in performance for values of noise between 0% and 10% and sparsity between 15% and 55%. Performance can be improved by removing variables that are either/or highly sparse or have very low ratio split. Finally, performance can be improved by reducing the number of variables associated with certain nodes.

Regarding application to real-world data, anticipated values of $|Nat|$ would be up to 195, $|Ind|$ would depend on the taxonomy chosen. For example the UNISIC has 21 and 419 top and bottom level categories respectively [UNSD, 2017]. As Cap is user defined, $|Cap|$ would entirely depend on the implementation. In Demonstration 2, $|Cap| = 1$.

There are two limitations of this method. The first is that it only considers binary data, not multi-category data as will be used for the real-world data. Further experiments could be used to bridge this gap, for instance either changing the simulated data to multi-categorical data, allowing the investigation of the effect on performance of the number of categories or changing the real-world data to binary data. A second limitation is that only part of the conceptual graph model is being demonstrated, further work would incorporate increasing numbers of nodes.

5.3 Demonstration 2: Web Scraping

As described in Section 5.1, the purpose of Demonstration 2 is, in part, to incorporate the web scraping step of the Research Framework. Using two data sources, three methods and corresponding results for web scraping are described in Sections 5.3.1 to 5.3.3, in accordance with Table 5.1. Sections 5.3.1 and 5.3.2 address scraping OECD data for input and target variables respectively. Section 5.3.3 addresses scraping IEA data for target variables. The discussion for each of the three methods is given in Section 5.3.4. The scraped data is used in each case study as shown in Table 5.2.

OECD data was chosen as it is a readily available structured dataset that covers official industrial statistics broken down by industry, nation and year. It provides the input data for both Case Studies and the target data for Case Study 2. The IEA data is also readily available, structured and represents the official source for data on national energy industries broken down by year. It provides the target variable of interest for Case Study 1. As structured datasets, these represent a relatively easy web scraping case.

Table 5.1: Sections describing data extraction methods.

	X (input)	y (target)
OECD	Section 5.3.1	Section 5.3.2
IEA	N/A	Section 5.3.3

Table 5.2: Use of data sources for Demonstration 2 Case Studies.

Data source	X (input)	y (target)
Case Study 1	OECD	IEA
Case Study 2	OECD	OECD

5.3.1 Method and Results: OECD Input Data

Method Both case studies source OECD data for their input data. OECD provides access to statistical data via an API. The user sends a string query and a response in the SDMX format is returned. The OECD API uses two types of protocol, the first returns data in

an XML format, SDMX-ML. The second returns it in a JavaScript Object Notation (JSON) format, SDMX-JSON. XML and JSON are established data exchange formats that are human readable-and-writable [OECD, 2016b,a]. Query responses need to be parsed in order to extract the data of interest. A high-level illustration of the method for scraping OECD data via the API is shown in Figure 5.14. Each step is expanded on below. Briefly, it begins with querying dataset availability, understanding the internal structure of each dataset, filtering and then downloading datasets.

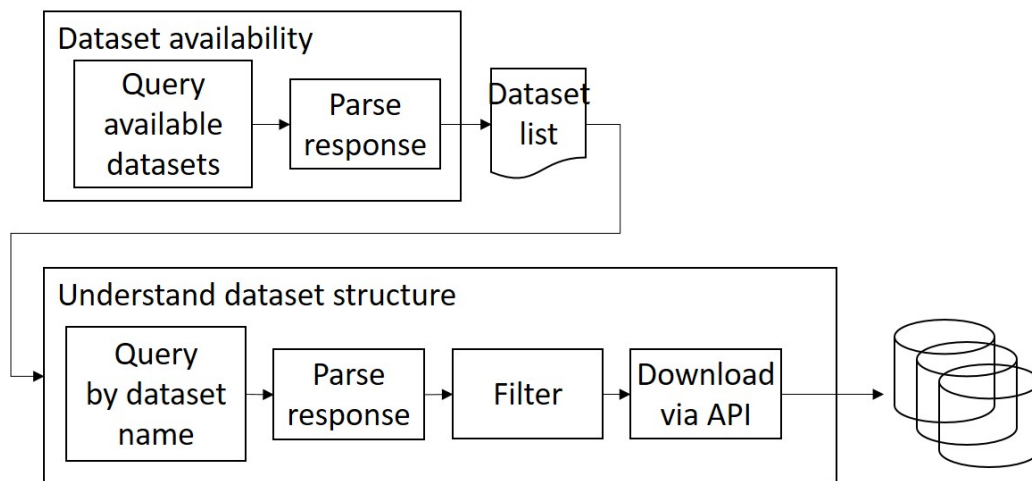


Figure 5.14: OECD data scraping high level method.

Dataset availability The first step was to understand what datasets were available via the OECD API. This was achieved via a string query and parsing the SDMX-ML response to extract the English natural language names of the datasets and their corresponding codes which were used in the subsequent query. The query is shown in Figure 5.15a and an extract of the response in Figure 5.15c. The pseudocode for parsing the response and an extract of the results are shown in Figure A.7 and Figure 5.15d respectively.

Understand dataset structure The second step was to understand which of the available datasets were ordered by industry name and nation. Note that all datasets were assumed to be time series. This was done by iterating through the list of dataset names from the previous step. For each dataset name, a query was formed and executed through the OECD API, parsing the response and testing for keywords. The datasets that satisfied the keyword search were downloaded via a final API query. The keywords were selected based on visual inspection

```
#Make SDMX-ML query using requests. Returns response object r.
URL = 'http://stats.oecd.org/RESTSDMX/sdmx.ashx/GetDataStructure/ALL/'
r = requests.get(URL)
```

(a) Querying OECD API and capturing SDMX-ML response variable r.



```
URL = 'http://stats.oecd.org/RESTSDMX/sdmx.ashx/GetDataStructure/ALL/' #Some vari
r = requests.get(URL) #Other variables specified via POST
print(r.text)
```

```
<message:Structure xmlns="http://www.SDMX.org/resources/SDMXXML/schemas/v2_0/st
ructure" xmlns:common="http://www.SDMX.org/resources/SDMXXML/schemas/v2_0/commo
n" xsi:schemaLocation="http://www.SDMX.org/resources/SDMXXML/schemas/v2_0/struc
ture http://www.sdmx.org/docs/2_0/SDMXStructure.xsd http://www.SDMX.org/resour
ces/SDMXXML/schemas/v2_0/message http://www.sdmx.org/docs/2_0/SDMXMessage.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:message="http://ww
w.SDMX.org/resources/SDMXXML/schemas/v2_0/message"><Header xmlns="http://www.SD
MX.org/resources/SDMXXML/schemas/v2_0/message"><ID>none<!--Copyright OECD - All
rights reserved
Terms and Conditions
http://www.oecd.org/termsandconditions/
Privacy Policy
http://www.oecd.org/privacy/
MyOECD
https://www.oecd.org/login
Contact Us
http://www.oecd.org/contact/
--></ID><Test>false</Test><Truncated>false</Truncated><Prepared>2017-09-21T14:
32:51</Prepared><Sender id="OECD"><Name xml:lang="en">Organisation for Economi
c Co-operation and Development</Name><Name xml:lang="fr">Organisation de coo
```

(b) Response object of OECD API SDMX-ML query.

(c) Code for querying OECD API for dataset structure and its corresponding response.

```
In [64]: print(result[0:10])
```

	DB Name	DB Name2
1	QNA	Quarterly National Accounts
1	PAT_IND	Patent indicators
1	SNA_TABLE11	11. Government expenditure by function (COFOG)
1	E078_MAIN	Economic Outlook No 78 - December 2005 - Annua...
1	ANHRS	Average annual hours actually worked per worker
1	QNA	Quarterly National Accounts
1	PAT_IND	Patent indicators
1	SNA_TABLE11	11. Government expenditure by function (COFOG)
1	E078_MAIN	Economic Outlook No 78 - December 2005 - Annua...
1	ANHRS	Average annual hours actually worked per worker

(d) Post parsing of response of the OECD query in Figure 5.15c, dataframe showing English natural language and codified names of datasets available via OECD API.

Figure 5.15: Stages of retrieving names of available datasets.


```

1 ,NATION,YEAR,"GDP per capita at current prices and current
  CommitmentsConstant Prices"
2 0,Australia,1970,4653.260456664199,
3 1,Australia,1971,4862.6214138824,
4 2,Australia,1972,5212.4772497435,
5 3,Australia,1973,5799.4248645217,
6 4,Australia,1974,6246.463219851,
7 5,Australia,1975,6928.930051797,
8 6,Australia,1976,7393.7503061209,
9 7,Australia,1977,7723.0410390882,
10 8,Australia,1978,8545.909772500601,
11 9,Australia,1979,9450.2736246481,
12 10,Australia,1980,10470.2848645244,
13 11,Australia,1981,11826.2837797806,
14 12,Australia,1982,11890.7316583669,
15 13,Australia,1983,12696.5249158399,
16 14,Australia,1984,13473.6481166301,
17 15,Australia,1985,14394.2155460445,
18 16,Australia,1986,14886.157508587901,
19 17,Australia,1987,15928.6943353024,
20 18,Australia,1988,16996.3378980837,
21 19,Australia,1989,17708.687670896103,
22 20,Australia,1990,17714.1041875078,
23 21,Australia,1991,18075.5223560605,

```

Figure 5.17: Result of parsing extracted database.

$$\begin{array}{l}
\text{All Donors, Total 1 : Part I - Developing Countries Total Commitments Constant Prices} \\
\hline
\text{GDP per capita at current prices and current PPPs, US dollars}
\end{array}
\quad (5.6)$$

5.3.3 Method and Results: IEA Target Data

Case study 1 uses IEA data for its target data. IEA data is available via the UK Data Service [UKDS, 2018]. A comma separated variable (csv) file of the required data was downloaded directly from the website and filtered for data addressing *Total of All Energy Sources* and *Total of Renewable Energy Sources*. An example of the filtered data is shown in Figure 5.18. The processing of the IEA data addressed conversion of the downloaded data to calculating the ratio given by Equation (5.7) by nation and by year [IEA, 2018]. This involved a number of standard transformations in order to prepare the data for subsequent processing and will not be expanded upon here. An example of the processed data is shown in Figure 5.19.

$$\frac{\text{Total Renewable Energy Sources, terajoules}}{\text{Total of All Energy Sources, terajoules}} \quad (5.7)$$

```

1 "LOCATION","Country","PRODUCT","Product","FLOW","Flow","UNIT","Unit","TIME","Time","Value","Flag Codes","Flags"
2 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of
oil equivalent","1990","1990",5074,,
3 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of
oil equivalent","1991","1991",5192,,
4 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of
oil equivalent","1992","1992",4728,,
5 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of
oil equivalent","1993","1993",5482,,
6 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of
oil equivalent","1994","1994",5382,,
7 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of
oil equivalent","1995","1995",5531,,
8 "AUS","Australia","PDPRODUCT016","Total of Renewable Energy Sources","PDFLOW007","Total Primary Energy Supply","PDUNIT001","kilotonne of

```

Figure 5.18: Filtered IEA data downloaded as a csv file from UK Data Service website.

```

1 ,Country,Year,ratio terajoule
2 1990 Albania,Albania,1990,0.2274339841830124
3 1990 Algeria,Algeria,1990,0.0009969359829293553
4 1990 Angola,Angola,1990,0.745197226327157
5 1990 Argentina,Argentina,1990,0.07074835293495428
6 1990 Armenia,Armenia,1990,0.0192953046129629
7 1990 Australia,Australia,1990,0.058739808453539544
8 1990 Austria,Austria,1990,0.20250510295193527
9 1990 Azerbaijan,Azerbaijan,1990,0.00703728803555638
10 1990 Bahrain,Bahrain,1990,0.0
11 1990 Bangladesh,Bangladesh,1990,0.5447575224026686

```

Figure 5.19: IEA data after pre-processing.

5.3.4 Discussion and Conclusion

The purpose of Demonstration 2, in part, is to show two examples of the use of real-world data within the Research Framework. The aim here was to scrape the required data for the two case studies and in doing so, demonstrate the data mining component of the Research Framework. Two data sources were scraped. The first applied automated web scraping of statistical data in the SDMX format via an API for a large dataset provided by the OECD. The second was a manual example applied to a smaller web dataset from the IEA. This experience highlighted some of the aspects of working with web data that are applicable for development of the Research Framework. Here three aspects of data are considered: size, cleaning, sparsity.

In this implementation, the data was downloaded to expedite down-stream processing and

for the larger OECD datasets an automated method was implemented for processing. While this approach was reasonable for this demonstration, and remains practical for low numbers of datasets, it is not scalable as it required a bespoke approach to parse each dataset. The nature of raw data is that it is messy and requires cleaning and formatting. This presents challenges for automated methods that compounds as more data types from multiple sources in multiple formats are introduced. The code developed for parsing the OECD SDMX data relied on the specific structure of the OECD dataset and its field names and is not able to be transferred to another SDMX dataset without adaptation. There are even challenges for automation for what should have been a relatively simple case. Despite the data being provided in an internationally recognised standard and curated by an international organisation, the dataset was reasonably messy such as the use of multiple naming conventions for the same field, e.g. *COU*, *LOCATION*, *COUNTRY*.

The final feature of the resultant cleaned and formatted dataset, is that it was sparse. It is reasonable to assume that this will be a feature of data regardless of its source or method of retrieval and extraction. Therefore downstream processes, here the ML algorithm, must be able to account for missing data. Solutions for web information retrieval and extraction at web scales for multiple data-types is a non-trivial task and an area of ongoing research.

5.4 Demonstration 2: Case Study 1

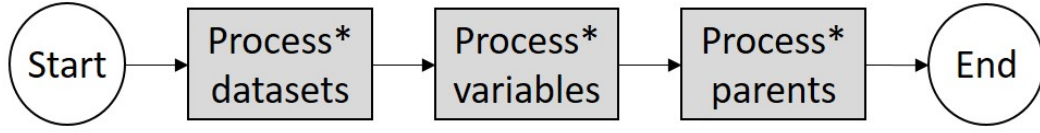
As described in Section 5.1 the purpose of Demonstration 2 is to incorporate data scraped from the web in order to demonstrate the entire Research Framework and part of the conceptual graph using statistical data for two case studies. Here the method, results and conclusion are described for the first case study which looks at one industry and one capability. The capability under study is a nation's ability to generate electricity from renewable sources. The commensurate industry under the International Standard Industrial Classification of All Economic Activities (ISIC) is *Electricity, gas, steam and air conditioning supply*.

The question is: what combination of industry variables is best for predicting this national capability? Combined with Section 5.3 it demonstrates the part of the conceptual graph model and Research Framework as illustrated in Figure 5.2. For the purpose of this demonstration $|Nat| \times |year|$ is equivalent to the number of samples from the perspective of the BBN ML algorithm. $|year|$ is the number of time periods in the dataset. Due to the use of statistical, rather than binary data, the definition of *Ind* and *Cap* are adapted from those described in Section 3.3.1 as follows where *var* indicates a variable related to the industry under study, and ratio is the target data given by Equation (5.7):

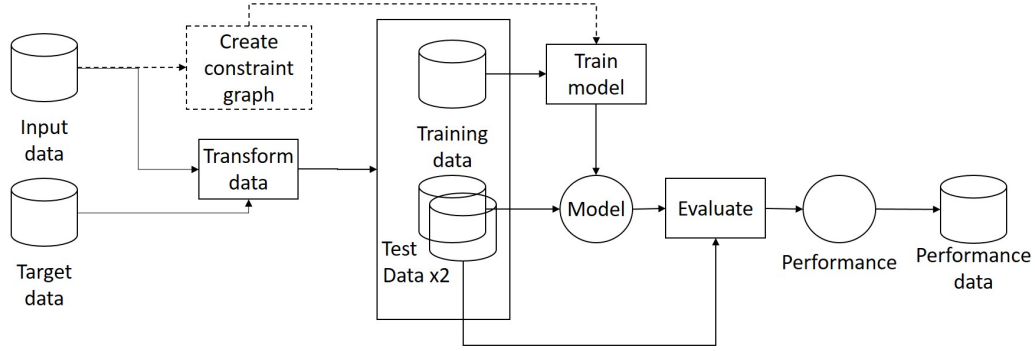
- $Ind = \{var_1, var_2, \dots, var_v\}$
- $Cap = \{ratio\}$

5.4.1 Method

A high level illustration of the method for Demonstration 2 is shown in Figure 5.20a and consists of three steps, process: datasets, variables and parents. These steps are described in more detail in Sections 5.4.1.1 to 5.4.1.3. Within each step the following occurs: transform data, train and evaluate model as shown in Figure 5.20b and described in detail in Sections 5.4.1.4 to 5.4.1.7 .



(a) Case study 1 high level method illustrating three processing steps with distinct objectives although they share the same method, * shown in more detail in Figure 5.20b. Shaded processes explained in more detail in Sections 5.4.1.1 through 5.4.1.3.



(b) Process from 5.20a shown in detail.

Figure 5.20: Demonstration 2 Method.

5.4.1.1 Process Datasets

The inputs to this step were 21 selected OECD datasets and the processed IEA ratio data which represent *Ind* and *Cap* respectively. This step iterates through the selected OECD datasets, training a model for each dataset and evaluating its performance. The outputs are a tree model and performance of the model for each OECD dataset. The parameters for model training were:

- Constraint Graph = No
- Algorithm = Chow-Liu
- Models per dataset = 1

5.4.1.2 Process Variables

The inputs to this step were the 9 OECD datasets from the previous step that produced models and the IEA ratio data which represent *Ind* and *Cap* respectively. The datasets were concatenated into one dataset and the model trained and evaluated. The outputs are a tree model and the performance of the model for the concatenated dataset. The parameters for model training were:

- Constraint Graph = $\langle Ind, Cap \rangle$
- Algorithm = Chow-Liu
- Models per dataset = 1

5.4.1.3 Process Parents

The inputs to this step were the parent variables of the ratio node and the IEA ratio data which represent *Ind* and *Cap* respectively. The outputs are an exact model and the performance of the model for the parent variables. The parameters for model training were:

- Constraint Graph = $\langle Ind, Cap \rangle$
- Algorithm = Exact
- Models per dataset = 1

Note that no exact solution was found for Case Study 1, meaning the algorithm could not find the structure of the BBN from the provided inputs.

5.4.1.4 Transform Data

The initial data processing varied slightly for the dataset, variables and parent processing steps before following a baseline procedure. The variations and baseline procedure are described below.

Variations The process dataset step iterated through the selected OECD datasets which formed the input data and then followed the baseline process. The process variables and parents steps iterated through the selected datasets, concatenating them into one dataset to form the input data. The process variables step then followed the baseline process for the single concatenated dataset while the process parents step filtered the data to the parent variables, found at the process variables step, then followed the baseline process.

Baseline Data Transform The baseline data processing is defined as follows:

- Concatenate input data with the IEA ratio dataset, the target data.
- Filter the data by industry type to only contain data on the *Electricity, gas, steam and air conditioning supply* industry.
- Pivot the data to group by nation and time period, with the input and target variables forming the columns of the dataframe, ensuring the target data forms the final column.
- Discretise the variable data. For the purpose of this demonstration, each variable was discretised into four categories/bins. Four categories were chosen as a balance between providing more information to the algorithm and reducing the number of training examples for each category. The range of the bins was chosen to evenly populate each bin with examples and was based on the range of the continuous data in each column. Values were assigned to each bin on the bases of bin minima $\leq x <$ bin maxima with the exception of the final bin which encompassed the maximum value of the variable bin minima $\leq x \leq$ bin maxima. An absence of data was represented by a fifth bin.
- Filter the data so that variables have a minimum of 10 examples of each category. This was to help ensure that when the data is split into a test and training set, examples of each category for each variable appear in each set. This was based on an outcome of Demonstration 1.
- As the bins were evenly populated the data could be split randomly into training and test sets while maintaining examples of each category in each set. Typically in a supervised ML problem, such as this, a 70/30 split of data into training and test sets respectively would be used. This was not possible in this case due to the BBN ML algorithm being increasingly unlikely to find a model that made predictions across all four categories as the size of the test set was reduced. Instead a 50/50 split was used. This split was chosen based on an experiment that varied the split until the model was likely to find an acceptable solution (predictions within all four categories) within a reasonable time frame.
- Create a second version of the test set, with the data in the final column, which corres-

ponds to the target variable, replaced with NaN.

5.4.1.5 Create Constraint Graph

The input of this step is the number of variables in the input dataset, $|Ind|$. $|Cap|$ was always set to 1 and represents the target (ratio) variable. The output is a set of triples describing the constraint graph and represents the part of the conceptual graph model being demonstrated $\langle Ind, Cap \rangle$.

5.4.1.6 Train Model

The inputs of this step are the optional constraint graph and training dataset. The output is a BBN model. As with Demonstration 1, the Pomegranate library is used, specifically the *BayesianNetwork.from_samples()* function.

5.4.1.7 Evaluate

The resulting trained model was then applied to the test set with NaN values and predictions made. If the model failed to make predictions across all four categories, the Train Model step was repeated until either 100 attempts had been made or a solution was found. When a model was found that made predictions across all four bins, these predictions were compared to the *ground truth* test data and the performance in terms of F_1 , precision and recall, were calculated using the Scikit learn library.

5.4.2 Results

5.4.2.1 Process Datasets

Result 1 Of the 21 OECD datasets nine found solutions, meaning the algorithm was able to learn the structure and prior probability tables for a BBN for nine of the input datasets.

Result 2 The performance of the resulting models is shown in Table 5.3. The metrics initially appear low (perfect performance is represented by scores of 1 for each metric). However due to the use of categorised data, a random model with four categories would get an average recall

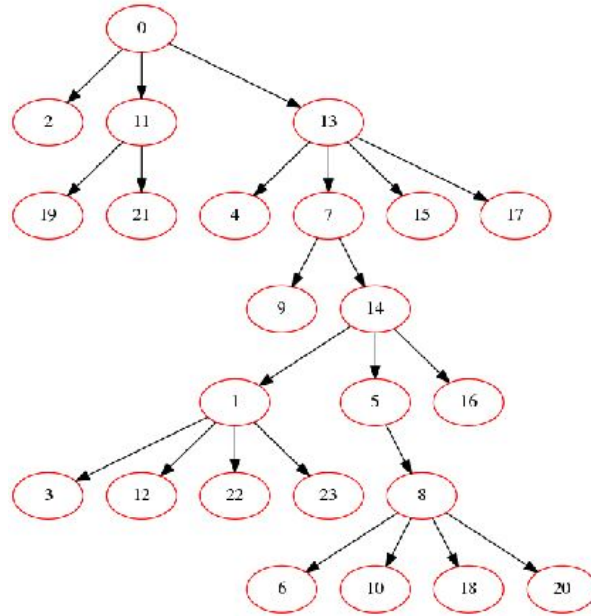
Table 5.3: Case Study 1 model performance.

Step	Database Code	av F_1	av precision	av recall
Eval. Dataset	AEA	0.186	0.576	0.298
	Inward Activity of Multinationals (AFA), Share in national total (manufacturing) (AFA_CALC_IN3)	0.121	0.337	0.260
	AFA by industrial sector (manufacturing) (ISIC Rev.3) (AFA_IN3)	0.144	0.343	0.269
	BERD by industry and by type of cost (ISIC Rev.4) (BERD_COST_ISIC4)	0.141	0.469	0.270
	BERD by industry and by source of funds (ISIC Rev.4) (BERD_FUNDS_ISIC4)	0.140	0.415	0.272
	BERD by industry and by source of funds (BERD_FUNDS)	0.207	0.350	0.304
	BERD by industry (ISIC Rev.4) (BERD_INDUSTRIY_ISIC4)	0.152	0.397	0.277
	BERD_INDUSTRIY	0.214	0.399	0.318
	Productivity by Industry (ISIC Rev.3) (PDBI)	0.162	0.308	0.272
Eval. Variables	Concatenated	0.203	0.533	0.305

score of 0.25. Therefore these models are all performing better than a random model.

Result 3 The dataset which resulted in the *best* model, by F_1 score, for predicting the target variable was the Business Enterprise R&D Expenditure (BERD) by industry (BERD_INDUSTRIY) dataset. This is an interesting intermediate result, indicating a relation between R&D and green energy production.

Example Output A typical output for a dataset is shown in Figure 5.21 for the Air Emission Accounts (AEA) dataset. The trained model is shown in Figure 5.21a. Nodes 0 to 22 represent a categorised industry variable in the dataset. Node 23 is the renewable energy ratio category. The performance of this model is in Figure 5.21b, showing the performance by category and the overall average performance.



(a) Example output of the Pomegranate algorithm learning the structure from training data for case study 1. Nodes 0 through 22 correspond to industry variables and node 23 corresponds to the energy ratio.

	precision	recall	f1-score	support
1	0.60	0.01	0.01	452
2	0.61	0.13	0.22	456
3	0.83	0.06	0.10	449
4	0.26	1.00	0.41	428
avg / total	0.58	0.29	0.18	1785

(b) The corresponding performance of the trained model on test data for Case Study 1.

Figure 5.21: Case Study 1 example model and performance output for an OECD dataset

5.4.2.2 Process Variables

Result 4 A BBN solution was found for the input data formed by concatenating the nine datasets from the process datasets step which had formed models. The performance of the learned model is shown in Table 5.3 as Concatenated.

Result 5 The following variables are the parents of the ratio node in the tree model and are originally from the BERD_FUNDS dataset:

- Total (funding sector), 2010 Dollars - Constant prices and Prices and Purchasing Power Parities (PPP). This is the total R&D expenditure for all funding sources (business enterprise, funds from abroad, and other national and government) [OECD, 2019f] converted to 2010 US Dollars and adjusted for price differences between countries [OECD, 2019d].
- Total (funding sector), PPP Dollars - Current prices. As above except reported at price of the reporting period [OECD, 2019b].
- Business enterprise, PPP Dollars - Current prices. The total R&D expenditure by business enterprises reported at price of the reporting period.

The first two variables are related, the same metric but with different units. This is not too surprising as the Chow–Liu algorithm was used which creates a tree based on a measure of *mutual information* (see Section 4.2.3 for the definition of mutual information). If variables have high mutual information, they will appear close together in the tree. The final variable, business enterprise, forms part of the funding sector, so again is related to the previous two.

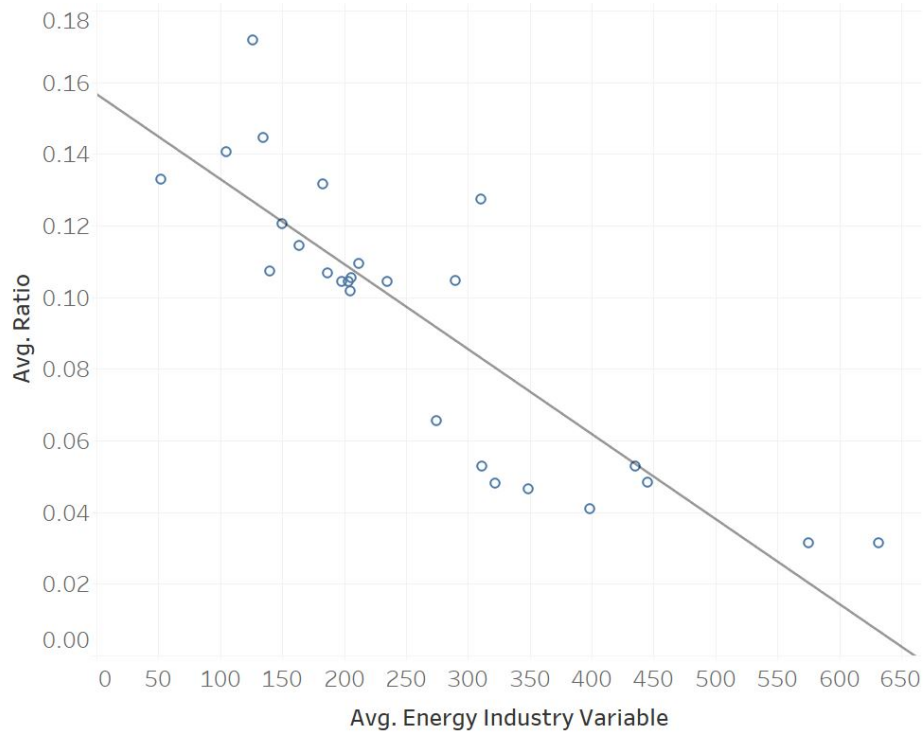
5.4.3 Discussion and Conclusion

The purpose of Demonstration 2 Case Study 1, combined with Section 5.3, is to incorporate the use of data scraped from the web in order to demonstrate the entire Research Framework for a subset of the conceptual graph model, as illustrated in Figure 5.2, using statistical data for the capability under study: a nation’s ability to generate electricity from renewable sources. By looking at statistical variables for the *Electricity, gas, steam and air conditioning supply* industry, the demonstration fulfilled its aims, with three variables being identified as providing

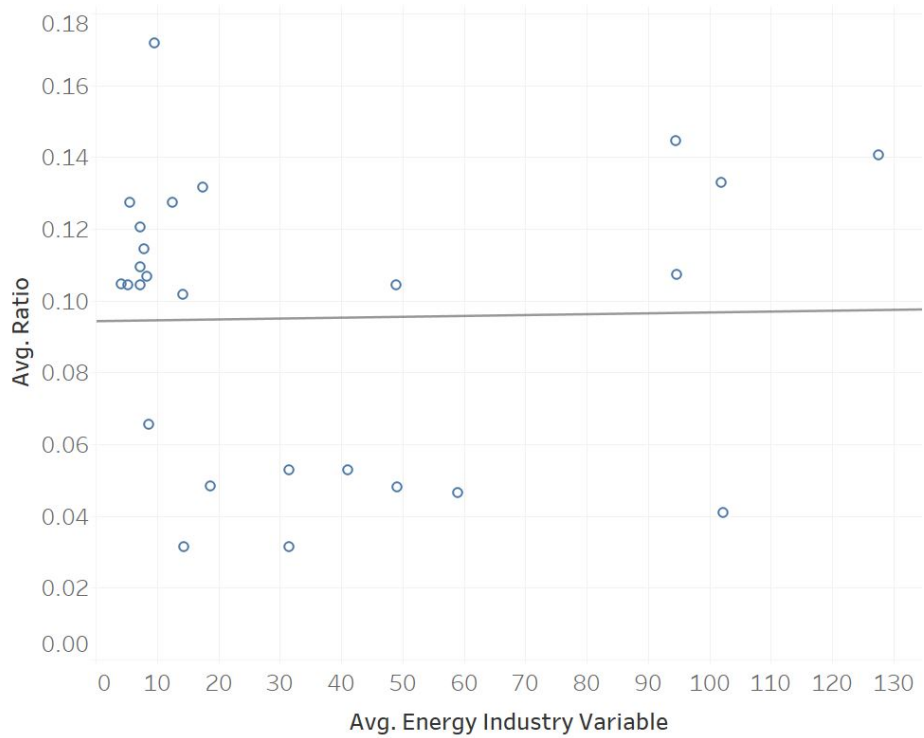
the best prediction of the ratio of renewable to total electricity generation. The performance of the resulting model had an average recall of 0.305, which is above the result of a random model for four categories of 0.25. The resulting variables were confirmed by looking at the plot of the ratio variable against the industry variable and calculating the p -value. One example is shown in Figure 5.22a, where the p -value is < 0.0001 indicating a high correlation. Comparing this with another variable not selected via the BBN ML (Sub-total government, PPP Dollars - Constant prices) is shown in Figure 5.22b where the p -value $= 0.911$ indicating no correlation.

Unexpectedly, the final step, understanding how the combination of the three variables contributed to the prediction of the target variable, using the exact algorithm to learn the *IndCap* matrix, was not tractable. This is thought to be due to the increasing number of combinations of values of Ind_n and Cap_n when each can take one of four values, and not having enough samples. As already mentioned, a 50/50 split for training and test data was used as a trade-off between the standard practice split and the likelihood of finding a solution. This suggests the split also needs to be further investigated as a tuning parameter and reinforces the finding from Demonstration 1 that the number of categories used should be investigated as a parameter.

The main limitation of this method is that only part of the conceptual graph model was incorporated into the demonstration. Further work would extend the nodes incorporated, although this would probably also generate a need to include a dimensionality reduction step in order to more efficiently understand which of the large number of potential variables should be used to represent each node.



(a) Plot of renewable to total electricity generation ratio and Business Enterprise PPP Dollars Constant Prices averaged over nations and time



(b) Plot of renewable to total electricity generation ratio and Sub-total government, PPP Dollars - Constant prices averaged over nations and time

Figure 5.22: Understanding correlation between industry variables and ratio.

5.5 Demonstration 2: Case Study 2

Here Demonstration 2 is continued with a second Case Study, looking at a nation's ability to transition from aid to internal industrial development. Unlike the first case study which focused on one industry, this Case Study looks across industries. The question is, what combination of industries contribute to national economic development and transition away from aid? Combined with Section 5.3, it demonstrates the part of the conceptual graph model and Research Framework as illustrated in Figure 5.2. For the purpose of this demonstration $|Nat| \times |year|$ is equivalent to the number of samples from the perspective of the BBN ML algorithm. The definition of *Cap* is the same as for Case Study 1 but the ratio is given by Equation (5.6). The definition of *Ind* is as follows where *var* indicates a variable related to the industry *ind*:

- $Ind = \{ind_1var_1, ind_1var_2, \dots, ind_1var_v, ind_2var_1, \dots, ind_2var_v, \dots, ind_ivar_v\}$

5.5.1 Method

The method for Case Study 1 was followed as described in Section 5.4.1 and illustrated in Figure 5.20. The details applicable to Case Study 2 are described below. In particular two variations were conducted which represented different selections of industry categories. Variation 1 addressed all available industry categories whereas Variation 2 excluded some.

5.5.1.1 Evaluate Datasets

The inputs to this step were 105 selected OECD datasets and the processed OECD ratio data. The reason for the difference in the number of selected datasets was due to assistance provided by Petras Saduikis who developed additional code to scrape and select OECD datasets.

5.5.1.2 Variation 1

Process Variables The inputs to this step were the 36 OECD datasets from the previous step which produced models and the OECD ratio data.

Process Parents The inputs to this step were the 105 parent variables of the ratio node and the OECD ratio data. Here 102 models were created and their average performance evaluated.

5.5.1.3 Variation 2

Process Variables The inputs to this step were the 36 OECD datasets and the OECD ratio data as before. However, in this variation the variables were filtered to exclude categories that were not strictly industry, such as Totals and Government.

Process Parents The inputs to this step were the 24 parent variables of the ratio node and the OECD ratio data. Here 102 models were created and their average performance evaluated.

5.5.2 Results

5.5.2.1 Evaluate Datasets

Result 1 Of the 105 selected databases, 36 found solutions. Meaning a BBN was learned for 36 of the input datasets.

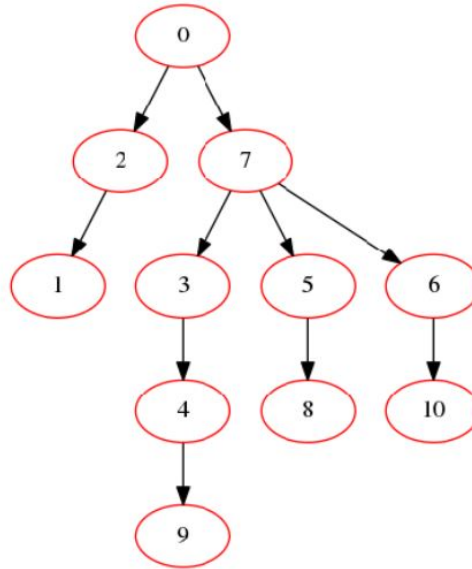
Result 2 The performance of the top 10, by average F_1 score, are shown in Table 5.4.

Result 3 The dataset which resulted in the *best* model, by F_1 score, for predicting the target variable was the 7A. Labour input by activity (ISIC Rev.4, Simplified National Accounts (SNA) 93) (SNA_TABLE7A_SNA93) dataset. This dataset addresses employment statistics broken down by industry [OECD, 2019e]. This is not too surprising a result given the target variable of aid to GDP.

Example Output A typical output for a dataset (the Financial balance sheets - non consolidated (Annual) dataset) is shown in Figure 5.23a. Nodes 0 to 9 represent a categorised industry variable in the dataset. Node 10 is the aid to GDP ratio category. The performance of this model is shown in Figure 5.23b, showing performance by category and the average performance. Of note is the low recall performance in three of the four categories. There is no immediate explanation for this as it varied which categories were the poor performers. It does however suggest an ensemble approach, one model per category, might be appropriate.

Table 5.4: Case Study 2 model performance. Note that the performance is for a single model, with the exception of Parents which is the average performance and standard deviation of 102 models.

Database Code	av F_1	av precision	av recall
Datasets			
SNA_TABLE7A_SNA93	0.529	0.540	0.529
SNA_TABLE_7_SNA93	0.507	0.590	0.534
Carbon Dioxide Emissions embodied in International Trade (IO_GHG.2015)	0.418	0.529	0.416
Productivity and Unit Labour Costs by main economic activity (ISIC Rev.4) (PDBI14)	0.413	0.430	0.443
Structural Analysis Database for Structural Analysis (ISIC Rev.3, SNA93) (STAN08BIS)	0.402	0.571	0.430
9. Fixed assets by activity and by type of product, SNA93 (SNA_TABLE9_SNA93)	0.400	0.403	0.405
13. Simplified non-financial accounts, SNA93 (SNA_TABLE13_SNA93)	0.396	0.600	0.424
13. Simplified non-financial accounts (SNA_TABLE13)	0.392	0.508	0.393
6A. Value added and its components by activity (ISIC Rev.4, SNA93) (SNA_TABLE6A_SNA93)	0.368	0.374	0.372
Annual institutional investors - Template B (PGI_71A)	0.366	0.497	0.420
Variation 1			
Concatenated variables	0.391	0.511	0.443
Parents	0.538, ± 0.039	0.551, ± 0.034	0.541, ± 0.034
Variation 2			
Concatenated variables	0.425	0.571	0.459
Parents	0.416, ± 0.035	0.565, ± 0.050	0.457, ± 0.027



(a) Example output of the Pomegranate algorithm learning the structure from training data for Case Study 2.

	precision	recall	f1-score	support
1	0.18	0.06	0.10	31
2	0.33	0.05	0.09	38
3	0.22	0.07	0.11	27
4	0.29	1.00	0.45	29
avg / total	0.26	0.28	0.18	125

(b) Performance of trained model on test data for Case Study 2.

Figure 5.23: Case Study 2 example model and performance output for an OECD dataset.

5.5.2.2 Variation 1 Process Variables

Result 4 A solution was found, meaning the algorithm was able to learn the structure and prior probability tables of a BBN from the 36 concatenated data sets. The performance of which is shown in Table 5.4 as Concatenated variables.

Result 5 There were 105 parent variables for the ratio node. The three closest parent nodes were:

- Total activity, Total population, national concept⁴, Persons. This is the total economic activity of the total population of a country [OECD, 2019g] measured in persons.

⁴No explanation was found regarding the meaning of *national concept*

- Annual Millions of national currency – current prices, Other, not elsewhere classified, Insurance corporations.
- Annual Millions of national currency –current prices , Financial assets, Insurance corporations and pension funds – consolidated.

Two variables from the Financial corporation sector⁵ and one addressing all areas of economic activity were identified. The total activity metric is less relevant to this investigation as the aim is to identify a set variables across industry sectors.

5.5.2.3 Variation 1 Process Parents

Result 6 Using as input the 105 parent variables found in the previous step, a BBN was learned. The resultant model consistently identified the following variable as the parent node:

- Persons Total population, national concept, Total activity.

It is not surprising that this variable, a measure of total value for economic activity, is a good predictor for the target ratio (aid to GDP). However, this approach has not identified important industry sectors. The method was adapted by removing variables not of interest, in particular totals and government metrics.

5.5.2.4 Variation 2 Process Variables

Result 7 Taking as input a concatenated dataset (as in 5.5.2.2) but with government and total metrics excluded, a BBN was found, the performance of which is shown in Table 5.4 as Concatenated variables.

Result 8 Once totals and government variables were removed, variables associated with the Financial and non-financial corporations were identified. There were 24 parent nodes for the ratio node. The three closest parent nodes were:

- Property income attributed to insurance policy holders (current prices), Financial corporations.

⁵Definitions of these variables could not be found

- Social contributions and benefits (current prices), Financial corporations.
- Distributed income of corporations (current prices), Non-financial corporations.

5.5.2.5 Variation 2 Process Parents

Result 9 Using as input the 24 parent variables found in the previous step, a BBN was learned. The resultant model typically identified the following variable as the parent node:

- Current prices Property income attributed to insurance policy holders, Financial corporations. This is defined by the OECD as ‘The investment income receivable by insurance enterprises on insurance technical reserves’ [OECD, 2019h]; ‘The financial corporation sector includes all private and public entities engaged in financial activities such as monetary institutions (including central banks), financial intermediaries, insurance companies and pension funds’ [OECD, 2019i].

Variation 2 identified one industry variable as a predictor of the target variable and therefore was not able to identify variables across industry sectors. The results highlight the difficulty of interpreting the significance of the identified variables, which are highly specific, with respect to the target ratio (aid to GDP). This suggests that the method should be further adapted to include only higher-level statistical measures as input, specifically measures of industry size/level of activity.

5.5.3 Discussion and Conclusion

The purpose of Demonstration 2 Case Study 2 is two-fold. The first is to provide assurance that the Research Framework is generally applicable to user defined domains of interest. The second is to build upon Case Study 1 by increasing the number of datasets and extending the number of industries with the aim of identifying both the industries and their variables which support predictions of the ratio of aid to national GDP per capita to address the question - What combination of industries contributes to national economic development and transition away from aid? Combined with Section 5.3, it demonstrates the part of the conceptual graph model and Research Framework as illustrated in Figure 5.2. The demonstration partially ful-

filled its aim, identifying the Financial Corporation sector as the best predictor of the ratio of aid to GDP. The performance of the resulting model had an average recall of 0.457, which is above the random model for four categories of 0.25.

Unlike Case Study 1, Case Study 2 did find solutions with the exact algorithm but, unexpectedly, the resultant graph only identified a single industry variable rather than looking across industry variables as had occurred with the simulated data. A potential reason for this is that the structure learning algorithm, MDL, favours simple structures over more complex ones and represents a limitation of the method used. Further work would explore alternatives to the MDL scoring function although this is not possible with the Pomegranate Library used in this research. The library would either have to be extended or an alternative library found. An alternative would be to adapt the method to encourage metrics from multiple industries to be included in the final BBN, perhaps by using the constraint graph. This Case Study also highlighted a potential dis-benefit of including very detailed variables, making it hard to interpret the results. Another potential route would be to use industry-level aggregate variables. Further work, as with Case Study 1, would also address extending the number of nodes of the conceptual graph model incorporated.

Chapter 6

Extensions

6.1 Introduction

The purpose of this chapter is to expand on two potential routes for further research: ML techniques that account for time-series data and the use of text as a data source combined with ontological techniques. These potential research avenues are described in Sections 6.2 and 6.3 respectively.

6.2 Extension 1: Treatment of Time

As already mentioned in Section 3.3.2, the edges and nodes can be considered dynamic and a lag is anticipated with respect to the time it takes for policy to affect industry and, from there, capability. Demonstrations 1 and 2 ignore the dynamic property of the graph. Here two potential methods for incorporating graph dynamics are examined from a ML perspective. The first method is Dynamic-BBN (DBN) ML and the second Long Short Term Memory Network (LSTM), an example of a deep neural net.

6.2.1 DBN

A DBN is a series of BBNs, each representing the state of the system at discrete points in time. Each BBN is linked to a neighbour by edges which represent the influence of the state of a node at a previous time-step. Often DBNs satisfy the first-order Markov Property that the state of

a node at time t is only dependent on the state of a node at time $t - 1$ as shown in Figure 6.1a. However, DBNs can also be considered n -order, where the states at time t depend on states at times $t - 1, t - 2$ to $t - n$ as shown in Figure 6.1b.

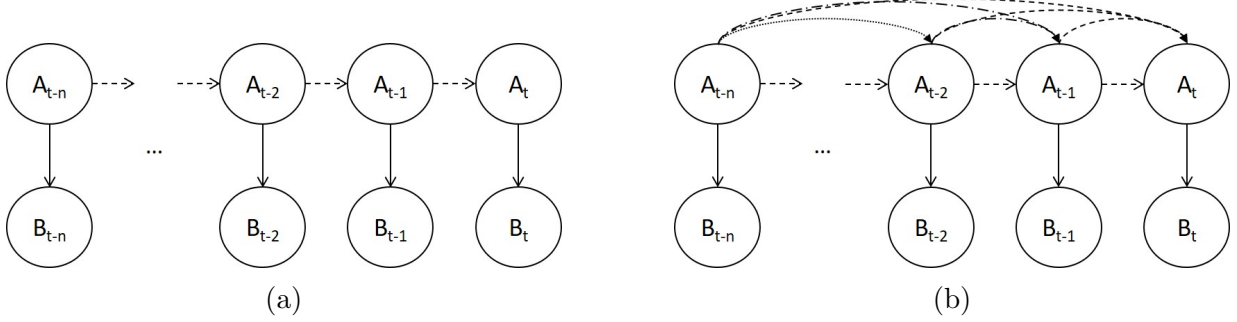


Figure 6.1: A static and dynamic n -order form of a BBN.

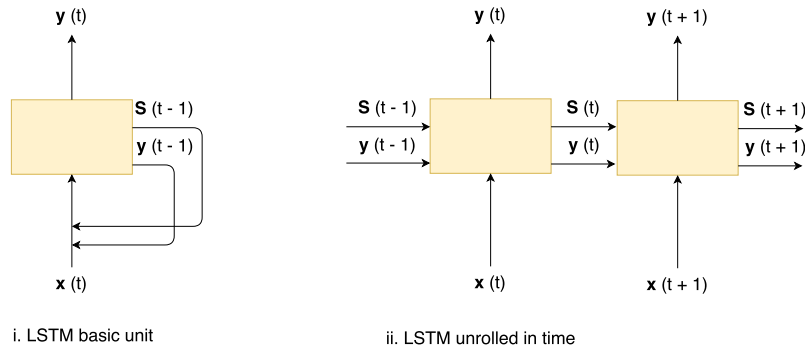
Applying DBN to the Research Framework A BBN can be reframed as a DBN using a constraint graph where nodes at t can only connect to other nodes at time $\geq t$ for an n -order DBN. There are two factors at work affecting computational complexity. The first is the number of nodes increasing by the number of time steps (the size of n in n -order), from N to $N \times T$ where N is the number of nodes and T is the number of time steps. This increases the existing computational complexity linearly. The second is the addition of constraints which reduces the search space, where a node at time t can only be connected with a directed edge to nodes where $T > t$. Overall this is promising for approximate approaches, but less so for exact which remain NP-hard.

6.2.2 LSTM Networks

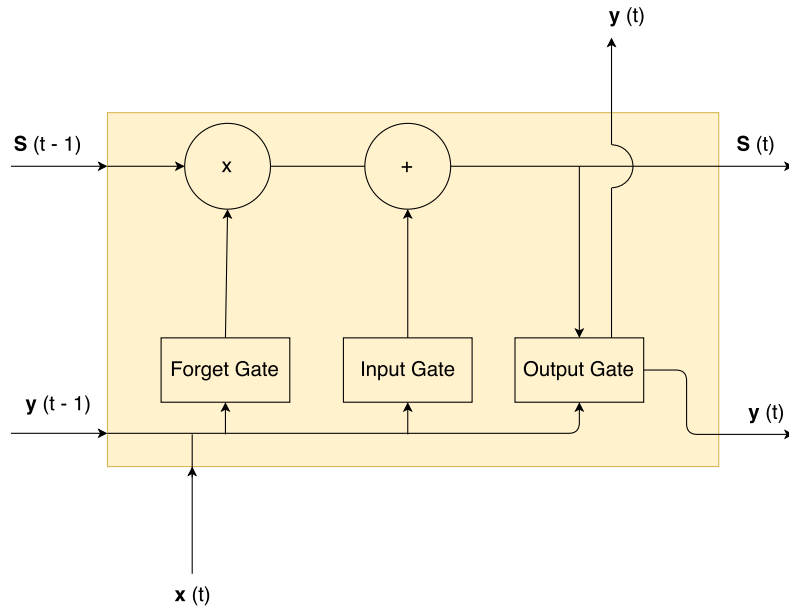
An LSTM network¹ is a type of recurrent neural network, recurrent because the output at time t becomes part of the input at time $t + 1$. The defining feature of LSTMs is the *memory cell* that consists of three gates which maintain the state matrix, $\mathbf{S}$, which is used to calculate the output $y(t)$ from the inputs of $X(t)$ and $y(t - 1)$. Many variants of LSTMs have been developed, but a basic unit is illustrated in Figures 6.2a and 6.2b. Training an LSTM model involves changing the weights applied to the inputs at each gate to minimise the error between

¹Subsequently referred to as LSTMs for brevity.

the predicted output and the known output.



(a) A basic unit of a LSTM (i), and the same unit unrolled in time (ii) [Olah, 2015].



(b) LSTM memory cell showing how the state matrix, S , is updated via forget and input gates, and in turn is used to update the output, y , via the output gate [Olah, 2015].

Figure 6.2: LSTM and memory cell.

Applying LSTM to the Research Framework As LSTMs are a family of ML algorithms, the process for training and evaluating models remains as shown in Figure 4.6. The overall method is as shown in Figure 5.3a, with the exception that the outer loop now represents the parameters of the LSTM which require *tuning* to find a model which has a similar performance for training and test data. These tuning parameters include the number of layers within the LSTM, the learning rate used and number of memory cells. The outer loop remains as before,

resulting in multiple repeats of model training for each parameter combination.

As a type of neural network model, a weakness of the LSTM in this application is there may not be enough data. Neural network ML algorithms typically require a large number of samples whereas in this case a sample is represented by a country, of which there are 200. The number of samples could be increased by splitting a single sample into multiple ones by time period, e.g. a single sample spanning 10 years could be split into two 5 year period samples. Another significant drawback is that LSTMs require dense data which either result in stricter performance requirements for data scraping² and/or would reduce the number of suitable samples for training.

6.3 Extension 2: Text as a Data Source

While Demonstration 2 of the Research Framework focused on statistical data, other types of data are of interest. Given the research challenge to focus on web sourced data, natural language text is an obvious area to investigate as a data source. Here the task of IE from Natural Language Text is described, with a focus on triple and n -ary extraction. This leads to discussion of the relationship between: the Research Framework, conceptual graph, ontologies, IE for automatic ontology population and reasoning with ontologies.

6.3.1 The IE Task on Natural Language Text

IE addresses the retrieval of relevant information from within an object, such as web pages. Of interest to this research is IE from web-based text via NLP. At the core of NLP is the ability to take unstructured text, extract interesting parts and output structured data. Methods include: pattern matching, computational linguistics, statistical learning and logical reasoning [ACM, 2013]. These methods enable machines to break down natural language and perform a range of useful IE tasks. Most NLP IE tasks are performed as a pipeline, with the output of one stage forming the input to the next and begins with processing of the retrieved object. Individual IE

²Or perhaps interpolation could be used.

tasks are built up to perform increasingly sophisticated extractions such as sentiment identification or question answering³. Here we examine the most relevant NLP task to the Research Framework, triple extraction.

Triple Extraction A *triple* is a simple graph of two entities connected by a relation and is broken down into the more basic tasks of Named Entity and Relation Extraction. Named Entity Extraction is the identification and classification of a word, or groups of words, associated with a type of entity such as a person, organisation or location. Relation extraction is the identification and classification of the relation between two entities [Stanford, 2017]. Triple extraction has attracted much attention and as a result there are many triple extractors freely available as part of NLP toolkits:

- Apache Lucene is an open source text search engine that takes text in various formats, which can be a web object, as input. Chunks of data within each document, such as a triple, are extracted and stored in an index [UCLA, 2016].
- University of Washington’s TextRunner is an example of a text extractor designed specifically for web objects and extracts triples in a fully automated way [Yates, Banko, Broadhead, Cafarella, Etzioni, and Soderland, 2007].
- Stanford University’s Natural Language Processing Group make some of their software open source. Their *CoreNLP* package integrates a number of their tools such as a parser, named entity recogniser and a co-reference identifier [Stanford, 2016a].
- The RDF EXtractor (T-REX) automatically extracts triples in RDF format in accordance with a user-specified schema of interest. Developed at the University of Maryland, T-REX has been applied to extracting data on social, cultural, political, economic and religious variables on 20 ethnic-cultural groups and event information from 80 web-based news sites [Albanese and Subrahmanian, 2007].

³Many IE techniques are also employed in document/web object retrieval but we will focus here on their application to IE.

N-ary Fact Extraction A triple can be considered a single *fact*. However many facts have multiple parts, so-called *n*-ary facts. These are facts which are incorrect or incomplete when represented as a triple. For example the following fact is time dependent: the Queen of England is called Elizabeth. [Akbik and Löser, 2012]

6.3.2 Ontologies

A commonly cited definition in terms of computer science for an ontology is ‘*a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents*’ [Gruber, 2006]. But here a more specific definition is adopted from Protegé⁴, the industry standard software for building ontologies, which describes an ontology as consisting of classes, slots and instances.

- Classes are used to categorise concepts and bound the domain of the ontology. Classes exist in a taxonomy, or hierarchy, with classes having subclasses. For example a class might be ‘book’, with sub-classes of ‘hard-back’, ‘paper-back’ and ‘e-(book)’.
- Slots describe the properties of a class. All subclasses share the properties of their parent class. For instance a slot for a book can be publication year. All subclasses of book, hard-back, paper-back and e-book also have publication years. Typically there are three types of slot: intrinsic, extrinsic and parts and relationships. Slots have facets which define the values the slot can take and if a slot can contain multiple values. For example the slot publication year would be a four figure integer between the values of 0000 and 9999 and may be allowed to take multiple values. A relationship slot for the book class could be *authored by* and could link to another class of *Author* which in turn would be a subclass of *Person*.
- Instances are unique occurrences of a class. Continuing the example an instance would be The Lord of the Flies. This instance would appear in all three subclasses. Its *authored by* slot would take the string ‘William Golding’ which would be an instance of the *Author* class which in turn takes on all the slots of the *Person* class [Noy and McGuinness, n.d.].

⁴The book example used is also adopted/adapted from Protegé materials.

Automatic Ontology Population The construction and population of ontologies manually is a time consuming task and thus the automation of ontology learning and population has attracted wide interest from researchers. In *An analysis of main solutions for the automatic construction of ontologies from text*, Rosario Girardi defines the following categories of approach to automatic ontology learning and population as: statistical, ML for classification and clustering, and linguistic [Girardi, 2016]. A discussion on the state of the art in automatic ontology construction is covered in Chapter 2.

Reasoning with Ontologies Ontologies have the innate ability to reason with their data, that is infer new knowledge from existing knowledge. In addition they apply the *open world* assumption where if something is not known, a response to a query would be ‘unknown’ rather than assumed to be ‘no’. Reasoning is incorporated in the ontology language, a standard format to codify knowledge, which incorporates reasoning rules typically based on first-order logic or description logic. Typical reasoning tasks include determining if one class is a member of another class, i.e. a subclass, or class consistency, that a class as defined can contain instances. This is a useful technique for populating an ontology where specific commonsense extractions have not occurred through automatic or manual techniques.

Further review of the literature is required to determine if probabilistic reasoning techniques, such as BBN ML as employed by this Research Framework, can be applied to infer missing data from existing data within an ontology (beyond the first-order/descriptive logic approach). The literature appears to limit the application of probabilistic reasoning to representing uncertainty with respect to facts, such as described by [Carvalho, Laskey, and Costa, 2017] with PR-OWL and applying it to manually constructed BBNs.

Representation of N-ary facts and time Two significant properties of ontologies from the perspective of this Research Framework are its ability to represent n -ary facts and time, although time associated with a fact can be considered a special case of an n -ary fact. As already mentioned, n -ary facts are facts with multiple parts. The W3C working group has proposed a solution to incorporate the representation of n -ary facts in OWL. One method is to

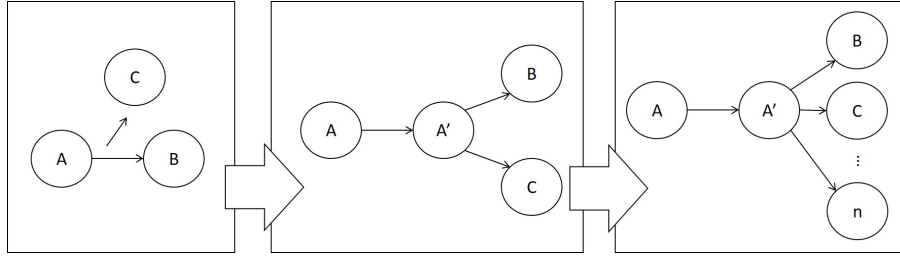


Figure 6.3: Example of extending triple to n-ary relations via use of intermediate class node with multiple slots.

introduce a new class with n slots as illustrated in Figure 6.3. [W3C, 2006]

A proposal to extend OWL 2 to be able to both represent dynamic properties of instances for multiple methods of representing time (qualitative and quantitative, as time points and intervals) by extending W3C's n -ary method above is made by Sotiris Batsakis, Euripides Petrakis, Ilias Tachmazidis, and Grigoris Antoniou. The advantage of their work is that it works within existing standards, meaning it is compatible with ontology reasoning approaches. [Batsakis, Petrakis, Tachmazidis, and Antoniou, 2017]

6.3.3 Application to Research Framework

Conceptual Graph Model and Web Scraping The conceptual graph model lends itself to being interpreted as an ontology, with the nodes and their sets becoming classes, edges becoming slots and the data becoming the instances as illustrated in Figure 6.4. The application of automatic ontology population techniques to an ontology defined by the conceptual graph model would be a neat solution for the data scraping aspect of the Research Framework, resulting in a structured dataset in the form of an ontology.

Reasoning In addition to applying ontologies to web scraping, a second application of ontologies to the Research Framework is reasoning with the data to infer new knowledge from existing data.

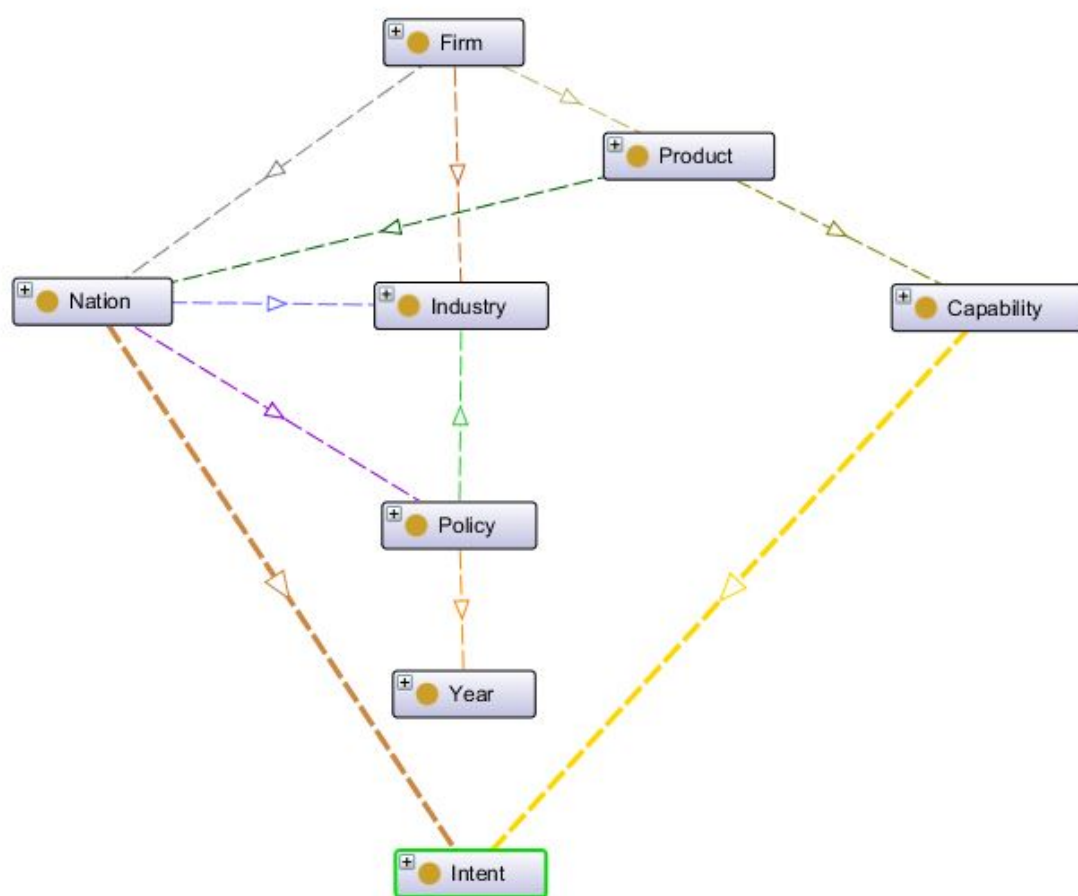


Figure 6.4: Conceptual Graph Model rendered as an ontology.

Chapter 7

Discussion and Conclusion

While discussion and conclusions for each contribution of this thesis have already been addressed individually within their respective chapters, this chapter presents the discussion and conclusion of the research as a whole in the context of its aims.

7.1 Summary of Achievements

The aim of this research was to develop and demonstrate a flexible framework to infer and track the strategic intention of nations from observations of their industrial base using web sourced data. This aim has been partially fulfilled by i) proposing a conceptual graph model (Chapter 3) based on a multi-disciplinary literature review and ii) accompanying Research Framework (Chapter 4) iii) demonstrated via simulated and real data (Chapter 5) which together form the three contributions of this thesis. The main results are:

- For simulated binary data: The BBN ML algorithm can be used to learn relations between national industry and national capability sets, demonstrating a method to determine $E(NatCap \mid NatInd)$. The experiments confirmed and revealed several design considerations. First that performance of the resulting model is sensitive to the number of members of some nodes but not others, here $|Ind|$ and $|Cap|$ respectively. Second the performance of the model can be improved by removing variables that are highly sparse and/or have low numbers of training examples for each category. Third, that in terms of prioritising data scraping, correctness of data should be prioritised over completeness.

- For real categorised statistical data, the breadth of the Research Framework was demonstrated from scraping statistical data from the web to training a BBN model using supervised ML and identifying the industry variables that help predict national capability for two case studies. The experiments also identified a potential limitation of the implemented BBN ML algorithm when used with large numbers of variables - that the scoring function, MDL, used to balance graph complexity with performance, results in graphs connecting the target variable with a single input variable.
 - Using data from the *electricity, gas, steam and air conditioning supply* sector, Case Study 1 sought to identify what combination of industry variables predicts the ratio of green energy to total energy generation. No combination was identified but three variables related to total sector funding and business enterprise formed the immediate three parent nodes to the target ratio as part of an approximate solution.
 - Using data from all available industry sectors, Case Study 2 sought to identify what combination of industry variables predicts the growth of national GDP and transition from state aid. Again no combination was identified but an exact solution was found. In the first variation the total population was identified and in the second, property income attributed to insurance policy holders, a variable from the finance sector.

7.2 Research Aims and Key Contributions

Exploit web data A core aim of this research was to exploit web-based data. This was demonstrated via automated scraping of a large statistical data in the SDMX format via an API from the OECD and manual scraping of a smaller statistical dataset from the IEA. This was consistent with previous work which often limits itself to specific web sources and a single type of data. This aim did not require the research to address the challenges of web scale data scraping nor the integration of multiple data-types and sources, as these areas in themselves could be the topic of multiple PhDs. Nonetheless, challenges were met and aspects of working with big data highlighted, even with the perceived *easier* case of using data in a standardised format, curated by a reliable source and readily accessible via an API, essentially the ideal from

the perspective of proponents of the semantic web.

While not strictly limitations given the research aim, the following highlights aspects which need to be considered to extend the framework to general web sources. First, scalability: the implementation relied on downloading datasets for onward processing and a bespoke approach was taken for each data source which limits the ready extension of the framework to other data sources. To fully exploit web sources future work should look to developing techniques to work with big data. Second, while every effort was made, the data used was not comprehensive, even within the OECD dataset. This was due to two reasons, the first that not all the data was available at the time of download and a practical self-imposed limit that datasets which were too large to download in a given time-frame were excluded. A requirement for a future Research Framework is to be able to update models and predictions as new data becomes available. Finally, the code used was bespoke to the OECD data and is not readily extendible to other SDMX sources. This highlights a limitation of the semantic web approach and points to future iterations of the Framework employing methods that are independent of semantic-web efforts.

With respect to the Framework, this suggests two potential approaches to sourcing data. The first is to choose to take a source-by-source approach, the second route to use semantic web independent technologies. In the first case, quality control is, in part, addressed via human curation of individual sources, but code for each source must be individually developed and as a result would not be web scalable, and thus will not benefit from a key property of the web, its size. The second semantic web independent route is felt to have greater potential for web scalability as it is independent of standards but requires additional quality control processes which is a non-trivial issue. Potentially one could support the other, with the latter approach being used to understand the individual implementations of a given standard and additionally account for variation within the same implementation. It is recommended that future work focuses on automatic methods for populating an ontology defined by the conceptual graph model. The integration of quality control and confidence in the data, so-called probabilistic ontologies,

could also form part of this direction of research. Finally, other data types, video, image and sound, could also be integrated as data sources as research into information extraction from these sources develops.

Conceptual Graph Model A conceptual graph was developed based upon an extensive multi-disciplinary literature review and forms the first contribution of this thesis. This graph was then partly implemented in demonstrations of the Research Framework, discussed below, validating three of the graph’s nodes: *Nat*, *Ind*, and *Cap*, and more specifically the ability to predict a nation’s industry (variables) from national capability. The presence of further nodes needs to be validated. The priority of future work should be integrating time and the *Int* node either/or defined as a change in *Cap* as a function of time or as a combination of capabilities. Additional nodes, *Pol*, *Pro* and *Co*. should also be demonstrated and their contribution to model performance quantified. Further work could also expand the scope of modelling capability, currently limited to material aspects, to include the full TEPIDOIL spectrum.

Research Framework A Research Framework was developed and represents both a key aim of this research and the second contribution of this thesis. The Framework was based upon an extensive multi-disciplinary literature review and has been fully implemented via demonstration using simulated and real-world data, validating the approach. The current limitation of this Framework is that it does not account for updating model and predictions based on new availability and updates of data and the implemented ML algorithm does not take into account the dynamic nature of the data. Two independent directions for future research present themselves. The first is to investigate supervised ML approaches for sparse time-series data that can readily update in light of new/revised data. The second is to investigate the potential of ontologies and associated automated population methods.

Demonstration The third contribution of this thesis was the demonstration of the Research Framework through the use of simulated data and two real-world case studies. The latter demonstration included the exploitation of web data, but this has been addressed previously and will only be briefly mentioned here as necessary.

Demonstration 1 Part of the Research Framework and graph model were demonstrated using simulated binary data. The purpose being to confirm that the BBN ML can produce a model to predict $E(IndCap \mid NatInd, NatCap)$ and understand the effect of $|Ind|$, $|Cap|$, $|Nat|$, percentage data sparsity and percentage data noise on the ML model’s predictive performance. These experiments confirmed the use of a BBN ML and, as expected, that increasing $|Nat|$ improved model performance and increasing percentage data sparsity and noise decreased model performance. A consequence of this is that performance can be improved by removing very sparse variables. Unexpectedly it showed that while smaller values of $|Ind|$ improved model performance, $|Cap|$ had no significant effect. This is thought to be due to the sparsity of the *NatCap* matrix and shows that it is important to understand in experiment design to which nodes the model performance is sensitive. In addition it was found that the BBN ML algorithm used was more robust to sparsity than noise, driving a requirement for the data scraping aspect of the Research Framework to prioritise data accuracy over completeness. When percentage sparsity and noise co-varied, a plateau in performance was found, this could be exploited in future iterations of research. Also the experiments found that surprisingly low values of ratio split could be tolerated, meaning fewer variables need be removed in order to improve performance. Finally, this implementation found that the traditional performance metrics for ML trained models, F_1 , was not a good indicator of performance and an additional metric, percentage correctness was developed in its stead.

Ultimately the use of simulated data proved a very good way to understand the Research Framework and it is highly recommended that further work continue to employ simulated data experiments. In particular the use of simulated data could help investigate the relationship between number of variables and losing the desired behaviour of linking multiple variable nodes to the target node, currently attributed to the behaviour of the MDL. Other areas of investigation could include investigating the effect on performance of the number of data categories, a current gap between the simulated and real-data experiments.

Demonstration 2 The Research Framework and part of the conceptual graph model were demonstrated using two case Studies with real-world categorised statistical data, the purpose

being to demonstrate the validity of the approach. Case Study 1 demonstrated the ability of the Framework to identify variables within an industry that predict capability and Case Study 2 extended this to identify the variables from across industries that predict capability. The results show areas for further analysis and hint at a limitation of the current technique, namely the ability to identify combinations of industry variables that are important for capability prediction. This is thought to be due to the tractability of the Exact algorithm as the number of industries and variables investigated increases and potentially the use of the MDL algorithm which balances graph simplicity with performance.

Future work needs to investigate the balance between the number of variables and the scoring function generating the desired behaviour to understand the combination of industry variables for predicting national capability. This could be done in conjunction with varying the number of categories of data used. It was initially thought that greater resolution is better but this does not appear to be the case when contrasting the simulated and real-data experiments. This is potentially advantageous for the data scraping aspect of the Framework as it would mean a lower data resolution requirement which may prove to be a less challenging for the IE task.

Machine Learning The framework employed a BBN ML algorithm for demonstration purposes and a very basic ML pipeline was used. As a result there are a number of opportunities for refinement including conducting cross-validation and dimensionality reduction. However, a significant limitation of the BBN is that it does not account for the time-series aspect of the data. Future work should therefore focus on contrasting the static approach with a dynamic one, and address the issue of data confidence. This would require changing the ML algorithm to one that accounts for time and confidence but also works well with highly sparse and low example datasets. Should future work incorporate an ontology into the framework, this invites researchers to also consider ontological reasoning techniques in order to populate (or enrich) instances.

7.3 Future Work

Based on the above, it is recommended that, rather than continuing to refine the current implementation, two significant adaptations should be pursued. The first being to adopt an ontological approach, using low recall automated ontology population techniques to scrape and structure text data. The second should be the further investigation of reasoning techniques to account for dynamic data and lag that is also robust to sparse data. Future work should also seek to demonstrate the full conceptual graph model, in particular intent.

Bibliography

- ACM. Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, San Francisco, CA, 27-28 October 2013. New York, NY, USA, 2013. ACM. URL <https://dl.acm.org/citation.cfm?id=2509558>.
- A. Akbik and A. Löser. KRAKEN: N-array facts in open information extraction. 2012. URL https://akbcwekex2012.files.wordpress.com/2012/05/16_paper.pdf.
- M. Albanese and V. Subrahmanian. T-REX: A domain-independent system for automated cultural information extraction. In D. Nau and J. Wilkenfield, editors, *Proceedings of the First International Conference on Computational Cultural Dynamics, College Park, MD, 27-28 August 2007*, pages 2–8, Menlo Park, CA, USA, 2007. AAAI, AAAI Press. URL https://www.researchgate.net/publication/229043961_T-REX_A_domain-independent_system_for_automated_cultural_information_extraction.
- C.W. Anderson. *Statecraft: Introduction to political choice and judgement*. John Wiley Sons, 1977.
- M. Babiker, E. Karaarslan, and Y. Hoscan. Web application attack detection and forensics: A survey. In A. Varol, M Karabatak, and C. Varol, editors, *Proceedings of the 6th International Symposium on Digital Forensic and Security, Antalya, Turkey, 22-25 March 2018*, volume 2018, pages 1–6. IEEE, 2018. URL <https://ieeexplore-ieee-org.iclibezp1.cc.ic.ac.uk/document/8355378/>.
- R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval: The concepts and technology behind search*. Pearson Education, London, UK, 2nd edition, 2011.

- D. Barbosa, H. Wang, and C. Yu. Shallow information extraction for the knowledge web. In C.S. Jensen, C. Jermaine, J. Lu, E. Tanin, and X. Zhou, editors, *29th International Conference on Data Engineering, Brisbane, Australia, 8-12 April 2013*, pages 1264–1267. IEEE, IEEE, 2013. URL <https://ieeexplore-ieee-org.iclibezp1.cc.ic.ac.uk/document/6544920/>.
- S. Barclay, R. Brown, C. Kelly, C. Peterson, L. Phillips, and J. Selvidge. Handbook for decision analysis, 1977. URL <http://files.eric.ed.gov/fulltext/ED153329.pdf>. Accessed: 30 May 2016.
- S. Batsakis, E. Petrakis, I. Tachmazidis, and G. Antoniou. Temporal representation and reasoning in OWL 2. *Semantic Web*, 8(6):pp.981–1000, 2017. URL <https://content-iospress-com.iclibezp1.cc.ic.ac.uk/articles/semantic-web/sw248>.
- A. Beck. Some aspects of the history of anti-pollution legislation in England 1819-1954. *Journal of the History of Medicine and Allied Sciences*, 14(10):pp.475–489, 1959. URL <https://doi.org/10.1093/jhmas/XIV.10.475>.
- D. Begg. *Economics*. McGraw-Hill Higher Education, 9th edition, 2008.
- M. Bergman. The deep web: Surfacing hidden value, 2001. URL <http://www.brightplanet.com/2012/06/the-deep-web-surfacing-hidden-value/>. Accessed: 30 May 2016.
- T. Berners-Lee. *Weaving the Web : the past, present and future of the World Wide Web by its inventor*. Texere, London, UK, 2000.
- T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5): pp.34–43, 2001. URL <http://dx.doi.org/10.1038/scientificamerican0501-34>.
- R.A. Best. Intelligence issues for congress, 2006. URL <https://fas.org/sgp/crs/intel/IB10012.pdf>. Accessed: 25 July 2018.
- P. Bianchi and S. Labory. Empirical evidence on industrial policy using state aid data. *International Review of Applied Economics*, 20(5):pp.603–621, 2006. URL <https://doi.org/10.1080/02692170601005556>.

- C. Bielza and P. Larranaga. Discrete bayesian network classifiers: A survey. *ACM Computing Surveys*, 47(1):pp.1–43, 2014. URL <https://dl-acm-org.iclibezp1.cc.ic.ac.uk/citation.cfm?doid=2620784.2576868>.
- BrightPlanet. Clearing up confusion, web vs dark web, 2014. URL <http://www.brightplanet.com/2012/06/the-deep-web-surfacing-hidden-value/>. Accessed: 26 April 2016.
- J Cadle, D Paul, and P. Turner. *Business Analysis Techniques: 99 essential tools for success*. British Computer Society, Swindon, UK, second edition. edition, 2014.
- R. Cailliau. A little history of the world wide web, c.1995. URL <http://www.w3.org/History.html>. Accessed: 30 May 2016.
- A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr, and T.M Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the 24th Annual Conference on Artificial Intelligence*, 2010. URL <http://rtw.ml.cmu.edu/papers/carlson-aaai10.pdf>.
- R.N. Carvalho, K.B. Laskey, and P.C.G. Costa. PR-OWL, a language for defining probabilistic ontologies. *International Journal of Approximate Reasoning*, 91:pp.56–79, 2017. URL <https://doi.org/10.1016/j.ijar.2017.08.011>.
- S. Choudhury, K. Agarwal, S. Purohit, B. Zhang, M. Pirrung, W. Smith, and M. Thomas. NOUS: Construction and querying of dynamic knowledge graphs. *Computing Research Repository*, abs/1606.02314, 2016. URL <https://arxiv.org/pdf/1606.02314.pdf>.
- C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):pp.462–467, 1968. URL <https://ieeexplore.ieee.org/document/1054142/>.
- Coolclips. Water sprinkler, 2018a. URL <http://search.coolclips.com/m/vector/vc062938/water-sprinkler/>. Accessed: 18 June 2018.
- Coolclips. Royalty-free license agreement-overview, 2018b. URL http://search.coolclips.com/support/License_Overview.aspx. Accessed: 18 June 2018.

- S. Donnan. Donald Trump to slap 25% tariff on steel imports, 2018. URL <https://www.ft.com/content/7b354ff8-1d73-11e8-aaca-4574d7dabfb6>. Accessed: 6 June 2018.
- G. Dosi, C. Freeman, R. Nelson, G. Silverberg, and L. Soete, editors. *Technical Change and International Trade*, chapter 19, pages 399–527. IFIAS Research Series; 6. Pinter, London, UK, 1988.
- R.G. Freedman and S. Zilberstein. Roles that plan, activity and intent recognition with planning can play games. 2018. URL https://people.cs.umass.edu/~freedman/publications/freedman_rolesPairAndPlanningPlayInGames.pdf.
- freepngimg. Weather picture, 2018. URL <http://www.freepngimg.com/png/23648-weather-picture>. Accessed: 18 June 2018.
- N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):pp.131–163, 1997. URL <https://doi.org/10.1023/A:1007465528199>.
- L. Garg. How to plot these data in other way to better show the efficacy of my method?, 2015. URL <https://stackoverflow.com/questions/31008676/how-to-plot-these-data-in-other-way-to-better-show-the-efficacy-of-my-method/31009326>. Accessed: 18 June 2018.
- M. Gaurav, A. Srivastava, A. Kumar, and S. Miller. Leveraging candidate popularity on twitter to predict election outcome. In *Proceedings of the 7th Workshop on social network mining and analysis, Chicago, IL, 11 August 2013*, pages 1–8, New York, NY, USA, 2013. ACM. URL <https://dl.acm.org/citation.cfm?id=2501038&dl=ACM&coll=DL>.
- R. Girardi. An analysis of main solutions for the automatic construction of ontologies from text. In J.E. Guerrero, editor, *International Conference on Web Intelligence, Omaha, NE, 13-16 October 2016*, pages 457–460. IEEE/WIC/ACM, IEEE Computer Society Conference Publishing Services, 2016. URL <https://ieeexplore-ieee-org.iclibezp1.cc.ic.ac.uk/stamp/stamp.jsp?tp=&arnumber=7817091>.
- R.Q. Grafton, W. Adamowicz, D. Dupont, H. Nelson, R.J. Hill, and S. Renzetti. *The Economics*

of the Environment and Natural Resources. Blackwell Publishing Ltd, Malden, MA, USA, 2008.

M. Graham, B. Hogan, R. Straumann, and A. Medhat. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):pp.746–764, 2014. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2382617.

R. Gribben. Scramble to save 13,000 airbus jobs, 2006. URL <https://www.telegraph.co.uk/finance/2936482/Scramble-to-save-13000-Airbus-jobs.html>. Accessed: 6 June 2018.

H. Guo and W. Hsu. A survey of algorithms for real-time bayesian network inference. In *Papers from the AAAI Workshop real-time decision support and diagnosis systems, Technical Report WS-02-15*. AAAI, 2002. URL <http://www.aaai.org/Papers/Workshops/2002/WS-02-15/WS02-15-001.pdf>.

M. Harbers, J.J. Meyer, and K. Van Den Bosch. Modeling agents with a theory of mind: theory-theory versus simulation theory. *Web Intelligence and Agent Systems: An International Journal*, 10(3):pp.331–343, 2012. doi: 10.3233/WIA-2012-0250.

A.O. Hirschman. *The strategy of economic development*. Yale studies in economics; 10. Yale University Press, New Haven, CT, USA; London, UK, 1958.

HM Government HMG. MODAF acquisition viewpoint, 2010a. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/38717/20100426MODAFacVViewpointV1_2_004U.pdf. Accessed: 26 May 2016.

HM Government HMG. MODAF operational viewpoint, 2010b. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/38711/20100426MODAFOVViewpoint1_2_004U__1_.pdf. Accessed: 26 May 2016.

HM Government HMG. MODAF strategic viewpoint, 2010c. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/38710/20100426MODAFStVViewpoint1_2_004U.pdf. Accessed: 26 May 2016.

HM Government HMG. National security strategy and strategic defence and security review, 2015. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/478933/52309_Cm_9161_NSS_SD_Review_web_only.pdf. Accessed: 7 August 2018.

HM Government HMG. Industrial strategy, 2017. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730043/industrial-strategy-white-paper-print-ready-a4-version.pdf. Accessed: 8 May 2018.

HM Government HMG. Open government licence for public sector information, 2018. URL <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>. Accessed: 5 June 2018.

IBM. What is Watson?, 2016. URL <http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html>. Accessed: 26 April 2016.

IBM. Durham police department cuts violent crime by 39 percent using insight into patterns of criminal activity, 2017. URL https://public.dhe.ibm.com/common/ssi/ecm/gp/en/gpc12346usen/ibmpublicsafety_casestudy_durhampd_final_GPC12346USEN.pdf. Accessed: 14 September 2018.

IBM. Watson health: Get the facts, 2018. URL <https://www.ibm.com/blogs/watson-health/watson-health-get-facts/>. Accessed: 14 September 2018.

International Energy Agency IEA. Energy data, 2018. URL http://stats.ukdataservice.ac.uk/index.aspx?r=125870\&DataSetCode=IEA_COAL_BA. Accessed: various dates 2018.

A.T. Ihler, J.W. Fisher, and A.S. Willsky. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6:pp.905–936, 2005. URL <http://jmlr.org/papers/volume6/ihler05a/ihler05a.pdf>.

InternetWorldStats. Internet world stats: Usage and population statistics, 2016. URL <http://www.internetworldstats.com/stats.htm>. Accessed: 24 May 2016.

- International Standards Organisation ISO. A conceptual model of architecture description, 2011. URL <http://www.iso-architecture.org/42010/cm/>. Accessed: 30 May 2016.
- M. Jänicke, M. Binder, and H. Mönch. Dirty industries: patterns of change in industrial countries. *Environmental and Resource Economics*, 9(4):pp.467–491, 1997. URL <https://doi.org/10.1023/A:1026497608363>.
- Japan International Cooperation Agency Institute for International Cooperation JICA. *Japan’s Experiences in Public Health and Medical Systems*. JICA, Japan International Cooperation Agency, Institute for International Cooperation, Tokyo, Japan, 2005. URL open-jicareport.jica.go.jp/pdf/11868221.pdf. Accessed: 7 August 2018.
- Z. John. Can artificial intelligence and online dispute resolution enhance efficiency and effectiveness in courts. *International Journal for Court Administration*, 8(2):pp.30–45, 2017. URL <https://doaj.org/article/3fce19c1530d4e1b87ec585c9061f3ec>.
- N.F. Johnson, M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, and S. Wuchty. New online ecology of adversarial aggregates: Isis and beyond. *Science*, 352(6292):pp.1459–1463, 2016. URL <http://science.sciencemag.org/content/352/6292/1459>.
- Japan Ministry of Defense JpnMoD. Strategy on defense production and technological bases, 2014. URL http://www.mod.go.jp/j/approach/others/equipment/pdf/2606_e_honbun.pdf. Accessed: 9 May 2016.
- Japan Ministry of Defense JpnMoD. Defense of Japan white paper, 2015. URL http://www.mod.go.jp/e/publ/w_paper/2015.html. Accessed: 9 May 2016.
- A. Karpathy. CS231n convolutional neural networks for visual recognition, 2018. URL <http://cs231n.github.io/convolutional-networks>. Accessed: 18 June 2018.
- D. Kellenberg. An empirical investigation of the pollution haven effect with strategic environment and trade policy. *Journal of International Economics*, 78(2):pp.242–255, 2009. URL <https://doi.org/10.1016/j.jinteco.2009.04.004>.

- S.L. Kendal and M. Creen. *Types of Knowledge-Based Systems*, chapter 2, pages 26–88. Springer, London, UK, 2007.
- S. Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science Journal*, 8:pp.159–172, 2005. URL <https://doaj-org.iclibezp1.cc.ic.ac.uk/article/6217bded416b49c2adb3dc676ea7459c>.
- A. Krzywicki, W. Wobcke, M. Bain, J. Calvo Martinez, and P. Compton. Data mining for building knowledge bases: techniques, architectures and applications. *The Knowledge Engineering Review*, 31(2):p.97–123, 2016. URL <http://search.proquest.com/docview/1780765114/>.
- D. Landes. *The Wealth and Poverty of Nations: Why some are so rich and some so poor*. Little, Brown and Company, London, UK, 1998.
- K. Leetaru. The GDELT Project, 2018. URL www.gdeltproject.org. Accessed: 26 September 2018.
- G. List, S. Lloyd, and J. Nicholson. *The National System of Political Economy Translated by S. S. Lloyd*. Longmans, Green and Co., London, UK, 1885. URL <http://socserv2.socsci.mcmaster.ca/~econ/ugcm/3ll3/list/national.html>.
- B. Liu and K. Chang. Editorial: Special issue on web content mining. *ACM SIGKDD Explorations Newsletter*, 6(2):pp.1–4, 2004. URL <https://dl-acm-org.iclibezp1.cc.ic.ac.uk/citation.cfm?doid=1046456.1046457>.
- B. Lundvall, editor. *National Systems of Innovation: Towards a theory of innovation and interactive learning*. Pinter, London, 1992.
- D. McDaniel. DM2 HTML 130326, 2011. URL <http://www.ideasgroup.org/dm2/>. Accessed: 6 June 2018.
- Ministry of Defence MoD. Defence industrial policy, 2002. URL http://webarchive.nationalarchives.gov.uk/20071204155926/http://www.mod.uk/NR/rdonlyres/25726BCE-8DD6-4273-BE8D-6960738BEE0A/0/polpaper5_defence_industrial.pdf. Accessed: 7 August 2018.

- Ministry of Defence MoD. Defence industrial strategy, 2005. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/272203/6697.pdf. Accessed: 7 August 2018.
- Ministry of Defence MoD. Ministry of Defence architecture framework, 2010. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/36757/20100602MODAFDownload12004.pdf. Accessed: 5 June 2018.
- Ministry of Defence MoD. Understanding and intelligence support to joint operations, 2011. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311572/20110830_jdp2_00_ed3_with_change1.pdf. Accessed: 26 May 2016.
- Ministry of Defence MoD. Ministry of Defence architecture frameworks, 2012a. URL <https://www.gov.uk/guidance/mod-architecture-framework>. Accessed: 26 May 2016.
- Ministry of Defence MoD. National security through technology: technology, equipment and support for UK defence and security, 2012b. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/27390/cm8278.pdf. Accessed: 7 August 2018.
- Ministry of Defence MoD. Industry for defence and a prosperous Britain: refreshing defence industrial policy, 2017a. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/669958/DefenceIndustrialPolicy_Web.pdf. Accessed: 7 August 2018.
- Ministry of Defence MoD. National shipbuilding strategy: the future of naval shipbuilding in the UK, 2017b. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/643873/NationalShipbuildingStrategy_lowres.pdf. Accessed: 7 August 2018.
- D. Movshovitz-Attias and W. W. Cohen. KB-LDA: Jointly learning a knowledge base of hierarchy relations and facts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

- Processing, Beijing, China, 26-31 July 2015*, pages 1449–1459. Association for Computational Linguistics, 2015. URL <http://aclweb.org/anthology/P15-1140>.
- North Atlantic Treaty Organisation NATO. NATO glossary of terms and definitions, 2013. URL http://www.academia.edu/10269177/AAP-6_NATO_Glossary_of_Terms_and_Definitions_2013_. Accessed: 7 August 2018.
- R. Nelson, editor. *National Innovation Systems: A comparative analysis*. Oxford University Press, New York, 1993. URL <GotoWoS>://WOS:A1994BB23N00012.
- M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):pp.11–33, 2016. URL <https://arxiv.org/pdf/1503.00759.pdf>.
- M. Nielsen. Using neural nets to recognise handwritten digits, 2017. URL <http://neuralnetworksanddeeplearning.com/chap1.html>. Accessed: 18 June 2018.
- M. Nimishakavi, U.S. Saini, and P. Talukdar. Relation schema induction using tensor factorization with side information. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, 1-5 November 2016*, pages 414–423. Association for Computational Linguistics, 2016. URL <https://aclweb.org/anthology/D/D16/D16-1040.pdf>.
- M. Nimishakavi, M. Gupta, and P. Talukdar. Higher-order relation schema induction using tensor factorization with back-off and aggregation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15-20 July 2018*, pages 1575–1584. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/P18-1146>.
- N.F Noy and D.L. McGuinness. *Ontology development 101: A guide to creating your first ontology*. n.d. URL https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
- S.P. O’Brien. Crisis early warning and decision support: Contemporary approaches and

thoughts on future research. *International Studies Review*, 12(1):pp.87–104, 2010. URL <https://academic-oup-com.iclibezp1.cc.ic.ac.uk/isr/article/12/1/87/1797253>.

Organisation for Cooperation in Economic Development OECD. Strategic industries in a global economy, 1991. URL <http://www.oecd.org/futures/38464821.pdf>. Accessed: 7 August 2018.

Organisation for Cooperation in Economic Development OECD. API documentation (SDMX-JSON), 2016a. URL <https://data.oecd.org/api/sdmx-json-documentation/>. Accessed: 10 October 2017.

Organisation for Cooperation in Economic Development OECD. API documentation (SDMX-ML), 2016b. URL <https://data.oecd.org/api/sdmx-ml-documentation/>. Accessed: 10 October 2017.

Organisation for Cooperation in Economic Development OECD. Total official flows by country and region, 2019a. URL https://stats.oecd.org/viewhtml.aspx?datasetcode=REF_TOTALOFFICIAL&lang=en. Accessed: 20 February 2019.

Organisation for Cooperation in Economic Development OECD. Glossary of statistical terms - c. current, 2019b. URL <https://stats.oecd.org/glossary/detail.asp?ID=510>. Accessed: 19 February 2019.

Organisation for Cooperation in Economic Development OECD. Gross Domestic Product, 2019c. URL <https://data.oecd.org/gdp/gross-domestic-product-gdp.htm>. Accessed: 20 February 2019.

Organisation for Cooperation in Economic Development OECD. Glossary of statistical terms - b. ppp, 2019d. URL <https://stats.oecd.org/glossary/detail.asp?ID=2205>. Accessed: 19 February 2019.

Organisation for Cooperation in Economic Development OECD. 7a labour input by activity, isic rev4, 2019e. URL https://stats.oecd.org/Index.aspx?DataSetCode=SNA_TABLE7A. Accessed: 21 February 2019.

- Organisation for Cooperation in Economic Development OECD. Glossary of statistical terms - a. total funding sector, 2019f. URL https://stats.oecd.org/Index.aspx?DataSetCode=RD_ACTIVITY. Accessed: 19 February 2019.
- Organisation for Cooperation in Economic Development OECD. 3. population and employment by main activity, 2019g. URL https://stats.oecd.org/Index.aspx?DataSetCode=SNA_TABLE3. Accessed: 21 February 2019.
- Organisation for Cooperation in Economic Development OECD. Glossary of statistical terms, 2019h. URL <https://stats.oecd.org/glossary/detail.asp?ID=2179>. Accessed: 19 February 2019.
- Organisation for Cooperation in Economic Development OECD. Financial corporations debt to equity ratio, 2019i. URL https://www.oecd-ilibrary.org/economics/financial-corporations-debt-to-equity-ratio/indicator/english_a3108a99-en. Accessed: 19 February 2019.
- C. Olah. Understanding LSTM networks, 2015. URL <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Accessed: 26 September 2017.
- M. Olama, G.O. Allgood, K.M. Davenport, and J. Schryver. A Bayesian Belief Network of threat anticipation and terrorist motivations. In E.M. Carapezza, editor, *Volume 7666 Proceedings of SPIE, Sensors and Command, Control, Communications and Intelligence Technologies for Homeland Security and Homeland Defense IX, Orlando, FL , 5-9 April 2010*, Bellingham, WA, USA, 2010. SPIE, SPIE.
- Office for National Statistics ONS. Gross domestic product: chained volume measures: seasonally adjusted, 2018. URL <https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/abmi/pgdp>. Accessed: 8 May 2018.
- D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):pp.515526, 1978. URL <https://doi.org/10.1017/S0140525X00076512>.

- V. Raghavan, A. Galstyan, and A.G. Tartakovsky. Hidden markov models for the activity profile of terrorist groups. *The Annals of Applied Statistics*, 7(4):pp.2402–2430, 2013. URL <https://www-jstor-org.iclibezp1.cc.ic.ac.uk/stable/pdf/23566469.pdf?refreqid=excelsior%3Ac1545fbbf8bdb2e7da2495fff94d0e9b>.
- C. Rhodes. Industrial policy 2010 to 2015, 2014. URL <http://researchbriefings.files.parliament.uk/documents/SN06857/SN06857.pdf>. Accessed: 19 May 2017.
- S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artificial Intelligence. Prentice Hall ; Pearson Education [distributor], Upper Saddle River, NJ ; London, UK, 3rd, international edition, 1995.
- R. Schrag, J. McIntyre, M. Richey, K. Laskey, E. Wright, R. Kerr, R. Johnson, B. Ware, and R. Hoffman. Probabilistic argument maps for intelligence analysis: Completed capabilities. In F. Bex, F. Grasso, and N. Green, editors, *Proceedings of the 16th Workshop on computational models of natural argument, New York, NY, 9 July 2016*, volume 1876, pages 34–39. CEUR, CEUR-WS, 2017. URL <http://ceur-ws.org/Vol-1876/paper07.pdf>.
- J. Schreiber. Bayesian network structure learning in pomegranate, 2018a. URL https://github.com/jmschrei/pomegranate/blob/master/tutorials/Tutorial_4b_Bayesian_Network_Structure_Learning.ipynb. Accessed: 15 June 2018.
- J. Schreiber. Pomegranate, 2018b. URL <http://pomegranate.readthedocs.io/en/latest/>. Accessed: 19 June 2018.
- SDMX.org. SDMX, 2018. URL <https://sdmx.org/>. Accessed: 7 June 2018.
- W. Shen. Memex, 2016a. URL <http://www.darpa.mil/program/memex>. Accessed: 26 April 2016.
- W. Shen. Xdata, 2016b. URL <http://www.darpa.mil/program/xdata>. Accessed: 26 April 2016.
- J.D. Singer. Threat-perception and the armament-tension dilemma. *The Journal of Conflict Resolution*, 2(1):pp.90–105, 1958. URL <https://www.jstor.org/stable/172848>.

- A. Smith and J. McCulloch. *An inquiry into the nature and causes of the wealth of nations*. Texts in culture. Manchester University Press, Manchester, UK, 1995.
- S. P. Somashekhar, M.J. Sepplveda, S. Puglielli, A. D. Norden, E. H. Shortliffe, C.R. Kumar, A. Rauthan, N.A. Kumar, P. Patil, K. Rhee, and Y. Ramya. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Annals of Oncology*, 29(2):pp.418–423, 2018. URL <https://www.ncbi.nlm.nih.gov/pubmed/29324970>.
- K. Sprague, F. Grippink, J. Manyika, L. Moodley, B. Chappuis, K. Pattabiraman, and J. Bughin. Offline and falling behind: Barriers to internet adoption, 2014. URL <http://www.mckinsey.com/industries/high-tech/our-insights/offline-and-falling-behind-barriers-to-internet-adoption>. Accessed: 27 April 2016.
- Reuters Staff. Egypt receives first of two french mistral warships, 2016. URL <http://www.reuters.com/article/us-france-egypt-deals-idUSKCN0Y01D0>. Accessed: 11 May 2017.
- Stanford University Stanford. Stanford CoreNLP Natural Language software, 2016a. URL <http://stanfordnlp.github.io/CoreNLP>. Accessed: 17 April 2016.
- Stanford University. Stanford. Memex human trafficking summary, 2016b. URL <http://deepdive.stanford.edu/showcase/memex>. Accessed: 17 April 2016.
- Stanford University Stanford. Stanford Named Entity Recognizer, 2017. URL <http://nlp.stanford.edu/software/CRF-NER.html>. Accessed: 5 August 2016.
- Statistica. Freedom house index: Internet freedom in selected countries 2015, 2015. URL <http://www.statista.com/statistics/272533/degree-of-internet-freedom-in-selected-countries/>. Accessed: 27 April 2016.
- N. Thapen, D. Simmie, C. Hankin, and J. Gillard. DEFENDER: detecting and forecasting epidemics using novel data-analytics for enhanced response. *PLoS One [Online]*, 11(5), 2016. URL <https://doi.org/10.1371/journal.pone.0169944>. Accessed: 30 May 2016.

- L. Thomas. France balks at delivering new warship to Russia, 2014. URL <http://uk.reuters.com/article/uk-ukraine-crisis-france-mistral-idUKKBN0GY1Z120140903s>. Accessed: 6 May 2016.
- P. Thottakkara, T. Ozrazgat-Baslanti, B.B. Hupf, P. Rashidi, P. Pardalos, P. Momcilovic, A. Bihorac, and Z. Xie. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS ONE*, 11(5), 2016. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883761/>.
- J. Tian. A branch-and-bound algorithm for MDL learning bayesian networks. In C. Boutilier and M. Goldszmidt, editors, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Stanford, CA, 30 June - 3 July 2000*, pages 580–588, San Francisco, CA, USA, 2013. Stanford University, Morgan Kaufmann. URL <https://arxiv.org/ftp/arxiv/papers/1301/1301.3897.pdf>.
- T. Tietenberg. *Environmental and natural resource economics*. Pearson series in economics. Pearson Education, Upper Saddle River, NJ, USA, 9th edition, 2012.
- A. Toniolo, T.J. Norman, A. Etuk, F. Cerutti, R.W. Ouyang, M. Srivastava, N. Oren, T. Droppe, J.A. Allen, and P. Sullivan. Supporting reasoning with different types of evidence in intelligence analysis. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems, Istanbul, Turkey , 4-8 May 2015*, volume 2, pages 781–789, Liverpool, UK, 2015. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS). URL <https://dl-acm-org.iclibezp1.cc.ic.ac.uk/event.cfm?id=RE146&tab=pubs>.
- D.J. Trump. Presidential proclamation adjusting imports of aluminium into the United States, 2018a. URL <https://www.whitehouse.gov/presidential-actions/presidential-proclamation-adjusting-imports-aluminum-united-states-4/>. Accessed: 6 June 2018.
- D.J. Trump. Presidential proclamation adjusting imports of steel into the United States, 2018b. URL <https://www.whitehouse.gov/presidential-actions/>

- presidential-proclamation-adjusting-imports-steel-united-states-4/. Accessed: 6 June 2018.
- University of California UCLA. Lucene, 2016. URL <http://oak.cs.ucla.edu/cs144/projects/lucene/>. Accessed: 1 August 2016.
- UK Data Service UKDS. UK Data Service, 2018. URL <https://www.ukdataservice.ac.uk/>. Accessed: various dates 2018.
- United Nations UN. UN Standard Products and Services Code, 2018. URL <https://www.unspsc.org/>. Accessed: 8 June 2018.
- United Nations Statistics Division UNSD. United Nations ISIC Rev. 4, 2017. URL <https://unstats.un.org/unsd/cr/registry/isic-4.asp>. Accessed: 14 September 2017.
- US Congress. The defence production act of 1950, as amended, 2009. URL http://www.fema.gov/media-library-data/20130726-1650-20490-5258/final_defense_production_act_091030.pdf. Accessed: 9 May 2016.
- US Government USGov. US Code, Title 10 Armed Forces, Subtitle A General military law, Part IV service, supply and procurement, Chapter 148 national defense technology and industrial base, defense reinvestment, and defense conversion, 1956. URL <http://uscode.house.gov/view.xhtml?path=/prelim@title10/subtitleA/part4/chapter148&edition=prelim>. Accessed: 9 May 2016.
- C. Vandeppeer. *Rethinking threat: Intelligence Analysis, intentions, capabilities and the challenge of non-state actors*. PhD thesis, University of Adelaide, 2011. URL <https://digital.library.adelaide.edu.au/dspace/bitstream/2440/70732/8/02whole.pdf>. Accessed: 1 June 2016.
- W3C. Defining n-ary relations on the semantic web, 2006. URL <https://www.w3.org/TR/swbp-n-aryRelations/>. Accessed: 31 July 2018.
- World Wide Web Consortium W3C. Rdf schema 1.1, 2014. URL <http://www.w3.org/TR/rdf-schema/>. Accessed: 30 January 2019.

- R.H. Wade. Return of industrial policy? *International Review of Applied Economics*, 26(2): pp.223–239, 2012. URL <https://doi.org/10.1080/02692171.2011.640312>.
- B. Warf. Geographies of global internet censorship. *GeoJournal*, 76(1):pp.1–23, 2011. URL <https://doi.org/10.1007/s10708-010-9393-3>.
- The World Bank WB. Gross domestic product - data, 2018. URL <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD>. Accessed: 8 May 2018.
- A. Weiss. Searching in a global environment: Finding information from and on foreign countries, regions and markets. *Business Information Review*, 31(4):pp.243–256, 2014. URL <http://bir.sagepub.com/content/31/4/243.abstract>. Accessed: 30 May 2016.
- Wikinista, 2018. URL <http://wikinista.wikispaces.com/file/detail/grass.png>. Accessed: 18 June 2018.
- Wikispaces, 2018. URL <http://atoms2xp-technology-presentation.wikispaces.com>. Accessed: 18 June 2018.
- P. Wu, M. Jackman, D. Abecassis, R. Morgan, H. De Villiers, and D. Clancy. State of connectivity 2015: A report on global internet access, 2015. URL <https://fbnewsroomus.files.wordpress.com/2016/02/state-of-connectivity-2015-2016-02-21-final.pdf>. Accessed: 30 May 2016.
- J. Xu, T.C. Lu, R. Compton, and D. Allen. Civil unrest prediction: A tumblr-based exploration. In W.G. Kennedy, N. Agarwal, and Yang S.J., editors, *7th International Conference on Social Computing, Behavioral-cultural modelling and prediction, Washington DC, 31 March - 3 April 2014, Part of the Lecture Notes in Computer Science book series*, volume 8393, pages 403–411, Switzerland, 2014. Springer, Cham.
- A. Yates, M. Banko, M. Broadhead, M. Cafarella, O. Etzioni, and S. Soderland. TextRunner: Open Information Extraction on the web. In *North American Chapter of the Association for Computational Linguistics Human Language Technologies 2007 Demonstrations, Rochester, New York, 23-25 April 2007*, pages 25–26, Madison, WI, USA, 2007. Associ-

ation for Computational Linguistics, Omnipress Inc. URL <http://www.cs.sjtu.edu.cn/~li-fang/textrunner.pdf>. Accessed: 5 August 2016.

C. Yuan, B. Malone, and X. Wu. Learning optimal bayesian networks using a* search. n.d. URL <http://url.cs.qc.cuny.edu/publications/Yuan11learning.pdf>.

C. Zhang, C. Ré, M. Cafarella, C. De Sa, A. Ratner, J. Shin, F. Wang, and S. Wu. Deepdive: declarative knowledge base construction. *Communications of the ACM*, 60(5):pp.93–102, 2017. URL http://delivery.acm.org/10.1145/3070000/3060586/p93-zhang.pdf?ip=155.198.30.43&id=3060586&acc=CHORUS&key=BF07A2EE685417C5%2EF5014A9D3D5CC2D9%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1536922555_4da55b1aa52bfd7ea98c489f05c26e88.

Appendix A

Pseudocode

INPUT: $|\text{Nat}|$, $|\text{Cap}|$, $|\text{Ind}|$, percentage sparsity and noise, number of repeats.

OUTPUT: Performance metrics: precision, recall, F_1 , percentage correctness.

METHOD:

Step 1: Initialise parameters

```
nat := (1, |Nat|, step_nat)
ind := (1, |Ind|, step_ind)
cap := (1, |Cap|, step_cap)
sparse := (0, sparse_max, step_sparse)
noise := (0, noise_max, step_noise)
repeat := (1, R, step_repeats)
```

Step 2: Iteration through parameters, inner and outer loop

```
for each nat, cap, ind do
  for each repeat do
```

Step 2a: Create constraint graph

```
G := create_constraint(ind, cap)
```

Step 2b: Create simulated data

```
NatInd, IndCap, NatCap := create_sim_data(nat, ind, cap)
```

Step 2c: Transform data

Transform 1, adding noise and sparsity

```
if noise > 0 then
```

```
  NatInd, NatCap, IndCap := add_noise(NatInd, NatCap, IndCap, noise)
```

```
if sparse > 0 then
```

```
  NatInd, NatCap, IndCap := add_sparse(NatInd, NatCap, IndCap, sparse)
```

Transform 2, concatenate NatInd and NatCap matrices

```
NatIndCap = [NatInd NatCap]
```

Transform 3, split data into training and test sets

```
Train, Test, Testmissing := split(NatIndCap)
```

Step 2d: Train model

```
model := BBN(Train, G)
```

Step 2e: Evaluate

```
percentate_correctness := percent_correct(model, IndCap)
```

```
prediction := predict(model, Testmissing)
```

```
prec := precision_score(Test, prediction)
```

```
recall := recall_score(Test, prediction)
```

```
F1 := F1_score(Test, prediction)
```

Figure A.1: Top level code investigating effect of variables on performance.

USING:

```
procedure create_constraint ( integers a, b)
  G := (V, E) where V ∈ {0, ..., a+b-1}, E ∈ <p, q> where
    p ∈ {0,...,a-1}, q ∈ {a,...,a+b-1}
  return(G)
```

Figure A.2: Algorithm for creating constraint graph.

USING:

```
procedure create_sim_data ( integers a, b, c)
  AB := random binary [1..a, 1..b]
  BC := random binary [1..b, 1..c]
  i := (1, a)
  j := (1, c)
  k := (1, b)

  for each i, j do
    if  $a_i b_k \geq b_k c_j$  for all k then  $a_i c_j = 1$ 
    else  $a_i c_j = 0$ 
  return(AB, BC, AC)
```

Figure A.3: Algorithm for creating simulated data.

USING:

procedure add_noise($A_{i \times j, noise}$)

Step 1: Create mask matrix of size A with $i \times j \times noise$ random elements equal to 1

$M := \text{zeros}[1... i, 1... j]$

until $x := i \times j \times noise$ do

random $m_{ij} := 1$

$x := \text{sum}(A)$

Step 2: Create random binary matrix of size A

$R := \text{random binary}[1... i, 1... j]$

Step 3: Replace elements of A with elements of R **in** accordance with M

for each i, j do

if $m_{ij} := 1$ then $a_{ij} := r_{ij}$

return(A)

procedure add_sparse($A_{i \times j, sparse}$)

Step 1: Create mask matrix of size A with $i \times j \times sparse$ random elements equal to 1

$M := \text{zeros}[1... i, 1... j]$

until $x := i \times j \times sparse$ do

random $m_{ij} := 1$

$x := \text{sum}(A)$

Step 2: Replace elements of A with NaN values **in** accordance with M

for each i, j do

if $m_{ij} := 1$ then $a_{ij} := \text{NaN}$

return(A)

Figure A.4: Algorithm for adding noise and sparsity to data.

USING:

procedure split ($A_{i \times j}$)

Step 1: Split matrix A randomly by row into Train and Test matrices **in** a given proportion .

Step 2: Create second test **set** with empty final column

$\text{Test}_{missing} := \text{Test}$

$\text{Test}_{missing}[, j] := \text{None}$

return(Train, Test, $\text{Test}_{missing}$)

Figure A.5: Algorithm for splitting data.

USING:

procedure percent_correct ($G, A_{i \times j}$) (model, IndCap)

Step 1: Convert graph G into adjacency matrix

$G_{matrix} := \text{zeros}[1 \dots i, 1 \dots j]$

for each **tuple** $\langle \text{node_1}, \text{node_2} \rangle$ **in** G **do**

$G_{matrix}[\text{node_1} - 1, \text{node_2} - 1] := 1$

Step 2: Identify where G adjacency matrix matches matrix A

$\text{compare} := G_{matrix} == A$

Step 3: Calculate percent correct

$\text{percent_correct} = \text{sum}(\text{compare}) / (i \times j)$

return(percent_correct)

Figure A.6: Algorithm for calculating percentage correctness.

```

#Declare global variables
limit = 1192 #Counter
result = None
temp2 = pd.DataFrame()

#Put results into dataframe
df = xmltodict.parse(r.text, item_depth=3, item_callback=getDbNames)

def getDbNames(path, item):
    global limit
    global result
    global temp2

    if 'KeyFamily' in path[-1]:
        x = path[-1][1]['id'] #This is the codified database name

        for desc in item['Name']:
            if desc['@xml:lang'] == 'en':
                y = desc['#text'] #This is the English natural language name

                temp = pd.DataFrame({'DB_Name': x, 'DB_Name2': y, index = [1]})

        #Counter.
        limit = limit-1
        temp2 = pd.concat([temp2, temp], axis = 0)

    if { limit > 0:}
        return(True)
    else:
        result = temp2
        return(False)

```

Figure A.7: Parsing response r to extract English natural language and codified names.

```

result_list = list()
URL_Part1 = 'http://stats.oecd.org/restsdmx/sdmx.ashx/GetSchema/'

#Iterating through database code names
for index, row in result.iterrows():
    temp = pd.DataFrame(row)
    URL_Part2 = str(row['DB_Name']) #dataset code

    #concatinating URL.
    URL = URL_Part1 + URL_Part2
    print(URL)

    #API query
    r = requests.get(url = URL)

#Parsing response
    if (r.status_code == 200):
        temp2 = pd.DataFrame({'DataSetType': []})
        xmldict.parse(r.text, item_depth=5, item_callback = get_dbStructure)

        #Testing for keywords
        test = test_Keyword(temp2)
        save_to_file(str(URL_Part2), temp2, test) #Saving the results so can check
        if test == True:

            #Extract databases with keywords via API query (not shown)
            save_name = URL_Part2 + '_obs.csv'
            save_name = save_name

            [obs, var] = Extract_Data(URL_Part2)
            if obs is not None:
                globals()[URL_Part2 + '_obs'] = obs
                result_list.append(URL_Part2 + '_obs')
                obs.to_csv(save_name)
                print(result_list)
            else:
                continue

    else:
        print(URL_Part2, 'Error_status_code')

```

Figure A.8: Iterating through available databases, filtering databases based on keywords and extracting using API query.

Appendix B

Copyrights

Table B.1: Copyright summary for images

Page	Label	Source	Copyright holder	Permission
47	Figure 3.1a, Modaf	[MoD, 2010]	Her Majesty's Government (HMG)	Government Open Source Licence [HMG, 2018]
47	Figure 3.1b, Modaf	[MoD, 2010]	HMG	Government Open Source Licence [HMG, 2018]
47	Figure 3.1c, TEPIDOIL	[MoD, 2010]	HMG	Government Open Source Licence [HMG, 2018]
48	Figure 3.2, DM2	[McDaniel, 2011]	D. McDaniel	Granted, See Figure B.1
63	Figure 4.1, grass	[Wikinista, 2018]	radwulf77	Creative Commons Attribution Share-Alike 3.0 Licence
63	Figure 4.1, weather	[freepngimg, 2018]	freepngimg.com	Requested 18 June, 14 August 2018 via online form.
63	Figure 4.1, sprinkler	[Coolclips, 2018a]	Coolclips.com	With attribution for non-commercial projects [Coolclips, 2018b]
66	Figure 4.2, handwriting	[Nielsen, 2017]	M. Nielson	Creative Commons Attribution Non-Commercial 3.0 Unported Licence
66	Figure 4.2, video icon	[Wikispaces, 2018]	Wikispaces	Creative Commons Attribution Share-Alike 3.0 Licence
66	Figure 4.2, neural net	[Karpathy, 2018]	A. Karpathy	Requested 18 June, 14 August 2018 via email, see Figure B.2.
66	Figure 4.2, spider chart	[Garg, 2015]	L. Garg	Granted, see Figure B.3.
119	Figure 6.2, LSTM cell	[Olah, 2015]	C. Olah	Granted, see Figure B.4.
124	Figure 6.3, n-ary relation	[W3C, 2006]	W3C	Requested 5 September, see Figure B.5.

RE: Permission from David McDaniel to reproduce adaptation of image.

bethm@silverbulletinc.com

Thu 9/6/2018 2:31 AM

To: Greenhough, Helen <h.greenhough15@imperial.ac.uk>;

Cc: Dave McDaniel <davem@silverbulletinc.com>;

Dear Helen,

That would be fine, and glad that you are considering using our material. Please reach out if you need any additional information. I have not been on LinkedIn lately, so sorry if there was a delay in response. We wish you much success with your thesis. By the way, David is an alum of William and Mary, too.

*Sincerely/Elizabeth McDaniel
CEO, Silver Bullet Solutions, Inc.
(703) 893-1423 (O)
(703) 862-1570 (M)
www.silverbulletinc.com*

-----Original Message-----

From: "Greenhough, Helen" <h.greenhough15@imperial.ac.uk>
Sent: Wednesday, September 5, 2018 8:53am
To: "bethm@silverbulletinc.com" <bethm@silverbulletinc.com>
Subject: Permission from David McDaniel to reproduce adaptation of image.

Dear Beth,

I am currently writing my PhD thesis and was using an image from Mr McDaniels work sourced from <http://www.ideasgroup.org/dm2/>
Having failed to reach Mr McDaniel via LinkedIn, I am now reaching out to you as your information is provided on the Silver Bullet website which I believe Mr McDaniels is the Principal Engineer and President.
Could you please pass this email on to Mr McDaniels?
Many thanks

Helen Greenhough

PhD Research Student
Imperial College, Department of Computing
South Kensington Campus, London, SW7 2AZ

Figure B.1: Permission for use of DM2 image, via email, from D. McDaniel c/o B. McDaniel.

Re: permission to reproduce image from your course

Greenhough, Helen

Tue 8/14/2018 4:27 PM

To: karpthy@cs.stanford.edu <karpthy@cs.stanford.edu>;

Hello Andrej,

Could I please trouble you regarding your permission to use an image of yours for my research?
The pertinent email is below.

Many thanks for your time.

Helen Greenhough

PhD Research Student
Imperial College, Department of Computing
South Kensington Campus, London, SW7 2AZ
+44 (0) 7402 037949

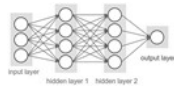
From: Greenhough, Helen
Sent: Monday, June 18, 2018 4:00:16 PM
To: karpthy@cs.stanford.edu
Subject: permission to reproduce image from your course

Dear Andrej,

I am currently writing my thesis and was using an image which I had found via a creative commons image search.

However after digging a bit deeper I have sourced it to the course at
<http://cs231n.github.io/convolutional-networks/>

The image I would like to use is below. May I have permission?



Helen Greenhough

PhD Research Student
Imperial College, Department of Computing
South Kensington Campus, London, SW7 2AZ
+44 (0) 7402 037949

Figure B.2: Request for permission for use of neural net image, via email, to A. Karpathy.

Re: Permission to use image in my thesis

Greenhough, Helen

Tue 6/19/2018 9:38 AM

Sent Items

To: Lakshay Garg <lakshayg@outlook.in>;

Many thanks Lakshay!

Helen Greenhough

PhD Research Student
Imperial College, Department of Computing
South Kensington Campus, London, SW7 2AZ
+44 (0) 7402 037949

From: Lakshay Garg <lakshayg@outlook.in>

Sent: Monday, June 18, 2018 4:20:24 PM

To: Greenhough, Helen

Subject: Re: Permission to use image in my thesis

Sure, feel free to use the image :)

From: Greenhough, Helen

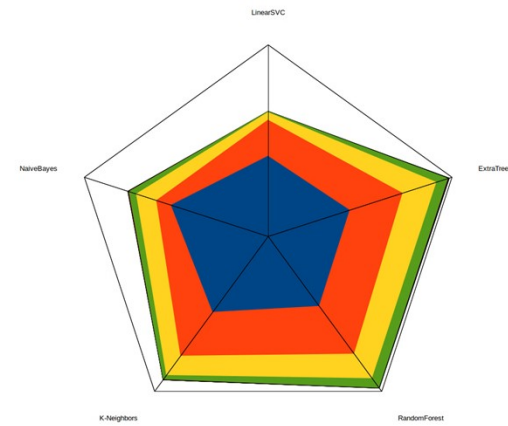
Sent: Monday, 18 June, 08:15

Subject: Permission to use image in my thesis

To: lakshayg@outlook.in

Hello Lakshay,

I planned on using an image



in my thesis. I had found it via a creative commons search but traced it to you at:

<https://stackoverflow.com/questions/31008676/how-to-plot-these-data-in-other-way-to-better-show-the-efficacy-of-my-method/31009326>

Could I please have your permission to reproduce this image in my thesis?

Many thanks

Helen Greenhough

PhD Research Student
Imperial College, Department of Computing
South Kensington Campus, London, SW7 2AZ
+44 (0) 7402 037949

Figure B.3: Permission for use of spider chart image, via email, from L. Garg.

Re: Permission to use image in PhD Thesis

Christopher Olah <christopherolah.co@gmail.com>

Tue 8/14/2018 5:03 PM

To: Greenhough, Helen <h.greenhough15@imperial.ac.uk>;

 1 attachments (135 KB)

pastedImage.png;

Of course! I'm very happy for my diagrams to be reused with attribution.

All the best with your thesis,

Chris

On Tue, Aug 14, 2018 at 8:34 AM Greenhough, Helen <h.greenhough15@imperial.ac.uk> wrote:

Dear Christopher,

Many thanks for your excellent blog on LSTMs.

I am writing to ask your permission regarding using an adaptation of your image in my thesis. The image I have adapted is (attached) and is from <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Many thanks. I look forward to hearing back from you.



Helen Greenhough

PhD Research Student

Imperial College, Department of Computing

South Kensington Campus, London, SW7 2AZ

+44 (0) 7402 037949

Figure B.4: Permission for use of LSTM cell image, via email, from C. Olah.

Permission to reproduce adaption of image.

Greenhough, Helen

Wed 9/5/2018 2:45 PM

To:copyright@w3.org <copyright@w3.org>;

 1 attachments (46 KB)

n-ary relations.png;

To whom it may concern,

I am currently writing my PhD thesis and would like to adapt an image from <https://www.w3.org/TR/swbp-n-aryRelations/>. The image is below.

Could I please have your permission to adapt this image for use in my PhD thesis?

Many thanks

Helen Greenhough

PhD Research Student
Imperial College, Department of Computing
South Kensington Campus, London, SW7 2AZ
+44 (0) 7402 037949

Figure B.5: Request for permission for use of n-ary relation, via email, to W3C.

