Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Structure and Evolution of Legislation Networks

by

Neda Sakhaee

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Science, the University of Auckland, 2020
Abstract

Laws have unique features that evolve in response to societies and environments. A full description of the complex legislation networks that underlie law-specific functions is still missing. We describe a document-document interaction map for the legislation network of New Zealand containing about 138000 highly reliable interactions between about 16000 laws. In this study, the global behaviour of legislation evolution emerges from statistical analyses of the resulting network, together with a novel stochastic model to explain the network generative processes. We observe a dynamic growth pattern of individual laws’ in-degree, providing evidence for a model of evolution for legislation networks. This and future legislation network evolution models should facilitate systematic approaches to understand legislative processes better and improve them.

Our proposed information extraction framework generates reliable dynamic networks of legislation by recognising distinct expressions in legal texts. We note the importance of data accuracy in network analysis and improved approximate string matching techniques. We demonstrate network data-sets with more than 98 percent precision and recall.

The linear growth in legislation network size reflects a fixed capacity for parliament to pass laws. Also, the exponential growth in their density explains improvements in the law drafting processes. The preferential attachment process for legislation is impacted as a result of node aging, node capacity in receiving new edges, and content limitations. Thus legislation networks don't show scale-free power-law properties. Our stochastic model of the legislation network evolution process explains their broad-scale to single-scale behaviour with a Lomax distribution.

**Keywords:** Legislation Network, Generative Model of Network Evolution, Citation Network, Information Extraction, Named Entity Recognition, Approximate String Matching, Stochastic Modelling, Lomax Distribution, Broad-Scale and Single-scale Networks.
Dedication

To:
Iman
and
Shayan,
you are the true meaning of love, happiness, kindness, and life.
I want to send my appreciation to my supervisor Dr Mark C. Wilson for the support of my PhD journey, for his patience, motivation, and knowledge. His supervision helped me during all steps of research and writing of this thesis. I would like to also appreciate my co-supervisor Associate Professor Golbon Zakeri, for her insightful comments and encouragement.

My sincere thanks also go to Professor Shaun Cameron Hendy and Te Pūnaha Matatini, who provided me with an opportunity to join their team and resourced my PhD study. Without their valuable support, it would not have been possible to conduct this research.

I am also grateful to the eResearch Platform and Service Lead, Martin Feller, for his unfailing support and assistance in accessing high capacity platforms. Without this help, this PhD research could take at least twelve months longer.

I gratefully acknowledge Sir Geoffrey Winston Russell Palmer who kindly contributed his insightful comments and valuable guidance.

The members of the Parliamentary Counsel Office have contributed immensely to my understanding of the law drafting process. I would like to acknowledge their close collaboration.

My special thanks go to Glenda Ryan for being the light in the darkest days of my life. I persevered and recovered, and it wouldn't have happened without your help.

Also, I thank my friends in Auckland Transport. In particular, I am grateful to Jeani Powell Smith for her support and encouragement.

Last but foremost, I want to thank my family: my true love fiancé and my beloved son who encouraged me to pursue my dreams and finish my thesis, my mom and dad for supporting me spiritually while writing this thesis and during my whole life.
# Table of Contents

Abstract ii

Dedication iii

Acknowledgements iv

Table of Contents v

List of Tables vii

List of Figures ix

1 Introduction 1

2 Pilot Studies, Legislation Network of New Zealand Current Acts 5
   2.1 Network Science .............................................. 5
   2.2 New Zealand Legislation Network .......................... 7
      2.2.1 Data ...................................................... 7
      2.2.2 Network Building ........................................ 8
      2.2.3 General Network Characteristics ...................... 10
   2.3 Applications .................................................. 13
      2.3.1 Centrality and Importance .............................. 13
      2.3.2 Community Detection ................................... 16
   2.4 Discussions .................................................. 18

3 Information Extraction Framework to Build and Study Dynamic Legislation Network 19
   3.1 The proposed Information Extraction framework ............ 22
      3.1.1 Text Canonicalization .................................. 23
      3.1.2 Named Entity Recognition and Relation Extraction .... 24
      3.1.3 Approximate String Matching .......................... 26
   3.2 Application .................................................... 28
      3.2.1 Dynamic Legislation Network Structure and the Important Acts .... 28
      3.2.2 Legislation Network Communities and Their Application .......... 32
# Table of Contents

3.3 Evaluation and Robustness .......................................................... 35
  3.3.1 Error Estimation, Precision, and Recall ................................. 37
  3.3.2 Robustness ................................................................. 40
3.4 Discussion ................................................................................. 42

4 Evaluation of Existing Network Generative Models for the Legislation Network 44
  4.1 Definitions and Methodologies ................................................... 45
    4.1.1 Network Structure, Dynamics, and Generative Models ............... 46
      4.1.1.1 Small-world Networks ........................................... 46
      4.1.1.2 Scale-free Networks and Preferential Attachment ............ 47
      4.1.1.3 Broad-scale Networks and Single-scale Networks ............ 49
    4.1.2 Real-world Network Evolution: Examples ................................ 50
      4.1.2.1 Scientific Citation Networks .................................... 50
      4.1.2.2 World Wide Web .................................................. 51
      4.1.2.3 Social Networks .................................................... 51
  4.2 Descriptive Statistics of Legislation Network Evolution .................... 52
    4.2.1 Legislation Network Size ............................................... 53
    4.2.2 Legislation Network Out-degree of New Nodes ....................... 56
    4.2.3 Legislation Network In-degree ........................................ 57
    4.2.4 Legislation Network Small-world Properties Evolution ............. 58
    4.2.5 Legislation Network Evolution and Preferential Attachment .... 59
    4.2.6 Legislation Network Evolution and Aging Constraint ............. 61
    4.2.7 Legislation Network Evolution and Cost Constraint ............... 63
    4.2.8 Legislation Network Evolution and Content Constraints .......... 64
  4.3 Evaluation of Existing Generative Models for Legislation Networks .......... 68
  4.4 Discussion ............................................................................. 70

5 Stochastic Generative Model for Network Evolution 72
  5.1 Definitions ............................................................................. 74
  5.2 Stochastic Modeling of Preferential Attachment .......................... 78
  5.3 Proposed Model ...................................................................... 81
  5.4 Evaluation of the Proposed Model ............................................ 86
  5.5 Discussion ............................................................................. 90

6 Conclusion ................................................................................. 92

Bibliography .................................................................................. 95
List of Tables

Table 2.1 Basic network characteristics for the NZAN, including the total number of nodes (Acts), total number of edges (references), and the average number of edges per node (average degree). ........................................... 9
Table 2.2 Network General Measures ............................................ 12
Table 2.3 Binary Reference Network, twenty most important acts .............. 15
Table 3.1 OCR and Text canonicalization result comparison ....................... 24
Table 3.2 Entities, types and examples ........................................... 25
Table 3.3 Examples of the Named Entity Recognition rules ........................ 25
Table 3.4 Relations example ..................................................... 25
Table 3.5 Approximate string matching example ................................ 26
Table 3.6 NZ Legislation Network, this chapter versus the previous chapters .... 28
Table 3.7 Overview of the network measures evolution .......................... 30
Table 3.8 Initial keywords in 12 subject-specific committees of New Zealand Parliament [80] ................................................................. 33
Table 3.9 Communities and their content similarity with the NZ parliament committees. The values in this table show the reference number of the cluster has the highest content similarity with the committees subjects at each point of time. The value in the parenthesis shows the content similarity percentage of the highly matched cluster with the parliament committees subjects. 34
Table 3.10 Errors, sensitivity and specificity ..................................... 38
Table 3.11 Node centrality Pearson correlation between the manipulated network and the original network .................................................... 42
Table 4.1 t-test result - preferential attachment ................................... 61
Table 4.2 t-test result - aging constraint ......................................... 63
Table 4.3 t-test result - high in-degree as a capacity cost constraint in the Top 0.1 % category ................................................................. 64
Table 4.4 Kruskal-Wallis test result - content constraint .......................... 66
Table 4.5 Overview of the network measures evolution .......................... 68
Table 4.6 $p(k)$ the in-degree probability distribution function and parameters of the existing generative models ........................................ 68
Table 5.1  Probability distribution functions, cumulative density functions, and log-likelihood functions of the three models fitted to the selected networks. . . 86

Table 5.2  Bayesian information criterion likelihood statistics for the three selected models, fitted by maximum likelihood to the five different periods of New Zealand legislation network (NZ-LG), scientific citation network of ACM papers (ACM SCN)[146], and social network of computer science scientists (SN-CS) [147]. Larger values of likelihood indicate the performance of models. 87

Table 5.3  Parameter estimates from the Lomax($\nu, \lambda$) model, to the selected networks, in by maximum likelihood (standard errors in parenthesis). The $p$-values of models are based on the Kolmogorov-Smirnov goodness of fitness test statistics [150] for Lomax($\nu, \lambda$). . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
# List of Figures

| Figure 2.1 | Sea Otters, Kelp Forest food web | 6 |
| Figure 2.2 | Citation network of papers proposing HIB theories | 7 |
| Figure 2.3 | Dataset Building Process | 9 |
| Figure 2.4 | New Zealand Current Acts Network (NZAN) | 10 |
| Figure 2.5 | Representations of the triangles in a directed network | 11 |
| Figure 2.6 | An example of the Katz prestige methodology applied to the Public Finance Act | 14 |
| Figure 2.7 | Louvain algorithm communities for the NZAN, shown in the full network (left) by colour and in a projected network of communities (right) where node size is proportional to the number of nodes inside each community | 16 |
| Figure 2.8 | Application of Community Detection | 17 |
| Figure 3.1 | Married Women Property Protection Act 1860 | 22 |
| Figure 3.2 | Legal Text IE Framework | 23 |
| Figure 3.3 | Precision and Recall comparison of the approximate string matching techniques | 26 |
| Figure 3.4 | Overview of the network structure evolution | 29 |
| Figure 3.5 | Time evolution of the top ten acts and the top subjects in the top twenty acts | 31 |
| Figure 3.6 | The trend of closeness centrality of the acts in each community | 36 |
| Figure 3.7 | Error Diagram | 37 |
| Figure 3.8 | Changes of the network diameter $d$ as the function of fraction $f$ | 40 |
| Figure 3.9 | Robustness of the top nodes as the function of fraction of manipulated edges $f$ | 41 |
| Figure 4.1 | An Erdős-Rényi random graph and a small-world legislation network with the same size | 46 |
| Figure 4.2 | Log-log complementary cumulative distribution of different standard distributions | 53 |
| Figure 4.3 | Legislation Network growth statistics | 54 |
| Figure 4.4 | Node addition model | 54 |
| Figure 4.5 | Edge growth model | 55 |
Figure 4.6  Average out-degree of new nodes ........................................ 56
Figure 4.7  Node’s in-degree growth .................................................. 58
Figure 4.8  The influence of preferential attachment on New Zealand legislation network for the selected periods. The boxplot compares the average in-degree of nodes based on their attachment story (1 if the node received at least a new edge, and 0 if no new nodes attached to the node at each time stamp (year)) ................................................................. 60
Figure 4.9  The impact of age as a constraint on New Zealand legislation network evolution for the selected periods. The boxplot compares the average age of nodes based on their attachment story (1 if the node received at least a new edge, and 0 if no new nodes attached to the node at each time stamp (year)) ................................................................. 62
Figure 4.10 The impact of node high in-degree as a capacity cost constraint on New Zealand legislation network evolution for the selected periods. The boxplot compares the average in-degree of nodes based on their attachment story, and also based on their position of in-degree. The position is considered as a categorical variable, nodes within the category of top 0.1% in-degree, nodes within the interval of 0.3%-top 0.1%, and nodes within the 99.7 in-degree percentile ................................................................. 65
Figure 4.11 The impact of node content as a constraint on the New Zealand legislation network evolution process for the selected periods. The pie-chart compares nodes attachment story based on the component that nodes belong. The node class as a categorical variable reflects the most significant weekly connected components of the network at each period. .................. 67
Figure 4.12 log-log complementary cumulative plot - observe shape of legislation network in-degree distribution .............................................. 69
Figure 5.1  The process of node addition - stochastic interpretation of preferential attachment process ................................................................. 79
Figure 5.2  Comparison between the distributions ........................................ 85
Figure 5.3  log-log complementary cumulative plot - Lomax distribution for the selected networks ................................................................. 88
Figure 5.3  log-log complementary cumulative plot - Lomax distribution for the selected networks - continued ................................................................. 89
Figure 5.3  log-log complementary cumulative plot - Lomax distribution for the selected networks - continued ................................................................. 90
Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 2—Pilot Studies, Legislation Network of New Zealand Current Acts
Pages: 6-18
Publication:

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Main Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>85</td>
</tr>
</tbody>
</table>

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Mark C Wilson</td>
<td>Initial discussions, proofreading</td>
</tr>
<tr>
<td>Dr Golbon Zakeri</td>
<td>Initial discussions, interpretation of results</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:
- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Mark C Wilson</td>
<td>[Signature]</td>
<td>19 June 2019</td>
</tr>
<tr>
<td>Dr Golbon Zakeri</td>
<td>[Signature]</td>
<td>24 June 2019</td>
</tr>
</tbody>
</table>

Last updated: 28 November 2017
Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. Please include one copy of this form for each co-authored work. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 2- Pilot Studies, Legislation Network of New Zealand Current Acts
Pages: 6-18
Publication:

| Nature of contribution by PhD candidate | Build the network structure, data gathering and preparation, write the algorithms, develop the codes, examine the data, visualisation, application analysis, main author of the paper. |
| Extent of contribution by PhD candidate (%) | 80 |

**CO-AUTHORS**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Mark C Wilson</td>
<td>Initial discussions, proofreading</td>
</tr>
<tr>
<td>Associate Professor Golbon</td>
<td>Initial discussions, interpretation of results</td>
</tr>
<tr>
<td>Zakeri</td>
<td></td>
</tr>
<tr>
<td>Professor Shaun Hendy</td>
<td>Initial conception; introduction</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Mark C Wilson</td>
<td></td>
<td>19 June 2019</td>
</tr>
<tr>
<td>Associate Professor Golbon</td>
<td></td>
<td>24 June 2019</td>
</tr>
<tr>
<td>Zakeri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professor Shaun Hendy</td>
<td></td>
<td>25 June 2019</td>
</tr>
</tbody>
</table>

Last updated: 28 November 2017
Co-Authorship Form

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

**Chapter 3 - Information Extraction Framework to Build and Study Dynamic Legislation Network**

**Pages 20 - 42**

**Publication:**
Information Extraction Framework to Build Legislation Network - submitted to Artificial Intelligence and Law Journal December 2018

<table>
<thead>
<tr>
<th>Nature of contribution by PhD candidate</th>
<th>Main Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent of contribution by PhD candidate (%)</td>
<td>85</td>
</tr>
</tbody>
</table>

### CO-AUTHORS

<table>
<thead>
<tr>
<th>Name</th>
<th>Nature of Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Mark C Wilson</td>
<td>General advice on approximate string matching and text analysis; proofreading</td>
</tr>
</tbody>
</table>

**Certification by Co-Authors**

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate’s contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

<table>
<thead>
<tr>
<th>Name</th>
<th>Signature</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Mark C Wilson</td>
<td>MCW Wilson</td>
<td>19 June 2019</td>
</tr>
</tbody>
</table>

*Last updated: 28 November 2017*
Chapter 1

Introduction

The impact of real-world graphs on our life is global; they are part of our environment. The commercial use of the Internet since 1989 offered valuable and quick data sharing methods and competitive storage. It fundamentally transformed the way we gather, assemble, share, and interpret data using network science methodologies. The revolution of the World Wide Web (WWW) in early 1990, a network of documents hosted across the internet, made it possible to navigate text in a new way; it allowed users to follow hyperlinks to source documents stored remotely [1]. Documents stored on the World Wide Web became accessible by automated web crawlers [2], which enabled readers to find documents anywhere that contain relevant keywords. Today, New Zealand current active legislation is available on the web [3], with hyperlinks that will allow users to follow cross-references between sections and subsections of acts. Plus, New Zealand historical legislation is accessible in scanned PDF format [4] that can be converted to hyperlinks. This raises appealing questions: *can we analyse legislation systems as complex citation networks?*

**What are the benefits of using network science to study the corpus of legislation?**

Citation networks generated from scientific papers are significant research objects of network science. In these networks, each node represents a paper, and each edge represents a citation from one document to another. Citation networks of scientific articles have been relatively well-studied [5, 6, 7, 8], but other application areas much less so. Analysis of legal documents using a citation network approach has a short history. One interesting application with a legal flavour is a study of cases citing majority opinions of the United States Supreme Court [9]. This study used network-based measures of importance to find the most legally relevant precedents. Another study proposes a semantics-based legal citation network viewer as a new tool for law professionals [10]. The network of French legal codes has been described [11], and another study compares several network representations of the corpus of United States Supreme Court decisions [12].
More recently, the corpus of European Union legislative documents has been described by Koniaris et al. as a citation network [13, 14]. They examined the structural analysis of static legislation networks based on the well-structured data available at EUR-Lex [15]. There are few available similar datasets that provide network information of legislation, and the majority of jurisdictions including New Zealand provide old laws in scanned PDF format. Thus there is an area that has not yet been fully explored about the methods of building dynamic legislation networks. Therefore, one of the goals of this study is to identify a reliable process to extract dynamic network information from old and poor-quality documents. We purpose this information extraction framework for legal documents, but one can extend the applications to the other types of old text files.

The development of ranking by importance algorithms such as PageRank [16] on the World Wide Web suggests another interesting question: using network science methodologies, can we rank the importance of particular parts of legislation? Without any network science hat, one could look at how often particular acts are used, or cited by the courts, for instance. This approach would have some benefit, but it instead may result in nominating acts based on their universality rather than their importance. It is the network science approach that we wish to explore in this work, while acknowledging the potential value of other approaches.

More recently many network science applications have focused on mathematically describing the generative models of evolutionary networks. In late 1990, the rediscovery of preferential attachment and growth mechanisms sought to explain the evolution of the World Wide Web with a scale-free structure [17]. Similar processes attempted to explain the origin of the biological, social and economic networks of competitive systems. The introduction of these models claimed that the scale-free networks are the consequence of self-organisation due to the local decisions made by the individual nodes, based on information that is biased toward the more precious nodes. Such models described the process of rich gets richer and resulted in an estimate of a power-law degree distribution for scale-free networks. These new network science developments suggest an interesting question for legal information systems: can we model and explain the evolution processes of the legislation network?

Two parts of the literature of network evolution have not yet been clarified. First, in the preferential attachment and following similar extensions, critical parameters of the model are assumed

---

In 1976 Price proposed a cumulative advantage process for scientific citation networks which is fundamentally equal to the Barabási-Albert preferential attachment model.
to be constant for elements in different periods. One example: the number of newly introduced
nodes by any new node at each time is a parameter that widely assumed to be fixed. However,
in practice, this parameter is observed to follow a power-law distribution [18][19], exponential
distribution [18][20], or Gaussian distribution [20]. Many researchers identified such gaps but
didn't answer them precisely. We, therefore, statistically analyse the presence of such discrep-
cancies for legislation networks and then we aim to address them in our proposed stochastic model
of network evolution.

In addition to the above issue, several real-world networks are observed to be scale-free and
follow an evolution pattern with power-law degree distribution. But the controversial subject in
the literature of network evolution is to address the growth model of non-scale-free networks.
Here, we establish a different approach to model scale-free networks and design statistical test
scenarios to examine their practicality. Our findings emphasise the structural difference in real-
world graphs and provide new theoretical explanations of these non-scale-free patterns.

In this thesis, we seek to accomplish a few tasks concerning the application of network science
to the study of legislation. First, in Chapter 2, we aim to persuade readers that legal authorities
and scientists should take advantage of network science more because of the significant concep-
tual approaches and analytical tools that network science provides. Specifically, network science
highlights the potential application of local connection patterns in defining and learning collect-
ive impacts of legislation. Chapter 2 also emphasises and begins to address the difficulty of
accessing all required data for dynamic network science studies of legislation corpus.

But as the science of networks elaborates, it guarantees to provide new approaches to the anal-
ysis of empirical data and new tools for modelling the expected results of legal change. Thus in
Chapter 3 we propose to assure a robust information extraction framework for building dynamic
Legislation Networks from legal documents. Unlike supervised learning approaches which re-
quire additional calculations, the idea in Chapter 3 is to apply information extraction method-
ologies by identifying distinct expressions in legislation documents and extract quality network
information. We highlight the importance of data accuracy in network analysis and improve ap-
proximate string matching techniques for producing reliable network data-sets with more than
98 per cent precision and recall.

For the most part, an edge from one law to another shows either that the later law builds upon
the elements of the earlier act, or that the new pieces of legislation are closely enough related
that the previous law was material to change. The legislation network thus provides a map of the
relationships between acts which can be explored using network science techniques. Chapter 3
also illustrates the application of a network approach to empirical data by describing the results of a network science study of New Zealand legislation network. The overall goal of the study is to gain a better understanding of the structure of unique relationships between the acts and to investigate behaviours to gain insight into legislation corpus and its relation to history and society.

In Chapter 4 of this thesis, we employ network science approaches to interpret the growth behaviour of the legislation network for the past few decades. In our study, the nodes are New Zealand current and historical acts, and the edges are references, citations, or amendments of one act by another. Legislation references indicate conceptual relationships between the laws. We have performed a detailed study of the evolution of the legislation network since 1840. In looking at how the network structure has evolved, we distinguish between possible explanations for the network growth behaviour.

Our analysis in Chapter 4 points us to conclude that increasing the number and size of legislative instruments alone are unlikely to explain how the legislation network has evolved in recent years. Thus we require more growth behaviour indications to model the network evolution mathematically. The number of edges introduced at each time and the growing behaviour of each node in-degree leads us to collect enough assumptions to fill the gap in the literature of generative network models.

Being the core of this thesis, Chapter 5 describes the possibility of applying a stochastic modeling approach to study network evolution. We first describe how the stochastic modelling concept is useful in describing the existing models of preferential attachment. We then propose our stochastic model based on proper assumptions around critical parameters that we observed in Chapter 4. For example, our approach to utilise empirical observations about the number of newly introduced edges at each time answers one of the essential gaps in the literature of network evolution.

We then discuss prospects for further study of the implications of the proposed stochastic model. Chapter 6 offers conclusions and returns to the extended point of this thesis that network theory and modelling have great potential to complement the analysis of old, present and future legislation.
Chapter 2

Pilot Studies, Legislation Network of New Zealand Current Acts

This chapter includes two papers published during the initial stages of the study. We presented the first paper in JURIX 2016 [21] to investigate the concept of legislation networks with an application focus on the New Zealand legislation corpus. The idea was to develop such a network from New Zealand current acts, compute relevant network science measures and study the relationship between the legislation network measures and the legal and political factors. For the second paper, we aimed at the audience of the New Zealand Law Community and published the article in the NZ Law Journal in 2017 [22]. The results and influence of these two publications motivated us to continue working on the subject and propose a solution to build dynamic legislation network from non-machine-readable legislation. As described below, we examine the corpus of legislative documents as a whole along with their interconnections. In the study we restrict ourselves to analysis of the relationships between documents as encoded by direct references and amendments, rather than the content of documents.

In this chapter, after a review of network science, we define the legislation network and clarify the data extraction and network building process for the current version of the New Zealand Act Network (NZAN) with examples. We discuss the advantages of building this network by reviewing some possible applications of network science measures and algorithms. Finally, we finish by discussing plans, obstacles to our research programme, and open questions.

2.1 Network Science

A network is a structure amounting to a set of objects in which some pairs of the objects are in some sense related. The objects correspond to mathematical abstractions called vertices, nodes
or points. Each of the related pairs of nodes is called an edge, arc or line [23]. Basic statistics associated with a network include number of nodes, number of edges, and average degree. The degree of a node is the number of edges connected to it. The average degree of the network is the mean degree of all nodes. So far we have only discussed undirected networks in which edges go both ways. The theory of directed networks is also well developed — one speaks then of in-degree (number of incoming edges) and out-degree. Directed networks will be more relevant to this study.

Figure 2.1 (taken from [24]) shows a simple application of network science to study a food web. Food webs describe which organisms eat which in an ecosystem, forming a directed network that is important for studying ecosystem resilience and sustainability. Visualising a network by drawing nodes and edges is often the first step in analysing a network. However when the network is large and complex, more quantitative analysis is required. For example, social network analysis examines the complex structure of relationships between social entities of people, but may also consider socially constructed entities such as groups, organizations, nation states, web sites, and publications.

![Figure 2.1: Sea Otters, Kelp Forest food web](image)

Figure 2.2 (based on [25]) shows the scientific citation relations among a small group of research papers, those proposing models of “Human Information Behaviour” (HIB). Each research paper will typically cite a small subset of the previous corpus of relevant work in a field (of the order
of 10-100 papers, depending on the field of research), but the resulting citation networks may consist of millions of papers.

![Citation network of papers proposing HIB theories](image)

**Figure 2.2:** Citation network of papers proposing HIB theories

### 2.2 New Zealand Legislation Network

The New Zealand legislation system includes several types of legislative documents such as Bills, Acts, Regulations and Case Laws, which results in a complex multilayer network. We focus here on one of the layers, the New Zealand Act Network (NZAN).

#### 2.2.1 Data

There are several different classes of legal documents. The main ones discussed here are Acts of Parliament: public, private, local, provincial, and imperial. This includes Amendment Acts whose purpose is to change a previous act (amendments to existing acts can also be made by other acts that are not officially titled Amendment Act). More than 8800 files in XML format were
collected from the NZ Government Legislation website [3] from a time range of a hundred years. Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. Required pieces of data are extracted from the XML files of the acts. Each file includes name, year, termination status, and links to other official documents.

There are two types of links in these documents as shown in Figure 2.3:

- Reference: If a document mentions another document in order to define a concept or to refer to a specific part of it, then the defined link is a reference.

- Amendment: If a document amends another document in order to add new information or change the current law, then the defined link is an amendment.

We encountered two problems in automating the process of data extraction. First the internal structure of the XML files is not completely consistent, as we would expect it to be in standardized documents. Second, our method of extracting links does not result in clean data, as the result of connectives and prepositions in titles of Acts. These unintended words required a multistage cleanup process.

### 2.2.2 Network Building

Using the dataset described above, the New Zealand Act Network (NZAN) can be built as follows:

- Nodes are New Zealand Acts (including Amendment Acts);

- Directed edges point from one document to another whenever the first refers to the second or amends it.

Edges can be binary or weighted. For example the Conservation Act 1987 refers to the Marine Mammals Protection Act 1978 in six different places. In this case the edge from the Conservation Act 1987 to the Marine Mammals Protection Act 1978 can either be considered as a binary edge with weight 1 or a weighted edge with weight 6.

There are other derived networks we can construct in the same way, and again each can be binary or weighted. The two of interest to us are as follows:

- Reference Network: Nodes are original documents, so this network excludes amendment act nodes by merging amendment nodes with original acts. Edges are caused by references, not amendments.
2.2. New Zealand Legislation Network

![Diagram of dataset building process]

**Figure 2.3: Dataset Building Process**

- Amendment Network: Nodes are all the original documents and amendments as nodes. Edges are only amendment links.

<table>
<thead>
<tr>
<th>Table 2.1: Basic network characteristics for the NZAN, including the total number of nodes (Acts), total number of edges (references), and the average number of edges per node (average degree).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
</tr>
<tr>
<td>Edges</td>
</tr>
<tr>
<td>Average Total Degree</td>
</tr>
</tbody>
</table>

The outcome of the above process for New Zealand Acts is the legislation network with the general specifications in Table 2.1. Figure 2.4 shows a small part of a visualisation of NZAN. A full visualisation of this network is available at [26].
2.2.3 General Network Characteristics

In this section we compute general network science measures of degree, average path length, clustering coefficient, and small world property [27, 28]. Unlike previous studies which applied methodology appropriate only for undirected graphs, we consider a legislation network as a directed graph (with cycles) for all the calculations. Here we define a few network science measures before we present the results.

Definition 2.2.1. The average path length of the network is the average of the lengths of all shortest paths between all pairs of nodes. The shortest path length between two nodes is the minimum number of edges that must be crossed to get from one node to the other.

Definition 2.2.2. In network science, the clustering coefficient measures the tendency of nodes in a network to cluster together [28] [27]. The clustering coefficient of node $i$ is defined as the ratio between all the possible triangles that include node $i$ and the number of all possible triangles that could be formed. In directed networks, triangles with edges pointing in different directions have different impacts in terms of the resulting pattern. Figure 2.5 presents all the possible scenarios as defined by Tabak et al. in 2014.

One pattern could be when there is a cyclical relation amongst node $i$ and any two of its neighbours as illustrated in Figure 2.5a and Figure 2.5b. The number of triangles with the cyclic pattern can be measured as $CC_{cyc}$ node $i$. 

Figure 2.4: New Zealand Current Acts Network (NZAN)
The middleman scenario happens when one of the incoming edges to node $i$ is from a neighbour that holds an outgoing edge to a third neighbour of node $i$ which has got a direct incoming edge from node $i$. Or similarly when one of the outgoing edges from node $i$ connects to a neighbour that holds an incoming edge from a third neighbour who also connects to node $i$ with an outgoing edge as we show in Figure 2.5c and Figure 2.5d. In both mentioned scenarios, node $i$ has middlemen and $CC_{mi,d}$ can measure the resulting triangles.

When node $i$ contains inward edges, then Figure 2.5e, and Figure 2.5f can measure the possible triangles $CC_{i,n}$. Similarly when node $i$ contains outgoing edges, then Figure 2.5g and Figure 2.5h can measure the possible triangles for $CC_{out}$. The total clustering coefficient of node $i$ is the sum of the four values that we explained. If we consider the network as an undirected graph, the clustering coefficient is smaller.

**Definition 2.2.3.** The Erdős-Rényi model [29], was the first method for generating **random graphs** with Poisson degree distribution [30], networks in which properties such as the number of nodes and edges are set in some random way.

In the Erdős-Rényi model, graph $G$ is constructed by connecting $n$ nodes randomly. Each edge is added to the graph with probability $p$ independent from every other edge.

**Definition 2.2.4.** Graphs are considered to have the **small-world property** if they have small average shortest-path lengths [31] and large clustering coefficients [28]. To examine the significance of small-world effect for real-world graphs, the best way is to compare their average path length...
and clustering coefficient with Erdős-Rényi random graphs of same size as we defined in Definition 2.2.3. According to Watts and Strogatz where the average shortest path length of the graph $L_p$ is approximately as small as $L$ in random graphs and clustering coefficient of the graph $C_p$ is significantly larger than $C$ in random graphs, then the graph is a small-world network [28][32].

With more exploration of small-world networks, a robust process was proposed to show how the hypothesis of a small-world network may be tested. The research resulted in the introduction of $\sigma$ based on the trade-off between high local clustering and short path length. [33].

$$\sigma = \frac{C_p}{L_p}$$

Based on the definition, graphs with small-world $\sigma$ significantly larger than 1 are small-world networks and are significantly different from random graphs.

<table>
<thead>
<tr>
<th>Table 2.2: Network General Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Nodes</strong></td>
</tr>
<tr>
<td>Nodes</td>
</tr>
<tr>
<td>Edges</td>
</tr>
<tr>
<td>CCcyc</td>
</tr>
<tr>
<td>CCmid</td>
</tr>
<tr>
<td>CCin</td>
</tr>
<tr>
<td>CCout</td>
</tr>
<tr>
<td>Average CC</td>
</tr>
<tr>
<td>Average Path length</td>
</tr>
<tr>
<td>Small world $\sigma$</td>
</tr>
<tr>
<td>Acyclic</td>
</tr>
<tr>
<td>in-degree 0, out-degree &gt; 0</td>
</tr>
<tr>
<td>in-degree &gt; 0, in-degree = 0</td>
</tr>
<tr>
<td>Isolated Nodes</td>
</tr>
</tbody>
</table>

1 In Table 2.2 RN is a random graph $G_{n,p}$ chosen according to the Erdős-Rényi model with a specific number of vertices $n$ and connection probability of $p$ chosen to match the network in question [34]. The indexes are calculated based on means from a sample of 100 graphs.
Table 2.2 illustrates the general measures for all six networks. As can be seen about one third of the edges are amendment, and two thirds of them are reference links. As illustrated, ACT and CITE networks have the small world property owing to the high clustering coefficient and low average path length compared to random networks. However AMEND is not as significantly a small world network comparing to CITE, and many of its nodes are isolated with 0 in-degree, and 0 out-degree.

2.3 Applications

2.3.1 Centrality and Importance

Centrality measures are standard network science tools for measuring “importance” of nodes [16]. There are many different centrality measures. We consider the most relevant for the kind of network we study here is the Katz prestige [35]. This measures the number of paths that lead to a node, with longer paths systematically weighted less than shorter ones. By using this measure we can identify the Acts which are procedurally important. Later studies may correlate the procedural and social importance of the Acts.

In the NZAN, the Katz prestige can be interpreted as the impact of changing a node on the whole system — acts referring to the altered act must themselves potentially be changed, and then acts that refer to these must be changed, etc.

Figure 2.6 shows an example of how Katz Prestige centrality methodology works to measure the importance of the Public Finance Act 1989. As can be seen in the network there are thousands of possible paths to reach this act from the other acts by reference links. Katz prestige uses the number of paths as well as their length to calculate the importance of the document. In this example the Katz prestige for the Public Finance Act is found to be 10.37 when these paths are computed for the whole network.

\footnote{At this stage, we study several types of networks derived from the data:

- ACT: Includes all nodes, and all edges.
- AMEND: Includes all nodes, but only Amendment edges.
- CITE: Includes only original Acts, and only reference edges, Here the amendment edges are excluded to remove direct loops, and amendment nodes are merged to their original Act node to avoid having isolated amendment nodes.

Edges can be considered as either binary (as above) or weighted. For example, the Conservation Act 1987 cites Marine Mammals Protection Act 1978 in six different places. In this case, the edge from Conservation Act 1987 to Marine Mammals Protection Act 1978 can be considered as a weighted edge with weight 6. We also build the weighted version of the three networks above and denote them with prefix W.}
Figure 2.6: An example of the Katz prestige methodology applied to the Public Finance Act.

Table 2.3 shows the top twenty acts based on Katz prestige. Bearing in mind that no analysis of content has been made, these acts certainly seem to be technically important. The fact that legislative drafters saw fit to introduce references to particular acts is the root of their importance. This is analogous to scientific citation networks, although in a purer form — scientific works are sometimes cited for social reasons or disagreements rather than because they are actually useful or relevant. It is clear that network science is telling us something useful about the NZ legal system. A completely self-taught legal scholar should presumably read and understand the above laws before spending time on less important ones.

On the other hand, Katz prestige is zero for 524 nodes. These nodes represent the acts that are not cited by any other act. The majority of such acts have at least one repealed section, e.g. Civil Service Act 1908, and they appear to be relatively obscure. We suggest that such acts are good candidates for initial examination if a cleanup of the corpus of acts is to be undertaken.

Importance scores of this type can be used to study social and political processes related to lawmaking. As an example, we analyse the binary reference network with Katz prestige to examine how the major party of government relates to the importance of legislation that is passed in Parliament.

Hypothesis: Labour and National-led governments pass legislation with different average Katz prestige. The Null Hypothesis is that the average Katz centrality of the acts passed during Na-
Table 2.3: Binary Reference Network, twenty most important acts

<table>
<thead>
<tr>
<th>Act</th>
<th>Rank</th>
<th>Katz Prestige Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Finance Act 1989</td>
<td>1</td>
<td>10.37</td>
</tr>
<tr>
<td>Criminal Procedure Act 2011</td>
<td>2</td>
<td>9.65</td>
</tr>
<tr>
<td>Summary Proceedings Act 1957</td>
<td>3</td>
<td>9.28</td>
</tr>
<tr>
<td>State Sector Act 1988</td>
<td>4</td>
<td>8.85</td>
</tr>
<tr>
<td>District Courts Act 1947</td>
<td>5</td>
<td>7.96</td>
</tr>
<tr>
<td>Crimes Act 1961</td>
<td>6</td>
<td>7.47</td>
</tr>
<tr>
<td>Companies Act 1993</td>
<td>7</td>
<td>7.43</td>
</tr>
<tr>
<td>Local Government Act 1974</td>
<td>8</td>
<td>7.4</td>
</tr>
<tr>
<td>Judicature Act 1908</td>
<td>9</td>
<td>7.1</td>
</tr>
<tr>
<td>Privacy Act 1993</td>
<td>10</td>
<td>6.79</td>
</tr>
<tr>
<td>Resource Management Act 1991</td>
<td>11</td>
<td>6.71</td>
</tr>
<tr>
<td>Official Information Act 1982</td>
<td>12</td>
<td>6.58</td>
</tr>
<tr>
<td>Land Transfer Act 1952</td>
<td>13</td>
<td>6.29</td>
</tr>
<tr>
<td>Government Superannuation Fund Act 1956</td>
<td>14</td>
<td>6.21</td>
</tr>
<tr>
<td>Local Government Act 2002</td>
<td>15</td>
<td>6.02</td>
</tr>
<tr>
<td>Regulations (Disallowance) Act 1989</td>
<td>16</td>
<td>5.7</td>
</tr>
<tr>
<td>Tax Administration Act 1994</td>
<td>17</td>
<td>5.23</td>
</tr>
<tr>
<td>Sentencing Act 2002</td>
<td>18</td>
<td>5.16</td>
</tr>
<tr>
<td>Ombudsmen Act 1975</td>
<td>19</td>
<td>5.04</td>
</tr>
<tr>
<td>Legislation Act 2012</td>
<td>20</td>
<td>4.95</td>
</tr>
</tbody>
</table>

tional party-led governments is equal to those passed during Labour-led governments. The Alternative Hypothesis is that the average centrality of the acts passed by National-led governments is not equal to those passed by Labour-led governments.

The average prestige of acts passed by National and Labour-led governments is found to be 0.043 and 0.068 respectively, and a 95% confidence interval for the difference between the two is $[-0.03, -0.0007]$. Thus $H_0$ is rejected, there is a statistically significant difference between the Katz prestige of each party’s acts, and on average Labour have passed more important acts among those acts currently in force. The interpretation of this is not completely clear, and we should note that during the time period under study National was in power 45 years and Labour 35.

For a full study, many different factors should be considered, such as the reform agenda of the particular government, or the majority available in parliament to pass legislation. In our Binary reference Network (in which amendment acts are merged with original acts), National accounts
for about two thirds of the laws. Study of the historical dataset of all laws, current or not, would be useful for exploring this further.

A key point to note here is that without building the entire network it is not possible to study such a hypothesis – such information cannot be obtained by purely local analysis of acts, because sophisticated measures of centrality require consideration of paths in the network of all lengths.

### 2.3.2 Community Detection

Networks science has developed a number of other concepts that have proved useful in the study of real-world networks. For instance, if the nodes can be easily grouped into subsets of nodes that are more densely connected internally than with the rest of the network (this is called modularity), then the network is said to have a community structure. Scientific citation networks form communities by research topic [36], and legislation networks form communities by law topic. By comparing several different community detection algorithms we found that the extended Louvain algorithm [37] for directed weighted graphs best resolved the community structure for the NZAN. Figure 2.7 illustrates how the Louvain algorithm resolves the New Zealand Act Network into communities of acts.

![Figure 2.7: Louvain algorithm communities for the NZAN, shown in the full network (left) by colour and in a projected network of communities (right) where node size is proportional to the number of nodes inside each community.](image)

To label the topics around which each community clusters, we used a keyword search. The list of keywords is generated systematically, but manually. The algorithm to create the list is to take a 10 percent random sample from each community, then find the most repeated words as the original
words in the list. Then between one to three synonyms for those main words are added to the list. After we create this list, in each community we search for the list of keywords that appear in at least 85% of its members, and don’t appear in more than 10% percent of the members in the remaining communities. As an example for the biggest community, ”Natural Heritage and Land”, a keyword list of 56 keywords is used, and the top ten keywords are: land, forest, resort, river, lake, harbour, bay, reserve, foreshore, and garden.

The second example in Figure 2.8 demonstrates how the communities identified in Figure 2.7 are correlated to New Zealand’s historical time periods and major party of the government.

![Figure 2.8: Application of Community Detection](image)

Figure 2.8a shows substantial differences in the types of acts passed in different historical periods. The spine plot shows that in the colony and self-government period (1840 to 1946), interrelated laws involving land and natural resources were predominant. During full independence (1947 to 1983), local government and Maori issues assume greater relative importance. In the restructuring time (1984 to date), due to the population increase, many more currently valid laws have been passed (unsurprisingly given that laws typically have a finite lifespan), and the range of topics covered shows the increasing complexity of society.

Figure 2.8b shows the activities of the major government parties in relation to passing the acts in different communities. As can be seen three first communities are the most interesting area for both National and Labour. The majority of Maori welfare legislation was created by the Labour party. The National party passed more acts in all other communities than the Labour party did. Note that these two parties had different numbers of years in government.
2.4 Discussions

We have high confidence that our data set of currently active laws, extracted from the New Zealand Legislation website [3], is complete. Adding automated content analysis to the set of tools we have used would give additional useful information. However, we are now in a position where comparative studies with other jurisdictions such as Canada could be undertaken. Further improvements in data quality will be necessary before comparative studies of the United Kingdom or Australia are possible. A non-profit organisation has published all Australian laws in plain text format since 2019 [38], but their bulk download is not available.

We also wish to study the historical evolution of the network, in order to model its growth, and correlate changes in the network with external political, technological and social events. The main barrier to this is the inadequacy of the historical data set. In the New Zealand Legislation website [3] there are some historical documents which are referred to by other documents but are not themselves available. Most of these expired documents are available in other resources such as NZLII [4] in scanned PDF format, but this format is not machine-readable. Our estimate of the extent of this missing historical data (which corresponds to missing nodes and edges in our network) is around twenty per cent of the current size of the network. Thus there is an area that has not yet been fully explored about the methods of building dynamic legislation networks, and we aim to address this gap in Chapter 3.

Legal informatics is more advanced in some jurisdictions than in New Zealand (note, however, that Australia does not even provide machine-readable laws on-line). The European Union has established EUR-lex [39] as a structured database of all EU laws. This live database is open access and facilitates customized user queries. A similar database of all Dutch national laws has been built in CEN MetaLex [40]. Australasian Legal Information Institute recently in 2019 provided open access to all Australian laws in plain text format. Given the NZ government's commitment to openness as expressed in the NZGOAL (New Zealand Government Open Access and Licensing) framework, it is desirable for NZ to follow suit. The support of the New Zealand legal community would be very useful in helping to make this a government priority.
Chapter 3

Information Extraction Framework to Build and Study Dynamic Legislation Network

This chapter concerns an Information Extraction process for building a dynamic legislation network from legal documents. The main body of this chapter is based on a paper submitted to the *Artificial Intelligence and Law Journal* in December 2018, which is currently under review [41]. Unlike supervised learning approaches which require additional calculations, the idea here is to apply Information Extraction methodologies by identifying distinct expressions in legal text and extract quality network information. The study highlights the importance of data accuracy in network analysis and improves approximate string matching techniques for producing reliable network data-sets with more than **98 percent precision and recall**\(^1\). We also discuss and challenge the values, applications, and the complexity of the created dynamic Legislation Network.

**Legislation Networks** were first introduced in 2015 [13] and we discussed them more in Chapter 2. These networks are essential to explore the relationship between legislation and societies’ evolution [21]. There are many obvious benefits from studying Legislation Networks [21, 22, 10, 42], but building these networks is not always a straightforward task, as only a few legislation systems provide machine-readable documents or structured databases [15, 43]. The majority of legislation systems supply legal documents in a human-readable format. For example, the New Zealand Parliamentary Council Office provides machine-readable XML files [3] only for the current active Acts, which constitute around 10% of the entire set of Acts. All other historical docu-

\(^1\)We will define the precision and recall measures later in Section 3.3.1
ments are scanned and supplied by a third party institute in Portable Document Format (PDF) [4].

For the extraction of information from legal documents, the first step is the conversion of images into text if the text is not available. This concept is well studied as **optical character recognition (OCR)** [44], and there are several techniques and tools developed to convert typewritten or handwritten images to text. OCR is the first step of the proposed framework, and for the case study, we selected ABBYY FineReader as our tool [45].

**Information Extraction (IE)** involves locating and extracting specific information from text [46]. Information Extraction assumes that in each text file there are one or more entities that are similar to those in other text documents but differing in the details [47].

IE approaches in the legal domain are considerably different from other knowledge areas because of the two main characteristics of legal texts. Legal documents exhibit a wide range of internal structure, and they often have a significant amount of manual editorial value added. One of the earliest information retrievals approaches for legal materials based on searching inside the document was proposed in 1978 [48]. Later works mainly used natural supervised learning techniques to retrieve the required data from legal texts, but with a substantial error [49] [50]. In the proposed framework of this study, several IE tasks are used, and more are described later in this section.

**Named entity recognition (NER)** is one of the main sub-tasks of IE. The goal of this task is to find each occurrence of a named entity in the text [51]. Entities usually include people, locations, quantities, and organisations but also more specific entities such as the names of genes and proteins [52], the names of college courses [53], and drugs [54]. In the New Zealand legislation corpus, entities could be the name of legislative documents such as Acts, Regulations, Bills, Orders, or Case-Laws [21]. In the case study which we explore in Section 3.2, the main required entities inside the text documents are the names of the New Zealand Acts.

The main traditional NER algorithm that identifies and classifies the named entities is statistical sequence modelling [51]. But there are other modern approaches based on combinations of lists, rules, and supervised machine learning [55]. While extracting the required information for Legislation Network, there are clear rules to identify the named entities, and the classification of the entities is not needed. Therefore the second NER approach is more appropriate and discussed further for the proposed framework.
The next IE task which we use in this chapter is to detect the relationships that exist among the recognised entities. This task is called relation extraction (RE) [51]. The earliest algorithm for relation extraction is the use of lexico-syntactic patterns [56]. This algorithm is still valid and widely used, but there are other algorithms introduced later, such as supervised learning [51] and bootstrapping [57][58]. Considering that legislation texts are well structured in New Zealand, it is assumed that there is a large collection of previously annotated material that can define the rules for classifiers.

Approximate string matching techniques find items in a database when there may be a spelling mistake or other error in the keyword [59]. This is becoming a more relevant issue for fast-growing knowledge areas such as information retrieval [60]. Various techniques are studied to address the identity uncertainty of the objects and briefly are reviewed in this study. These techniques could be distance based, token-based, or a hybrid model of the distance and token based models.

Damerau-Levenshtein metrics are the main approximate string matching techniques to address the distance functions between the two strings [61] [62]. The most famous function in this category is edit-distance, and it is defined as the minimum number of changes required to convert one string into the other [62]. Several alternative distance functions to the edit-distance have been proposed such as q-gram and maximal matches [63].

The next set of techniques are token based or probabilistic object identification methods adapted for string matching tasks [64][65] [59]. Jaccard similarity and cosine similarity are common token based measures widely used in the information retrieval community[64]. Hybrid techniques combine distance-based and token-based string matching measures such as Jaro-Winkler [66]. All of the string matching algorithms have been developed by filtering and bit-parallelism approaches.

The fastest algorithms use a combination of filters to discard most of the text by focusing on the potential matches. Hybrid models significantly improve precision and recall reducing the error the range 0.1 to 0.2 [60].

Network inferences require high accuracy of data [67]. In this study we examine various string matching techniques for the Legislation Network and comparing the results, we use a hybrid model of Jaccard similarity and edit-distance as proposed in the next section.

The main contribution of this chapter is the proposed Information Extraction framework, which engages several processes and enables the researcher to have access to the network information from historical documents. This framework makes it possible to study legislation networks as
An Act to protect the Property of Married Women. [28th September, 1860.]

WHEREAS it is expedient to amend the Law relating to Property acquired by Married Women deserted by their husbands, be it therefore enacted by the General Assembly of New Zealand, in Parliament assembled, and by authority of the same, as follows:

I. The Short Title of this Act shall be, the “Married Women’s Property Protection Act, 1860.”

II. A Wife deserted by her husband may at any time after such desertion apply to a Resident Magistrate or to Justices of the Peace in Petty Sessions, for an Order to protect any Money or Property she may acquire by her own lawful industry, and property which she may become possessed of, after such desertion, against her husband or his creditors or any person claiming under him; and such Resident Magistrate or Justices if satisfied of the fact of such desertion, and that the same was without reasonable cause, and that the wife is maintaining herself by her

Figure 3.1: Married Women Property Protection Act 1860

dynamic graphs. In this chapter, the case study covers all Acts in New Zealand legislation corpus including historic, expired, repealed and consolidated Acts as at the end of September 2018. This comes to a set of 23870 PDF files of which about 87% are in scanned image format.

Figure 3.1 shows a sample image of an average quality scanned PDF document. The proposed framework suggests a high-performance procedure to derive network information from such inadequate quality documents. In the following sections, we use examples and the experimental results to illustrate the framework, its performance, and its potential applications.

In this section, we discussed a summary of the required Information Extraction processes and methodologies. In the next section we will present our proposed framework and provide various examples. Then we examine the case study analysis and study the application of the proposed framework. Next, we design several experiments to evaluate the accuracy of the extracted information and to explore the robustness of Legislation Network. The chapter finishes with a quick review of the novelty and the importance of discovering the time-varying behaviour of the Legislation Network.

3.1 The proposed Information Extraction framework

In this section, we discuss the Information Extraction framework to build the Legislation Network. Figure 3.2 depicts the overview of the proposed framework.
3.1. The proposed Information Extraction framework

The process starts with the conversion of non-machine-readable files to text by using OCR available tools. This step is relatively straightforward but could be time-consuming, considering the number of documents in the study. As mentioned earlier, in the case study, we use the tool named ABBYY FineReader [45]. The average accuracy of this step is just above 80 percent and implies the need for a typos analysis step that initiates our proposed approach in Section 3.1.3.

### 3.1.1 Text Canonicalization

The next step in the proposed framework is text canonicalization [68]. There are several required tasks to convert all of the text files into a unique format, so the rules can be defined more easily while running the Information Extraction tasks. The text canonicalization step could be implemented via different approaches depending on the text style and language. In this section, we suggest some of the common tasks and describe two potentially required tasks.

In the case study our designed system transfers all letters to Lowercase. This transition applies a level of consistency across the text documents and the Information Extraction rules. In the experiments, we also suggest a system to replace Special Characters with generic tags in the text. The only character which is not replaced is the parenthesis, as it is often used in the title of the legislation. We also suggest a generic text canonicalization task to replace multiple spaces with one space.
Apart from the general text canonicalization steps, other potential corrections shape the text to make it a better input for the Information Extraction process. The first is to remove text margins that OCR mistakenly merges to the main body of the text. Figure 3.1 includes examples of these margins that might impact the Information Extraction rules and result in an error. As an example, in the case study, the phrase short title is often used in the margin, and OCR merges it to the nearest part of the text. This might impact the named entity recognition task, so the system removes this phrase and many of its possible misspelled forms from the text. The next recommended text canonicalization step is to resolve the misspelling issues for the keywords used in the Information Extraction process. As an example, in this study, Acts are the main entities, and the rule to recognise them uses the keyword Act. So the system corrects some of the possible misspelling forms of the word -act-. To provide a better explanation of this step, Table 3.1 provides an example referring to the third paragraph of the Figure 3.1 image. As can be seen, the text canonicalization converts the text to a more straightforward unique structure prepares it for the next steps of named entity recognition and relation extraction. Typo resolution is not expected at this step, being covered under the last step via Approximate String Matching.

### Table 3.1: OCR and Text canonicalization result comparison

<table>
<thead>
<tr>
<th>OCR</th>
<th>Text canonicalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>I~ The Short Title of this Act shall be, the ((Married Short Title, 'Women's Property Protection Act, 1860.</td>
<td>I the short title of this act shall be the ((married womens propertyprotectio act 1860</td>
</tr>
</tbody>
</table>

#### 3.1.2 Named Entity Recognition and Relation Extraction

As explained, the text canonicalization step normalises the text files to a unique format and prepares them for in-depth information extraction steps. To extract the network node information, we suggest a combined Named Entity Recognition approach which engages rules and supervised learning. To identify the rules, it is required to review a sample set of documents. The sample size is not necessarily significant, but a stratified sampling approach is suggested to eliminate the impact of time-period style and the author's writing style.

Table 3.2 shows examples of the entities in the case study based on the Acts recognition rules. Acts are the core of the New Zealand legislation system, and, as explained before, the case study only considers the Acts. In the case study, we used the stratified sampling method, so the strata are five different periods in a range of more than 200 years. We reviewed a total number of 55 text files, and several clear rules engaging a set of keywords and lists were built to identify the named
3.1. The proposed Information Extraction framework

Table 3.2: Entities, types and examples

<table>
<thead>
<tr>
<th>Type</th>
<th>Tag</th>
<th>Sample</th>
<th>Canonicalized text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>YR</td>
<td>1860</td>
<td>the short title of this act shall be the (married women property protection act 1860</td>
</tr>
<tr>
<td>Act</td>
<td>ACT</td>
<td>married women property protection act 1860</td>
<td>the short title of this act shall be the (married women property protection act 1860</td>
</tr>
</tbody>
</table>

entities. Table 3.3 provides examples of the rules in each stratum and \( y \) represents the year in which the Act is commenced.

Table 3.3: Examples of the Named Entity Recognition rules

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Keyword example</th>
<th>Rule example</th>
<th>Sample document</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y &lt; 1850 )</td>
<td>ordinance</td>
<td>an [keyword] to any phrase of [act name] [date]</td>
<td>Police Magistrates Act 1841</td>
</tr>
<tr>
<td>1850 &lt; ( y &lt; 1900 )</td>
<td>short title, shall be</td>
<td>the [keyword] of this act [keyword] the [act name] [year]</td>
<td>Customs Tariff Act 1873</td>
</tr>
<tr>
<td>1900 &lt; ( y &lt; 1950 )</td>
<td>amend, consolidate</td>
<td>an act to [keyword] any phrase of the [act name] [date]</td>
<td>Mining Act 1926</td>
</tr>
<tr>
<td>1950 &lt; ( y &lt; 2000 )</td>
<td>meaning, section</td>
<td>same [keyword] as in [keyword] [any number] of the [act name] [year]</td>
<td>Copyright Act 1962</td>
</tr>
<tr>
<td>2000 &lt; ( y )</td>
<td>act</td>
<td>this [keyword] is the [act name] [year]</td>
<td>Social Security Act 2018</td>
</tr>
</tbody>
</table>

Alongside recognising the entities, to extract the network edge information, we suggest a rule-based Relation Extraction approach considering that legislation texts are contextually structured. To identify the rules, we intend to use the same sample set, which we used for the Named Entity Recognition. From the case study, we observed that the style of writing legislation had changed considerably over time, so the sampling approach is critical to minimise the impact of various text styles. By reviewing the sample files, there is a large collection of previously annotated material that can define the rules for relation classifiers.

For the case study, as explained, we reviewed a total number of 55 text files, and we built several classifier rules engaging a set of keywords to identify the relations between the named entities. Table 3.4 summarises the entity relation list for the case study and provides examples. This sug-

Table 3.4: Relations example

<table>
<thead>
<tr>
<th>Relation</th>
<th>Type</th>
<th>Canonicalized text</th>
<th>Sample document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>TIT</td>
<td>the short title of this act shall be the (married women property protection act 1860</td>
<td>Married Women Property Protection Act 1860</td>
</tr>
<tr>
<td>Citation</td>
<td>CIT</td>
<td>within the meaning of section 5 of the companies act 1993</td>
<td>Trade Marks Act 2002</td>
</tr>
<tr>
<td>Amendment</td>
<td>AMD</td>
<td>section 25.1b amended, by section 5.2 of the trade marks amendment act 2005</td>
<td>Trade Marks Act 2002</td>
</tr>
<tr>
<td>Partial Repeal</td>
<td>PRP</td>
<td>section 5(1) repealed, by section 4(8) of the trade marks amendment act 2011</td>
<td>Trade Marks Act 2002</td>
</tr>
<tr>
<td>Repeal</td>
<td>FRP</td>
<td>acts repealed. 1860, No. 9 the married women property protection act, 1860.</td>
<td>Married Women Property Protection Act 1880</td>
</tr>
</tbody>
</table>

gested process can be generalised for any other case study in Legislation Network building pro-
cess considering that legislation texts are coherently structured. This ensures that there is a large collection of previously annotated material that can define the rules for entity recognition and relation classifiers.

3.1.3 Approximate String Matching

Named entity recognition identifies the Acts and relation extraction recognises the relationship between them. So these two steps result in an initial version of the node list and the edge list of the intended Legislation Network. However testing this network shows that the extracted data is unreliable with an average error rate of 12 %\(^2\), so another step is required to resolve typo issues and imperfect entities. This poor-quality data requires an approximate string matching step. To run this step, we require two main components, the technique and the correct pattern. Table 3.5 provides an example which shows the first match as the output of the proposed approximate string matching technique.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Extracted entity</th>
<th>First match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid Model</td>
<td>married womens property protection act 1860</td>
<td>married women property protection act 1860</td>
</tr>
</tbody>
</table>

As mentioned earlier, after the implementation of different approximate string matching techniques, we design and propose a hybrid model of Jaccard and Edit-Distance methods. Algorithm

\[ \text{Positive Predictive Value} = \text{Precision} \]

\[ \text{False Negatve Rate} = 1 - \text{Recall} \]

\[ \text{Edit-Distance} \]

\[ \text{Jaccard Index} \]

\[ \text{Hybrid Model} \]

\[ 87\% 88\% 89\% 90\% 91\% 92\% 93\% 94\% 95\% 96\% 97\% 98\% 99\% 100\% 0\% 2\% 4\% 6\% 8\% 10\% 12\% \]

\[ \text{Figure 3.3: Precision and Recall comparison of the approximate string matching techniques} \]

\(^2\)To estimate this error rate, we used a cluster sampling method to choose ten sets of 30 entities randomly. By manual check of the samples, we observed the rate of incorrectly matched entities.
1 shows the proposed hybrid model, and Figure 3.3 compares the results of the hybrid model with Edit-Distance and Jaccard techniques in terms of precision and recall of the approximate string matching step. To run this comparison, we used a stratified sampling technique with different periods being the groups\(^3\).

**Algorithm 1** Approximate String Matching of Legislation

```plaintext
1: procedure LEGISLATION NAME MATCHING
2:     string1 ← Extracted legislation name
3:     masterlist ← Open Legislation Title Master List
4:     j ← 1
5:     tline ← The first line of masterlist.
6:     GetOut ← 0
7:     while < GetOut ≠ 0 > and < tline ≠ 0 > do
8:         string2 ← tline.
9:         m(j) ← Jaccard(string1, string2)
10:        n(j) ← EditDistance(string1, string2)
11:        if < m(j) = 0.5 > or < n(j) = 0 > then
12:            GetOut
13:        j ← j + 1.
14:        tline ← The next line of masterlist
15:    close;
16:    [x1, I1] ← max(m)
17:    [y1, I2] ← min(n)
18:    if y1 is smaller than or equal to 5 then
19:        match ← I2
20:    else if x1 is bigger than 0 then
21:        match ← I1
22:    close;
```

The graph shows the error rates in each time sample of documents based on the chosen approximate string matching model. For example, for the documents commenced prior to 1850, the first marker point at each graph line shows the false negative error and the precision of the chosen approximate string matching method. As can be seen, samples from this oldest groups of acts show a higher error rate regardless of the approximate string matching method, and Edit-Distance performs slightly better for the old documents comparing to the Jaccard index. In summary, the proposed hybrid model performs significantly better than the other two methods for all documents regardless of their age with less than two percent of false-negative error and average precision of more than 98 percent.

In the case study, the pattern that we used for the approximate string matching step is the list of all NZ Acts provided by NZLII [4]. In case of not having access to such a master list, the typo resolution could be more time-consuming. Approximate string matching considerably improves the quality of the extracted information, result in reliable edge list and node list. Later in this chapter, we explore the evaluation of the final extracted data set and the robustness of the network. The robustness study proves the value of a high performing approximate string matching technique which improves the data quality significantly.

### 3.2 Application

The proposed Information Extraction framework resolves the historic data limitation in previous studies [21][22] and results in a large and reliable dynamic network data set, which is called LegiNet and is available at [69]. This dynamic [70] and the complex network has a very intersecting range of characteristics and behaviours. To maintain the subject consistency of this chapter, a more in-depth analysis of network behaviours are delayed to the future chapters. In this section, we examine general network science characteristics of the case studied network, and present an overall view of the structural and node importance evolution.

Table 3.6 compares the produced network based on the Information Extraction process with the earlier versions of the network that we built in 2016 with parsing of limited available XML resources. As illustrated the network size and structure is significantly changed comparing to its earlier versions.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>Average degree</th>
<th>Average CC$^4$</th>
<th>Average path-length</th>
<th>Network type</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study</td>
<td>16385</td>
<td>137751</td>
<td>8.407</td>
<td>0.216</td>
<td>4.873</td>
<td>dynamic</td>
</tr>
<tr>
<td>Previous studies</td>
<td>3856</td>
<td>33884</td>
<td>8.878</td>
<td>0.39</td>
<td>3.569</td>
<td>one snapshot</td>
</tr>
</tbody>
</table>

#### 3.2.1 Dynamic Legislation Network Structure and the Important Acts

Figure 3.4 and Table 3.7 capture the overall evolution of the Legislation Network in New Zealand from 1267 to the second quarter of 2018. To visualise the data, we used a network force-directed approach. In the layouts in Figure 3.4$^5$, each node is placed depending on their connection to

$^4$The average clustering coefficient (CC) is calculated based on the assumption that the network is directed using the approach that discussed in Chapter 2, Definition 2.2.2

$^5$The average clustering coefficient (CC) is calculated based on the assumption that the network is directed using the approach that discussed in Chapter 2, Definition 2.2.2
Figure 3.4: Overview of the network structure evolution
3.2. Application

the other nodes. As can be seen, references between the Acts first appear in the 1840s, but the data-set visually looks like a graph since the 1850s and it gets denser from the 1870s. As can be seen in Table 3.7 the graphs show some small-world properties from the 1860s with $\sigma > 1$ and small-world property of the graphs is significant from 1970s comparing to 50 random graphs. As illustrated overlay, the network gets denser, and the average degree is growing. We observe more significant clusters during the most recent decades, which can be seen in Figure 3.4. These clusters could be the outcome of housekeeping activities such as edge and node removal, or it could be the result of sophisticated referencing approach in the legal drafting process. Both of the above hypotheses should be examined in future studies.

Based on the network structure information provided in Table 3.7 and Figure 3.4, six different periods are chosen for the centrality evolution analysis. Figure 3.5 captures the time evolution of the top 10 nodes and the most frequent words 7 in the top 20 nodes based on Katz prestige centrality measure as defined in Section 2.3.1.

As mentioned earlier, prior to the 1860s, the graphs don't show significant small-world properties. The visual presentation in Figure 3.4a to Figure 3.4c also reflects that the network can be considered as a random graph during this period. So the Katz centrality degree distribution is nearly a uniform distribution in these periods and is excluded from Figure 3.5.

Figure 3.5a shows the most important nodes with the impression that Land was the most critical law subject back at that time. In the next selected period the network shows small-world prop-

---

**Table 3.7: Overview of the network measures evolution**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>28</td>
<td>148</td>
<td>315</td>
<td>939</td>
<td>1945</td>
<td>4712</td>
<td>5473</td>
<td>6292</td>
<td>8622</td>
<td>11940</td>
<td>15524</td>
<td>16199</td>
</tr>
<tr>
<td>Number of edges</td>
<td>0</td>
<td>12</td>
<td>64</td>
<td>1252</td>
<td>4756</td>
<td>14851</td>
<td>18767</td>
<td>24538</td>
<td>44859</td>
<td>70863</td>
<td>121019</td>
<td>130969</td>
</tr>
<tr>
<td>Average degree</td>
<td>0</td>
<td>0.081</td>
<td>0.203</td>
<td>1.333</td>
<td>2.445</td>
<td>3.152</td>
<td>3.429</td>
<td>3.900</td>
<td>5.203</td>
<td>5.820</td>
<td>7.796</td>
<td>8.085</td>
</tr>
<tr>
<td>Average path length</td>
<td>0</td>
<td>1</td>
<td>1.046</td>
<td>2.605</td>
<td>3.592</td>
<td>8.061</td>
<td>7.514</td>
<td>8.301</td>
<td>6.164</td>
<td>5.554</td>
<td>5.051</td>
<td>4.927</td>
</tr>
<tr>
<td>Directed CC</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.007</td>
<td>0.12</td>
<td>0.13</td>
<td>0.133</td>
<td>0.143</td>
<td>0.161</td>
<td>0.193</td>
<td>0.213</td>
<td>0.212</td>
</tr>
<tr>
<td>Small-world $\sigma$</td>
<td>NA</td>
<td>0</td>
<td>0.447</td>
<td>1.165</td>
<td>15.587</td>
<td>75.173</td>
<td>82.519</td>
<td>80.122</td>
<td>24.239</td>
<td>16.131</td>
<td>195.837</td>
<td>208.084</td>
</tr>
</tbody>
</table>

---

5We used Gephi [71] for these visualisations.

6The small-world $\sigma$ is calculated by comparing clustering coefficient and average path length of each network to 50 equivalent random network with same average degree as defined in Chapter 2, Definition 2.2.4

7To find the frequent words, we used Textalyzer Python module. The frequent prepositions, conjunctions and articles are excluded from the analysis.
3.2. Application

Figure 3.5: Time evolution of the top ten acts and the top subjects in the top twenty acts
erties, and as can be seen in Figure 3.5b the centrality measure shows a higher kurtosis with the word Council being the most frequent topic in the legal domain.

Similarly, the other graphs reflect the change in the network structure and highlight the relationship between the laws and the socio-economic requirements of the country. In the current decade, with the new sets of legislation being introduced and referenced to the older documents, the centrality measure is increased compared to the previous decade, and the hot legal topics show a change which could be a good reflection of the society’s needs.

3.2.2 Legislation Network Communities and Their Application

In this section, we aim to show that the community structure in legislation network gives us insights concerning society. Our approach is to run an appropriate community detection algorithm on the New Zealand dynamic legislation system and interpret the results in the shape of social, political and economic changes.

In Chapter 2, we used Louvain algorithm for directed graphs to detect the communities of the New Zealand current active acts network. The dynamic network that we built based on our proposed framework is more substantial and more complex. In Figure 3.4 we used a similar method to identify the clusters. In this chapter, we tend to review the clustering results of Louvain algorithm for the dynamic graph.

We start with a few definitions of measures and methods that we use in this section. Then we explore the results for the New Zealand dynamic legislation network. Using the results, we then correlate legislation communities with the statistics reflecting the development of the country published by the United Nation Statistics Department [72].

**Definition 3.2.1.** In network science, the average length of the shortest path of a node also called closeness measure, is viewed as the average distance from the node to any other node in the network. Closeness measure describes a node’s ability to access all other nodes of the network independently [73] [74] [75]. A node with low closeness measure is highly dependent on other nodes intermediaries to connect other nodes in the network, so it is more dependent on intermediates to connect to the rest of the network. Consider the directed network \( G(N_t, R_t) \), define a path from \( i \in N_t \) to \( j \in N_t \) as a varying series of nodes and edges, starting with \( i \) and ending with node \( j \). The sum of the number of edges in a path is the length of that path. Consider the minimum length of any paths connecting nodes \( i \) and \( j \) as \( d(i, j) \) which indicates the distance between nodes \( i \) and \( j \). Applying this definition, \( d(i, i) = 0 \) for any \( i \in N_t \), and the closeness measure \( CL_i \) of each node can
be obtained as [76]:

\[ CL_i = \frac{1}{\sum_{j \in N_i} d(i, j)} \]

While studying legislation networks, the closeness measure of each node implies the strength of that part of the law to relate to the whole legislation corpus. Any law with a higher value of closeness measure is expected to have more connection to the entire body of the legislation.

**Definition 3.2.2.** The content similarity is the degree of similarity between two documents or texts, based on their contents or keywords [77] [78]. We use this measure to compare the content of acts in each detected community with the groups of keywords (and their synonyms) presented in Table 3.8. Using this approach, we can label each cluster at each point in time. Assume that \( \xi_{it} \) is the total number of observed keywords in all acts in the detected cluster \( i \) at time \( t \). If \( Z_j \) shows the total number of keywords in the committee subject \( j \), then for each each detected cluster \( i \) at time \( t \) we can define the content similarity ratio \( S_t(i, j) \) as:

\[ S_t(i, j) = \frac{\xi_{it}}{Z_j} \]

**Definition 3.2.3.** A cluster or a community is a set of nodes that share similar features. Different methods of community detection for graphs are distinguished from each other based on their decision on the notion of similarity and the shared features [79].

**Table 3.8:** Initial keywords in 12 subject-specific committees of New Zealand Parliament [80]

<table>
<thead>
<tr>
<th>Social Development and Community</th>
<th>Defence, Security and Innovation</th>
<th>Finance and Expenditure</th>
<th>Science and Education</th>
<th>Administration and Regulatory Affairs</th>
<th>Justice</th>
<th>Primary Production</th>
<th>Transport and Infrastructure</th>
<th>Energy and Environment</th>
<th>Health and Ageing</th>
<th>Māori and Indigenous Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>social development</td>
<td>business development</td>
<td>local policy</td>
<td>Legislative services</td>
<td>education</td>
<td>human rights</td>
<td>agriculture</td>
<td>transport</td>
<td>customs</td>
<td>conservation</td>
<td>health</td>
</tr>
<tr>
<td>social housing</td>
<td>tourism</td>
<td>tourism</td>
<td>Prime Minister</td>
<td>training</td>
<td>justice</td>
<td>housing security</td>
<td>safety</td>
<td>defence</td>
<td>climate change</td>
<td>medicine</td>
</tr>
<tr>
<td>income support</td>
<td>Crown minerals</td>
<td>revenue</td>
<td>State services</td>
<td>employment</td>
<td>courses</td>
<td>racing</td>
<td>infrastructure</td>
<td>disinvestment</td>
<td>climate change</td>
<td>medicine</td>
</tr>
<tr>
<td>women</td>
<td>commerce</td>
<td>banking</td>
<td>statistics</td>
<td>immigration</td>
<td>crime</td>
<td>fisheries</td>
<td>energy</td>
<td>foreign affairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>children</td>
<td>consumer protection</td>
<td>superannuation</td>
<td>internal affairs</td>
<td>industrial relations</td>
<td>police</td>
<td>forestry</td>
<td>building</td>
<td>trade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>young people</td>
<td>research</td>
<td>insurance</td>
<td>social service</td>
<td>health and safety</td>
<td>corrections</td>
<td>land</td>
<td>construction</td>
<td>area's control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>seniors</td>
<td>science</td>
<td>Expenditure</td>
<td>local government</td>
<td>accident compensation</td>
<td>Crown legal services</td>
<td>women's affairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific peoples</td>
<td>innovation</td>
<td>public audit</td>
<td>regulations</td>
<td>Cabinet</td>
<td>criminal law</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other communities</td>
<td>intellectual property</td>
<td>financial performance</td>
<td>legal</td>
<td>and emergency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arts</td>
<td>broadcasting</td>
<td>economic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>culture and heritage</td>
<td>communications</td>
<td>finance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sport and recreation</td>
<td>information technology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voluntary sector</td>
<td>trading standards</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As we explained earlier, in this section, we use the Louvain algorithm for directed graph [37] to identify the communities in New Zealand dynamic legislation network. For this case study, we choose six snapshots to run the community detection algorithm. We chose them based on the dates of the statistics that the United Nation repository provides.
3.2. Application

Table 3.9 shows the results of community detection step. The table also provides the content similarity ratio of the cluster with Max($S_t$) and the parliament committees subjects at each point in time.

**Table 3.9**: Communities and their content similarity with the NZ parliament committees. The values in this table show the reference number of the cluster has the highest content similarity with the committees subjects at each point of time. The value in the parenthesis shows the content similarity percentage of the highly matched cluster with the parliament committees subjects.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Services and Community</td>
<td>0 (0.83)</td>
<td>13 (0.81)</td>
<td>22 (0.73)</td>
<td>11 (0.76)</td>
<td>11 (0.77)</td>
<td>5 (0.81)</td>
</tr>
<tr>
<td>Economic Development, Science and Innovation</td>
<td>8 (0.75)</td>
<td>10 (0.74)</td>
<td>15 (0.72)</td>
<td>17 (0.75)</td>
<td>2 (0.74)</td>
<td>12 (0.70)</td>
</tr>
<tr>
<td>Finance and Expenditure</td>
<td>14 (0.68)</td>
<td>5 (0.70)</td>
<td>18 (0.68)</td>
<td>5 (0.81)</td>
<td>13 (0.69)</td>
<td>6 (0.71)</td>
</tr>
<tr>
<td>Governance and Administration</td>
<td>2 (0.91)</td>
<td>18 (0.90)</td>
<td>0 (0.91)</td>
<td>15 (0.91)</td>
<td>20(0.90)</td>
<td>10 (0.89)</td>
</tr>
<tr>
<td>Education and Workforce</td>
<td>1 (0.88)</td>
<td>3 (0.85)</td>
<td>3 (0.84)</td>
<td>16 (0.85)</td>
<td>17 (0.80)</td>
<td>1(0.84)</td>
</tr>
<tr>
<td>Justice</td>
<td>4 (0.94)</td>
<td>1 (0.91)</td>
<td>16 (0.90)</td>
<td>7 (0.91)</td>
<td>1 (0.92)</td>
<td>2 (0.90)</td>
</tr>
<tr>
<td>Primary Production</td>
<td>11 (0.87)</td>
<td>8 (0.79)</td>
<td>1 (0.79)</td>
<td>4 (0.80)</td>
<td>0 (0.78)</td>
<td>3 (0.79)</td>
</tr>
<tr>
<td>Transport and Infrastructure</td>
<td>17 (0.84)</td>
<td>7 (0.85)</td>
<td>9 (0.80)</td>
<td>6 (0.80)</td>
<td>16 (0.83)</td>
<td>8 (0.82)</td>
</tr>
<tr>
<td>Foreign Affairs, Defence and Trade</td>
<td>15 (0.89)</td>
<td>20 (0.87)</td>
<td>12 (0.91)</td>
<td>3 (0.88)</td>
<td>6 (0.89)</td>
<td>11 (0.89)</td>
</tr>
<tr>
<td>Environment</td>
<td>20 (0.87)</td>
<td>2 (0.89)</td>
<td>13 ( 0.87)</td>
<td>0 (0.92)</td>
<td>5 (0.92)</td>
<td>18 (0.89)</td>
</tr>
<tr>
<td>Health</td>
<td>6 (0.77)</td>
<td>0 (0.75)</td>
<td>17 (0.79)</td>
<td>12 (0.78)</td>
<td>4 (0.78)</td>
<td>14 (0.79)</td>
</tr>
<tr>
<td>Māori</td>
<td>12 (0.92)</td>
<td>4 (0.96)</td>
<td>8 (0.89)</td>
<td>19 (0.90)</td>
<td>14 (0.88)</td>
<td>17 (0.92)</td>
</tr>
</tbody>
</table>

As can be seen, the content similarity of detected communities on average is more than 80 per cent, reflecting that the recognised communities can explain the relationship between legislation with acceptable confidence. Next, we intend to associate the evolution of these significant communities with socio-economic factors. We choose three clusters of Justice, Environment, and Primary Production for this analysis.
Figure 3.6 shows the trend of closeness centrality of the acts in each community and provides an overview of the direction of two socio-economic factors trend based on United Nations statistics for New Zealand.

We explained that in this section, we use the closeness measure of each act to show its inclusiveness and relevance to the whole body of legislation. In Figure 3.6a as we see that the closeness measure of the acts in Justice cluster increases over time.

Considering the decrease in the total number of thefts at the national level, we can confirm the integrity and performance of this community of laws to protect society from burglary. But at the same time, the rising trend of total sexual violence at the national level signifies that the legislative body has not been moving towards protecting people from sexual abuse.

Figure 3.6b suggests interesting results for the environment cluster. The inclusiveness of the acts in this cluster has been increasing recently, and it seems to relate to the decrease in carbon dioxide emission, but not comprehensive enough to protect threatened species of New Zealand.

In Figure 3.6c we can observe a considerable increase in the closeness of acts on Primary Production cluster. The trend might be related to the introduction of the Resource Management Act in 1991, which consolidated isolated legislation together. As can be seen in the graph, there is an increase in both arable and protected lands during the same period that indicates the performance of this cluster.

These insights are a few examples of what network science can suggest for studying legislation networks. The extended similar analysis may help legislators and authorities to improve law drafting processes. In the next section, we aim to explain the importance of data reliability in network science studies, and we claim that our framework to extract the network data from documents is exceptionally robust.

### 3.3 Evaluation and Robustness

In this section, the performance of the proposed framework is discussed. As explained in the previous section, the main goal of the study is to extract the information to build the legislation network. The framework includes Named Entity Recognition, Relation Extraction, and Approximate String Matching jointly to extract the network's node and edge information. In this section, we evaluate the proposed framework and calculate the related errors. The familiar metrics of recall and precision measures are used to evaluate the system. High precision means that the frame-
3.3. Evaluation and Robustness

Figure 3.6: The trend of closeness centrality of the acts in each community
work returns substantially more relevant results than irrelevant ones, while high recall means that the process returns most of the relevant results. At the end of this section, we assess the impact of the identified errors on the network structure and examine the robustness.

### 3.3.1 Error Estimation, Precision, and Recall

In the proposed Information Extraction process, Named Entity Recognition is combined with Approximate String Matching to recognise, validate and optimise the entities (the nodes and the edges). Figure 3.7 illustrates the occurrence of the false-positive and false-negative errors in this process and helps in studying the robustness of the network.

**Figure 3.7: Error Diagram**
If the Entity Recognition process finds the entities, then there is a possibility that the Approximate String Matching process fails to find the right match. A type I error $\alpha_1$ occurs when the approximate string matching process fails to find the right match. From one side, this issue contributes to the false-positive error because it adds invalid entities to the output. These invalid entities impact the accuracy of the node list and the edge list of Legislation Network. To estimate the $\bar{\alpha}_1$ in the case study, a cluster sampling method is used to choose ten sets of 30 entities randomly. By manual check of the samples, the rate of incorrectly matched objects is observed. A Kolmogorov-Smirnov test suggests that the estimated error of $\bar{\alpha}_1$ has a normal distribution with the parameters in Table 3.10.

$\beta_1$ also occurs when the approximate string matching system picks a wrong match for the entities. This issue can contribute to the false-negative error because those entities that are wrongly matched to other entities are missing from the data set. The estimation methodology and the estimated value of $\bar{\beta}_1$ are equal to that of $\bar{\alpha}_1$ as indicated in Table 3.10.

$\beta_2$ is measuring the Information Extraction rules’ performance. If it fails to recognise entities, then those entities are missed, and it results in another type of false-negative error. The estimation process for $\bar{\beta}_2$ is different from the previous two errors, and it is harder to address. For the case study, a sample set of 30 text files are randomly chosen using cluster sampling method. Then all of the extracted entities for each document is compared to the actual entities in a human involving process. The list of missing entities is categorised into two parts: caused by a typo, or caused by insufficient rules to recognize the entity. The rate of missing entities caused by weak or missing rules is calculated for each document and denoted by $\bar{\beta}_2$. The Kolmogorov-Smirnov results through all of the 30 documents show that $\bar{\beta}_2$ has a normal distribution with the parameters in Table 3.10.

### Table 3.10: Errors, sensitivity and specificity

<table>
<thead>
<tr>
<th>Measure</th>
<th>$\bar{\alpha}_1$</th>
<th>$\bar{\beta}_1$</th>
<th>$\bar{\beta}_2$</th>
<th>$\bar{\beta}_3$</th>
<th>$\bar{\alpha}$</th>
<th>$\bar{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.0160</td>
<td>0.0160</td>
<td>0.0012</td>
<td>0.0007</td>
<td>0.0160</td>
<td>0.0179</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0012</td>
<td>0.0012</td>
</tr>
<tr>
<td>KS significance</td>
<td>0.9345</td>
<td>0.7156</td>
<td>0.8756</td>
<td>0.8435</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

95% confidence interval

(0.0156, 0.0164) (0.0175, 0.0183)
3.3. Evaluation and Robustness

\( \beta_3 \) addresses the error when typos cause a problem for recognising the entities. The estimation process is very similar to that of \( \beta_2 \). A sample of 30 text files is collected. Then the rate of missing entities caused by OCR typos is calculated for each document and addressed as \( \hat{\beta}_3 \). The Kolmogorov-Smirnov results through the selected documents show that \( \hat{\beta}_3 \) has a normal distribution with the parameters in Table 3.10. In the sample, it is observed that the typos that cause the entity recognition failure are only numeric typos. For example, OCR might produce an error and convert 1987 to l987 by misspelling number 1 to letter l. Then the Information Extraction rules are impacted to recognise l987 as a year, so the entity is missed.

As Figure 3.7 shows \( \alpha_1 \) is the only false positive error which contributes to the \textbf{overall false positive error} of the system. Table 3.10 captures \( \bar{\alpha} \), assuming that \( \bar{\alpha} \) estimates the overall type one error. To estimate the \textbf{overall false negative error} of the system, \( \beta_1, \beta_2, \) and \( \beta_3 \) are considered as mutually exclusive events. From Figure 3.7 it is also clear that the intersections of each two of these errors are empty, so they are independent. Table 3.10 shows \( \hat{\beta} \), the estimated value for the overall false negative, or the type II error.

To calculate the Precision and Recall, Equation 3.1 and Equation 3.2 are used.

\[
\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} = \frac{1 - \bar{\alpha} - \hat{\beta}}{1 - \hat{\beta}} \tag{3.1}
\]

\[
\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} = \frac{1 - \bar{\alpha} - \hat{\beta}}{1 - \bar{\alpha}} \tag{3.2}
\]

Referring to the above equations and Table 3.10, we observed the Precision of 98.37\% and Recall of 98.18\%. These outcomes suggest the high performance of the Proposed Information Extraction framework that results in high data reliability of the output Legislation Network.

As explained in section 3.1.3, the proposed hybrid Approximate String Matching technique substantially reduces the errors. It is essential to mention that at the earlier stages of the study by using the classic string matching techniques, the error rates were considerably higher, and the accuracy of the network was questionable. A time-consuming examination process engaging manual checks was applied to propose the hybrid model, which resulted in an impressive performance and high precision and recall. The improvement involved a lot of efforts and time but resulted in accuracy and confidence in Legislation Network studies.
3.3. Evaluation and Robustness

3.3.2 Robustness

With a coherent understanding of the errors, it is essential to study the robustness of the network to the error. The robustness study proves the importance of the data accuracy, which supports the value of the proposed hybrid model for the approximate string matching. In this section to study the network robustness, we use diameter and three major centrality measures.

To understand the diameter robustness of the network, attack and failure analysis is required. As discussed earlier, the legislation network in general shows small-world characteristics. So it is expected to observe a reasonable error tolerance of the network as the result of random failures, but vulnerability as the result of attacks\[81\].

To study the network robustness to node failures, the method is to randomly remove a fraction of nodes \( f \) and recalculate the diameter of the network \( d \).

To study the network robustness to attack by removing a fraction \( f \) of the largest nodes\[8\] and observe the change to the diameter \( d \). The results of both failure and attack to the nodes are captured in Figure 3.8.

The observed tolerance to failures and the vulnerability to attacks shows that the connectivity is provided by a few highly connected nodes, and the majority of nodes have only a few edges. As can be seen the vulnerability to the attacks starts immediately after removing a small fraction \( f = 0.3\% \) of the highly connected nodes. This scenario of attack is highly unlikely in the Legislation Network considering the high Precision and Recall of the proposed data extraction process.

\[8\] Based on their connectivity (total degree)
As discussed in the previous studies [21] [22], the most relevant centrality measure for the Legislation Network is the Katz second prestige measure. In recent studies, the reliability of different centrality measurements against network manipulation has been addressed [82] [83], but Katz prestige centrality is not much discussed. In this chapter, we study the Katz centrality, betweenness centrality, and degree centrality robustness of legislation network against edge deletion error.

To address the robustness, four major measures of accuracy that proposed in [82] and [83] are used. These measures are Top 1, Top 3, Top 10 per cent, and the Pearson correlation to compare the centrality measures between the actual network and the manipulated network.

![Robustness of the top nodes as the function of fraction of manipulated edges](image)

**Figure 3.9:** Robustness of the top nodes as the function of fraction of manipulated edges $f$

We consider the error level as a specific percentage value from the set of 1%, 5%, 10%, 20%, that is relative to the number of manipulated edges from the actual original network. Figure 3.9 shows the results of the different Centrality measure as the function of the fraction of manipulated edges $f$. For each fraction level, the test is repeated for 100 times, and the graphs show the average of all sampled sets. Table 3.11 shows the Pearson correlation between the nodes central-
ity in the manipulated network and the original network when 10% of the edges are randomly deleted.

**Table 3.11:** Node centrality Pearson correlation between the manipulated network and the original network

<table>
<thead>
<tr>
<th>Measure</th>
<th>Katz prestige centrality</th>
<th>Betweenness centrality</th>
<th>Degree centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significance (p-value)</td>
<td>2.2(\times)10^{-16}</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.939</td>
<td>0.948</td>
<td>0.67</td>
</tr>
</tbody>
</table>

The pattern and level of robustness of the three selected centrality measures considered in this paper are not as similar, as suggested in [82]. In-degree centrality shows more fragility comparing to betweenness and Katz measures. This difference could be related to the network topology, as suggested by [83]. The results also confirm the findings of [83] [82] that accuracy declines monotonically with an increase in error.

As can be seen in the graph, the Katz centrality is relatively robust to the edge deletion when less than 20 per cent of the network structure is touched. The graphs indicate a moderate fragility when the network structure is hugely manipulated. For example, the removal of 20 per cent of the edges somehow impacts the in-degree centrality. However, in more than 90 per cent of these extreme samples, the top 1 node in the manipulated network is a member of the top ten per cent of nodes in the original graph.

The results imply that centrality measures on Legislation Network are quite robust under small amounts of error (such as 5 per cent or under) and to some extent fragile under bigger data errors. So the reliability of the network information is essential for in-depth network studies. As explained earlier the precision and the recall of the proposed Information Extraction process is above 98 per cent, so it is reasonable to compute the centrality measures when studying the Legislation Network.

### 3.4 Discussion

In this chapter, we focused on time as a fundamental attribute in understanding and analysing legislation. The legislation network concept has been discussed in recent years, but the importance of having access to historic laws was never discussed much. The chapter underlined the value of studying legislation as dynamic networks and proposed a new Information Extraction process to achieve a highly accurate Legislation Network. The performance of the data extraction framework is examined, as compared to the previous studies and proved to be considerably high.
This work contributed to the literature of network Information Extraction from old documents and insisted on the value and applications of the dynamic legislation network. The proposed process can be used not only in the legal domain but also in various research areas involving documented knowledge, facts, and cases.

Analysing a dynamic legislation network is a novel approach to understand the underlying process behind the generation of the laws and the evolution behaviour of such systems. This subject is fascinating but mathematically complicated. So it will be discussed in the next two chapters.
Chapter 4

Evaluation of Existing Network Generative Models for the Legislation Network

In the previous chapter we proposed a framework to develop dynamic networks of legislation. In this chapter we are interested in studying the time evolution of such networks. We aim to classify any of the existing models of network evolution that help to explain the generative function of legislation network. Traditional generative models for graphs represent a particular family of graphs and do not learn the generative model from the observed data. For example Price’s cumulative advantage model [84], exponential random graphs [85], Barabási-Albert’s preferential attachment model [37], stochastic Kronecker model [86], and stochastic block models [87] are all designed to test certain graph properties, such as degree distribution or diameter against a proposed model. Our approach in this chapter is different as we challenge ourselves to learn generative patterns from the observed data of graphs.

In recent decades, the topology of networks has been studied by modern graph theory to emphasis on the applications of network science [88]. These modern explorations first developed by Erdős who produced a strong network of collaboration in the 1950s. Later he also developed a systematic way of generating artificial networks. As we defined it in Definition 2.2.3, the Erdős-Rényi model [29], was the first method for generating random graphs with Poisson degree distribution [30], networks in which properties such as the number of nodes, edges, and degree distribution are set in some random way. The Erdős-Rényi model is still used because of its simplicity, for example in the context of diffusion and spread [89].

Perhaps the next impressive progress in the history of modern network science was the introduction of small-world networks defined by the Watts-Strogatz model [28]. The motivation of this
model was to show that many real-world networks do not follow the Erdős-Rényi model [17]. The concept of small-world graphs differentiated real-world networks from random graphs by considering local clustering that we explained in Definition 2.2.2 and short path length that we discussed in Definition 2.2.1 between nodes [90].

The study of small-world networks, which is about finding patterns and orders in clusters of chaos, opened up the idea of studying network scale and degree distribution. The revolution of network science with the focus on dynamic evolution rather than topological structure started from studying scale-free networks [91]. The concept of network evolution or generative models later extended in the shape of broad-scale and single-scale models to distinguish and explain the dynamic behaviour of many real-world graphs [20][92]. Changing the study approach from topology to statistical mechanics of networks revealed that the degree distribution of real-world graphs significantly deviates from a Poisson distribution as observed in random graphs [93].

In this chapter we aim to examine generative models that may represent the dynamic evolution behaviour of legislation network based on the observed data. In the literature of generative graph models, there are many different methods and approaches each having their advantages and limitations. The underlying assumption behind studying the network evolution in this piece of work is that if the processes that constructed the network are correctly captured, then network topology evolution can be obtained.

The chapter starts with an overview of existing models in Section 4.1. Then in Section 4.2 to understand growing behaviours of Legislation Networks, a few sets of descriptive-statistics analysis of their evolution processes are captured. After evaluating the observed models in Section 4.3, the chapter ends by discussing the most appropriate models, and their limitations in presenting legislation networks growing processes.

### 4.1 Definitions and Methodologies

In this section, we present the required terminology and background that is used throughout the chapter. We give the explanations for network structure theoretics that describe various small-world networks dynamics and evolution. We also provide examples of real-world graphs to understand the applications of the proposed models. Many real-world network examples are discussed in Section 4.1.2 including generative modeling of scientific citation networks, World-Wide Web, biological networks, and networks of social interactions.
4.1.1 Network Structure, Dynamics, and Generative Models

Since the introduction of the small-world [28] and scale-free [91] properties, generative models for complex networks have become a more interesting subject. Recently three classes of small-world graphs have been identified [20]. Scale-free networks are the most popular one in the literature, whose degree distribution follows a power law regime [37]. In many real systems, there are different constraints which limit the development of connected nodes in a network. These constraints may result in a cutoff of the power law connectivity distribution or make it completely change to a fast decaying tail, such as exponential or Gaussian [94]. The section discusses all these possible scenarios of network structure evolution in detail.

4.1.1.1 Small-world Networks

A commonly argued attribute of real-world networks is the small-world effect [31] in which a small number of moves between nodes is required to connect any two nodes in the graph. In 1998, Watts and Strogatz proposed a model to generate graphs with small average shortest-path lengths and large clustering coefficients. They also called any network with these characteristics a small-world network [28]. In general, it discovers how tightly attached a network is.

![Figure 4.1](image-url): An Erdős-Rényi random graph and a small-world legislation network with the same size.
Figure 4.1 visually compares the structure of Erdős-Rényi random graph and a small-world graph with both having the same size.

The presence or absence of the small world property is a metric which one may use to categorise observed networks and to explain their behaviour. There are various ways to measure the small world property, which have somewhat different significance. Dorogovtsev and Mendes tried found the distribution of path length to examine the small-world effect of graphs [95]. Latora and Marchiori introduced the concept of efficiency and claimed that small-world networks are both globally and locally efficient [96]. Klemm and Eguíluz proposed that the logarithmic increase of the shortest path length with system size is the main characteristic of the small-world effect [97].

The most common and robust measure is the small-world $\sigma$ that we presented in Definition 2.2.4. This definition is based on the trade-off between high local clustering and a short average path length and suggests the small-world effect existence for the values of $\sigma$ significantly larger than one [32]. Earlier in Table 3.7 we used the same approach and observed the small-world effect for New Zealand legislation network in various periods.

4.1.1.2 Scale-free Networks and Preferential Attachment

**Scale-free** networks developed to represent the process of network growing in which new nodes connect preferentially to the more highly connected nodes [17].

**Preferential attachment**, also known as *rich get richer effect*, describes a process for some variable such as wealth in which is distributed among some individuals according to how much they already have. It means that the individuals who are already “rich” in that variable receive more than those who are not [98].

In network science such a definition results in a process in which nodes with high degree are more likely than others to gain extra edges as nodes join the network. The application of a preferential attachment process in network growth modeling can lead to a scale-free degree distribution in which some nodes obtain a very large number of edges while others remain relatively unconnected.

**Definition 4.1.1.** The preferential attachment model proposed by Barabási-Albert is based two main principles [93]:

- **Growth**: starting with a small number of nodes, $k_0$, at each time step, a new node with exactly $k_0$ edges is added to the graph.
• **Preferential attachment**: the new node connects to node $i$ with probability proportional to the degree of node $i$.

**Remark 4.1.1.** This definition results in a nonuniform probability by which a new node attaches to an existing node of the network and that probability increases with the number of connections of that node. The Barabási-Albert preferential attachment model results in a **power-law** in-degree distribution with shape parameter $\alpha = 2$.

**Definition 4.1.2.** [99] A continuous random variable $K$ is assumed to have a **power-law distribution** with shape parameter $\alpha$ and scale parameter $k_0$, if the probability density function of $k$ is given by:

$$f(x) = \begin{cases} \frac{\alpha k_0^\alpha}{x^{\alpha+1}} & k \geq k_0 \\ 0 & k < k_0. \end{cases}$$

Thus the Barabási-Albert preferential attachment model results in a degree distribution that decays in a power-law form and leads to a scale-free network structure [93]. Their model predicts that the degree distribution of the network is independent of time, and subsequently independent of the system size, indicating that, despite the networks continuous growth, it reaches a stable scale-free state [37].

In summary, scale-free networks generated by the Barabási-Albert process have the following behaviours:

• Their in-degree distribution follows a power-law regime.

• Their out-degree is constant across nodes and time.

• Universally for all individual nodes, their in-degree grows sub-linearly with $k_i(t) = k_0(t_i/t)^{0.5}$, where $k_i(t)$ is the in-degree of node $i$ at time $t$ and $t_i$ is the time of birth of node $i$ [100].

Further studies challenged the relevance of such properties by testing a significant number of real-world networks. Broido and Clauset [101] observed that out of nearly a thousand real-world networks, only 4% are scale-free. Krapivsky et al. explained that the out-degree of nodes follows either a power-law distribution or an exponential distribution in many networks [18]. Amaral et al. observed exponential and Gaussian distributions for the nodes out-degree [20], and Leskovec et al. concluded similar results [19].

Discussions around the **broad-scale** and **single-scale** networks answered some of these discrepancies around the scale of the network, but none of them provided a transparent mathematical model to explain the behaviours. Also, the out-degree distribution challenge remains unan-
answered and not many researchers were interested in studying the in-degree growth of the individual nodes mainly because of the data availability of complex dynamic networks.

Later in Section 4.2, we provide descriptive statistics of legislation network around these behaviours. Then in Section 4.3, we test the existence of scale-free structure and power-law degree distribution in legislation networks. In the next section we review the existing ideas around the broad-scale and single-scale network structures.

4.1.1.3 Broad-scale Networks and Single-scale Networks

Broad-scale networks are defined by a power-law degree distribution followed by a sharp cutoff [20]. In the literature of network science, there are many real-world networks that do not follow a power law regime for all degree size, specifically for nodes with large degrees. The degree distribution of broad-scale networks is proposed to be estimated as a power-law distribution with exponential truncation for large degrees [20] [102].

These networks are called broad-scale networks as they are more homogeneous than scale-free networks. The observed exponential truncation in power-law behaviour of broad-scale networks is sometimes justified by the small size of networks [103], or high network robustness [104]. But it is mainly explained through processes such as the addition of edges limited by node aging [20], connection costs for nodes with very large degree [20], and content filtering [94] [105]. These processes suggest that the changes of node characteristics through the network evolution, restrict the preferential attachment mechanism in broad-scale networks.

Pieces of evidence in the literature imply that in significant scenarios of processes resulting in cost and content constraints, the power-law degree distribution and the scale-free characteristics of the network might completely disappear. These networks are called single-scale networks and defined by a degree distribution with a fast decaying tail, such as exponential or Gaussian [94]. In this structure most individual nodes have a similar number of connections, leading to a narrow degree distribution [92].

To understand the impact of the identified processes of node aging, cost and content on network dynamics, in Section 4.2, we examine their correlation with the attachment probability and summarise the results in Table 4.5. Then in Section 4.3, while we test the existence of scale-free behaviour, we also explore if the topology of legislation network tends to a broad-scale or single-scale practice.
4.1.2 Real-world Network Evolution: Examples

Many methods and mechanisms suggested different network evolutionary processes for various types of networks. Among them, the creation of functionally in protein networks [106], self-organization processes in World Wide Web [91], and preferential attachment in scientific citation networks [107] have been shown to result in scale-free networks, while individual capacity in social ties [108], evolutionary drift [109], and edge acceleration by age social networks [19] have been explained to contradict the power-law.

To contribute to the concept of network evolution, we require to understand the assumptions that resulted in these contradictory models. In this section, we aim to review and understand the proposed generative models to describe the evolution behaviour of popular real-world graphs. By the end of this section, we understand whether the existing models can universally explain the generative models of real-world networks evolution.

4.1.2.1 Scientific Citation Networks

Citation networks generated from scientific papers are significant research objects of network science; In these networks, each node represents a paper, and each edge represents a citation from one document to another. Studying and analysing these networks provides insights on understanding the evolution and spread of academic findings [110]. Empirical studies of citation networks evolution explained the dynamical pattern of citation accumulation as the mechanism of cumulative advantage proposed by Price in 1976, and preferential attachment offered by Barabási-Albert in 1999, all suggested the scale-free nature of these networks [17].

In 2000, Dorogovtsev and Mendes presented a method to determine the broad-scale behaviour of scientific citation network of high energy physics papers as the result of node aging [111]. The aging phenomenon of citation behavior has been explored more by other researchers [112, 113, 107]. The examples of in-degree distributions of scientific citation networks studied up to this point all modelled power-law with fixed exponents. In 2010, Peterson et al. proposed a model to address the fixed power-law parameter problem by suggesting a combination of one direct process and one indirect method. The first relates to a new paper citing an old article randomly, and the second relates to an indirect mechanism of copying edges from the existing nodes. Their proposed model results in a power-law degree distribution with varying exponent [114].

Later, Xie et al. proposed a directed geometric graph model to explain citation networks with exponential growth of nodes, non-scale-free structure, and geometric zones caused by aging [115]. So they could explain the power-law tails of in-degree distributions of citation networks by the
in-homogeneous influences of papers. Their work illustrated an essential constraint in citations, content relativity of papers, which resulted in broad-scale network design [115]. As explained for scientific citation networks, there is strong evidence of preferential attachment and scale-free shape. But recent studies imply that if those mechanisms apply more broadly, they are considered restricted or even controlled by domain-specific processes [101].

4.1.2.2 World Wide Web

The evolution of World Wide Web is discussed in several studies using different sub-graphs of WWW, and its incoming and outgoing link distributions are proposed to follow a power-law distribution which indicates the scale-free nature of these networks [100, 91]. Stochastic dynamical growth approach is also used to model World Wide Web generative models. All of these models assume the network growth based on different appearance times of sites and different growth rates of sites [116, 100]. World Wide Web communities provide a clear view of consistent development of the WWW structure [117].

Broder et al. suggested that in the World Wide Web, the number of links grows faster than the number predicted by the power-law distribution [118]. In 2002 another model introduced the impact of out-going edges evolution but again the model resulted in a scale-free power-law distribution of in-degree [119]. The majority of network evolution studies that focused on the World Wide Web suggested a scale-free behaviour of these networks. But Amaral and Ottino in 2003 proposed that specific sub-graphs of WWW show content constraints that might result in a broad-scale structure [94].

4.1.2.3 Social Networks

Generative models of network evolution is a controversial subject for social networks. On the one hand, researchers supported the assumption that explaining social networks with preferential attachment mechanism is a challenging subject [120]. In 2001, Jin et al. proposed a model to distinguish evolution behaviour of social networks from WWW and citation networks. While the evolution of the web appears to follow the preferential attachment process, their proposed model identified three methods for friendship and edge addition that separated the evolution process for social networks significantly from a power-law distribution [121].

Similarly, Newman inferred the degree distribution of co-authorship networks of scientists, follows a broad-scale model with exponential cutoff instead of universal power-law [122]. Some later studies supported the assumption that explaining social networks with preferential attachment mechanism is a challenging subject [120].
On the other hand, other researchers accept the scale-free structure of social networks. Albert et al. in 2002 implied that collaboration networks are scale-free and that the network evolution explained by preferential attachment [93]. In 2007, Mislove et al. presented the power-law, small-world, and scale-free properties of four popular online social networks [123]. Broido and Clauset in their very recent studies in 2019 suggest that half of social networks lack any direct or indirect evidence of scale-free structure and insisted on the need to more in-depth stochastic investigation in network evolution area [101].

4.2 Descriptive Statistics of Legislation Network Evolution

Explaining the network topology helps to understand the design principles of networks and therefore provide some evidence into the dynamical evolutionary processes that generated them. To understand these processes, many researchers discussed evolutionary models that lead to the generation of scale-free networks. Such models undoubtedly are oversimplifications of reality. But the underlying assumptions of these models are valuable, as they can contribute to our knowledge of network evolution.

This section explores information from legislation network topology using stochastic tools to explain the underlying processes behind the network evolution. Before we present the result, we define a method in Definition 4.2.1 to study how often the observed data is above a particular estimated distribution.

**Definition 4.2.1.** Complementary cumulative distribution function shows how often the random variable $X$ is above a particular level. If $F_X(x)$ is the cumulative distribution function of variable $X$, then the complementary cumulative distribution function is $P(X > x) = 1 - F_X(x)$ [124].

Figure 4.2 shows examples of complementary cumulative distribution functions for different distributions in log-log scale.

In statistics, a complementary cumulative distribution graph in log-log scale is often used for the testing of a hypothesis that the observed data fit a theoretical distribution.

It is essential to understand the processes of node and edge addition, and the impact of node attributes such as aging, cost and content effects. In this section, we consolidate these scenarios by applying descriptive statistics.
4.2 Descriptive Statistics of Legislation Network Evolution

4.2.1 Legislation Network Size

New Zealand legislation system as a dynamic network incorporates a timeline from 1267 with Statute of Marlborough being the first authorised law in the country. The English Laws Act 1858 declared that English acts were effective in NZ from 14 January 1840. This act ceased to operate in 2008 following the introduction of the Property Law Act. The timescale from the 1300s to the 2010s appears to be a wide range, but as illustrated in Figure 4.3 and also discussed in Figure 3.4, English law was applied retrospectively, so the legislation network started to shape and grow from somewhere between the 1840s and the 1850s.

Apart from the apparent increase in the number of nodes in Figure 4.3a and edges in Figure 4.3b since the 1840s, Figure 4.3c suggests notable growth in the average degree of nodes meaning the network also gets denser. This increase in the average degree could relate to the change in the style of the legislation in terms of the need for more cross-referencing, or it might be an indication of growth in document length. The significant increase in the clustering coefficient in Figure 4.3d indicates the growth of small-world characteristics of the legislation network. The graph also highlights the distinction of the network from random graphs in various periods.

To have a closer look at the growth dynamics of nodes, Figure 4.4 provides details of two separate curve fittings. The coefficient of determination in Figure 4.4b is considerably higher than the one in Figure 4.4a suggesting the better result of a linear growth than an exponential growth for the
4.2. Descriptive Statistics of Legislation Network Evolution

(a) Number of nodes  
(b) Number of edges

(c) Average Degree  
(d) Directed graph clustering coefficient

Figure 4.3: Legislation Network growth statistics

Despite the observed linear node addition process, the results in Figure 4.5 implies the superiority of exponential curve fittings for the number of edges. As can be seen, the coefficient of
determination in Figure 4.5a is 99.65 % which is significantly higher than 88.24 % in Figure 4.5b, suggesting the better result of an exponential growth than a leaner growth for the number of edges.

The exponential edge growth of linearly growing nodes reflects that acts in average are getting longer and more abundant in terms of their connection to the rest of legislation corpus. This scaling behaviour of legislation network suggests a change in society that reinforced the density of laws. To be explicit, the more the network grows, the new nodes bring more edges to the system meaning that for citation or amendment purposes the new laws require to refer to a more significant portion of existing legislation. This idea is more explored in Section 4.2.2 by studying the average out-degree of new nodes.

Figure 4.5: Edge growth model
4.2. Descriptive Statistics of Legislation Network Evolution

4.2.2 Legislation Network Out-degree of New Nodes

As both Figure 4.6a and Figure 4.6b illustrate, there is no meaningful growth model for the out-degree of new nodes with both coefficients of determination being less than 80%.

Next, we examine the tendency of each new node to add new edges to the network. This information is essential in terms of understanding the growth behaviour of the system. For example, in the original Barabási-Albert preferential attachment model, the number of edges introduced at each time step is considered to be a fixed value for each node and at all periods [100]. In the later extensions of that model to include the impact of age [95], capacity cost [20], and content relevancy [94], the out-degree of the new nodes also was considered to be fixed. In this section, using the observed data, we examine whether this assumption is valid for legislation networks.

(a) Exponential fitting

(b) Linear fitting

(c) Distribution fitting. The vertical dotted line reflects the assumption of having a fixed number of newly introduced edges.

Figure 4.6: Average out-degree of new nodes
Figure 4.6c shows the log-log complementary distribution function of the observed number of edges introduced at each time step for our studied legislation network. The hypothesis that we want to test is:

- $H_0$: The observed number of edges introduced at each time step is constant, meaning the data fits any parallel line to the dotted line in Figure 4.6c.
- $H_1$: The observed number of edges introduced at each time step follows a probability distribution.

The results suggest that for legislation network, it is a very unrealistic assumption to consider a fixed value for the out-degree of the new nodes in the generative modelling of network evolution.

But Figure 4.6c shows that the out-degree of new nodes is a random variable and a memory-less time-independent exponential distribution can explain its behaviour. The results suggest that the process of generative modelling of network evolution requires a more appropriate assumption around the out-degree of new nodes. Thus we meet this requirement in Chapter 5 as one of the main contributions of the proposed generative model.

### 4.2.3 Legislation Network In-degree

In this section, we aim to understand the individual node in-degree growth behaviour. Figure 4.7 shows the in-degree of nodes changing pattern of the individual nodes. Each line in the graph represents a node.

At first glance, there are two outliers with short lives and extremely high in-degree. The first one is the *Acts and Regulations Publication Act 1989*, and the second one is *Legislation Act 2012*. The context of both acts suggest that they joined the network for a specific reason of major refining parts of legislation, they got mentioned by many acts because of their modified role, and then both acts were expired or repealed after they achieved their mission.

A similar graph of real-world systems rarely exists in the literature due to lack of data. But from the preferential attachment model, as we explained in Section 4.1.1.2, we can consider a sub-linear growth for nodes in-degree. For legislation network that we built using our proposed framework in Chapter 3, all the temporal node information is available to run such detailed analysis. The observation in Figure 4.7 indicates a common growth pattern in the in-degree of individual nodes.
There seems to be a common practice of increase and then stability for the laws in terms of their in-degree. The observed pattern of increase, then stability is substantially different from what we expect to see from a preferential attachment process. This behaviour might be an indication of a barrier for preferential attachment such as aging, capacity, or content as we introduced in Section 4.1.1.3. We will investigate this assumption further in Section 4.2.6, Section 4.2.7, and Section 4.2.8.

### 4.2.4 Legislation Network Small-world Properties Evolution

We defined the small-world effect of networks in Definition 2.2.4. As discussed in Section 3.2, Table 3.7, the studied legislation network has small-world properties from the 1860s with $\sigma > 1$. It is also discussed that the small-world property of the graphs is significant from the 1970s comparing to 50 random graphs.

In Section 4.2.5 to Section 4.2.8, we run statistical analysis to understand the influence of preferential attachment, age, capacity cost, and content relevancy mechanisms on the evolution.
4.2. Descriptive Statistics of Legislation Network Evolution

behaviour of the legislation network. To design such tests, we require define a variable to describe the change in in-degree of nodes at each time step.

**Definition 4.2.2.** Assume that the $k_i(t)$ is the in-degree of node $i$ at time $t$. We define the variable attachment story, $\delta_i(t)$ for node $i$ at time $t$ as:

$$\delta_i(t) = \begin{cases} 1 & k_i(t) > k_i(t-1) \\ 0 & k_i(t) = k_i(t-1). \end{cases}$$

Thus we use the attachment story of nodes to describe the change in their in-degree at each time step.

In the rest of Section 4.2, to provide enough descriptive statistics of network evolution and avoid replication of similar results, we select five periods.

4.2.5 Legislation Network Evolution and Preferential Attachment

In this section, we assess the existence of preferential attachment process on Legislation Network evolution. The goal is to test the hypothesis that node in-degree at each time results in an advantage for its in-degree increase. Assume that $\overline{k}_{\delta=1}$ is the average in-degree of nodes with $\delta(t) = 1$ meaning that their in-degree increases from time $t-1$ to time $t$. Also assume $\overline{k}_{\delta=0}$ is the average in-degree of nodes with $\delta(t) = 0$. The hypothesis that we want to test is:

$$\begin{cases} H_0 : \overline{k}_{\delta=1} = \overline{k}_{\delta=0} \\ H_1 : \overline{k}_{\delta=1} \neq \overline{k}_{\delta=0}. \end{cases}$$

To test this hypothesis, we require to use a statistic to compare the means of two independent samples, so we use Student's $t$ statistic.

Figure 4.8 shows the results of the test visually in a boxplot and Table 4.1 provides the $t$-test statistics to compare the $k_{\delta=1}$ with $k_{\delta=0}$ for all nodes. According to the results, for all the selected periods the in-degree of nodes that gained new incoming edges at time $t$ is significantly higher than the in-degree of nodes that didn't receive any new incoming edge. As an example, in the most recent period, the in-degree of nodes that gained new incoming edges is on average, 34.47 more than that of nodes that didn't receive any new in-degree.

4.2. Descriptive Statistics of Legislation Network Evolution

Figure 4.8: The influence of preferential attachment on New Zealand legislation network for the selected periods. The boxplot compares the average in-degree of nodes based on their attachment story (1 if the node received at least a new edge, and 0 if no new nodes attached to the node at each time stamp (year).

So in the studied legislation network, we observed a process of preferential attachment. But based on the growth-stability pattern of the individual nodes’ in-degree that we explored in Section 4.2.3, we don’t expect to see a pure scale-free power-law degree distribution. As we dis-
4.2. Descriptive Statistics of Legislation Network Evolution

Table 4.1: t-test result - preferential attachment

<table>
<thead>
<tr>
<th></th>
<th>H₀: ( \bar{k}<em>{\delta=1} = \bar{k}</em>{\delta=0} )</th>
<th>p-value</th>
<th>t</th>
<th>df</th>
<th>( \mu_{diff} )</th>
<th>( \sigma_{diff} )</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>before 1850</td>
<td>1E-4</td>
<td>11.88</td>
<td>144</td>
<td>5.19</td>
<td>0.43</td>
<td>4.32 , 6.04</td>
<td></td>
</tr>
<tr>
<td>1850 - 1900</td>
<td>1E-3</td>
<td>10.71</td>
<td>3598</td>
<td>3.30</td>
<td>0.31</td>
<td>2.70 , 3.901</td>
<td></td>
</tr>
<tr>
<td>1900 - 1950</td>
<td>5E-4</td>
<td>26.33</td>
<td>7293</td>
<td>11.18</td>
<td>0.42</td>
<td>10.35 , 12.01</td>
<td></td>
</tr>
<tr>
<td>1950 - 2000</td>
<td>1E-3</td>
<td>26.30</td>
<td>12889</td>
<td>12.54</td>
<td>0.48</td>
<td>11.60 , 13.48</td>
<td></td>
</tr>
<tr>
<td>2000 - current</td>
<td>2E-4</td>
<td>44.85</td>
<td>15534</td>
<td>34.47</td>
<td>0.77</td>
<td>32.96 , 35.98</td>
<td></td>
</tr>
</tbody>
</table>

As discussed in Section 4.1.1.2, the degree of individual nodes in a scale-free network always growth sub-linearly. Thus we require to explore more factors that might have resulted in limiting the preferential attachment process as we will discuss in the next three sections.

4.2.6 Legislation Network Evolution and Aging Constraint

This section evaluates the impact of node aging on Legislation Network evolution. The goal is to test the hypothesis that node aging results in a constraint for its in-degree increase. As explained before at time \( t \), \( \delta_i \) is a binary variable of attachment story which \( \delta_i = 1 \) if there is an increase in the in-degree of node \( i \) from time \( t - 1 \), otherwise \( \delta_i = 0 \). Variable \( \tau_i \) shows the age of node \( i \) at time \( t - 1 \). The idea is to compare the means of two independent samples \( \delta = 1 \) and \( \delta = 0 \) in terms of their age using Student’s \( t - test \). The hypothesis that we test in here is:

\[
\begin{aligned}
H_0: & \bar{\tau}_{\delta=1} = \bar{\tau}_{\delta=0} \\
H_1: & \bar{\tau}_{\delta=1} \neq \bar{\tau}_{\delta=0}.
\end{aligned}
\]

Table 4.2 captures the null hypothesis and the test results.

According to the results, from 1850 onward, the age of nodes that gained new incoming edges at time \( t \) is significantly lower than the age of nodes that didn't receive any new incoming edge. For example, in the most recent period, the acts that gain a new in-degree at a timestamp are on average 39.27 years younger than the ones that experience no change in their in-degree. As Table 4.2 reflects, there was no significant aging impact before 1850.

Figure 4.9 and the test results in Table 4.2 demonstrate the impact of aging in Legislation Network evolution model which suggest that legislation networks might follow broad-scale networks evo-
4.2. Descriptive Statistics of Legislation Network Evolution

Figure 4.9: The impact of age as a constraint on New Zealand legislation network evolution for the selected periods. The boxplot compares the average age of nodes based on their attachment story (1 if the node received at least a new edge, and 0 if no new nodes attached to the node at each time stamp (year).)

The result can help us to explain the evolution behaviour of individual nodes in the legislation network with considering age as a constraint. We will also examine the cost
4.2. Descriptive Statistics of Legislation Network Evolution

Table 4.2: t-test result - aging constraint

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>t-test result - aging constraint</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0: \tau_{\delta=1} = \tau_{\delta=0}$</td>
<td>$p-value$</td>
<td>$t$</td>
<td>$df$</td>
<td>$\mu_{diff}$</td>
</tr>
<tr>
<td>before 1850</td>
<td>0.231</td>
<td>-0.62</td>
<td>144</td>
<td>-41.82</td>
</tr>
<tr>
<td>1850 - 1900</td>
<td>1E-4</td>
<td>-13.84</td>
<td>3600</td>
<td>-16.85</td>
</tr>
<tr>
<td>1900 - 1950</td>
<td>2E-5</td>
<td>-19.06</td>
<td>7566</td>
<td>-23.69</td>
</tr>
<tr>
<td>1950 - 2000</td>
<td>1E-4</td>
<td>-23.55</td>
<td>14579</td>
<td>-27.03</td>
</tr>
<tr>
<td>2000 - current</td>
<td>1E-4</td>
<td>-24.26</td>
<td>15914</td>
<td>-39.27</td>
</tr>
</tbody>
</table>

constraint as being another constraint of preferential attachment in legislation networks evolution.

4.2.7 Legislation Network Evolution and Cost Constraint

In this section, we examine the impact of cost constraint on legislation network evolution process. The aim is to test if nodes with very high in-degree have a lower chance of gaining new edges because of the existence of a limited capacity.

To establish this test, we design a categorical variable of node position which is:

- Top 0.1%, if the in-degree of node $i$ at time $t - 1$ is within the highest 0.1 per cent of the in-degree values at time $t - 1$.
- 0.2%, if the in-degree of node $i$ at time $t - 1$ is larger than 99.7 percentile of the in-degree but below the top 0.1 per cent highest in-degree values at time $t - 1$.
- 99.7%, if the in-degree of node $i$ is below the 99.7 percentile of in-degrees at time $t - 1$

The hypothesis is to see if nodes in these categories have different behaviour in terms of their attachment story $\delta$. If nodes in the top group intent to not receive new edges because of reaching a specific limit, then there exist an observation of cost constraint.

With a closer look at the Top 0.1% nodes in the clustered boxplot of Figure 4.10, during 1900-1950 and also after 2000, there is no significant difference between the in-degree of nodes with $\delta_i = 1$ and the ones with $\delta_i = 0$. Their $t$-student results in Table 4.3 also confirms the visual observations. This effect suggests that the preferential attachment is constrained for the top nodes during 1900-1950 and from 2000.

It is not so challenging to explain cost constraint in the network of world airports [20], but achieving such results for legislation network might sound surprising. With a closer look at some ex-
4.2. Descriptive Statistics of Legislation Network Evolution

amples, this complicated observation might be an evidence of an evolution process that disadvantages large but not comprehensive legislation due to a need for introducing more significant laws. For example, the Crimes Act 1962 with a very large in-degree of 526 in the year 2005, suddenly experienced a lower rate in obtaining new edges. Six years later, the introduction of the Criminal Procedure Act 2011 explains the need for a structural change which might have resulted in costing constraint for the Crimes Act 1962.

The result guides us to explain the evolution behaviour of individual nodes in the legislation network by considering nodes’ capacity cost constraint in gaining new edges. This pattern indicates that legislation network at some periods shows a single-scale effect with degree distribution closer to an exponential distribution rather than a power-law. We will also examine the content constraint as being another constraint of preferential attachment in legislation networks evolution.

<table>
<thead>
<tr>
<th>Table 4.3: t-test result - high in-degree as a capacity cost constraint in the Top 0.1 % category</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: \bar{k}<em>{\delta=1} = \bar{k}</em>{\delta=0}$ for top nodes</td>
</tr>
<tr>
<td>before 1850</td>
</tr>
<tr>
<td>1850 - 1900</td>
</tr>
<tr>
<td>1900 - 1950</td>
</tr>
<tr>
<td>1950 - 2000</td>
</tr>
<tr>
<td>2000 - current</td>
</tr>
</tbody>
</table>

4.2.8 Legislation Network Evolution and Content Constraints

In this section, we explore the content of nodes impact on the attachment process of legislation network evolution. The plan is to test if nodes from different contents have significantly different chances of gaining new edges because of the existence of a content constraint.

To establish this test, we design a categorical variable of node content based on the most significant weekly connected components of the network. The hypothesis to test is to see whether nodes in these categories have different behaviour in terms of attachment story $\delta_i$.

2From 1850 onward there are nine considerably large weakly connected components, but before 1850 due to a small size and low connectivity of the graph, there were only seven weakly connected components observed.
Figure 4.10: The impact of node high in-degree as a capacity cost constraint on New Zealand legislation network evolution for the selected periods. The boxplot compares the average in-degree of nodes based on their attachment story, and also based on their position of in-degree. The position is considered as a categorical variable, nodes within the category of top 0.1% in-degree, nodes within the interval of 0.3%-top 0.1%, and nodes within the 99.7 in-degree percentile.

If nodes in these groups intent to receive new edges with significantly different probability, then there exist and observation of a content constraint. Assume that \( J_t \) shows the number of groups
at time $t$. Assume $F(j)$ is the ratio of the number of nodes with $\delta = 1$ to the number of nodes with $\delta = 0$ in the category $j$ at time $t - 1$. Then the hypothesis is:

$$H_0 : F(1) = F(2) = ... = F(J_t)$$

The alternative hypothesis is that at least nodes in one category stochastically dominates one other group in terms of their chance of receiving new edges. The Kruskal-Wallis $H$ [125] is the proper statistic to test this hypothesis of comparing more than two nonparametric categories.

The Kruskal-Wallis $H$ - test results in Table 4.4 confirms that within 95% confidence level legislation network evolution after 1950 experiences a content constraint.

<table>
<thead>
<tr>
<th>Period</th>
<th>$H$</th>
<th>$p$-value</th>
<th>$df$</th>
</tr>
</thead>
<tbody>
<tr>
<td>before 1850</td>
<td>0.38</td>
<td>0.073</td>
<td>6</td>
</tr>
<tr>
<td>1850-1900</td>
<td>11.12</td>
<td>0.031</td>
<td>8</td>
</tr>
<tr>
<td>1900-1950</td>
<td>64.25</td>
<td>0.002</td>
<td>8</td>
</tr>
<tr>
<td>1950-2000</td>
<td>66.79</td>
<td>0.001</td>
<td>8</td>
</tr>
<tr>
<td>2000-current</td>
<td>166.13</td>
<td>0.002</td>
<td>8</td>
</tr>
</tbody>
</table>

With a closer look at the clustered pie-chart of Figure 4.11, apart from the first period in which there is no significant difference between the attachment story of nodes in majority of the components, a content constraint appears to exist for other periods.

By observing the existence of content constraints on the generative process of New Zealand legislation network evolution, we can expect a tendency to single-scale network structure.

Thus we observed that the age, cost, and content constraints significantly impact the process of preferential attachment in legalisation network, and we would not expect to find a power-law scale-free structure for these networks. We will explore this expectation more in Section 4.3.

The outcomes also indicate that the newly introduced edges at each time do not necessarily attach to the other nodes with a pure preferential attachment, so their distribution between the existing nodes might follow a more single-scale distribution such as exponential.
4.2. Descriptive Statistics of Legislation Network Evolution

Figure 4.11: The impact of node content as a constraint on the New Zealand legislation network evolution process for the selected periods. The pie-chart compares nodes attachment story based on the component that nodes belong. The node class as a categorical variable reflects the most significant weekly connected components of the network at each period.
4.3 Evaluation of Existing Generative Models for Legislation Networks

In this section, we aim to observe the existing generative models of network evolution on legislation network based on in-degree distribution provided in the literature. In Section 4.2, we observed that there are indications of preferential attachment, but also there are significant signs of age, cost and content constraints. Table 4.5 summarises our findings on the generative processes in legislation network evolution and guides us towards identifying appropriate existing models to explain the network evolution.

Table 4.5: Overview of the network measures evolution

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-world $\sigma$</td>
<td>0.54 (NO)</td>
<td>30.63 (Yes)</td>
<td>65.51 (Yes)</td>
<td>78.89 (Yes)</td>
<td>158.34 (Yes)</td>
</tr>
<tr>
<td>Preferential attachment</td>
<td>1E–4 (Yes)</td>
<td>0.001 (Yes)</td>
<td>5E–4 (Yes)</td>
<td>0.001 (Yes)</td>
<td>2E–4 (Yes)</td>
</tr>
<tr>
<td>Aging impact</td>
<td>0.231 (No)</td>
<td>1E–4 (Yes)</td>
<td>2E–5 (Yes)</td>
<td>1E–4 (Yes)</td>
<td>1E–4 (Yes)</td>
</tr>
<tr>
<td>Cost impact</td>
<td>3E–5 (No)</td>
<td>2E–4 (No)</td>
<td>0.146 (Yes)</td>
<td>0.001 (No)</td>
<td>0.311 (Yes)</td>
</tr>
<tr>
<td>Content impact</td>
<td>0.073 (No)</td>
<td>0.031 (Yes)</td>
<td>0.002 (Yes)</td>
<td>0.002 (Yes)</td>
<td>0.002 (Yes)</td>
</tr>
</tbody>
</table>

To achieve an accurate assessment of the existing evolution models on legislation network, Table 4.6 represents the mathematical forms of in-degree distribution developed in the literature for each model.

Table 4.6: $p(k)$ the in-degree probability distribution function and parameters of the existing generative models

<table>
<thead>
<tr>
<th>Model</th>
<th>$p(k)$</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferential attachment</td>
<td>$p(k) \sim 2k_0^{1/2}k^{-\gamma}$ [93]</td>
<td>$k_0 \sim$ constant (average out-degree of new nodes) $\beta = 1/2, \gamma = 1/2 + 1 = 3,$</td>
</tr>
<tr>
<td>Preferential attachment and aging</td>
<td>$p(k) \sim 2k_0^{1/2}k^{-\gamma}$ [30]</td>
<td>$k_0 \sim$ constant (average out-degree of new nodes) $\beta = c(1 \sim \alpha), \gamma = 1/3, \alpha =$ aging power (constant), $c =$ constant</td>
</tr>
<tr>
<td>Preferential attachment and cost and content constraints</td>
<td>NA</td>
<td>NA $^4$</td>
</tr>
</tbody>
</table>

$^3$The small-world sigma $\sigma$ is calculated by comparing clustering coefficient and average path length of each network to 50 equivalent random network with same average degree as suggested by [33]

$^4$Although in the literature the impact of cost and content constraints on in-degree distribution of many graphs was observed, no mathematical model was provided. Previous studies decided to stop developing models and only fit the closest possible distributions.[20][94][92]
4.3. Evaluation of Existing Generative Models for Legislation Networks

Figure 4.12: log-log complementary cumulative plot - observe shape of legislation network in-degree distribution.

Table 4.6 shows that the existing scale-free and broad-scale models mathematically are based on the Barabási-Albert original power-law model with variation in the parameters. For single-scale models, there is no mathematical form presented, and the ideas developed visually and with descriptive tests. So there is an undoubted gap in the literature to cover all three classes of small-world networks mathematically. We will contribute to the research by answering this gap.
4.4 Discussion

For a valid judgment of a model, a careful analytical and experiential study of its properties and its parameters is necessary. Such empirical studies not only provide insight into why specific features occur but also suggests ideas for improvement.

In this chapter, the focus was to run in-depth analytical experiments and explore the attachment process of legislation network evolution and compare the behaviour with the literature of network evolution. The results of this chapter suggested that attachment behaviour in legislation network evolution might be affected by the processes of node aging, cost of having very high in-degree, and content limitation. The results remarked that we should not expect a scale-free structure from legislation network, but lack of enough mathematics in the literature of non-scale-free networks did not allow us to fit a practical model. Our findings verify that there is a need for new theoretical explanations of these non-scale-free patterns as also recommended recently by Broaido and Clauset [101].

In many research areas, often network evolution assumptions are adjusted so the method results in scale-free models [126]. For example, in the literature and the models discussed in this chapter, the value of out-degree of new nodes always considered to be constant. Another finding of this chapter is that the out-degree of new nodes in legislation network is not constant and instead follows an exponential regime. We also discovered that despite scale-free networks, the distribution of these newly introduced edges between the existing nodes is not purely a prefer-
ential attachment with a power-law, and it is rather close to a single-scale distribution such as an exponential distribution.

While studying real-world networks, it is extensively essential to understand and describe the topology and the developing behaviour of networks precisely and not to limit to simplistic assumptions. In the next chapter, we focus on these two main findings of this chapter and propose a model that answers both.
Chapter 5

Stochastic Generative Model for Network Evolution

To study the evolution of a dynamic system, if its output can be specified uniquely from inputs together with defined initial conditions, then deterministic modelling is satisfactory. But when there are scenarios of uncertainty, a stochastic approach is preferable. If one can assume that everything relevant that happens in the system is well understood and correctly specified in the formula(s) and fixed parameters, then deterministic methods result in a reliable model\[127\]. But in many real-world systems, the assumption that there are precisely known quantitative relationships between the components of the system is not necessarily valid.

After completing a comprehensive set of descriptive statistics on New Zealand legislation network evolution processes in Chapter 4, we observed that the in-degree distribution of such a network is not consistent with a scale-free structure. Also no existing mathematical model of network evolution fits the data. To overcome this issue in this chapter, we present a stochastic process and define a new mathematical model of network evolution.

In preferential attachment modelling, it is assumed that the probability of attachment is proportional to node in-degree. In Price's cumulative advantage model [84], the Barabási-Albert model [17] and later extensions of preferential attachment [95, 128], all analysis on network degree, diameter, and clustering coefficient is based on the assumption of preferentially spreading new edges between existing nodes using fixed values for variables.

For example, there are unrealistic assumptions in the literature around the determined number of new nodes added at each time, the fixed number of edges per new node, and the chosen probability of removing existing edges at each time. These unreliable assumptions result in biased conclusions such as a power-law degree distribution.
Modelling directed graphs using the same approach results in a uniform distribution of out-degree which contradicts the numerical findings of a power-law distribution [18, 19], exponential distribution [18, 20], and Gaussian distribution [20]. Such a paradox suggests the possibility of oversimplified modelling assumptions.

We found in the previous chapter that the out-degree of new nodes in the legislation network is not constant and instead follows an exponential regime. We also observed that the in-degree of individual nodes follows a common growth pattern, increasing through a period and then staying stable. We saw that the distribution of these newly introduced edges between the existing nodes is not purely based on a preferential attachment process with a power-law regime, but it is rather close to a single-scale distribution such as an exponential distribution.

In this chapter, we use such findings to propose our stochastic model and fill the gaps as we explained in Chapter 1.

The aim of Chapter 5 is stochastic modelling of the evolution of dynamic directed networks. We start by reviewing many definitions that are used to explain the model. To make a sensible comparison between the existing models and the proposed model, we also describe the preferential attachment model in stochastic language so the reader can follow and compare the assumptions in both models. Finally, the chapter includes testing the proposed model on five NZ legislation networks, a scientific citation network, and a social network, and comparing its performance to that of existing models.

By the end of this chapter, we answer the issue of unrealistic assumptions around the fixed number of newly introduced edges in the literature. Moreover, we show that our model results in a Lomax degree distribution that can explain a wide range of scale-free, broad-scale and single-scale network structures.
5.1 Definitions

As mentioned, continuous-time stochastic processes are used to model network evolution. This section covers definitions and examples relevant to the proposed model.

**Definition 5.1.1.** A **stochastic process** is a set of random variables representing statistical values of a system evolving [129]. We denote the process by $X(t)$ with parameter $t \in T$ representing time.

**Remark 5.1.1.** Uncertainty appears in many areas of knowledge, so there is a significant application domain for stochastic processes. In network science, the statistical mechanics of different networks is studied to describe network properties, but the application of stochastic processes to model network evolution is often ignored. There are a few studies that to some extent consider stochastic processes to study network size growth [130], preferential attachment [107], or transition rate of social networks [131] [132].

But very well known methodologies of network evolution have usually considered static model parameters [93] rather than stochastic processes, which resulted in biased conclusions [101]. The story behind the biased results in the literature of network evolution is more explained in Corollary 5.2.1.1.

**Definition 5.1.2.** A continuous random variable $X$ is assumed to have an **exponential distribution** with rate parameter $\lambda > 0$ if the probability density function of $x$ is given by:

\[
    f(x) = \begin{cases} 
        \lambda e^{-\lambda x} & x \geq 0 \\
        0 & x < 0.
    \end{cases}
\]

Then the cumulative density function of $x$ is $F(x) = 1 - e^{-\lambda x}$ and the mean of variable $X$ is $\frac{1}{\lambda}$ [124].

**Remark 5.1.2.** In mathematical modelling, one simplifying assumption that is often made is to consider certain random variables to be exponentially distributed. The exponential distribution is memory-less, which makes it easy to analyse and attractive for modelling [124].

**Remark 5.1.3.** An exponentially distributed process is observed to explain the out-degree of nodes in scientific citation networks [133]. As mentioned in Section 4.4, we found that the exponential distribution is a good approximation to model the process of edge addition in legislation networks. Also, we noted that we can estimate the distribution of the newly introduced edges between the existing nodes with an exponential regime. Later in Theorem 5.3.1, we will use these findings to
propose our stochastic model.

**Definition 5.1.3.** A **Poisson Process** is a continuous-time stochastic process $N(t) : t \geq 0$ with rate $\lambda > 0$ if:

- $N(0) = 0$.
- $N(t)$ represents the number of events that have occurred up to time $t$.
- The times between events are independent and distributed with an Exponential distribution of rate $\lambda$. [124].

**Proposition 5.1.1.** If $N(t)$ is a Poisson process with rate $\lambda > 0$, the random variable describing the number of events per unit of time follows a Poisson distribution with the expected value of $\lambda$ [124].

**Proposition 5.1.2.** If $N(t)$ is a Poisson process with a rate of $\lambda > 0$, then the age of the Poisson events is an exponential variable with the parameter $\lambda$ [134].

**Remark 5.1.4.** Queuing theory often models the arrival of events using the Poisson process [135]. In network science, a Poisson process is often used to explain the process of node addition to the graph. For example, the node addition process in collaboration networks [130], and scientific citation networks [115] were explained using Poisson processes.

**Definition 5.1.4.** A continuous random variable $Y$ is assumed to have a **gamma distribution** with shape parameter $n > 0$ and rate parameter $\lambda > 0$ if the probability density function of $y$ is given by:

$$f(y) = \text{Gamma}(n, \lambda) \begin{cases} \frac{\lambda^n y^{n-1} e^{-\lambda y}}{\Gamma(n)} & y \geq 0 \\ 0 & y < 0. \end{cases}$$

The scale parameter of the gamma distribution is $\frac{1}{\lambda}$ and the mean of variable $Y$ is $\frac{n}{\lambda}$ [124].

**Proposition 5.1.3.** Let $X_1, \ldots, X_n$ be independent and identically distributed exponential random variables having rate parameter $\lambda$ (scale parameter $\frac{1}{\lambda}$). Then the random variable $Y = X_1 + \cdots + X_n$ has a gamma distribution with shape parameter $n$ and rate parameter $\lambda$ [136].

**Proposition 5.1.4.** If $\alpha$ is a constant and $X$ is a distributed gamma variable having the shape parameter $n$ and the scale parameter $\theta$, then $Y = \alpha X$ is a gamma distributed variable with the shape

\[1\Gamma(n) \text{ is the usual gamma function defined by } \Gamma(n) = \int_0^\infty x^{n-1} e^{-x} \, dx.\]
5.1. Definitions

parameter $n$ and the scale parameter $\theta$ [136].

**Definition 5.1.5.** A continuous random variable $Y$ is assumed to have an **inverse gamma distribution** with shape parameter $n > 0$ and scale parameter $\theta > 0$ if the probability density function of $y$ is given by [124]:

$$f(y) = \begin{cases} \frac{\theta^n y^{-n-1} e^{-\frac{\theta}{y}}}{\Gamma(n)} & y \geq 0 \\ 0 & y < 0. \end{cases}$$

**Proposition 5.1.5.** If $X$ is a gamma variable with the shape parameter $n$ and scale parameter $\frac{1}{\lambda}$, then $\frac{1}{X}$ is distributed as an inverse gamma variable with the shape parameter $n$ and scale parameter $\theta = \frac{1}{\lambda}$ [136].

**Definition 5.1.6.** A continuous random variable $X$ is assumed to have a **Pareto distribution** with shape parameter $\alpha > 0$ and scale parameter $x_0 > 0$ if the probability density function of $x$ is given by: [99]$^2$

$$f(x) = \text{Pareto}(x_0, \alpha) = \begin{cases} \frac{\alpha x_0^\alpha}{x^{\alpha+1}} & x \geq x_0 \\ 0 & x < x_0. \end{cases}$$

Then the mean value of variable $X$ for $\alpha > 1$ is $\frac{\alpha x_0}{\alpha-1}$ and its cumulative density function is $F(x) = 1 - \left(\frac{x}{x_0}\right)^\alpha$.

**Remark 5.1.5.** The distance between the highest values in a Pareto distribution function is consistently large compared with the highest values in an exponential distribution function [136]. A biased process of edge allocation, such as a preferential attachment process may result in a large distance between the high values of degree distribution that results in a Pareto regime. In our proposed model in Section 5.3, we aim to propose a more unbiased process of edge allocation, which doesn’t necessarily result in a high distance between the high values of in-degree distribution. To explain such an unbiased edge allocation process, we use an exponential probability density function.

**Remark 5.1.6.** The **Pareto distribution** is often called the **power-law distribution**. [137].

$^2$The mean value of a Pareto random variable for $\alpha \leq 1$ is infinity.
Remark 5.1.7. The logarithmic cumulative density function of a Pareto distribution is linear [137]. Often researchers graph the logarithmic complementary cumulative density against the actual data to test the power-law characteristics.

Remark 5.1.8. The power-law distribution is observed in a very diverse range of events. For example city populations [137], the usage frequency of words in speeches [138], in-degree of scientific papers [139], and in-degree of WWW [17] is recognised to follow a Pareto regime [137]. The preferential attachment process in network science, also known as the cumulative advantage, and the rich-get-richer process is considered to be the reason behind observed power-law distribution of events in many subjects [103]. Other similar processes may result in a power-law distribution, such as accelerating growth model [140], additive multiplicative fitness [141], and our interpretation of preferential attachment model in Section 5.2.

Definition 5.1.7. The probability density function of a random variable is scale-invariant, self-similar, or scale-free when its shape does not change if the variable is multiplied by a common factor [137]. Let \( f(x) \) be the probability density function of variable \( X \), then \( f(x) \) is scale-free if there exists a choice for \( \Delta \) in the form of:

\[
f(\beta x) = \beta^\Delta f(x).
\]

We explained scale-free networks more in Section 4.1.1.2.

Corollary 5.1.0.1. Any variable with power-law (Pareto) distribution is scale-free.

\[
f(\beta x) = \frac{\alpha x_0^\alpha}{(\beta x)^{\alpha+1}}
= \beta^{-(\alpha+1)} f(x).
\]

Remark 5.1.9. In network science, the degree distribution of many real-world networks is assumed and discussed to follow a power-law regime. For example networks of scientific papers [84], biological networks [106], and the World Wide Web [20] were all assumed to have a power-law degree distribution. This idea controversially suggested scale-free network properties of many graphs with several historical studies supporting the scale-free networks [93], and later reviews criticising the concept with a considerable amount of testing [101].
Remark 5.1.10. The Pareto distribution type II which is a shifted Pareto distribution is also called the Lomax distribution. Lomax distribution is a Pareto distribution which allows zero [142].

Definition 5.1.8. A continuous random variable $X$ is assumed to have a Lomax distribution with shape parameter $n > 0$ and scale parameter $\lambda > 0$ if the probability density function of $x$ is given by:

$$f(x) = \text{Lomax}(n, \lambda) = \begin{cases} \frac{n \lambda^n}{(x + \lambda)^{n+1}} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

The mean value of variable $X$ for $n > 1$ is $\frac{\lambda}{n - 1}$ and $F(x) = 1 - (1 + \frac{x}{\lambda})^{-n}$ is its cumulative density function [142].

Proposition 5.1.6. If $X$ is a Pareto variable with the shape parameter $n$ and scale parameter $\lambda$, then $X - \lambda \sim \text{Lomax}(n, \lambda)$ [143].

Remark 5.1.11. Based on Section 4.1.1.2, and Corollary 5.1.0.1, any variable with Lomax distribution is not scale-free.

5.2 Stochastic Modeling of Preferential Attachment

From our findings in Chapter 4 about the generative models of legislation network evolution, we know that such graphs do not follow a scale-free preferential attachment evolution process. But it is essential to review such models as it helps us to address the issues better when we propose our stochastic model.

In this section, we use a stochastic approach to interpret a preferential attachment growth process as we defined in Section 4.1.1.2 and explain the evolution of dynamic directed networks. The node addition process in Figure 5.1 depicts our analysis of a simple preferential attachment network growth model and Theorem 5.2.1 explains it formally.

Using the queuing theory standard assumption [134], we assume that nodes join the network following a memory-less Poisson process as we defined in Definition 5.1.3. This assumption results in a memory-less nodes’ age distribution and time-independent nodes’ in-degree distribution.
5.2. Stochastic Modeling of Preferential Attachment

Theorem 5.2.1. Let $N(t)$ be a Poisson process of node addition to the graph with the rate parameter $\nu$. Let $\Phi_i(t)$ be the age of node $i$ at observation time $t$.

Let $K$ be the random variable of nodes' in-degree. Suppose that $K_i(t)$ is the in-degree of node $i$ at observation time $t$.

Let $\pi > 0$. Consider a simple preferential attachment process where the in-degree growth is proportional to the in-degree of the node at observation time $t$: that is for all $i \in N(t)$, $\frac{\partial K_i(t)}{\partial t} = \pi K_i(t)$.

Let also assume that there is a constant $k_0 > 0$ such that for all for $i \in N(t)$, $K_i(0) = k_0$.

Then regardless of the observation time, $K$ follows a Pareto distribution with the shape parameter $\frac{\nu}{\lambda}$ and the scale parameter $k_0$.

Proof. We know that $\frac{\partial K_i(t)}{\partial t} = \pi K_i(t)$, and $K_i(0) = k_0$. So for nodes at their age $\Phi$:

$$K = k_0 e^{\pi \Phi}.$$ 

We aim to find the in-degree probability density function $f_K(k)$ or. We start by finding the cumulative density function $F_K(k)$ as follows:

Figure 5.1: The process of node addition - stochastic interpretation of preferential attachment process

Refer to Definition 5.1.3 for more details about a Poisson Process.
5.2. Stochastic Modeling of Preferential Attachment

\[ F_K(k) = 1 - P(K \geq k) \]
\[ = P(k_0 e^{\pi \Phi} \geq k) \]
\[ = 1 - P \left( e^{\pi \Phi} \geq \frac{k}{k_0} \right) \]
\[ = 1 - P \left( \Phi \geq \frac{1}{\pi} \ln \frac{k}{k_0} \right) \]
\[ = 1 - \left( 1 - P \left( \phi \leq \frac{1}{\pi} \ln \frac{k}{k_0} \right) \right) \]

Based on Proposition 5.1.2, \( \Phi \sim \text{Exp}(\nu) \), and based on Definition 5.1.2, the cumulative density function of age is \( F_{\Phi}(\phi) = 1 - e^{-\nu \phi} \). So:

\[ F_K(k) = 1 - \left( 1 - \left( 1 - e^{-\frac{\nu}{\pi} \ln \frac{k}{k_0}} \right) \right) \]
\[ = 1 - \left( \frac{k}{k_0} \right)^{-\frac{\nu}{\pi}} \]

Therefore referring to Definition 5.1.6 the random variable of nodes in-degree, \( K \), has a Pareto or power-law distribution with the shape parameter \( \frac{\nu}{\pi} \) and the scale parameter \( k_0 \).

\( \square \)

As a corollary to Theorem 5.2.1, the Barabási-Albert original model and the majority of succeeding studies can be considered as simplified cases of the Pareto distribution.

**Corollary 5.2.1.1.** Consider Theorem 5.2.1, and assume a specific scenario in which \( \nu = 1 \) and \( \pi = \frac{1}{2} \). Then:

\[ P(k) = 2 \frac{k_0^2}{k^3} \]

which reflects the Barabási-Albert original preferential attachment model that we defined in Definition 4.1.1 [93].

**Corollary 5.2.1.2.** The majority of Barabási-Albert extended models which result in scale-free effects are simplified variations of the generic model proposed in Theorem 5.2.1, the only difference being the assumptions affecting the scale parameter and the shape parameter of the Pareto model.
5.3 Proposed Model

For example in Krapivsky et al. model \( \nu = 2 \) [128], in Dorogovtsev et al. accelerating model \( \nu + 1 = 3 \) [30], in Barabási et al. model \( \nu + 1 = 1.5 \) [93], and in Krapivsky-Render model \( \frac{\nu}{\lambda} = \frac{1}{r} \) with \( r \) being the probability of edge redirection [18].

**Paradox 5.2.1.1.** We defined a preferential attachment process in Section 4.1.1.2. One of the main assumptions in Preferential Attachment modelling both in the literature and in Theorem 5.2.1 is the constant value of out-degree. But the out-degree of real-world networks displays variation [20] [19] which contradicts with the above assumption. In the proposed model in Theorem 5.3.1 we address this issue, and this is one of the main contributions of the chapter.

**Paradox 5.2.1.2.** Real-world networks are frequently claimed to be scale-free, but in many research domains, for example, in biology, this assumption is confirmed to be a myth [144] [101] [145]. Thus, the presence of scale-free behaviour in real-world networks is questionable.

Across 1000 social, biological, technological, transportation, and information networks, only less than 36 of them (4%) are proved to follow a scale-free structure, and in 88% of those networks some alternative mathematical models could explain the data considerably better than a power-law scale-free model [101].

Such studies suggest that no particular general mechanism can explain the diversity of degree structures found in real-world networks. So further studies of statistical difference to generate new insights about the structure of complex systems are still needed and supported by the proposed model in this chapter.

5.3 Proposed Model

In the previous section, we interpreted a preferential attachment network evolution process. We explained in Corollary 5.2.1.2 that such a stochastic interpretation resulted in a fundamental model that represents many scale-free network structures in the literature.

In this section, we review our observations in Chapter 4 about the generative models of legislation network evolution. Our findings suggested many other factors impact the preferential attachment process in legislation networks, so we require an evolution model that can present their non-scale-free growth behaviour. We observed that exponentially distributed variables might represent out-degree distribution and the in-degree distribution in legislation network.

Based on these findings and information provided in Paradox 5.2.1.1 and Paradox 5.2.1.2, we propose our generic model in this chapter. It demonstrates a stochastic process for network evo-
5.3. Proposed Model

We explain our assumptions and propose the model in Theorem 5.3.1 which based on two main principles that are substantially different from those of Theorem 5.2.1.

First is the assumption that the out-degree of nodes at their introduction is exponentially distributed. By contrast in Theorem 5.2.1 this variable was set to a fixed value $k_0$.

Second is that in contrast to the assumption of a simple preferential attachment process in Theorem 5.2.1, in our model in Theorem 5.3.1, we consider a more general distribution of in-degree.

In this section we denote $f_X(x) = P(x)$ as the probability density function, $E[X]$ as the expected value, and $F_X(x)$ as the cumulative density function of the variable $X$. Also in the following theorem by using the term “unbiased”, we mean a non-preferential attachment process in which we give no advantage to the nodes in receiving the new edges. We explained such a process more in Remark 5.1.5. Using the queuing theory standard assumptions [134], we assume that nodes join the network following a memory-less Poisson process as we defined in Definition 5.1.3, and we assume a memory-less exponential out-degree distribution. These assumptions result in time-independent nodes’ in-degree distribution as we explain more in Theorem 5.3.1.

**Theorem 5.3.1.** Let $G(N_t, R_t)$ be a directed network which grows following a stochastic process. Let $N(t)$ be a Poisson process of node addition to the graph with the rate parameter $\nu$.

Let $M \sim \text{Exp}(\lambda)$ be an exponential random variable of nodes' out-degree at its introduction with the rate parameter of $\lambda$. Note that the out-degree of nodes never changes after their introduction. Let $U$ be the random variable giving the mean out-degree of introduced nodes per unit of time.

Let $K$ be the random variable representing in-degree of nodes.

Referring to Remark 5.1.5 consider an unbiased network evolution process in which $U$ derives the in-degree distribution of nodes, $K$. So $K$ given $U$ is exponentially distributed with its expected value derived from $U$: that is $f_{K|U} \sim \text{Exp}\left(\frac{1}{U}\right)$, and $E[K|U] = U$.

Then $K$ follows a Lomax distribution with the shape parameter $\nu$, and the scale parameter of $\frac{\nu}{\lambda}$.

**Proof.** We know that $M \sim \text{Exp}(\lambda)$. Denote $S = \partial(R_t)$, the total out-degree of introduced nodes per unit of time. Referring to Proposition 5.1.3:

$$f_S(s) = \sum_{i=1}^{\nu} M_i = \sum_{i=1}^{\nu} \text{Exp}(\lambda) \sim \text{Gamma}(\nu, \lambda)$$
Referring to Proposition 5.1.1 the expected number of nodes per unit of time is $v$. Referring to Proposition 5.1.4:

$$ f_U(u) = \frac{f_S(s)}{v} = \frac{\text{Gamma}(v, \lambda)}{v} = \text{Gamma} \left( v, \frac{\lambda}{v} \right) $$

We denote $W = \frac{1}{U}$, From Definition 5.1.5 and Proposition 5.1.5:

$$ f_W(w) = \left( \frac{\chi}{\lambda} \right)^v w^{v+1} e^{-\left( \frac{\chi}{\lambda} \right)w} \frac{1}{\Gamma(v)} $$

And $f_{K|W}(k|w) \sim \text{Exp}(w)$.

Then the marginal distribution of $K$ can be obtained:

$$ f_K(k) = \int_{w=0}^{\infty} f_{K|W}(k|w) f_W(w) \, dw $$

$$ = \int_{w=0}^{\infty} we^{-wk} \left( \frac{\chi}{\lambda} \right)^v w^{v+1} e^{-\left( \frac{\chi}{\lambda} \right)w} \frac{1}{\Gamma(v)} \, dw $$

$$ = \left( \frac{\chi}{\lambda} \right)^v \frac{1}{\Gamma(v)} \int_{w=0}^{\infty} w^{v+2} e^{-\left( \frac{\chi}{\lambda} + k \right)w} \, dw $$

$$ = \frac{\chi^v \Gamma(v+1)}{\Gamma(v) \left( \frac{\chi}{\lambda} + k \right)^{v+1}} \int_{w=0}^{\infty} \left( \frac{\chi}{\lambda} + k \right)^{v+1} w^{v+2} e^{-\left( \frac{\chi}{\lambda} + k \right)w} \, dw $$

$$ = \frac{v \left( \frac{\chi}{\lambda} \right)^v}{(k + \frac{\chi}{\lambda})^{v+1}} $$

which is a Lomax distribution.

\[ \square \]

**Corollary 5.3.1.1.** Referring to Definition 5.1.8,

$$ P(k) = \frac{v \left( \frac{\chi}{\lambda} \right)^v}{(k + \frac{\chi}{\lambda})^{v+1}} $$

is a Lomax distribution with the shape parameter $v$, the scale parameter $\frac{\chi}{\lambda}$, mean $\frac{v}{\lambda(v-1)}$, and median $\frac{\chi}{\lambda} \left( \frac{1}{2} - 1 \right)$.

**Corollary 5.3.1.2.** The proposed generative model in Theorem 5.3.1 illustrates that the distribution of in-degree and out-degree of networks can be obtained from the network growth parameters $v$ and $\lambda$. 

Corollary 5.3.1.3. The proposed generative model in Theorem 5.3.1 is not scale-free. But as we explained in Proposition 5.1.6 the power-law distribution is a specific case of Lomax distribution, so with specific assumptions, the proposed Lomax distribution can represent scale-free networks. For specific combinations of \( \nu \) and \( \lambda \), the proposed model in Theorem 5.3.1 could get very close to different types of distributions, for example, exponential, or power-law regimes. The results in the next section cover some of these scenarios. Figure 5.2 shows the variation of the proposed model for in-degree with different values of \( \nu \), and \( \lambda \).
5.3. Proposed Model

(a) $\lambda = 0.5$

(b) $\lambda = 1$

(c) $\lambda = 2$

(d) $\lambda = 20$

Figure 5.2: Comparison between the distributions
5.4 Evaluation of the Proposed Model

As can be seen:

- For large values of \( \nu \), the Lomax distribution gets closer to the exponential distribution. From the graphs, it can be observed that for \( \nu > 30 \), the distribution is considerably close to the exponential distribution.

- For \( \nu = 3 \) and the values of \( \lambda \) that are significantly larger than 1, the proposed Lomax distribution can get close to power-law distribution, but with a meaningful distance. A very extreme scenario of network evolution is required to meet the above conditions. It means that the expected value of induced nodes per unit of time is three, and each of those introduced nodes needs to add significantly less than 1 edge to the graph.

- the smaller the \( \lambda \), the larger the distance between power-law and the Lomax distribution.

5.4 Evaluation of the Proposed Model

This section implements an evaluation of the Lomax(\( \nu, \nu \lambda \)) model with power-law and exponential distributions. We selected three different networks for this section: New Zealand legislation network in five different periods, a scientific citation network of ACM papers [146], and a social network of computer scientists [147]. Table 5.1 shows the cumulative distribution functions \( F(k) \) and the probability density functions \( p(k) \) of the three models considered.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( p(k) )</th>
<th>( F(k) )</th>
<th>log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lomax(( \nu, \nu \lambda ))</td>
<td>( \frac{\nu^\nu}{(k+\nu \lambda)^{\nu+1}} )</td>
<td>( F(k) = 1 - \left(1 + \frac{k}{\nu \lambda}\right)^{-\nu} )</td>
<td>( \log \left( \prod_{i=1}^{\beta} \frac{\nu^\nu}{(k_i+\nu \lambda)^{\nu+1}} \right) = \beta \log \nu + \log \lambda - (\nu + 1) \sum_{i=1}^{\beta} \log(1 + \frac{k_i}{\nu \lambda}) )</td>
</tr>
<tr>
<td>Exp(( \lambda ))</td>
<td>( \lambda e^{-\lambda k} )</td>
<td>( 1 - e^{-\lambda k} )</td>
<td>( \log \left( \prod_{i=1}^{\beta} \lambda e^{-\lambda k_i} \right) = \beta \log \lambda - \lambda \sum_{i=1}^{\beta} k_i )</td>
</tr>
<tr>
<td>Power-law(( \alpha, k_0 ))</td>
<td>( \frac{\alpha^{\frac{k_0}{\lambda}}}{\Gamma\left(\frac{k_0}{\lambda}\right)} )</td>
<td>( 1 - \left(\frac{k_0}{\lambda}\right)^{\alpha} )</td>
<td>( \log \left( \prod_{i=1}^{\beta} \frac{\alpha^{\frac{k_i}{\lambda}}}{\Gamma\left(\frac{k_i}{\lambda}\right)} \right) = \beta \log \alpha + \log k_0 - (\alpha + 1) \sum_{i=1}^{\beta} \log(k_i) )</td>
</tr>
</tbody>
</table>

Those models are compared using the Bayesian information criterion, which measures log-likelihood for the models based on their parameters [148] [149]. Bayesian information criterion is defined as \( BIC = \log L - \frac{p}{2} \log \beta \) in which \( L \) is the log-likelihood of the model evaluated at the maximum likelihood estimates, \( p \) is the number of parameters, for example for Lomax(\( \nu, \lambda \)) distribution \( p = 2 \), and \( \beta \) is the number of observed data [148]. Table 5.2 shows the BIC statistics obtained, from the three selected models, corresponding to the selected networks.
5.4. Evaluation of the Proposed Model

Table 5.2: Bayesian information criterion likelihood statistics for the three selected models, fitted by maximum likelihood to the five different periods of New Zealand legislation network (NZ-LG), scientific citation network of ACM papers (ACM SCN) [146], and social network of computer science scientists (SN-CS) [147]. Larger values of likelihood indicate the performance of models.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lomax</td>
<td>-1002.610</td>
<td>-7767.220</td>
<td>-1163.210</td>
<td>-13398.000</td>
<td>-1238.300</td>
<td>4.960</td>
<td>166894.560</td>
</tr>
<tr>
<td>Exponential</td>
<td>-648.259</td>
<td>-19936.880</td>
<td>-18234.600</td>
<td>-36284.500</td>
<td>-48229.300</td>
<td>-21163.370</td>
<td>-12434.510</td>
</tr>
<tr>
<td>Power-law</td>
<td>90.047</td>
<td>-10176.761</td>
<td>-15966.400</td>
<td>-23049.700</td>
<td>-30411.700</td>
<td>-30411.700</td>
<td>-1118.010</td>
</tr>
</tbody>
</table>

Lomax($\nu, \lambda$) distribution presents the largest values of Bayesian information criterion statistics for New Zealand legislation network after 1850.

As the Table 5.2 illustrates the proposed Lomax distribution also explains the degree distribution of the selected scientific citation network and the chosen social network with a significantly high Bayesian information criterion value comparing to the other models. A similar interpretation can be obtained from Figure 5.3.

Table 5.3 shows the corresponding parameter estimates and their standard errors from the Lomax($\nu, \lambda$) distribution.

Table 5.3: Parameter estimates from the Lomax($\nu, \lambda$) model, to the selected networks, in by maximum likelihood (standard errors in parenthesis). The $p$-values of models are based on the Kolmogorov-Smirnov goodness of fitness test statistics [150] for Lomax($\nu, \lambda$).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu$</td>
<td>0.368 (0.045)</td>
<td>0.427 (0.021)</td>
<td>1.156 (0.029)</td>
<td>1.561 (0.012)</td>
<td>1.754 (0.014)</td>
<td>5.105</td>
<td>1.014</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.920 (0.052)</td>
<td>1.859 (0.049)</td>
<td>1.349 (0.040)</td>
<td>2.724 (0.039)</td>
<td>2.317 (0.034)</td>
<td>0.18</td>
<td>2.086</td>
</tr>
</tbody>
</table>

Model $p$-value: 0.011 0.021 0.032 0.041 0.013 0.006 0.014

The adequacy of the Lomax($\nu, \lambda$) distribution is shown using log-log complementary cumulative density function plots. Figure 5.3 and Figure 5.3 shows the plots obtained from different networks. In summary, Lomax($\nu, \lambda$) distribution can be useful for modelling many different types of real-world networks. It presents the best BIC statistics of the selected models. Graphically it gives a reasonable description of the datasets.
5.4. Evaluation of the Proposed Model

(a) NZ legislation network before 1850

(b) NZ legislation network 1850-1900

(c) NZ legislation network 1900-1950

Figure 5.3: log-log complementary cumulative plot - Lomax distribution for the selected networks
5.4. Evaluation of the Proposed Model

(d) NZ legislation network 1950-2000

(e) NZ legislation network 2000-current

(f) ACM papers citation network [146]

Figure 5.3: log-log complementary cumulative plot - Lomax distribution for the selected networks - continued
5.5 Discussion

In Chapter 5, we summarised the main research contribution of the study towards network evolution. In the history of network evolution models, the theory of preferential attachment and scale-free networks is the leading theory that understood and followed by many researchers. But recent studies clarify the limitations of such methods. Only four per cent of real-world networks are observed to show scale-free characteristics and fit in the preferential attachment models [101]. Despite similar earlier processes, the proposed model in this chapter suggested a stochastic process for network evolution based on stochastic sub-processes of node and edge addition. Also, the proposed method does not rely on the determined allocation of the new edges to the existing nodes, and it instead considers the stochastic process for the in-degree distribution between the nodes.

In this chapter, we focussed on modelling network evolution processes in terms of degree distribution. To do that, we first confirmed the preferential attachment assumptions and concepts using stochastic modelling. Then we identified the limitations of such a deterministic process which led us to propose a new stochastic process for network evolution resulting in Lomax distributions, with two parameters, which directly relate to network growth rate and network size. We explained that our proposed generative model in Theorem 5.3.1 illustrates that the distribution of in-degree and out-degree of networks can be obtained from the network growth rates \( \nu \) and \( \lambda \). We showed that the proposed model covers a genuine family of networks. We provided some particular examples and tested them. We considered New Zealand Legislation Network in five
different periods, scientific citation network of ACM papers, and social network of a computer scientist. We showed that the proposed stochastic process well explains the chosen Networks. And finally, we showed analytically and graphically the competence of the Lomax distribution in comparison with other known distributions in the literature.

We proposed a model that satisfied two contributions and improvements. We first answered the gap of unrealistic assumption around the fixed number of newly introduced edges. We also suggested a stochastic process that resulted in a degree distribution with realistic parameters that can represent a full variety of scale-free, broad-scale and single-scale network evolution processes.

The novelty around the effective use of stochastic modelling alongside the answers to the two fundamental paradoxes in the literature of generative models makes this chapter the heart of the study. The results are promising.
Chapter 6

Conclusion

We began this study by arguing that network science is very useful for studying legislation systems by offering proper concepts, empirical methods, and modelling techniques that enable us to explain legislation systems components and behaviours.

Following such motivation, in the next part of this study, we claimed a reliable way with more than 98 per cent precision and recall to build dynamic legislation networks. We proposed a framework which includes several Information Extraction processes such as optical character recognition, named entity recognition and approximate string matching. We showed that we can make reliable networks from old and poor-quality documents. We emphasised the importance of data quality for the analysis of dynamic networks. Thus we were challenged to propose a method to minimise the errors by performing time-consuming but highly valuable Information Extraction tasks. The results were promising and ensured significant enhancements in data quality. Surely the improvement in data quality was one of the main contributions and deliverables of the work.

We provided examples of network science approaches which produced insights on history and performance of New Zealand legislation system. We noted that the change in the network structure highlights the relationship between the laws and the socio-economic requirements of the country. We pointed out that in the current decade, the hot legal topics show a change which could be a good reflection of the society’s needs. We investigated the knowledge that network properties provide to, and we noted its benefit for smarter legal processes. For example, the inclusiveness of nodes in the communities of laws could reflect the performance of legislative documents to protect the environment, resources and society.

We started the second half of this thesis with discussions on how network-based analysis provides insights into the generative processes of network evolution. We observed linear growth in
network size as an indication of a fixed capacity for parliament to pass laws, and exponential
growth of network density as an indication of reforms in the quantity of legislation and quality of
cross-referencing. Consistent with earlier statistical studies in citation networks, social networks
and biological networks, we found that the probability that existing legislation will receive a new
edge is higher for the nodes with larger in-degree. But other factors might impact this preferential
attachment and hold the network structure distinct from a scale-free power-law regime. We
pointed out that for a legislation network, age of the nodes might negatively impact the probability of their in-degree increase. We observed such a scenario for some periods of New Zealand legislation network evolution. For example, since year 2000, the acts which receive new references at each point of time are on average 39.27 years younger than the ones with no increase in their in-degree at the same point in time.

Surprisingly, we saw that in legislation network, the cost constraint might happen. For example, there exists a cost constraint for the nodes with very large in-degree to gain new edges in New Zealand legislation network during 1900-1950 and since 2000. This complicated dependence is evidence of an evolution process that disadvantages long but not comprehensive legislation due to a need for introducing more fundamental laws. For example, the Crimes Act 1962 with a very large in-degree of 526 in the year 2005, suddenly experienced a lower speed of receiving new edges compared to the previous decade. Six years later, the introduction of the Criminal Procedure Act 2011 can explain the need for a fundamental change.

We reported some preliminary results on observing content constraint in the attachment process from the new nodes to the existing ones. We suggested that this dependency is a fruitful means to understand the evolution of law communities over time.

After a comprehensive set of descriptive statistics on New Zealand legislation network evolution processes, we observed that the in-degree distribution of such networks is unlikely to follow a scale-free structure. There was no existing mathematical model of network evolution to fit the data. To overcome this issue, we presented a stochastic process and defined a new mathematical model of network evolution.

We found that the in-degree of individual nodes follows a universal growth pattern, increasing gradually and reaching a steady state. We also reported an exponential distribution of the out-degree of newly introduced nodes. We used these observations to define and establish our assumptions of the proposed model appropriately.

By our proposed model, we made two contributions and improvements. First, we answered the gap of unrealistic assumption around the fixed number of newly introduced edges. Second, our
model resulted in a Lomax degree distribution with practical parameters that can explain a wide range of scale-free, broad-scale and single-scale network structures. We claim that our model moves the history of explaining generative models of legislation network evolution one step forward.

Another subject in which a network science approach shows relevance for legislation network is to study the betweenness measure[74] and observe the impact of removing hubs in legislation networks. Network-based measures of a local structure may be able to distinguish between the expired and repealed acts. In this work, we provided only a small sample of the application of network science to legislation systems. Much remains in application studies of network science in the legal domain. We used a community detection method in this work to explain the applications of legislation networks, but there is much more to in the literature of community detection methodologies for directed dynamic graphs. The comparison between legislation networks of different jurisdictions is another interesting study as a signal of differences between the societies.

The proposed stochastic process well described the degree distribution of the studied legislation networks, a social network, and a citation network. A similar stochastic model can be studied for scenarios of edge rewiring, which remains for future studies.


[77] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.


[140] Sergey N Dorogovtsev, José Fernando F Mendes, and AN Samukhin. “WWW and Internet models from 1955 till our days and the popularity is attractive principle”. In: *arXiv preprint cond-mat/0009090* (2000).


