Visualisation of the New Zealand Web Topology

Wai Loong Tham

Department of Computer Science

University of Auckland

June 2015

A dissertation submitted in partial fulfillment of the requirements
for the degree of Master of Professional Studies in Data Science

Abstract

This research documents an attempt to depict the topology of the New Zealand web. It describes the use of Scrapy and Python scripts to harvested websites and linkages between them, and the use of a visualisation application (Gephi) to portray the network structure. Some communities were found and these were given semantically meaningful labels. The contribution of this research comes from the provision of a pictorial view of the New Zealand online landscape and the sharing of some experience from having carried out this exercise. The results should be of interest equally to those who have or are considering having an online presence in the New Zealand environment.

Key words: Network analysis, New Zealand web topology, linkages, community, Scrapy, Gephi.

Acknowledgement

I would like to record my gratitude to the following people for providing me with valuable help in completing this research:

Professor Sebastian Link
Dr. Stefan Schliebs

Data Science is a discipline that straddles both Computer Science and Statistics. In the research journey, I have benefitted from the knowledge and advice from the following people in the Department of Computer Science and Department of Statistics:

Department of Computer Science: Damir Azhar, Professor Gill Dobbie, Dr. Michael Dinneen, Ann Cameron, Adriana Ferraro, Dr. Yun Sing Koh, Dr. Patricia Riddle, Professor Rober Amor, Associate Professor Nevil Brownlee, Paul Denny, Professor Bob Doran, Associate Professor Beryl Plimmer, Dr. Gerald Weber, Robyn Young.

Department of Statistics: Associate Professor Ross Ihaka, Professor David Scott, Dr. Steffen Klaere, Associate Professor Ilze Ziedins, Professor Alan Lee, Dr. Andrew Balemi, Dr, Brendon Brewer, Associate Professor Brian McArdle, Professor James Curran, Marie Fitch, Dr. Mark Holmes, Associate Professor Paul Murrell, Associate Professor Rachel Fewster, Associate Professor Russell Millar, Sampath Fernando, Professor Thomas Lumley, Dr. Thomas Yee, Nancy Wong.

Visualisation of the New Zealand Web Topology

Content

Chapter 1 Introduction 1

Chapter 2 The Science and Art of Network Analysis 2

Nodes and Linkages 3
Theory of Linkages 4
Communities 6
Getting weblinks by scrapping 7
Degree, Centrality, Betweenness, and Closeness 7
Communities detection 9
Hubs and authorities 11
Visualising the weblinks using Gephi 12

Chapter 3 Research Design 14

Collect - Getting weblinks 15
Clean - Cleaning data and obtaining stable linkages 16
Connect - Portraying the results 17

Chapter 4 Results 18

Processing of data files 18
Modularity 22
In-Degree 23
Out-Degree 24
The New Zealand web topology 25

Chapter 5 Conclusion 35

Summary of findings and further work 35

References 37

Appendix The New Zealand web topology 39

Chapter 1

Introduction

With the advent of the Internet, information is readily available to those who require it. However, with the proliferation of websites, the user is overwhelmed with choices in relation to the sources and quality of information. It then becomes a case of which website one trusts to obtain the required information.

This project documents a first attempt to map out the sources of information about things New Zealand. The research was instigated by the question: If we want some information about some aspects of this country, where would one reach out to find it? An analogy to the physical road map is useful here: if we have an equivalent road map for the web in New Zealand, it would be much easier for the user to get to the required place easier. The impetus for this research comes from an article on development of the French national web archive [1] [2]. That article chronicles the expansion of sources for archiving purposes, from print publication to sources available on the web. In this research, the aim is to discover the location of sources of information about things New Zealand and the websites that provide this information. Are there any prominent websites that guide us in our quest for online information. Taking all these together, the aim is to provide a **depiction of the New Zealand web topology**.

This thesis is divided into a few logical sections. Chapter 2 provides the background in network analysis and understanding of graphs and web linkages. Following that, the research design for this study is set out in Chapter 3. It chronicles the approach chosen for obtaining data, cleaning the data, and portraying the results. Chapter 4 contains results, a picture of the New Zealand web topology, and a collection of semantically meaningful communities that were discovered. Chapter 5 concludes with a discussion of the findings in this research and some thoughts on further study in this area.

Chapter 2

The Science and Art of Network Analysis

" That two individuals on the opposite sides of the world, and with little in common, can be connected through a short chain of network ties – through only *six degrees* - is a claim about the social world that has fascinated generation after generation."

Duncan J Watts (2003) *Six Degrees: The science of a connected age,* [12] p. 299

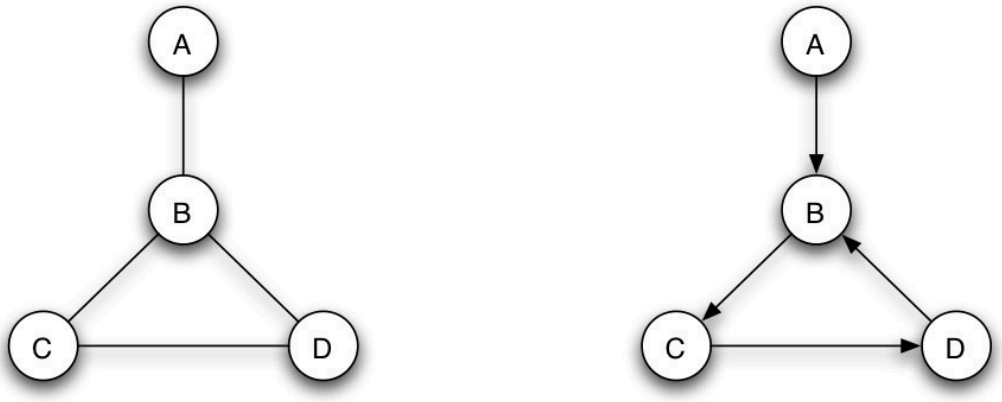
We live in the age of Big Data and connections. The volume, variety and velocity in which data and information arrives from the Internet has increased phenomenally in this age. Data becomes information when it is processed in a carefully thought out manner. Our lives are made much more meaningful and productive by who we know and the people we associate with. On the digital front, we connect and communicate unseen and it is this linkage that enhances our lives. Social network analysis - the science of studying connections - has been established a number of years now and has started with the analysis of social ties.

The advent of the Internet provided the opportunity to apply that established discipline in an area that is still evolving. Rather than seeing things as they are, a more informative way is to relate the things. Graphs - which use nodes and linkages to show affinity in things - help reveal the structure and relationships in data. A knowledge of those relationships help us understand how things work and why they work in a certain manner. By summarising and condensing the data, graphs give us an informative overview of the topology to reveal prominent structures. Just like having a road map guides us as we journey from one destination to another, a knowledge of the topology of the web provides guidance on where to go for information. It has been said that a picture is worth a thousand words, and graphs provide this function in relation to weblinks. Brath and Jonker [5] (p.39) put it this way: "visually spatial relationships are more important than link clarity".

This chapter provides material on the important concepts in network analysis. As will be soon clear, the skills required for a good analysis involves both scientific knowledge and an artistic touch. The important concepts relating to nodes, linkages, communities, and network metrics form important knowledge in this area. Artistic skills come to the fore when laying out the network to show the prominent features.

Nodes and linkages

The basic components of a network analysis are the nodes and linkages between the nodes. Early researches concentrated on social networks and the tools for analysis were borrowed from the work done in graph theory in mathematics. Much of the language used in social network analysis has been adapted from this discipline. Thus, in a hyperlinked environment, we would have a website as a node the links between the websites would be an arc in graph theory language. The linkages between nodes can be either directed or undirected. In a directed link, the linkage starts from a source website and extends a path to the target website. This is more informative compared to an undirected link which is often used to signify mutual relationships such as siblings and penpals in social network analysis.



A 4-node Undirected graph A 4-node Directed graph

Figure 2.1 Undirected graph and Directed graph (Adapted from [3] p.24)

In mathematics, graph theory encompasses an area that covers networks and its analysis. A graph, consisting of vertices and arcs, is synonymous to a network. Borrowing the ideas from that established discipline, researchers in social networks refer use nodes to refer to vertices, and links, relationships, or ties to refer to the connections between nodes. In this research, the language used in social networks will be adopted.

Weblinks by themselves are not very interesting. Grouped together by how they are linked to each other, a more interesting picture emerges. The links are voluntary, and the linkages provide evidence of usefulness or authoritativeness of information. Tightly linked websites form communities, and graphs provide a valuable means of revealing the relationships in communities. The web, which has content and hyperlinks as its elements, has been described as a self-organising decentralised architecture. Those largely uncoordinated links between websites are done in a voluntary manner, leading Barabasi [13] to describe it as a "web without a spider". It is discovery of the topology of this spiderless web in New Zealand that is of interest to in this research.

Theory of Linkages

Two Hungarian mathematicians, Paul Erdos and Alfred Renyi, established the seminal notion of Random Networks [7]. Based on two simple assumptions - that nodes are connected to others at random, and the network is undirected - their model has been used to describe and understand phenomenon such as the spread of diseases through social contact. The constant probability of a link between two nodes and the independence of links means that we can use the binomial probability to derive the average degree of the network from a calculation of binomial expected value. Thus, a network with 100 nodes and a fixed probability of 0.2 of a link between any two nodes would have an average degree of 20. The main insight from this network model is that nodes are just a few hops away as the network grows in size. However, the manner in which the average degree is calculated makes it very unlikely for the prediction of hubs where such nodes have many more connections than the norm.

Watts and Strogatz gave us the "small-world" networks [10]. These networks contain properties from regular and completely random models. Such networks, containing clusters and short path lengths between nodes, describe electricity grid and social networks well. In short, the authors introduce the idea that there are two sets of forces, one which holds members of a cluster together and a more macro force which governs nodes of the network.

The weakness of the Erdos and Renyi model were addressed by Barabasi and Albert by incorporating growth and preferential attachment [13]. These models describe networks in a better manner by allowing more nodes over time and that new nodes are more likely to attach to existing nodes that already have more linkages to other nodes in a "rich-get-richer" phenomenon. These power law networks, which exhibit a highly skewed distribution where a few networks have very high probability of linking and the majority have low probability, have been found to model networks such as journal citations very well. Crucially, such directed networks predict the emergence of hubs, based on its principle that the probability of a link is proportional to the number of links it has already acquired. The distribution is scale free and its exponent dictates the steepness of the curve.

The above description suggests that hubs are responsible for the dissemination of information, with the other nodes playing a minor role. Granovetter [8] first put forward the importance of weak ties in getting things done in the network. His demonstrated that weak ties (co-workers and aquaintances) are more useful than strong ties (family members and close friends) in securing employment. A more recent study [14] rediscovered that those numerous nodes in a network with weak links play a more important role in the relaying of online information than was initially thought. Perhaps the weak ties is related to the relatively shortness of paths in small-world environments, one where the path is not as long from one starting point to an unfamiliar destination. The Kevin Bacon study and the six degrees of separation [12] between anyone on earth provides proof that we are indeed connected to each other along relatively few hops (traversals between nodes along directed linkages).

The concept of preferential attachment became accepted in the community. Essentially, this "rich-gets-richer' idea says that nodes that have a higher number of links are more likely to receive more future links. The power law is associated with this idea: it says that a few websites get most of the links and the majority of the existing websites get only a few links.

Plotted on a number of websites and number of links axes, the bell curve is strongly skewed to the left. Along with this comes the idea of hubs and authorities. Hubs are websites that have numerous links, making it a crucible for the accumulation of information. Authorities are websites that connect to the hubs to enable their users to have access to the credible information.

Communities

As more websites link to each other, various topologies emerge. Often a group of websites have tight links among themselves, leading to the creation of cliques and communities. Members of a community are more likely to receive information from their constituent members before those that are outside the community. The study of communities involves the analysis of interest of the membership and their relationships. An understanding of the web structure is useful in improving methods of accessing the vast array of information on the web [reference].

The process of visualisation by Brath and Jonker [5](p.52) will guide us in our work. That process consists of the following:

- Collect and clean the data
- Review statistics and rearrange the layout of nodes and edges to reveal components, clusters and useful information
- Add visual attributes (labels, sizes, colours, line thickness) to enhance understanding
- Zoom, select, filter, annotate, and explain

Getting weblinks by scrapping

The raw material in an analysis of the web is the collection of web links. In doing so, there are two major decisions: a source which provides a list of relevant websites, and what to collect.

Scrapping is the name given to the harvesting of weblinks using an electronic spider which traverses the relevant portion of the web. We used a Python plug-in, appropriately called Scrapy, for this purpose. Typically, we would release Scrapy and it would follow links for a few hops. The websites it encounters would be recorded in a directed edge relationship: source website to target website. A graph visualisation application, Gephi, is used to portray the links in a picture to help us discern hubs and communities.

Here are some of the common network metrics that will be used to describe and analyse our network:

Degree, Centrality, Betweenness, and Closeness

This provides a measure of importance of nodes in a network. In an undirected network, the degree measure might be sufficient. This measures how many links the node has with other nodes. In a directed network, where it is important to identify the source and target of the link, measures like In-degree (the number of incoming links to the node) and Out-degree (number of links that emanate from the node) would be more useful than just the Degree measure.

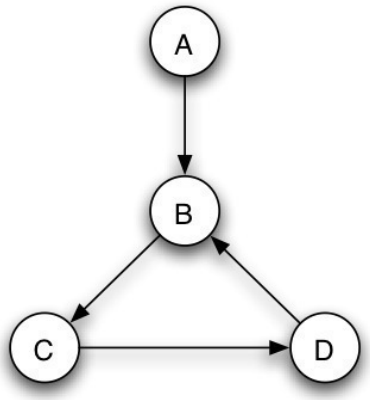


Figure 2.2 Node B has In-degree of 2 and Out-degree of 1 (Adapted from [3] p.24)

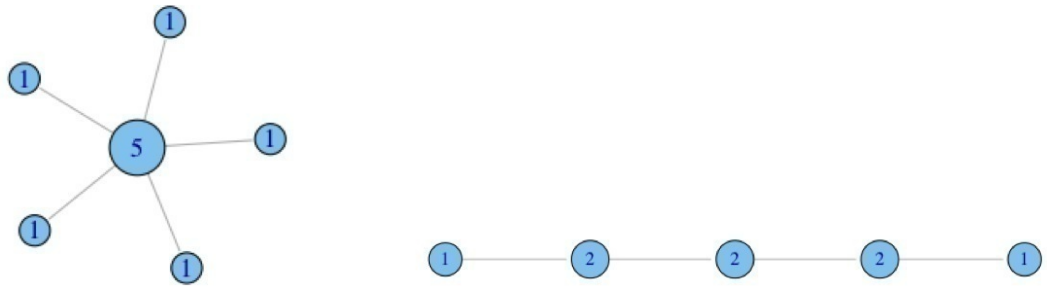
The various measures of degree merely provides an idea of the number of links each node has. It does not tell if the node is 'close to the action' in the network. In order to tell how central a node is, a betweenness measure would be useful. It captures the notion of 'brokerage' - acting as the middle man - in a network to bridge connections between other nodes. If a node occupies a position in the network that makes it easier for one node to link to another, it has a higher value of centrality than one that occupies a position in the periphery of the network. Closeness is another notion for centrality, and it depends on the proximity to other nodes in the network. A node with a high level of closeness measure indicates that it occupies a position that it can easily access the other nodes in the network following directed linkages. The closeness measure of a node is very much enhanced if it is linked to a hub in the network.

Figure 2.3 Network for illustrating the Closeness measure (Adapted from [21])

The two networks above will be used to illustrate the Closeness measure. It is clear why the numbers in the circles that indicate their degrees (connections to other nodes) do not provide a sense of closeness. The closeness of a node is calculated by summing the number of hops it requires to visit all the other nodes in the network divided by the number of nodes in the network excluding itself, and taking the reciprocal of this value. For example, the node on the extreme left would have a Closeness measure of 0.4 (total of 10 hops to visit all other nodes divided by 4, and taking the reciprocal of this result). The node on its immediate right would have a Closeness measure of 0.57 (the reciprocal of a total of 7 hops divided by 4 neighbours). Taking this one step further, the node in the middle would have a Closeness measure of 0.67. Thus, the node on the extremes of the network, by virtue of having a lower Closeness measure, is deemed to be further away from the action in the network.

Communities detection

Modularity is a measure used in the detection of non-overlapping communities [6]. It is based on graph partitioning principles which attempts to delineate a network into tightly-knit groups which we recognise as communities. Numerically, the modularity metric is calculated based on the number of links in the network that are close to each other for them to be categorised into groups less the expected number of links that one could expect in a network with randomly placed links. Just like the sun and the planets in our solar system, those planets are seen to belong to this solar system rather than orbiting aimlessly in space. It can be thought of as a significance test for community detection. As modularity values range from −1 to +1, positive modularity values indicate the presence of community structure, and large positive values closer to +1 for this metric provides an encouraging sign of close-knit communities to be found in the network [6].



Network Alpha Network Beta

Figure 2.4 Illustration of Centralization measure (Adapted from [21])

There are many implementations of modularity but these are all based on the Centralization measure. The two networks above provides an illustration of Centralization score, that is, whether other nodes surround a central node. The numbers in the nodes denote the degrees of those nodes. Normalising the degrees by diving by the number of neighbours less 1 would give the Centrality measure of that node. The Centralization of the network is calculated by summing the difference between the maximum centrality in the network and the centrality of each node, and normalising it by diving by (N −1)(N-2) where N is the number of members in the network. It is very similar to the calculation of standard deviation, and the Centralization

measure captures the inequality of centrality in the nodes [21]. For the two networks above, Network Alpha would have a higher Centralization measure (1.0) compared to that of Network Beta (0.167). Thus the Centralization measure would give a more certain indication of the presence of a community in Network Alpha.

There are many ways of detecting communities. Traditionally, the number of expected communities would be established and the application would find this number of communities. In recent years, the trend has been to allow the network to 'reveal' the community structure naturally [21]. This is the approach adopted in this research. We have no idea of the number of communities we would find, or if any at all. Fortunately, Gephi works on the principle of discovering communities naturally.

Most of the algorithms for discovering communities rely on breaking certain high betweenness link to create tightly-knit groups.

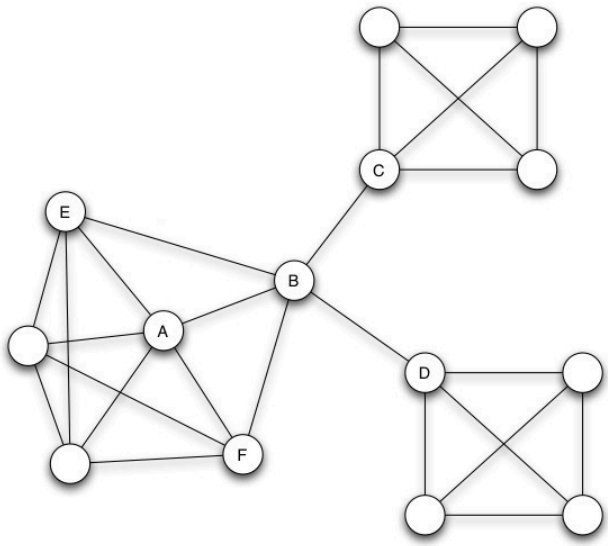


Figure 2.5 Illustration of Community detection (Adapted from [3] p.65)

In the diagram above, if links between B-C and B-D are severed, we would have three communities.

Gephi operationalises this concept by allowing us to separate the nodes in the network by a number of metrices, modularity being one of them. It uses the algorithm developed by Blondel

et.al. [23] which seeks to partition the network into highly cohesive subunits based on modularity optimization.. For visualisation purposes, members of the delineated communities can be coloured similarly to distinguish them from members of other communities, making it easy to identify the location of the community and its constituent members. Further, in a graph with numerous edges between nodes, Brath and Jonker suggested that the edges be removed from the graph for easy identification of the community members [5].

Hubs and authorities

In a directed network, certain websites make more than average linkages to other sites. We call these sites hubs [4], and act as large directories to guide information search. There are other websites that receive more than average linkages from other sites, mainly because these contain authoritative pages of information. These sites are known as authorities, giving the idea that their contents are relevant and credible [4]. The diagram below provides a pictorial representation of hub and authorities.

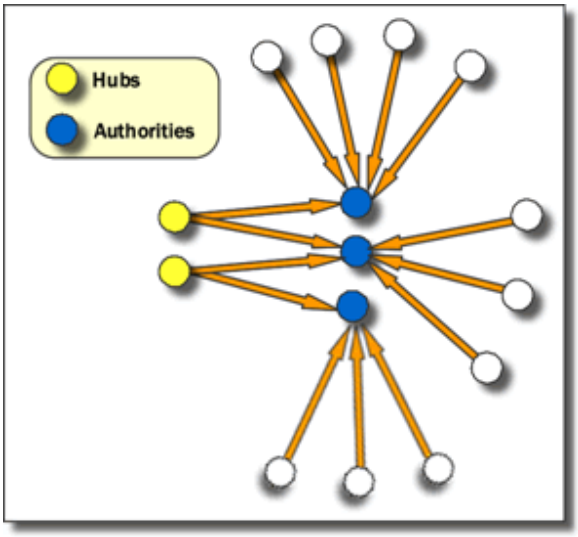


Figure 2.6 A diagram showing hubs and authorities (Source: [19])

Kleinberg [16] put it in this way: “Hubs link heavily to authorities but hubs themselves have a few incoming links and authorities may well not link to other authorities”. Thus, knowing the network structure provides much useful information about the linked environment [17].

Seeing that hubs have many outlinks and authorities many inward links, a naive way to detect these hubs and authorities is to use the out-degree and in-degree measures respectively. However, this would ignore the quality of the links.

To remedy the problem, Kleinberg [16] provided an iterative way of rating the websites in a linked environment in terms of how well they perform as hubs and authorities. This idea behind this measure is that quality of links is more important than the quality. For example, a citation of an article from a well-known author imparts much more valuable credence than numerous citations from students. His algorithm, Hyperlink-Induced Topic Search (HITS), assigns scores to indicate which websites would be regarded as hubs and which as authorities. Knowing this guides a user to know where to go in order to find information on the Internet.

Google's PageRank uses a similar algorithm in its effort to index pages to provide results for a search query. Its iterative algorithm progressively assigns weightings and measures to webpages to evaluate its relevance and popularity in answering queries [18]. It should be noted that Kleinberg's HITS algorithm places more emphasis on getting relevant webpages on a particular topic for the query than considerations of efficiency [16] whereas Google's PageRank algorithm emphasizes efficiency in its effort to pre-produce an index in anticipation of queries.

Visualising the weblinks using Gephi

Just knowing that one website has a link to another is useful but not very informative. A better picture would emerge if we know what other websites are linked to that same one, as the previous illustrations have shown. To do this, a graphical portrayal of the linkages is required [3][5][11]. In this study, Gephi is used for this purpose.

A graph refers to the pictorial representation of connected things and their relationships with each other. Humans have evolved to perceive and comprehend information visually. By displaying the results in picture rather than a table, visualisation reveals relationships in the data that is not easily discernible otherwise [5][16][17]. Whereas one would need to draw the graph by hand previously, the availability of more powerful computers with graphics capability and

software applications makes the task of developing such graphs much more convenient. Graphing has been thrust to the fore in this age of Big Data to provide an overview and to display the richness in the data. Truly, a pictorial graph is worth a thousand megabytes, to update a popular saying.

Various users have written plug-ins for Gephi to enhance its visualisation ability. Brath and Jonker [5] advised the use of colours, varying the size of nodes to indicate importance according to some chosen metric, using a layout that places nodes in relation to their affinity to other nodes, and the use of labels to identify the prominent nodes. In terms of positioning of the nodes, they adovocated the use of force-directed layouts which "pulls together nodes that are connected and may push apart nodes that are not connected" [5] (p.99). Heeding their advice, the following layout algorithms will be tried in Gephi: Force Atlas, Frutherman Reingold, Yifan Hu, and OpenOrd. Some artistic skills is required for this aspect of network analysis, relying on judgement and taste. The chosen layout is the one that best segment our network to permit us to look at it from various perspectives to view its structure. Also, the shedding of links to leave just the nodes behind enables us to place nodes that are strongly affiliated with a hub close together to create 'solar-system'-like maps to illustrate the discovered communities [5]. This method also allows the easy identification of members of each community.

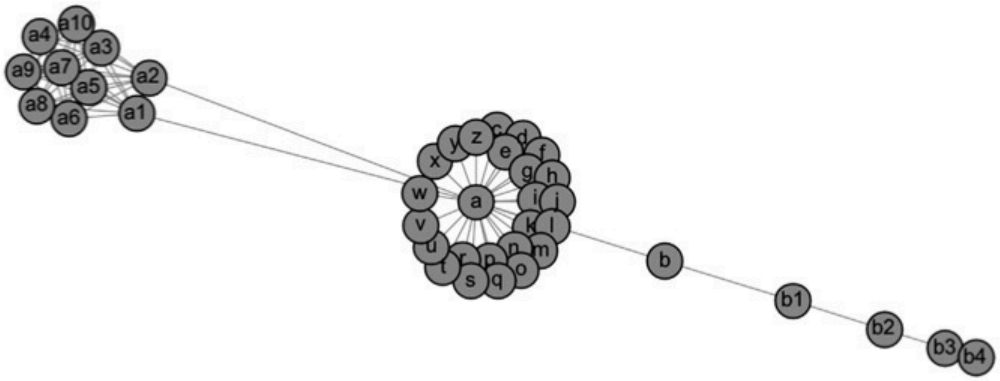


Figure 2.7 An example of a force-directed layout (Adapted from [11] p.51)

This chapter has provided the important aspects of network analysis and visualisation. These ideas will be used in the research design in the next chapter.

Chapter 3

Research Design

This chapter documents the thinking behind the experiments and how it was carried out.

The experiments were carried out on a Mac computer with a high definition screen for viewing graphs of networks. The specifications for that computer is as follows: Mac OS X version 10.8.5, 2.5 GHz Intel Core i5 processor, 4 GB of 1333 MHZ DDR3 memory, and AMD Radeon HD 6750 M 512 MB graphics.

Brath and Jonker [5] provided a series of steps to follow to proceed from raw data to the graphical output. These guidance steps can be succinctly summarised as the 3Cs: Collect, Clean, and Connect. In the Collect step, consideration is given to the source of the data and the best way of extracting it. The next step involves the Cleaning of the collected data. This entails looking for completeness of data and taking steps to rectify any inconsistencies in the data. The final step, Connect, involves turning the data into a graphical form. Depending on the graphical application that is used, one must prepare the data in a format that is readily imported for use.

Collect - Getting weblinks

The raw material for this research is New Zealand websites. We need a tool for harvesting these websites and Scrapy (http://scrapy.org) was chosen for this purpose. As Scrapy (version 0.24.4) is Python-based , it was used in relation to Python (version 2.7.9) to collect and process the harvested websites data. Scrapy was chosen for its popularity as a web harvesting tool and because it is open-source software. In addition, using a Python library provides a good balance between ease of development and the flexibility required for a web crawler. It uses a mechanism based on XPath - a language for expressing paths in an XML document for navigational purposes [24]. This property facilitates Scrapy to extract a complete subtree of the website it visits.

A start source - or seed - is needed for collecting those websites. The quality of harvested websites depends on the seed selection [5]. After analysing a few sources, we decided on http://www.nzpages.co.nz/. This "Everything New Zealand" site contains a directory of New Zealand websites with various website categories, playing a similar role to the Yellow Pages for phone numbers.

Using a specially written Python script, the spider was tasked to identify all web linkages in the seed. This harvested list - stored in *collector.csv* - will serve as the starting point for the crawler to visit other websites.

Using the list above, two crawls were made. The first crawl follows each link at a time, attempts to reach a different website, and returns a list of the directed linkages (represent by a set of source and target URLs) . The second crawl repeats the task of the first crawl and records the parsed data in a different *visitor* file. There is a reason why two crawls were chosen when one would suffice. In previous trials of the crawling, many impermanent links were encountered and these were associated with advertisements which complicated the depiction of the web. Kleinberg mentioned the problems with advertisement links but did not offer a solution for solving it [17]. The two-crawl model remedies this problem. When the results of the two crawls are processed, any linkages that are not present in the both the crawls are considered as unconfirmed edges, and these are discarded. Linkages that appear in both crawls are permanent non-advertisement links and these end up in the *confirmed_edges* file.

The task of finding linkages in the two crawl was performed by a Python script that was written for this purpose.

It is worth re-iterating that for each linkage, the spider was to obtain links to a website located outside the originating domain. In relation to this, the spider is allowed to make a maximum of 15 hops from the starting point and record the locations it encountered Also, in order to obey the rules of any websites that prohibit web crawling, a flag (ROBOTSTXT_OBEY = True) was inserted in the *settings* file to prevent Scrapy from harvesting from a website that specifically prohibits this exercise.

Clean - Obtaining stable linkages and Cleaning the data

There are two stages of crawling enable us to get rid of those non-permanent advertising links. The two visitor data files were combined, duplicates removed, and cleaned to produce a data file which contains a list of confirmed-edges, that is, linkages from one website to another. A number of cleaning processes were involved. These include:

Mapping web links to their domain names. For example, links such as http://www.stuff.co.nz/page_abc.html and http://www.stuff.co.nz/page_xyz.html would be reduced to http://www.stuff.co.nz. Doing this would reduce the number of nodes and should improve the quality of the data input file.

Removal of non-sensical linkages. Previous trial runs of the crawl brought back links to phone numbers and other broken links. These would be removed to improve to the quality of the resulting data file.

Connect - Portraying the results

Gephi will be used to produce our network graphs. In addition to putting the linkages in a picture, Gephi has facilities for calculating network statistics and partitioning the universe of linkages into communities. Useful network statistics like degree, modularity, HITS, and PageRank is calculated for our network. This provides us with an idea of how tightly the linkages are formed. Partitioning helps to discern communities, and Gephi's ability to colour nodes belonging to the same community makes such discovery easier.

Once the data is imported into Gephi, the following procedures were carried out to generate information about the network:

- Record number of nodes and edges;
- Calculate average degree to determine how well connected are the nodes;
- Calculate modularity to determine if communities are present;
- Calculate Kleinberg's Hyperlink-Induced Topic Search (HITS) metric to determine if there are hubs and authorities present.
- Portray the communities, if any, using the layout facilities in Gephi;
- Identify the members of the prominent communities and give a semantically meaningful name to each community.

In laying out the network, various layout algorithms will be employed to see which one portrays the network and its characteristics best. Segmenting the nodes of the network using colours will also be used. Filtering techniques, such as the exclusion of nodes and linkages below a chosen threshold, will be kept in mind to remove clutter from the network to enable it to reveal its underlying characteristics.

Chapter 4

Results

Processing of data files

From the start source (or seed), 1,166,103 starting linkages were harvested. This formed the starting point for the two subsequent crawls to start from each of these nodes and follow each directed link until it reaches a domain that is different from the one it began.

The crawler was allowed to run till completion. The two crawls took about two days to complete. Here is an indication of the sizes of files that were involved:

collector	226 MB
visitor1	9 GB
visitor2	23 GB

The two visitor files - containing results from each of the two crawls - were then processed to extract permanent linkages. The Mac computer used for this processing suffered from a low RAM memory capacity. The smaller visitor file took about a day and a half to be processed, and the other bigger file took more than two days. At the completion of processing, there are two categories of edges: confirmed edges and unconfirmed edges. Confirmed edges were retained while unconfirmed edges (resulting from impermanent advertising links) were discarded.

Here are the statistics from processing the two visitor files from the two crawls:

	Rows	Unconfirmed edges	Confirmed edges
visitor1	63,400,000	159,323	104,311
visitor2	126,100,000	232,337	204,985

Table 4.1 Statistics relating to the processing of the two crawls

After processing the visitor files, a confirmed_edges data file (about 10 MB in size) was obtained. This comma-separated file was then converted into a .gdf file for use in Gephi.

The resulting network has a total of 148,683 nodes. The following shows some of the statistics obtained in Gephi:

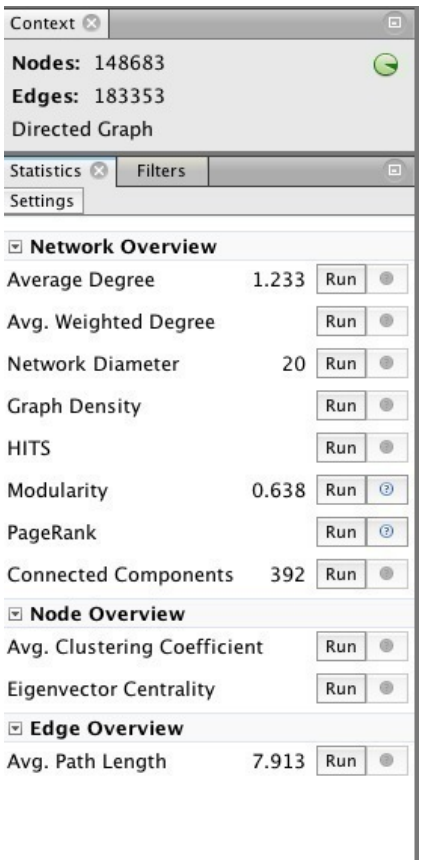


Figure 4.1 Initial statistics relating to the network

This network has 148,683 nodes and 183,353 edges. This means that the average degree is very low, showing that each node is connected to less than two others on average. The fragmented nature of the network is confirmed by the Modularity (at less than 1.0) and the high number of connected components. On average, it takes about 8 hops to reach another website.

An analysis of the nodes in relation to their top-level domains reveal the following:

	Count	Percentage
.nz	32,385	21.78%
.com	106,916	71.91%
.org	6,751	4.54%
Others	2,631	1.77%
	148,683	100.00%

Table 4.2 Analysis of top-level domain nodes

Slightly more than a fifth of the websites are based in New Zealand. In comparison, the French web harvest exercise reported about 36% of their websites were hosted on .fr top level domain [2]. Looking further into the .com domains, we found the vast majority related to *tumblr* and special interest websites that have only one linkage to it. These nodes complicate the layout of the network and a decision was made to remove these. In addition the small percentage of .org and other top-level domain name nodes were also removed, leaving us with only the .nz nodes. We justify this by relying on the statement that web users in New Zealand overwhelmingly trust .nz sites compare to other top-level domains [15]. The remaining nodes were further cleaned by removing nodes that consisted of phone numbers, leaving us with a more manageable 32,049 nodes. The statistics relating to this pruned network shows improvements in many areas and this is shown below:

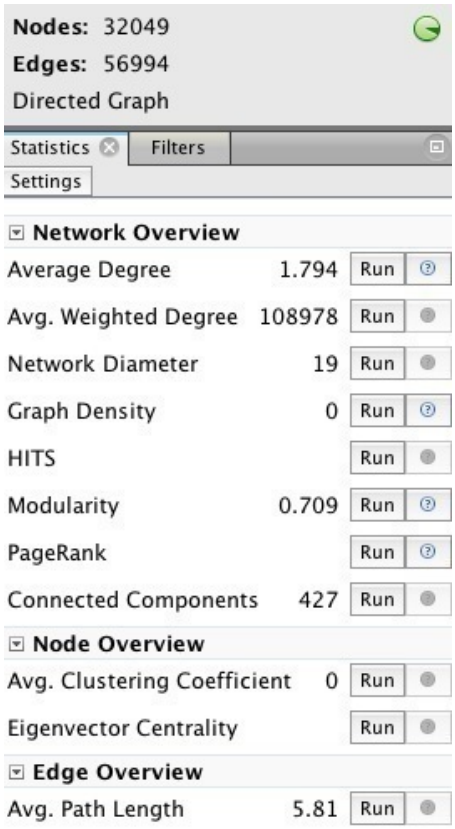


Figure 4.2 Statistics relating to pruned network

The average degree and Modularity has increased. Also, the average path length shows that it requires fewer hops to move from one node to another. The number of connected components has increase. Gephi, using Tarjan's depth-first search algorithm [22], reported 427 weakly connected components and 29,797 strongly connected components. The fact that there are many more strongly connected components and weakly connected ones contributed to the relatively high Modularity score. Thus, it is more probable that some distinct communities would be found in our network.

The statistics relating to HITS and PageRank which measure the importance of websites that act as hubs and authorities, were also calculated for the nodes.

Modularity

In total, there are 351 Modularity classes where each Modularity class contains nodes that have the same modularity measure. The increase in modularity in the pruned network suggest that we are more likely to discover distinct communities compared to the initial network. This means that we can go ahead to perform some community detection.

Before doing community detection, a test is undertaken to determine if the modularity exhibited by the websites in this network follows the power law. If we were to detect a few distinct communities, we would expect to see a few of modularity classes having a high membership and the majority having low ones. The following plot provides some evidence that preferential attachment could be working in this network.

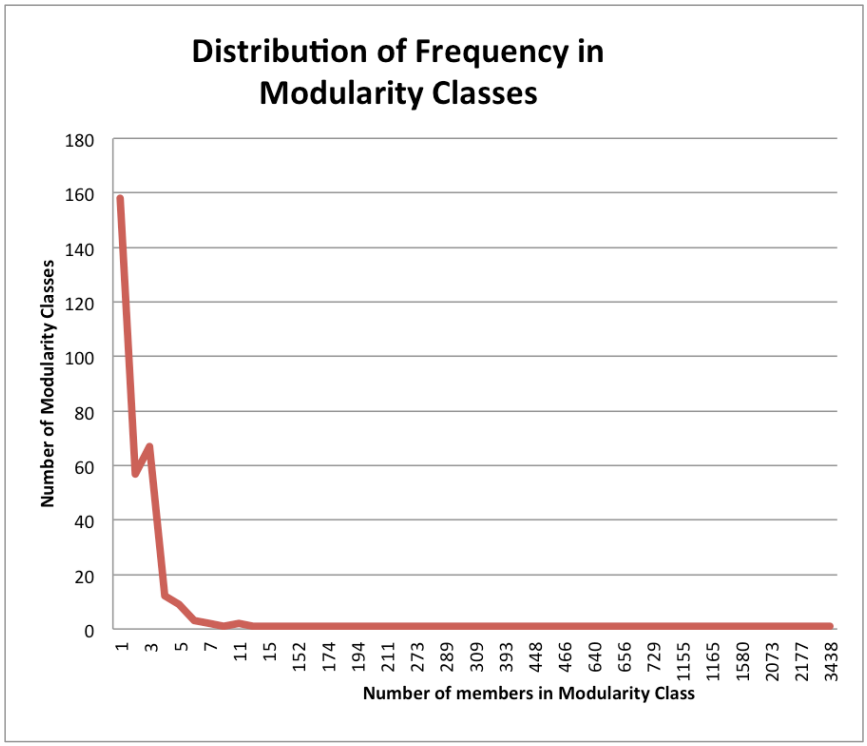


Figure 4.3 Distribution of frequency in Modularity Classes

First, an analysis will be made of those websites that exhibit high in-degree and high out-degree to reveal those websites that can be considered as potential authorities and hubs respectively.

In-Degree

Websites that have a high in-degree receive a proportionately larger number of links from other websites compared to the average. These can be considered as authorities. This research revealed the following websites that have a high in-degree measure:

Nodes	In-Degree
www.stuff.co.nz	190
www.nzherald.co.nz	171
www.legislation.govt.nz	161
newzealand.govt.nz	155
www.google.co.nz	126
www.nzta.govt.nz	105
tvnz.co.nz	105
www.business.govt.nz	103
www.doc.govt.nz	99
www.trademe.co.nz	95
www.ird.govt.nz	91
www.minedu.govt.nz	89
www.ccc.govt.nz	88
www.tripadvisor.co.nz	86
www.beehive.govt.nz	85
www.immigration.govt.nz	81
www.health.govt.nz	81
www.dia.govt.nz	77
www.radionz.co.nz	76
www.mfe.govt.nz	75
www.justice.govt.nz	74
www.dbh.govt.nz	73
www.stats.govt.nz	70
www.3news.co.nz	70
www.dol.govt.nz	68
www.avatar.co.nz	65
www.newzealand.govt.nz	64
www.nzqa.govt.nz	64
www.webgenius.co.nz	63

Table 4.3 Nodes with high In-degree measures

Websites with high in-degree are predominantly news and information portals. Also, there are many government department websites in this list. Due to their service providing nature, it is of no surprise that many other sites have links to these.

Out-Degree

The websites that link to other websites often can be considered as hubs. The following websites reported a high out-degree measure:

Nodes	Out-Degree
www.nzpages.co.nz	4118
www.new-zealand-focus.co.nz	1863
www.gopher.co.nz	1547
yellow.co.nz	1390
www.ohbaby.co.nz	817
www.kidspot.co.nz	603
www.tourism.net.nz	577
www.health.govt.nz	569
www.autobase.co.nz	447
www.grownups.co.nz	433
harcourts.co.nz	405
www.mfe.govt.nz	367
www.consumer.org.nz	346
www.thebigidea.co.nz	308
ccc.govt.nz	304
ecan.govt.nz	300
www.careers.govt.nz	282
sumnercommunity.co.nz	282
www.canterbury.ac.nz	278
www.scoop.co.nz	274
www.avatar.co.nz	255
www.nzrentacar.co.nz	234
www.grabone.co.nz	233

Table 4.4 Nodes with high Out-degree measures

Hubs obtain their information from authorities. It is heartening to see that our seed page - nzpages - scores highly in this metric. These websites can be considered as directories for users to find information online. The fact that Yellow Pages, Tourism, Consumer Institute, and various local and central governmental agencies are found here means that consumers trust these websites to point them to other sources of information.

It is heartening to discover that nodes do not appear in the both list of websites that have high in-degree and out-degree. This means that nodes either act as directory of information or sources of information, but not both roles.

The New Zealand web topology

Various attempts were made to portray the nodes and their edges using Gephi. The layout was a time-consuming task and it was run overnight. The layout below shows the structural relationship within the New Zealand web developed using PageRank as the partition criteria and Force Atlas 2 as the layout algorithm:

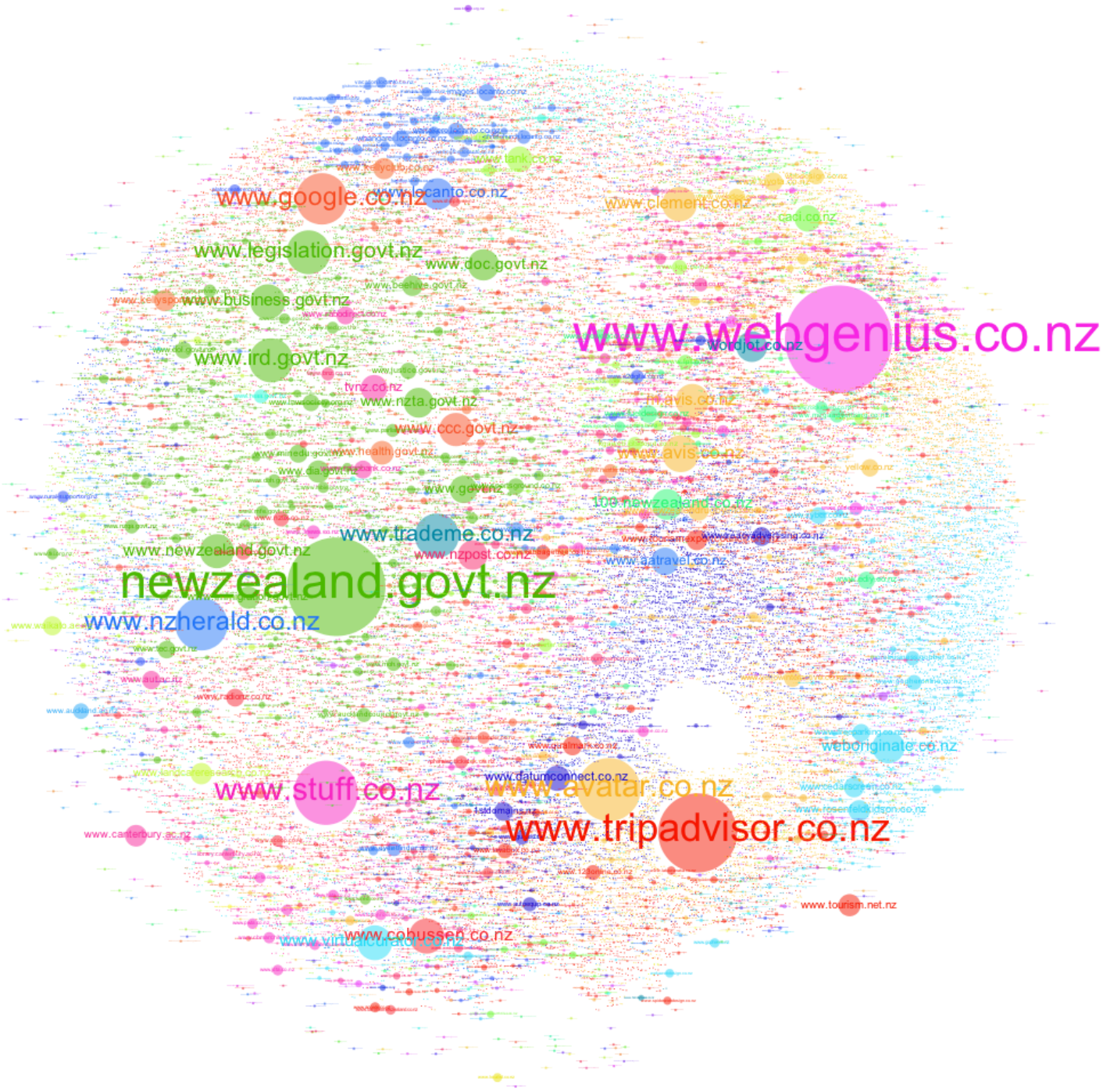


Figure 4.4 An overview of the New Zealand web topology

Another view of the topology is included in the Appendix.

Using the advice of removing linkages and the use of colour to identify each community in order to provide a better picture of the prominent websites and their satellite constituents [5], the picture above depicts a bird's-eye view of the New Zealand web topology. There are several prominent websites and these are list in terms of their significance below:

Circle colour	Radius	Website
#f725e3	140.000000	www.webgenius.co.nz
#4bb800	128.547640	newzealand.govt.nz
#f51600	103.059170	www.tripadvisor.co.nz
#f725c1	83.660570	www.stuff.co.nz
#f7b325	82.336220	www.avatar.co.nz
#2576f7	68.516800	www.nzherald.co.nz
#f74e25	67.191956	www.google.co.nz
#0086a3	57.568455	www.trademe.co.nz

Table 4.5 Prominent websites in the New Zealand web topology

Each of these prominent websites will be analysed in turn. Using a selection of the websites in the group, an attempt is made to give each group a semantically meaningful label.

The *webgenius* group

Circle colour	Radius	Website
#f725e3	140.000000	www.webgenius.co.nz
#f725e3	16.927338	www.platocreative.co.nz
#f725e3	16.058693	www.bni.co.nz
#f725e3	8.867868	www.tabletalker.co.nz
#f725e3	8.250752	www.displaytactix.co.nz
#f725e3	8.192051	www.mckenzieheating.co.nz
#f725e3	7.898613	carbonzero.co.nz
#f725e3	7.779071	www.business-card-holder.co.nz
#f725e3	7.779071	www.donation-box.co.nz
#f725e3	7.779071	www.ballot-box.co.nz
#f725e3	7.778534	www.kitchensrevamped.co.nz
#f725e3	7.770991	www.buoy.co.nz
#f725e3	7.770991	www.capitalchemdry.co.nz
#f725e3	7.770991	www.youthquest.co.nz
#f725e3	7.770991	www.furnitureupholstery.co.nz
#f725e3	7.770991	www.murdochjames.co.nz
#f725e3	7.770991	www.maxicarpetservices.co.nz
#f725e3	7.770991	www.kapitirunforyouth.co.nz
#f725e3	7.770991	www.septictankservicesauckland.co.nz
#f725e3	7.770991	equilibriuminteriors.co.nz
#f725e3	7.770991	www.zebrano.co.nz
#f725e3	7.770991	www.kungfucatering.co.nz
#f725e3	7.770991	www.raygas.co.nz
#f725e3	7.770991	www.actionservices.co.nz
#f725e3	7.620122	www.postcard-stand.co.nz
#f725e3	7.620122	www.brochureholder.co.nz

Table 4-6 Top nodes in the webgenius group

Webgenius is in the business of web marketing for clients. The websites in this group captures the essence of provision of goods and services. None of these businesses are household names yet. Let us call this the **SME Web Marketing** group.

There are 179 nodes in this group.

The *newzealand.govt* group

Circle colour	Radius	Website
#4bb800	128.547640	newzealand.govt.nz
#4bb800	57.279415	www.ird.govt.nz
#4bb800	57.043606	www.legislation.govt.nz
#4bb800	45.491320	www.business.govt.nz
#4bb800	44.479786	www.newzealand.govt.nz
#4bb800	40.738870	www.doc.govt.nz
#4bb800	38.870350	www.nzta.govt.nz
#4bb800	35.675370	www.govt.nz
#4bb800	29.283707	www.immigration.govt.nz
#4bb800	28.996143	www.dia.govt.nz
#4bb800	26.874165	www.beehive.govt.nz
#4bb800	23.664427	www.minedu.govt.nz
#4bb800	23.000097	www.tec.govt.nz
#4bb800	22.002827	www.dol.govt.nz

Table 4-7 Top nodes in the newzealand.govt group

A number of government department websites are included in this group. We can rightly call this the **Governmental Services** group.

This is the largest group with 3,103 nodes in it.

The main website in this group was at least twice as large as the next one. This and the layout of the members of this group point to the fact that this is likely to be a very cohesive group.

It should be noted that not all government departments and agencies are put into this group. For example, the Accident Compensation Corporation (ACC) was put in a group headed by the Ministry of Health. Other prominent members of that group included Mighty Ape and Fishpond, which are two on-line retailers. Since the Force Atlas algorithm tends to position strongly connected nodes together and push unconnected nodes apart [5][11], the placement of the two government agencies in the same group as the two big on-line retailers could mean that the emphasis has been placed on their commercial, rather than administrative, nature.

The *tripadvisor* group

Circle colour	Radius	Website
#f51600	103.059170	www.tripadvisor.co.nz
#f51600	28.870773	www.tourism.net.nz
#f51600	26.626358	www.tourismexportcouncil.org.nz
#f51600	24.784336	www.qualmark.co.nz
#f51600	19.694174	www.123online.co.nz
#f51600	19.519947	www.lavabox.co.nz
#f51600	18.996765	www.cabbagetree.co.nz
#f51600	13.192824	www.accede.co.nz
#f51600	11.960484	www.hairylemon.co.nz
#f51600	11.310234	www.rippon.co.nz
#f51600	11.296299	www.lakewanaka.co.nz
#f51600	10.806097	www.spiderwebdesign.co.nz
#f51600	10.751515	www.macstudio.co.nz
#f51600	10.744162	www.tomahawk.co.nz
#f51600	10.580923	www.jasons.co.nz
#f51600	10.570612	www.skyline.co.nz
#f51600	9.705892	www.queenstownnz.co.nz
#f51600	9.649014	www.fullers.co.nz
#f51600	9.290692	www.nzmotels.co.nz

Table 4-8 Top nodes in the tripadvisor group

There is a feeling of travel and adventure among the members of this group. Let this be the **Tourism** group.

There are 2,176 nodes in this group.

The size of the tripadvisor node dwarfed the other nodes in this group. Also, there was a large number of nodes in this group and most of these were located far away from that main node. The position of the prominent node at one end of the graph tells us that it is likely to have influence over those nodes in its vicinity (likely to the ones listed above) and not much over those nodes that are scattered far away.

The *stuff* group

Circle colour	Radius	Website
#f725c1	83.660570	www.stuff.co.nz
#f725c1	17.727957	www.3news.co.nz
#f725c1	15.711847	support.webhost.co.nz
#f725c1	14.636813	www.cbmarchitects.co.nz
#f725c1	14.416871	www.spark.co.nz
#f725c1	14.350423	www.vodafone.co.nz
#f725c1	14.216652	www.hkta.co.nz
#f725c1	14.216652	www.fultonross.co.nz
#f725c1	14.052268	www.jdta.co.nz
#f725c1	14.052268	www.bkb-ta.co.nz
#f725c1	13.883309	www.bgb-ta.co.nz
#f725c1	13.846001	www.averyarchitects.co.nz
#f725c1	13.846001	www.pwta.co.nz
#f725c1	12.431014	www.deloitteprivate.co.nz
#f725c1	9.791134	givealittle.co.nz
#f725c1	9.005492	www.radiolive.co.nz
#f725c1	8.848031	www.tv3.co.nz
#f725c1	8.533081	www.benefitz.co.nz
#f725c1	8.447660	www.mags4gifts.co.nz
#f725c1	7.965690	www.chchcasino.co.nz

Table 4-9 Top nodes in the stuff group

Similar to those members in the webgenius group, the members of this group are involved in the provision of goods and services. Unlike the webgenuis group, those included in this group tend to be larger in size and more established businesses. To differentiate this group from the other, we call this the **Large Businesses** group.

There are 870 nodes in this group.

The is a handful of modest sized nodes close to the Stuff node. Also, the majority of the relatively small number of member in this group are located close to that dominant node at the bottom left of the graph. This gives the impression of a relatively cohesive group in their own niche.

The *avatar* group

Circle colour	Radius	Website
#f7b325	82.336220	www.avatar.co.nz
#f7b325	48.045920	www.avis.co.nz
#f7b325	45.157696	www.clement.co.nz
#f7b325	43.587852	m.avis.co.nz
#f7b325	26.097002	www.newzealandrugbyinfo.co.nz
#f7b325	12.828092	www.rankers.co.nz
#f7b325	12.795945	www.specialeyes.co.nz
#f7b325	12.305384	www.redbus.co.nz
#f7b325	11.321493	www.kaiser.co.nz
#f7b325	11.307264	www.welman.co.nz
#f7b325	10.882342	www.greatnewzealand.co.nz
#f7b325	10.221101	www.energise.co.nz
#f7b325	9.963923	www.dnc.org.nz
#f7b325	9.675239	www.webfactor.nz
#f7b325	9.097757	www.metservice.co.nz
#f7b325	9.018968	www.canterburysports.co.nz

Table 4.10 Top nodes in the avatar group

The Avatar group is involved in web design and web marketing. The organisations in this group are mainly service providers. This group can be known as the **Large Service Providers**.

There are 1,155 nodes in this group.

The relatively small number of nodes in this group and the location of a small group of nodes close to the dominant node give the impression that there is a small active core of interacting nodes, while the other nodes have weaker ties to this core. The smaller size of the dominant node and the large number of nodes in this group points to the idea of a niche group.

The *nzherald* group

Circle colour	Radius	Website
#2576f7	68.516800	www.nzherald.co.nz
#2576f7	35.893066	www.aatravel.co.nz
#2576f7	20.097620	www.aa.co.nz
#2576f7	14.775651	www.tianz.org.nz
#2576f7	12.944094	www.aawebbuilders.co.nz
#2576f7	12.570397	www.odt.co.nz
#2576f7	9.046557	advertising.nzme.co.nz
#2576f7	8.927506	www.newstalkzb.co.nz
#2576f7	7.350001	locations.aa.co.nz
#2576f7	7.183907	shop.aa.co.nz
#2576f7	6.847120	www.iheartradio.co.nz
#2576f7	6.814543	www.cbre.co.nz

Table 4.11 Top nodes in the nzherald group

The New Zealand Herald is a long-established regional newspaper that has added an online presence in recent years. Its portal, with its rich news articles, has attracted a number of regional service providers. This is the **Online Advertising** group.

There are 729 nodes in this group.

The next five largest node in this group are located quite far away from the main node in this group. There seem to be a subgroup of nodes at the top and on the right hand side of the graph. It paints a picture of a group with several factions in it.

The *google.co.nz* group

Circle colour	Radius	Website
#f74e25	67.191956	www.google.co.nz
#f74e25	43.513550	www.ccc.govt.nz
#f74e25	12.928269	www.energywise.govt.nz
#f74e25	10.878243	www.restauranthub.co.nz
#f74e25	10.531792	www.yamaha-motor.co.nz
#f74e25	10.209825	www.givealittle.co.nz
#f74e25	9.363172	resources.ccc.govt.nz
#f74e25	8.804418	www.creativenz.govt.nz
#f74e25	8.632647	www.waimakariri.govt.nz
#f74e25	8.541503	www.eveve.co.nz
#f74e25	8.431177	grants.nzct.org.nz
#f74e25	7.783763	www.watersafety.org.nz
#f74e25	7.532321	www.logicstudio.co.nz
#f74e25	7.049424	www.mammothweb.co.nz
#f74e25	6.877582	www.redcross.org.nz
#f74e25	6.828199	www.cdc.org.nz
#f74e25	6.713446	www.metroinfo.co.nz
#f74e25	6.648601	www.nzct.org.nz
#f74e25	6.647950	www.christchurch.org.nz
#f74e25	6.636889	www.bluebridge.co.nz
#f74e25	6.610492	www.marinehub.co.nz
#f74e25	6.538670	www.fscl.org.nz
#f74e25	6.538587	www.nrfa.org.nz
#f74e25	6.441649	www.lionfoundation.org.nz

Table 4.12 Top nodes in the google.co.nz group

The google.co.nz group comprises a diverse range of organisations in its membership. These are organisations that are well-known in New Zealand who wanted to extend their physical presence to the Internet to endear themselves to the digital generation. These are members of the **Striving-to-be-relevant** group.

There are 1,160 nodes in this group.

The graph shows that there are not many member nodes close to the dominant node in this group. The location of the dominant node, at the top of the graph, puts it in a location far from the action. The scattering of member nodes throughout the graph suggests that this is a group consisting of diverse interests.

The *trademe* group

Circle colour	Radius	Website
#0086a3	57.568455	www.trademe.co.nz
#0086a3	40.109955	wordjot.co.nz
#0086a3	7.787227	www.holidayhouses.co.nz
#0086a3	6.563261	www.travelbug.co.nz
#0086a3	5.995778	www.netactive.co.nz
#0086a3	5.469744	www.campaignforwool.co.nz
#0086a3	5.207296	gliding.co.nz
#0086a3	5.158544	www.fatweb.co.nz
#0086a3	5.137285	www.duopublishing.co.nz
#0086a3	5.130191	pear.co.nz
#0086a3	5.114618	www.ctisales.co.nz
#0086a3	4.966476	www.claas.co.nz
#0086a3	4.565187	www.paterson.co.nz
#0086a3	4.535798	routestoprofit.co.nz
#0086a3	4.346567	cobwebs.co.nz
#0086a3	4.345383	lakehouse.smartfx.co.nz
#0086a3	4.021345	www.isuzu.co.nz
#0086a3	3.719951	my.lifedirect.co.nz
#0086a3	3.669320	www.netprophet.co.nz
#0086a3	3.627817	www.findsomeone.co.nz
#0086a3	3.609032	www.motorweb.co.nz
#0086a3	3.461438	www.nzfurniture.co.nz
#0086a3	3.453895	designdirect.co.nz
#0086a3	3.453895	www.howpres.org.nz
#0086a3	3.453895	www.pocketmachines.co.nz
#0086a3	3.453895	www.onefreeday.co.nz

Table 4.13 Top nodes in the trademe group

Trademe operates an online platform for selling goods and services through auctions. The members of this group tend to be heavy users of the auction site to sell their goods and services. This is the **Cyber-traders** group.

There are 288 nodes in this group.

The next largest node in this group was located far away from the dominant node. There are relatively few members in this group and none of them are significant in size after the first two.

Chapter 5

Conclusion

This research started by the consideration of the French national web archive project [1] and asking if an equivalent web archive exists for New Zealand. The research question was further refined to the development of a web topology to determine the producers and sources of information that has a New Zealand flavour.

The material in this report documents the process of embarking on the project and the results that were obtained. The novelty contributed by this research lies in the area of obtaining good quality web linkage data for analysis. It has been long known that impermanent advertisement links between websites often complicate and confound network analysis [21]. The two-crawls process used in this research is an attempt to overcome that problem. By not including those one-off advertisement linkages, the quality of data is enhanced and the network analysis made less demanding.

Summary of findings and further work

The results indicate that the New Zealand web is fragmented with low average degree. The high modularity score prompted further analysing of this directed network, and some prominent websites were found among the large number of communities. Eight prominent groups and their membership have been analysed and given semantically meaningful labels. The fact that the distribution of modularity of the nodes obeys the power law indicates that preferential attachment could be working in this network.

This first attempt to create a picture of the New Zealand web used just one visualisation tool – Gephi. It was chosen largely based on familiarity with the application and a good balance between easy of development and flexibility. The communities discovered were the result of

using the PageRank algorithm built into that application. It would be interesting to use other community discovering applications and algorithms to isolate the main groups in the New Zealand web.

The network topology generated in this research is a one-off snapshot of the New Zealand web carried out in 2015. Linkages come and go and it would be good for interested parties to pick up the momentum and continue to generate topologies on a regular basis. Trends from such exercises would be useful as input for the development of digital policy of this country.

This research has concentrated on those websites that have a .nz as their top-level domain name. We based out decision on the *.nz Domain Name Report 2014* [[15] which found that users in New Zealand said they trusted .nz domains more than any other top-level domains. In doing so, we have left out a few of the New Zealand websites that are hosted overseas. In particular, those with a .com top-level domain name has not been captured in this research. It would be good for some future research to sieve out those .com websites that relate to New Zealand and include these for analysis.

In conclusion, this research has provided a visualisation of the New Zealand web topology. Prominent websites have been highlighted and their constituency analysed to characterise these groups. We now have a better understanding of the Internet road map in New Zealand and the main ports of call in this hyperlinked environment.

References

[1] S Abiteboul, G Cobena, J Masanes, and G Sedrati (2002) A First Experience in Archiving the French Web, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2002, 2458,1-15, Elsevier.

[2] France Lasfargues, Clement Oury, and Bert Wendland (2008) Legal Deposit of the French Web: Harvesting strategies for a national domain, *International Web Archiving Workshop*, September 2008, Conference paper, Aarhus, Denmark.

[3] David Easley and Jon Kleinberg (2010) *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, Cambridge University Press.

[4] Jon Kleinberg and Steve Lawrence (2001) The Structure of the Web, *Science, New Series*, 294(5548), 1849-1850.

[5] Richard Brath and David Jonker (2015) *Graph Analysis and Visualization: Discovering business opportunities in linked data*, John Wiley & Sons, Inc.

[6] M E J Newman (2006) Modularity and Community Structure in Networks, *Proceedings of the National Academy of Science of the United States*, 103(23), 8577-8582.

[7] Paul Erdos and Alfred Renyi (1959) On Random Graphs, *Publicationes Mathematicae*, 6, 290-297.

[8] Mark S. Granovetter (1973) The Strength of Weak Ties, *American Journal of Sociology*, 78(6), 1360-1380.

[9] Albert-Laszlo Barabasi and Reka Albert (1999) Emergence of Scaling in Random Networks, Science, 286(5439), 509-512.

[10] Duncan J. Watts and Steven H. Strogatz (1998) Collective Dynamics of Small-World Networks, *Nature*, June 1998, 393, 440-442.

[11] Jennifer Golbeck (2013) *Analyzing the Social Web*, Elsevier Inc.

[12] Duncan J Watts (2003) *Six Degrees: The science of a connected age*, W. W. Norton & Company.

[13] Albert-Laszlo Barabasi (2013) *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, Basic Books, New York.

[14] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic (2012), The Role of Social Networks in Information Diffusion, *Proceedings of ACM WWW*, April 16–20, 2012, 10 pages, Lyon, France

[15] *.nz Domain Name Research 2014*, Colmar Brunton on behalf of .nz Registry Services.

[16] Jon M. Kleinberg (1999) Hubs, Authorities, and Communities, *ACM Computing Surveys*, 31(4es), Article No. 5.

[17] Jon M. Kleinberg (1999) Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, 46(5), 604-632.

[18] Sergey Brin and Lawrence Page (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of the Seventh International World Wide Web Conference*, April 1998, 30(1-7), 107-117.

[19] Social Network Analysis website, Available from http://www.fmsasg.com/socialnetworkanalysis/. Accessed 1 June 2015.

[20] Don Tapscott (2015) *The Digital Economy: Rethinking Promise and Peril in the Age of Networked Intelligence* – 20th Anniversary Edition, McGraw-Hill Education.

[21] Lada Adamic (2015) Lecture notes from Social Network Analysis, Coursera MOOC course, University of Michigan.

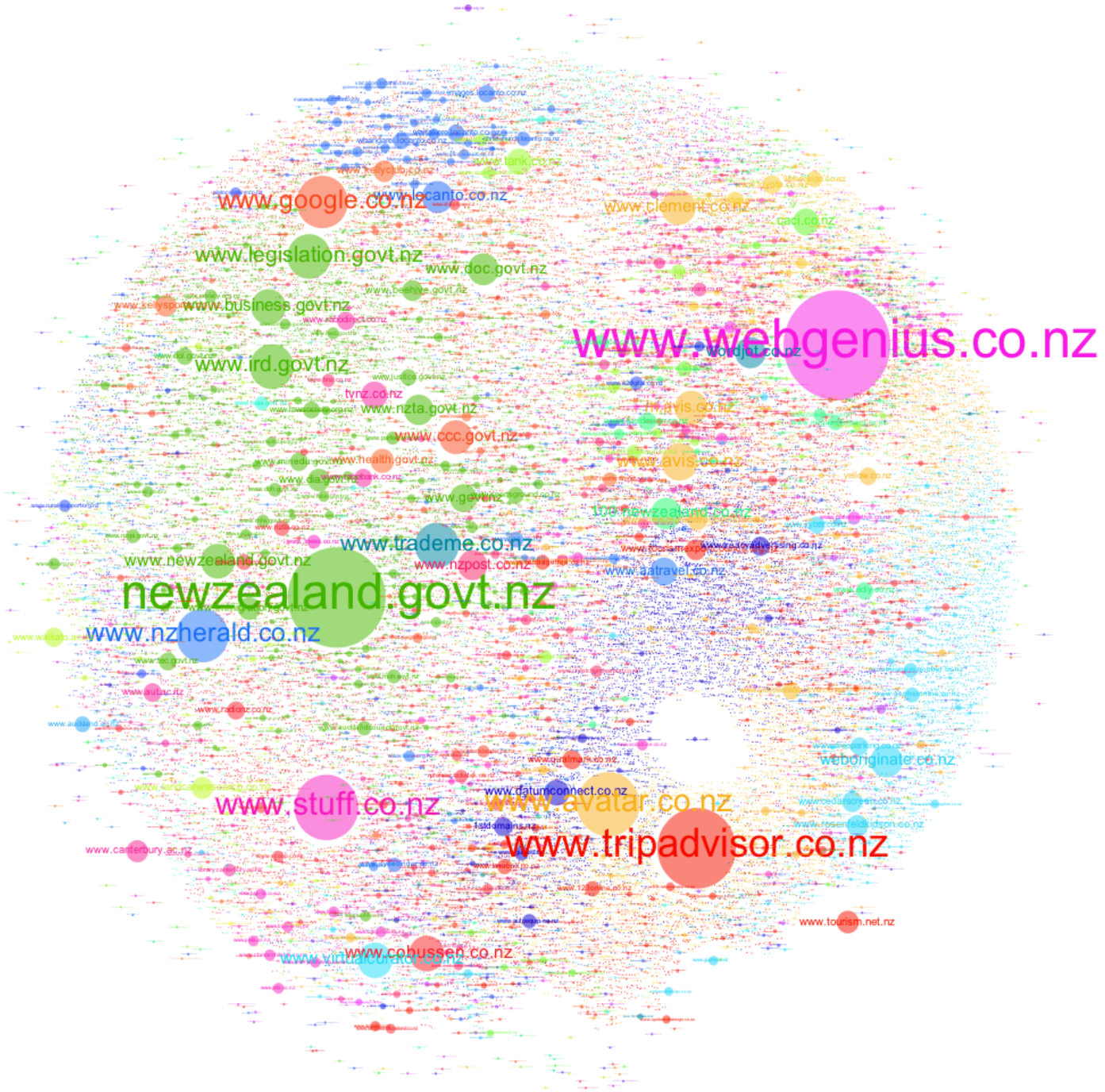
[22] Robert Tarjan (1972) Depth-First Search and Linear Graph Algorithm, *SIAM Journal of Computing*, June 1972, 1(2), 146-160.

[23] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008) Fast Folding of Communities in Large Networks, *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 12 pages.

[24] Serge Arbiteboul, Ioana Manolescu, Phillippe Rigaux, Marie-Christine Rousset, and Pierre Senellart (2011) *Web Data Management*, Cambridge University Press

Appendix

The New Zealand Web topology



The New Zealand Web Topology