DEGREES OF SEPARATION IN THE NEW ZEALAND WORKFORCE: EVIDENCE FROM LINKED EMPLOYER-EMPLOYEE DATA

Nairn W. MacGibbon

Statistics New Zealand

Abstract

Recently published analysis of messages sent over the Microsoft instant-messaging network has shown that the old maxim of six degrees of separation is not far from the truth. The idea is that, on average, you are connected by no more than six links to all other 6.7 billion people on Earth. These links can be through blood, friendship or an acquaintance – you know someone who is friends with someone whose sister is married to someone ... and so on. Using Statistics NZ's Linked Employer-Employee Dataset (LEED), this maxim is tested on a network of wage and salary earners in New Zealand. The average shortest path between employees is derived, together with a range of measures which describe characteristics of this unique view of the New Zealand labour market network.

Introduction

The idea that the world is a small place has been around for a long time. In the 1960's the psychologist Stanley Milgram, who is perhaps best known for his controversial work on obedience behaviour (Milgram, 1963), conducted an experiment to determine the average path length for social networks of people in the United States of America (Milgram, 1967). His findings indicated that on average people in the United States were separated by 5.5 friendship links. Since then, the idea of 'six degrees of separation' has achieved widespread recognition, and has given rise to plays, films[1], and games.[2] The idea has also achieved a measure of perceived prominence amongst mathematicians with the 'Erdos Number' (Goffman, 1969).

On a more serious note, the small-world network phenomenon has been the subject of a number of studies, and has implications for a range of issues from the robustness and efficiency of transportation and power networks to models of neural networks (Watts and Strogatz, 1998). More recently, attention has focused on the Internet with a recent study (Leskovec and Horvitz, 2008) of the Microsoft Messenger instant-messaging (IM) network finding that the average path length among Messenger users is 6.6.

There is also a significant body of work endeavouring to explain how knowledge is created and diffused through collaborative networks. Knowledge creation occurs when new information is integrated into the network or when the existing information within the network is recombined in new ways. A long line of research emphasizes the latter method, suggesting that the creation of new knowledge is most often the result of novel recombinations of known elements of knowledge, problems, or solutions (Schilling and Phelps, 2004). Much of this work has focused on patent registration data to proxy collaboration and knowledge. Investigations indicate that the existence of a tie is found to be associated with a greater probability of knowledge flow, with the probability decreasing as the path length (geodesic) increases (Sing, 2005).

Work has also been done to estimate measures of human capital by making use of linked employer-employee data from the US (Abowd, Lengermann, and McKinney, 2003).

In New Zealand, there is considerable interest in anything that can help productivity in general and labour productivity in particular. Given that knowledge creation and diffusion can be said to enhance efficiency and performance, and that employee networks can be an enabler of this diffusion, understanding the characteristics of the New Zealand labour market is an important first step in developing initiatives to enhance the performance of the New Zealand economy.

Using Statistics NZ's Linked Employer-Employee Dataset (LEED), I constructed an approximation of a 'knowledge network' of wage and salary earners in New Zealand. The network spanned the period 1999–2008 in an initial attempt to understand the structure and characteristics of a network view of the New Zealand labour market.

This paper is set out as follows. Section 2 describes the data source used, the definition, and the assumptions underlying the network created from this data. It also provides some base metrics describing the size of the network. In section 3, selected characteristics of the network are presented and section 4 concludes with some observations of possible interpretations and implications from this initial investigation.

Description of the data

LEED uses existing administrative data drawn from the taxation system, together with business data from Statistics NZ's Business Frame (BF). The LEED dataset is created by linking a longitudinal employer series from the BF to a longitudinal series of Employer Monthly Schedule (EMS) payroll data from Inland Revenue. The LEED initiative follows the successful development of similar datasets by a number of European and North American countries such as the US, France, Sweden and Germany.

LEED covers all individuals ('employees') who receive income from which tax is deducted at source. These payments are made by organisations that are registered with Inland Revenue. Note that the LEED data includes social assistance payments such as paid parental leave, student allowances, benefits, pensions and ACC payments, although these are excluded from the quarterly measures. For confidentiality purposes, some individuals are withheld from the data provided to Statistics NZ by Inland Revenue.

In LEED, the employer is the geographical unit or physical location of the business rather than the administrative reporting unit. For example, a nationwide retail chain may have one Inland Revenue reporting unit covering all of its retail branches. In LEED, each branch is considered to be a distinct employer. This approach has been taken to allow regional statistics to be produced. It also ensures that LEED is comparable with similar international statistics.

Network definition

In constructing the 'knowledge network' of wage and salary earners, a knowledge relationship is presumed to exist between two individuals if they both worked at the same geographic place of employment at the same point in time. This is clearly only a proxy for a knowledge relationship, as many workplaces are large and there is no guarantee that the people who share a common workplace do in fact know each other. Consequently, a time threshold has been imposed, so that the two individuals must have shared the same geographic place of employment for a continuous span of at least three months.

The network constructed is limited to wage and salary earners, and as such excludes self-employed individuals and those solely in receipt of social assistance benefits (such as ACC, Unemployment Benefit, and NZ Superannuation). The time threshold imposed also potentially excludes a subset of individuals who are engaged in seasonal or transitory short-term employment.

In network terms, each wage and salary earner is considered to be a 'node' and undirected 'arcs' are found to exist between two nodes where they have shared the same geographic place of employment for a continuous span of three months. Once created, an arc endures even after the employment relationship ceases to exist.

The network was created from the monthly LEED data spanning the period April 1999 to May 2008.

Size and algorithm performance

The knowledge network which forms the basis of this initial study was derived from a base monthly employer-employee dataset containing approximately 306 million records. A total of 2,724,725 nodes (employees) were in the network, with slightly more than 678 million undirected arcs existing between them.

Running algorithms to determine the characteristics this large network required significant computational resource and time. Table 1 provides a summary of the final run-times[4] of the various algorithms that were run (the results of which are discussed in the next section).

Characteristics of the network

For most of the analysis of the network, the focus is directed at the largest connected component in the network – the largest subset of the network where all nodes are able to be connected to one another through varying numbers of steps (arcs).

Size and distributions

Figure 1 shows the cumulative distribution of the degrees in the network, for both the largest connected component and the entire network. As can be seen, the cumulative distributions are virtually indistinguishable, with 90 percent of people having up to approximately 1,200 connections (for the largest connected component, the number of degrees was 1,201 while for the entire network it was 1,196). What is also evident is that the tail of the distribution is quite extended (hence the use of the logarithmic scale on the degrees axis), with the maximum number of degrees being just over 19 thousand.

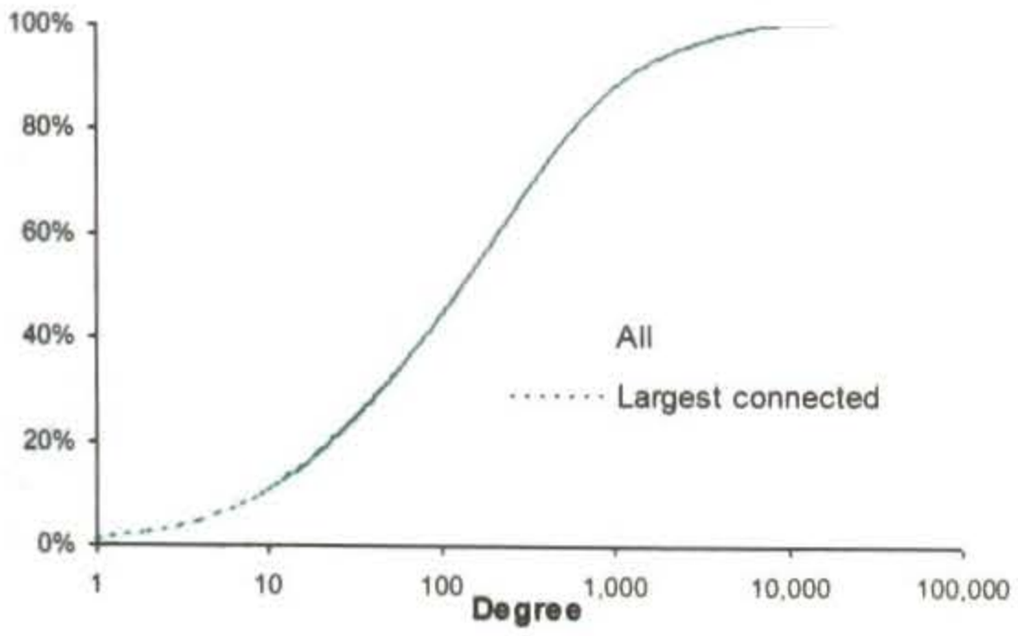
Table 1: Algorithm run-times

Algorithm	Number of observations / iterations	Run-time (hours)
Creation of base network dataset	Nodes = 2,724,745 Arcs = 678,178,460	9.5
Connected components	35,434 identified sub-components	10.5
Shortest paths	1 million pairs	57.0
Network core	Iteration through 7,200 K-cores	66.6
Network strength (random)	52 steps of 50,000 node removals	7.3
Network strength (ordered)	52 steps of 50,000 node removals	6.6
Strength – largest connected component (random)	52 steps of 50,000 node removals	103.5
Strength – largest connected component (ordered by degree)	26 steps of 100,000 node removals	59.5
Strength – largest connected component (ordered by number of workplaces)	52 steps of 50,000 node removals	73.5
Clustering coefficient	Sample of 10,000 nodes	13.4

One percent of people have in excess of 5,800 degrees, indicating that they had worked (for a continuous three-month period with this number of people) over the 1999–2008 time-span. This relatively large number is a function of number of factors. Firstly, the geographic unit structure on the BF can result in many large employers, such as some universities and district health boards, having a small number of (and in some cases a single) geographic units associated with them. Due to the way that the network has by necessity been defined, all of the people working at these large employers have been 'connected'.

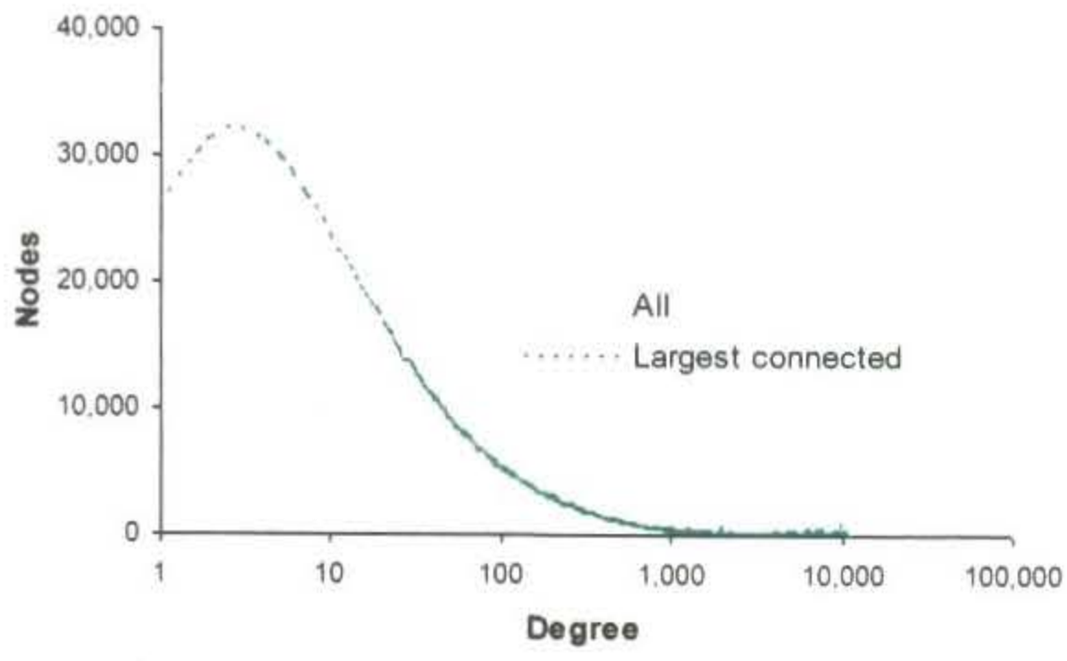
Secondly, there are a number of institutions who employ a large number of individuals and remunerate them through the EMS system, but the employees are in reality only working part-time and often only occasionally. An example of this would be a university paying a large number of student tutors for their 1–3 hours work a week over the course of an academic year. All of these student tutors are treated as being indistinguishable from the full-time teachers and other staff at the university, and so connections are established between all of them.[4]

Figure 1: Cumulative Distribution of Degrees



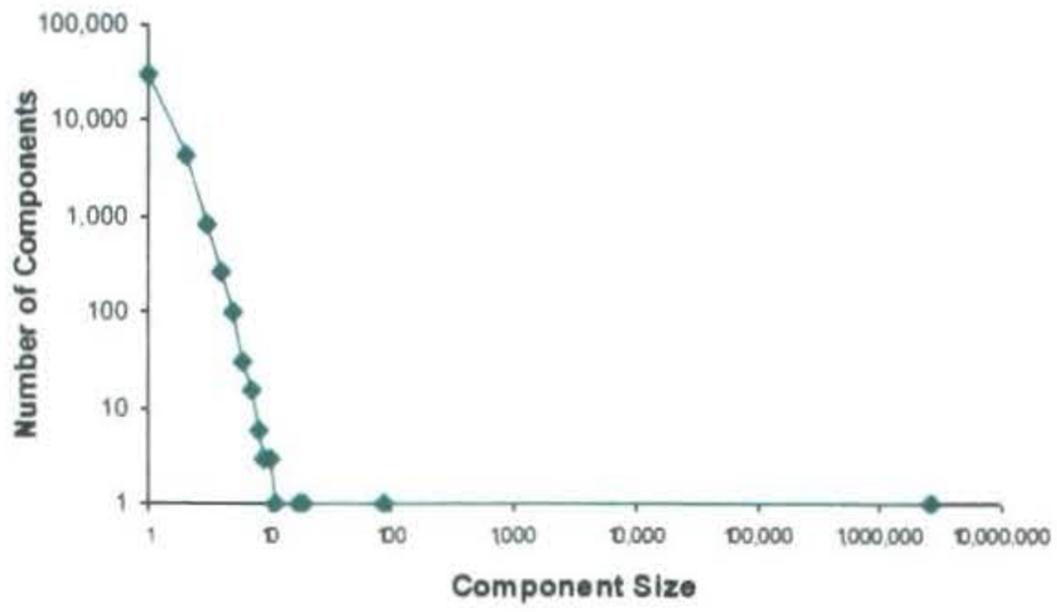
While the cumulative distributions of the entire network and the largest connected component appear largely indistinguishable, the distribution presented in Figure 2 shows that the subset of employees who are not part of the largest connected component are those with a small number of degrees. Seventy percent of this subset (approximately 9,200 people) have no connections to any other employee. These people will be employees in single-employee firms who have not worked with anyone else over the time-span.[5]

Figure 2: Distribution of Degree



The distribution of the various sub-components of the network is presented in Figure 3 and in Table 2.

Figure 3: Component Distribution



No analysis has been undertaken at this stage on the nature and characteristics of the 1.6 percent of employees who are not part of the largest connected component (sometimes referred to in network literature as 'isolates').[6]

The largest connected component in this employee network accounts for 98.4 percent of the nodes (employees) in the network. This indicates a high degree of connection, and corresponds to previous work on the LEED dataset which indicated that 99.7 percent of firms were able to be connected through observed worker-firm matches (Maré and Hyslop, 2006).[7] All further analysis of the network in this paper is for this largest connected component.

Table 2: Distribution of network sub-components

Component size	Number of components
1	29,925
2	4,248
3	826
4	268
5	104
6	31
7	15
8	6
9	3
10	3
11	1
17	1
19	1
86	1
2,681,725	1

Shortest paths

The average shortest path was calculated for a random selection of parings of employees from the largest connected component of the network. As a first step, 100 employees were selected at random (without replacement) from the universe of nodes in the largest connected component. They were defined as being the 'source' set of nodes. As a second step, a further 10,000 employees were selected at random (again without replacement) from the remaining set of nodes. They were defined as the 'target' set of nodes. Finally, the shortest path was calculated for each source-target pairing, giving total of 1 million unique pair shortest-path observations.[8]

Results from this sample indicate that the average shortest path between two randomly selected employees is 3.63 (with a sample error of 0.08 at the 95 percent confidence interval), while the mode and median of the distribution are both 4.

By definition, the distributions shown in Figures 4 and 5 are left-censored at 1 (a shortest path of zero is not possible, since the random nodes were selected without replacement), and right-censored at a number one less than the total number of nodes in the network (ie, 2,681,724). In practice, the largest path length observed in the sample of 1 million random pairings was 8.

It is possible that there are longer path lengths existing in the network. By running a second version of the shortest path algorithm over five random 'source' nodes and matching to 1,000,000 random 'target' nodes the longest observed shortest path was 11.

Figure 4: Distribution of Shortest Paths

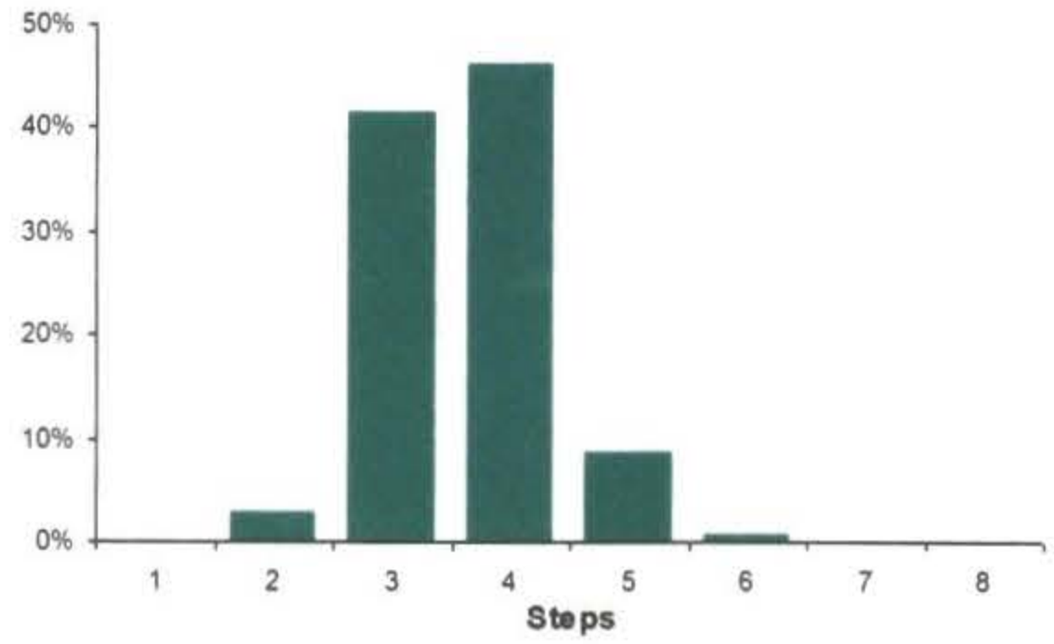
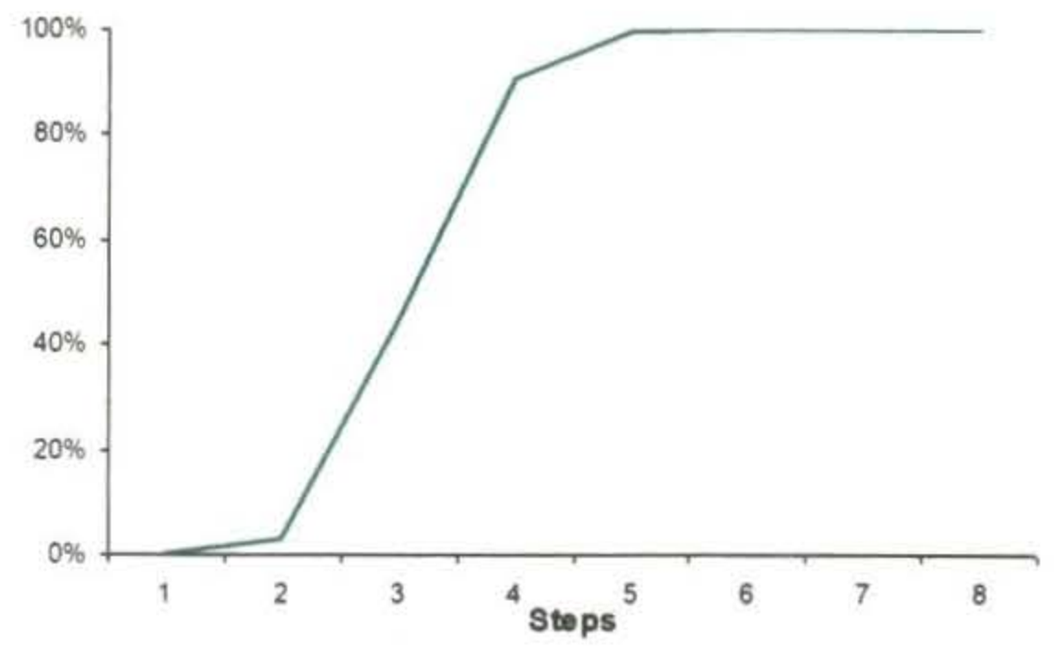


Figure 5: Cumulative Distribution of Shortest Paths



Therefore, there can be said to be four degrees of separation on average between employees in the New Zealand workforce.

Table 5 Distribution of shortest paths[9]

Steps	Proportion (%)	Cumulative proportion
1	0.02	0.02
2	2.81	2.84
3	41.60	44.43
4	46.14	90.58
5	8.72	99.30
6	0.65	99.95
7	0.05	99.99
8	0.01	100.00

Network cores

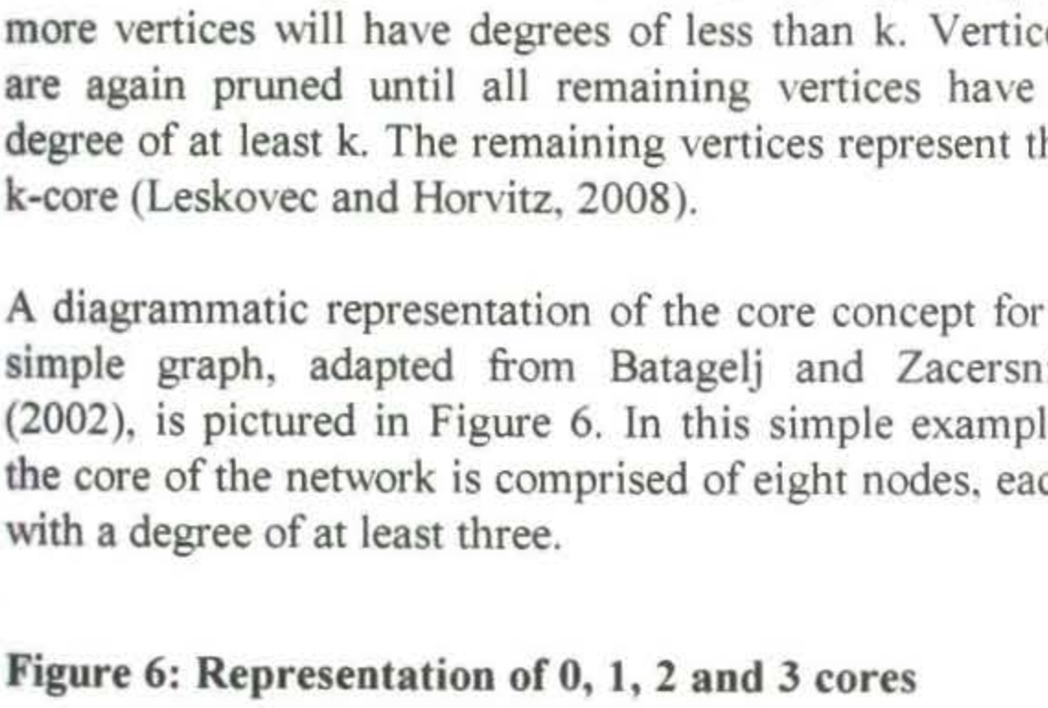
Another way of looking at the connectivity of a network is to examine the k-cores of the network (Leskovec and Horvitz, 2008). A generalization of the notion of network cores was presented by Batagelj and Zacersnik (2002) as follows:

Let G = (V, L) be a simple graph. V is the set of vertices and L is the set of lines (edges or arcs). We will denote n = V and m = L. A subgraph H = (C, |LC|) induced by the set $C \subseteq V$ is a k-core or a core of order k iff $\forall \square v \subseteq C$: degH (v) $\geq$k and H is a maximum subgraph with this property. The core of maximum order is also called the main core.

The k-core of a graph is obtained by deleting from the network all vertices of degree less than k. This process will decrease degrees of some non-deleted vertices, so more vertices will have degrees of less than k. Vertices are again pruned until all remaining vertices have a degree of at least k. The remaining vertices represent the k-core (Leskovec and Horvitz, 2008).

A diagrammatic representation of the core concept for a simple graph, adapted from Batagelj and Zacersnik (2002), is pictured in Figure 6. In this simple example, the core of the network is comprised of eight nodes, each with a degree of at least three.

Figure 6: Representation of 0, 1, 2 and 3 cores



Figures 7 and 8 plot the distribution of the number of nodes (employees) in a core of order k. The distribution has a very long tail, with the largest core comprised of one person with at least 19,618 connections. Since the k-core algorithm for the employee network has been run on the largest connected component, there are by definition no cores of zero.

This large tail (and large number of connections) is in part due to the structure of the LEED data, whereby some large employers (such as universities and District Health Boards) are represented by a single geographic place of employment. The distribution of cores decays relatively quickly up to around a k-core of 2,900 which is comprised of approximately 100,000 employees. Over half a million employees (525,615) have at least 600 connections.

Figure 7: Distribution of K-cores

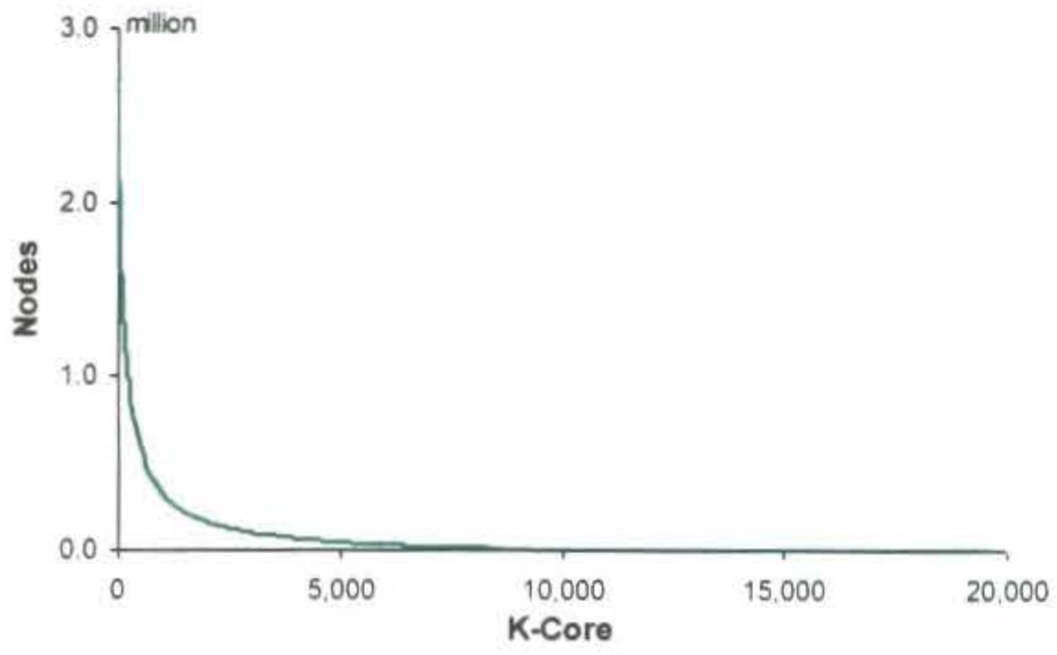
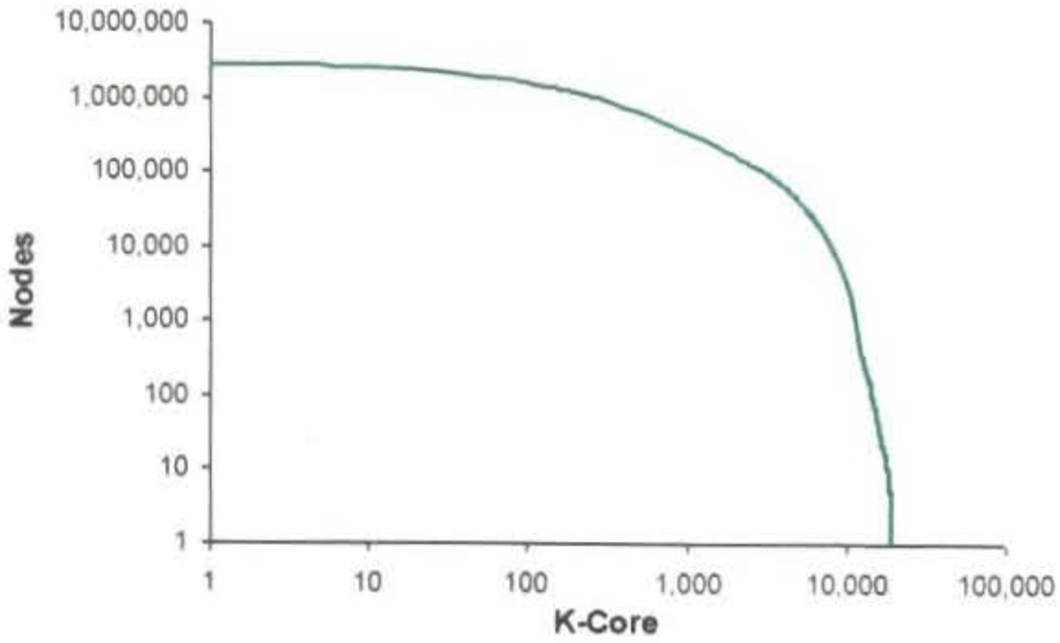


Figure 8: Distribution of K-cores (log)



Strength of the network

Another way of looking at the characteristics of the network is to consider how connected the network remains as it is subjected to 'attacks'. Albert et al (2000) observe in their study that complex communication networks display a surprising degree of robustness: although key components regularly malfunction, local failures rarely lead to the loss of the global information-carrying ability of the network. They find that such networks display an unexpected degree of robustness, the

ability of their nodes to communicate being unaffected even by unrealistically high failure rates. However, error tolerance comes at a high price in that these networks are extremely vulnerable to attacks (that is, to the selection and removal of a few nodes that play a vital role in maintaining the network's connectivity).

The recent study on the Instant Messenger network confirmed this phenomenon, observing that removing a few high-degree nodes can have a dramatic effect on the connectivity of the network (Leskovec and Horvitz, 2008).

For the labour market network under consideration here, the strength of the network was tested using a similar method to that employed by Leskovec and Horvitz. Nodes were progressively deleted from the network[10] and the relative size of the largest remaining connected component was observed (ie, the proportion of the remaining network represented by the largest connected component). While the relative size of the largest observed connected component accounts for more than half the remaining network, the sub-component is definitely the largest connected component (by definition). When the largest observed connected component accounts for less than half the network, there remains a possibility that there exists a larger, unobserved, component. Multiple iterations[11] of the search to find the largest connected component were conducted, with the largest connected component being returned.

Nodes were progressively deleted under two different scenarios. Firstly nodes were chosen for deletion completely at random, to test the effect of error on the network connectivity. Secondly, nodes were chosen for deletion on the basis of their connectivity, with the nodes displaying the greatest connectivity (ie, with the greatest degree) deleted in descending order of preference.

Figure 9: Size of Largest Component

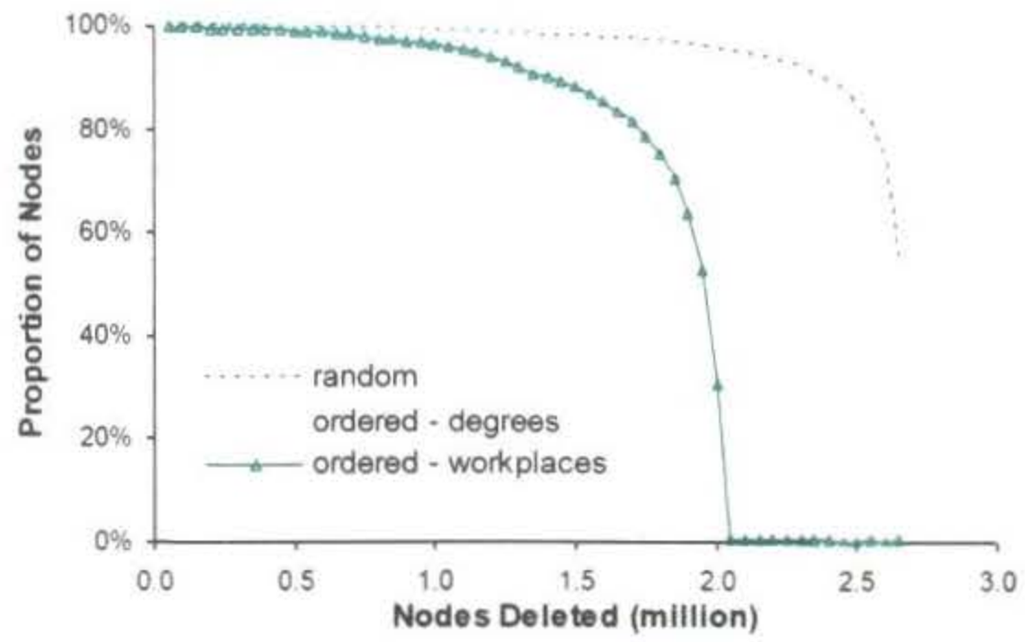


Figure 9 shows the size of the largest connected component (expressed as a proportion of the remaining nodes). The results for the random deletion of nodes return broadly similar results to the Instant Messenger (IM) study, with the largest connected component remaining after virtually all the nodes had been deleted accounting for 56 percent of the remaining network. This compares with the IM study where the largest connected component of the remaining network using random node deletion accounted for just under half the remaining network. One point of difference in our results is that there is relatively little decay in the 'connectedness' of the network until over 90 percent of the nodes have been removed (even after removing 90 percent of the nodes from the network the largest connected component still accounted for approximately 90 percent of the remaining network). This compares with the IM study, where the decay in connectedness was much more linear. This indicates a greater degree of connectedness in the employee network, making it more robust to error.

The differences in observed network strength between this study and the IM study are even more pronounced when looking at the ordered deletion of nodes (ie, 'attacks'). In the IM study, deletion of nodes in an ordered manner (based on number of connections) resulted in a relatively rapid decay in the connectedness of the network. After half the nodes had been deleted, the largest remaining connected component in the IM study accounted for just over 10 percent of the remaining nodes. This compares with the employee network, where even after half the nodes have been removed, the largest connected component still accounted for approximately 98 percent of the remaining nodes. It is only after 70 percent of the nodes have been deleted that the network 'connectedness' begins to decline, which it does so rapidly.

Table 3: Distribution of employees by number of workplaces attended

Workplaces attended	Number of employees
1	651,171
2	690,448
3	541,779
4	366,570
5	220,463
6	117,090
7	55,940
8	23,693
9	9,277
10	3,391
11	1,215
12	390
13	147
14	72
15	23
16	17
17	7
18	12
19 +	20

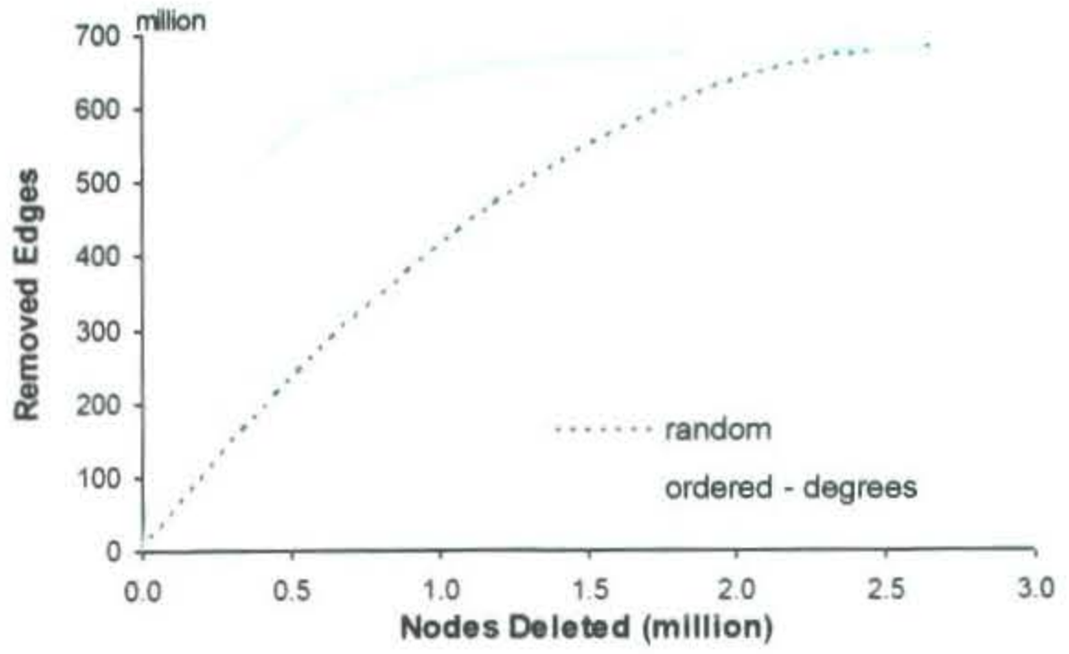
Another method of ordering the nodes for deletion was considered, whereby the employees were deleted in order of the number of distinct geographic locations (workplaces) they had worked at over the time-span considered. People who move between geographic locations serve as the 'bridges' between clusters of employees at different geographic locations, and play a key role in determining the breadth as well as the depth of the largest connected component of the network. Not surprisingly, the decay in the network connectedness was more pronounced when the number of workplaces was used as the ordering criteria. The distribution of employees by number of workplaces they were engaged in over the time-span is presented in Table 3.

Table 4: Relative size of largest connected component

Nodes deleted (million)	Ordered deletion (degrees)	Ordered deletion (workplaces)	Random deletion
0.1	1.00	1.00	1.00
0.2	1.00	1.00	1.00
0.3	1.00	1.00	1.00
0.4	1.00	0.99	1.00
0.5	1.00	0.99	1.00
0.6	1.00	0.99	1.00
0.7	1.00	0.98	0.99
0.8	1.00	0.98	0.99
0.9	0.99	0.97	0.99
1.0	0.99	0.97	0.99
1.1	0.99	0.96	0.99
1.2	0.99	0.94	0.99
1.3	0.98	0.92	0.99
1.4	0.98	0.90	0.98
1.5	0.97	0.88	0.98
1.6	0.96	0.86	0.98
1.7	0.94	0.82	0.97
1.8	0.91	0.75	0.97
1.9	0.86	0.64	0.96
2.0	0.78	0.31	0.96
2.1	0.71	0.01	0.95
2.1	0.62	0.01	0.95
2.2	0.50	0.01	0.94
2.2	0.34	0.01	0.94
2.3	0.17	0.00	0.93
2.3	0.05	0.00	0.92
2.4	0.00	0.00	0.91
2.4	0.00	0.00	0.90
2.5	0.00	0.00	0.88
2.5	0.00	0.00	0.85
2.6	0.00	0.00	0.82
2.6	0.00	0.00	0.74
2.7	0.00	0.00	0.56

Figure 10 shows the number of edges that are removed from the network under the two scenarios. As expected, the removal of nodes on a random basis results in removal of edges in a linear manner, while the ordered removal of nodes based on degrees removes edges more quickly.

Figure 10: Removed Edges



These results would seem to indicate that the employee network is not only resilient to error, but is also relatively resistant to attack. What this means is that even if the most connected people in the network were to disappear (for example, though emigration) the connectedness of the network would not be unduly compromised.

Clustering coefficient

The observation that the employment network is highly connected is further illustrated by the relatively high clustering coefficient (0.59)[12] which is observed for the network.

A clustering coefficient is a measure of the transitivity of a network. It represents how close the immediate neighbors of a node are to being a clique (ie, a complete graph). The clustering coefficient for a node is calculated as the number of links that exist between the immediate neighbours of the node, divided by the total number of connections that could possibly exist between these neighbours (Watts and Strogatz, 1998). Table 4 provides some examples of both average shortest path lengths and clustering coefficients for observed networks for comparison.

Table 5: Examples of clustering coefficients and average shortest path length

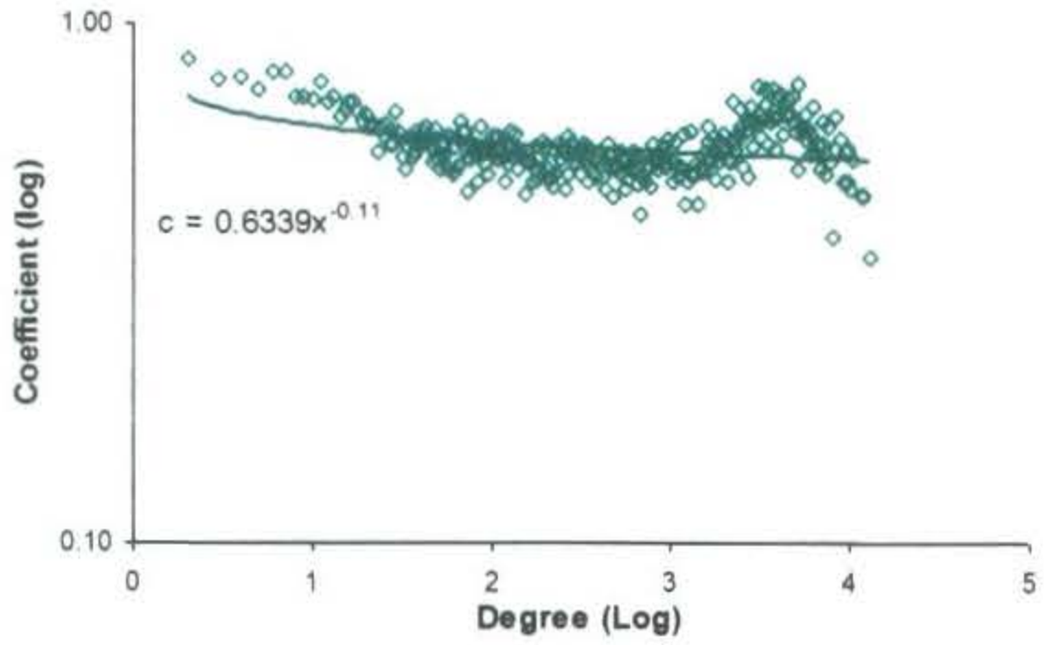
Network	Average shortest path length	Average clustering coefficient
IM network	6.6	0.137
Power grid	18.7	0.080
Film actors	3.65	0.79
C. elegans	2.65	0.28
Employee network	**3.63**	**0.59**

A high clustering coefficient is to be expected, since the definition of the network (i.e. that people are connected

if they have shared the same workplace at the same point in time) tends to enforce a significant degree of transitivity onto the network.

Figure 11 shows that the clustering coefficient decays as the degree of the node increases, although the rate of decay is relatively small (the employee network decays with exponent -0.11 compared with the IM study where the coefficient decayed with exponent -0.37).

Figure 11: Cluster Coefficient



Recent research on the role of networks in knowledge creation indicates that clustering appears to play a valuable role in the transfer and assimilation of information between nodes and that highly clustered networks (i.e. those that have a high degree of 'bandwidth') have an inherent advantage in knowledge creation (Schilling and Phelps, 2004).

Concluding remarks

The network of wage and salary earners in New Zealand displays characteristics of a 'small-world network' in that it is a sparse network with relatively short average path length, together with a high degree of clustering.

The structure of the network lends itself to the efficient creation and transfer of knowledge, and the network itself is relatively robust to both error and attack.

This paper provides a very first (and simplistic) analysis of the New Zealand labour market for wage and salary earners, exploiting the unique opportunities the LEED dataset provides.

The structure and characteristics of the network have possible implications for policy analysis and developments aimed at improving workforce productivity through understanding and enhancing the knowledge creation-enabling nature of the network.

Future possible areas of work in understanding the network of wage and salary earners in New Zealand include:

- understanding the characteristics of the employees who are not attached to the largest connected component of the network
- exploring the temporal dynamics of the network (i.e. analysing how the network develops over time)
- implementing some refinements in the specification of the network, such as through introducing the idea of variably-weighted arcs (based on time spent working together for example) to proxy the strength of the ties
- examining differing subsets of the network, based on employee demographic characteristics (for example, region, age, sex, industry of employment).

A similar analysis could also be undertaken from a firm perspective, where the nodes represent firms, and the arcs represent the movement of workers between firms over time

Acknowledgements and notes on the data

Particular thanks are due to a number of colleagues at Statistics NZ; in particular Les Cochran for his technical assistance in creating and managing what was a truly massive dataset and Mike Doherty for feedback on calculating sample errors.

The opinions, findings, recommendations and conclusions expressed in this report are those of the author. They do not purport to represent those of Statistics NZ, who take no responsibility for any omissions or errors in the information contained here.

Access to the data used in this study was provided by Statistics NZ under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act

1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person or firm. The tables in this paper contain information about groups of people so that the confidentiality of individuals is protected.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act. These tax data must be used only for statistical purposes, and no individual information is published or disclosed in any other form, or provided back to Inland Revenue for administrative or regulatory purposes. Any discussion of data limitations or weaknesses is in the context of using the Linked Employer-Employee Dataset (LEED) for statistical purposes, and is not related to the ability of the data to support Inland Revenue's core operational requirements. Careful consideration has been given to the privacy, security and confidentiality issues associated with using

tax data in this project. Any person who had access to the unit record data has certified that they have been shown, have read and have understood Section 87 (Privacy and Confidentiality) of the Tax Administration Act. A full discussion can be found in the LEED Project Privacy Impact Assessment paper, available on the Statistics NZ Website.

Any table or other material published in this report may be reproduced and published without further license, provided that it does not purport to be published under government authority and that acknowledgement is made of this source.

Notes

1. John Guare's 1990 play "Six Degrees of Separation" which was later made into a film.

2. The game "Six Degrees of Kevin Bacon".

3. The data was processed on a dual 3.66 GHz machine with 8 gigabytes of memory, running Windows Server 2003 operating system and SQL Server 2000 database.

4. A possible future refinement in defining the network could be to impose a minimum monthly earnings threshold to endeavour to exclude part-time / casual employees from the network.

5. It is important to note that since this network is restricted to wage and salary earners who are paid through the EMS system, any working proprietors are excluded from consideration. See Kelly (2003) for more information on the LEED dataset.

6. Moxley and Moxley (1974).

7. This study was based on observer worker-firm matches and interactions from the LEED dataset over the period April 1999 – March 2005.

8. Initially, the SAS Netflow procedure was trialled to determine the shortest path; however, the scale of the arc dataset meant that run-times were prohibitive given the hardware and memory available. The final algorithm used was breadth-first search variation of Dijkstra's algorithm (Dijkstra, 1959) implemented in SQL, as were all other algorithms.

9. Based on 1 million random pairings.

10. Given the size of the network, and the computational resources required, nodes were deleted in 'batches' of 50,000 (for the random test) and 100,000 (for the ordered test).

11. The algorithm to find the largest connected component was run (at each step in the deletion process) for a random 50 employees.

12. The clustering coefficients for each degree are calculated as the average of the observed coefficients for 10,000 randomly selected nodes.

References

Abowd, A., Lengermann, P. and McKinney, K. (2003), The Measurement of Human Capital in the U.S. Economy. *U.S. Census Bureau LEHD Technical Paper* TP–2002–09.

Albert, R., Jeong, H. and Barabasi, A-L. (2000) Error and attack tolerance of complex networks. *Nature*, 406:378.

Batagelj, V. and Zaversnik, M. (2002), Generalized Cores. *Journal of the ACM*, Vol. V, No. N, Month 20YY, Pages 1–8.

Dijkstra, E. (1959), A note on two problems in connexion with graphs. *Numerische Mathematik* S. 269–271.

Goffman, C. (1969), What is Your Erdos Number?. *American Mathematical Monthly*, **76**: 791.

Kelly, N. (2003), Prototype Outputs Using Linked Employer-Employee Data. *Statistics New Zealand* (available at www.stats.govt.nz).

Leskovec, J. and Horvitz, E. (2008), Planetary-Scale Views on Large Instant-Messaging Network. *World Wide Web (WWW)* 2008.

Maré, D. and Hyslop, D. (2006), Worker-Firm Heterogeneity and Matching: An analysis using worker and form fixed effects estimated from LEED. *Statistics New Zealand*, (available at www.stats.govt.nz).

Milgram, S. (1963), Behavioural Science of Obedience. *Journal of Abnormal and Social Psychology*, **67**: 371–378.

Milgram, S. (1967), The Small World Problem. *Psychology Today*, **2**, 60–67.

Moxley, R. and F. (1974), Determining Point-Centrality in Uncontrived Social Networks. *Sociometry* **Vo. 37** No. 1, 122–130.

Schilling, M. and Phelps, C. (2004), Small World Networks and Knowledge Creation: Implications for multiple levels of analysis. *New York University – Department of Management and Organizational Behaviour and University of Washington – Department of Management &*

Organization Working Paper Series, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=385022.

Singh, J. (2005), Collaborative Networks as Determinants of Knowledge Diffusion Patterns. *Management Science*, **51(5)**: 756–770.

Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**:440–442.

Author

Nairn W. MacGibbon
Senior Research Statistician
Statistics New Zealand
Statistics House
The Boulevard
Harbour Quays
PO Box 2922
Wellington 6140
Nairn.MacGibbon@stats.govt.nz