Australian Bureau of Statistics

1352.0.55.151

Research Paper

A New Analytical Platform to Explore Linked Data

www.abs.gov.au

New Issue

Research Paper

A New Analytical Platform to Explore Linked Data

Chien-Hung Chien and Andreas Mayer

Transformation Projects Branch

Methodology Advisory Committee

4–5 June 2015, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 10 AUG 2015

ABS Catalogue no. 1352.0.55.151

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Ric Clarke, Director, Statistical Informatics Section, on Canberra (02) 6252 6736 or email <analytical.services@abs.gov.au>.

A NEW ANALYTICAL PLATFORM TO EXPLORE LINKED DATA

Chien-Hung Chien and Andreas Mayer
Transformation Projects Branch

EXECUTIVE SUMMARY

The ABS is exploring semantic web techniques as one possible solution to overcome some of the challenges in the information age. Some examples include:

- increasing user demand for more timely and sophisticated statistical products;
- exploring efficient data management systems for data integration which are flexible and cost effective;
- maintaining conceptual coherence from many data sources; and
- researching new methods for the analysis and visualisation of networks in Big Data to unlock new insights.

This paper describes a prototype Graphically Linked Information Discovery Environment (GLIDE) created using semantic web techniques to better manage statistical information. The Semantic Web framework provides an alternative approach to data representation, linking and retrieval that can unlock the full potential of interconnected and multi-dimensional datasets. Instead of organising datasets in a structured row-column tabular form, the Semantic Web approach models information in the form of a network of entities and relationships. The relationships are given strong computable semantics by precisely specifying their logical properties in a machine-interpretable format. This allows computers to understand these relationships to easily explore multiple data dimensions to identify interesting data patterns and analyse the network structure of data.

This paper demonstrates one analytical application of the GLIDE by using it to derive network statistics and create models to distinguish true firm deaths from spurious ones. The ABS has an established process for identifying firm exits, but is not able to distinguish the type of exit – whether it is due to restructuring, merger/takeover or a genuine death.

The analytical results have shown that it is important to account for spurious death for statistical production. This is because failure to account for spurious firm deaths can result in continuing enterprises being incorrectly classified as firm deaths and as a result it can affect the statistical quality from the perspectives of survey frame and accuracy of the statistics.

This paper considers both multilevel and Bayesian Networks (BNs) models. Our approach applies the BNs method within a statistical framework. We have shown that BNs can handle observations with missing variables in the test data. This paper does not intend to compare both methods on the prediction outcomes. It clearly shows that it is important to incorporate network information for modelling purposes. This leads to the prediction outcomes improved substantially for both models, reaching a 95% accuracy rate.

We conclude that the semantic web is a useful approach for statistical purposes and that network analysis can be used to effectively distinguish true and spurious firm deaths.

QUESTIONS FOR THE COMMITTEE

1. Are the modelling techniques and approach proposed in this paper appropriate?
2. Is the proposed approach appropriate?
3. Are there other methods which should be considered for the research question?
4. What other considerations should be looked at for the GLIDE?
5. What future statistical techniques and capability should the ABS (or Methodology Division) consider ?

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

DISCLAIMER

The results of these studies are based, in part, on tax data supplied by the Australian Taxation Office to the Australian Bureau of Statistics under the Taxation Administration Act 1953, which requires that such data is only used for the purpose of administering the Census and Statistics Act 1905. Legislative requirements to ensure privacy and secrecy of this data have been adhered to. In accordance with the Census and Statistics Act 1905, results have been confidentialised to ensure that they are not likely to enable identification of a particular person or organisation.

Any discussion of data limitations is in the context of using the prototype linked employer-employee database for the productivity analysis project undertaken by the ABS. The discussion is not related to the ability of the data to support the ATO's core operational requirements. Any findings from this paper are not official statistics and the opinions and conclusions expressed in this paper are those of the authors.
The ABS takes no responsibility for any omissions or errors in the information contained here.

A NEW ANALYTICAL PLATFORM TO EXPLORE LINKED DATA

Chien-Hung Chien and Andreas Mayer
Transformation Projects Branch

ABSTRACT

The advancement of technology, new methods and emerging data sources have presented both opportunities and challenges to the ABS. While Big Data provides new business opportunities for statistical production, there remain some challenges the ABS needs to overcome. The ABS is exploring semantic web techniques as one possible solution to overcome some of these challenges. This paper describes a prototype Graphically Linked Information Discovery Environment (GLIDE) created using semantic web techniques to better manage statistical information. This paper demonstrates one analytical application of the GLIDE by using it to derive network statistics and create models to distinguish true firm deaths from spurious ones. The ABS has an established process for identifying firm exits, but is not able to distinguish the type of exit – whether it is due to restructuring, merger/takeover or a genuine death. This paper uses multilevel and Bayesian Network models to distinguish true and spurious firm deaths by incorporating network statistics. It is important to account for spurious deaths for statistical production to ensure data quality. The model results also perform much better after incorporating network statistics. We conclude that the semantic web is a useful approach for statistical purposes and that network analysis can be used to effectively distinguish true and spurious firm deaths.

1. INTRODUCTION

The advancement of technology, new analytical methods and emerging data sources have presented both opportunities and challenges to the ABS. While Big Data provides new business opportunities[1] for statistical production, there remain some challenges facing the ABS with the advent of Big Data, which include:

- increasing user demand for more timely and sophisticated statistical products (with finer granularity) to meet analytical needs;
- exploring efficient data management systems for data integration which are flexible and cost effective. There is an increasing need to integrate and analyse data from semi- and un-structured sources to produce relevant and timely statistics (AGIMO, 2013);
- maintaining conceptual coherence from many data sources; and
- researching new methods for the analysis and visualisation of networks in Big Data to unlock new insights.

The use of Big Data[2] collected for non-statistical purposes for statistical production is not new for the ABS. Business Activity Statement data has been used in ABS business collections.[3] Transactional data has been incorporated into the consumer price index to improve its quality (ABS, 2013). More recently, the ABS has created a prototype Graphically Linked Information Discovery Environment (GLIDE) using tax records and ABS Business Register data. This provides an example of the analytical value of data integration from different sources. The long term vision of a GLIDE would be to integrate with other administrative data, survey data (person and business level), Census data or even more unstructured data for statistical production to answer complex policy and research questions. However, as data structures become more complex and multi-dimensional, data integration becomes difficult using relational databases. Semantic web techniques allows for a flexible data structure, reusable classifications and standards, easy data exploration, network analysis, and automated machine reasoning on the graphical databases.

This Methodology Advisory Committee (MAC) paper describes the prototype GLIDE created using semantic web techniques. This prototype GLIDE, which contains multi-dimensional and detailed microeconomic data on persons and firms, is a good showcase to demonstrate how semantic web techniques can be used to better manage statistical information and exploit new Big Data sources. It exploits the

[1] They include: advances in information technology which have lowered data collection, storage, and processing costs; new sources of data and improved access to existing big data sets; the development of creative and powerful new methods to exploit 'Big data'; and opportunities for data integration to unlock new insights.

[2] i.e. sets of large business transactional and government administrative collected data.

[3] e.g. Economic Activity Survey and Business Longitudinal Database.

analytical potential of the prototype GLIDE through network analysis as well as considering statistical modelling and machine learning techniques to answer the following research question:

- Can we distinguish true and spurious firm death events from by analysing the network connections in the prototype GLIDE?

The main objectives of this MAC paper are:

- discussing the importance of the statistical issue and describing the prototype GLIDE;
- using multilevel modelling and Bayesian network techniques to detect true births and deaths.[4] This is to propose a method to improve the quality of the ABS statistics; and
- summarising the empirical findings by comparing the differences in the industry labour productivity dispersion across industries before and after adjusting for the true firm births and deaths.

The paper is structured as follows. Section 2 discusses the proposed methodologies used to detect true firm births and deaths. Section 3 discusses the prediction results and compares the labour productivity dispersion before and after adjusting for spurious deaths. Section 4 concludes and proposes future research directions. A description of the GLIDE can be found in the Appendix.

[4] These methodologies are also useful for the ABS business survey profiling activities on the non-profiled business population.

2. METHODOLOGY

This section discusses the statistical problem and methods used to distinguish true and spurious firm births and deaths.

2.1 The importance of accounting for true firm deaths

Firms can enter and exit the economy for a number of reasons – administrative, financial, structural etc. – but not all entries and exits reflect true births of new enterprises and deaths of old enterprises.

We follow the OECD definitions to distinguish true firm deaths from spurious ones, with the main difference in terms of whether the event involves another firm.

“A death amounts to the dissolution of a combination of production factors with the restriction that no other enterprises are involved in the event. Deaths do not include exits from the population due to mergers, take-overs, breakups and restructuring of a set of enterprises. It does not include exits from a sub-population resulting only from a change of activity” (OECD Eurostat, 2008).

The statistical problems associated with spurious firm births and deaths have been well documented in the literature. Failure to identify these as merely firm entries and exits (rather than true births and deaths) can result in continuing enterprises being incorrectly classified as firm deaths. This can be caused by a business reregistering with a new ABN for business reasons (Fabling *et al.*, 2008). Alternatively, new entrants may or may not indicate genuine births of new businesses. These mere entrants may be created due to business restructuring, merger acquisition or even renaming of a firm. For example, a firm selling off one of its operations may be seen as the simultaneous contraction of a large firm and birth of a smaller one (Dixon *et al.*, 2011, Criscuolo *et al.*, 2014).

Job changes at the firm level without corrections may include not only the true changes, but also spurious increases or decreases. This can bias both job and worker flow statistics based on administrative records (Seyb, 2003). Haltiwanger *et al.* (2010) highlighted this issue and suggested the importance of tracking entries, exits and continuing firms to correct for spurious births and deaths. They analysed the Census Bureau’s Longitudinal Business Database and found that previous studies showing a strong negative relationship between firm size and job growth was a product of statistical fallacies.

2.2 Challenges with measuring true firm births and deaths

Criscuolo *et al.* (2014) pointed out that, depending on the context, there is no one right concept of the measurement of entries and exits. Some challenges include:

- the definition of a true birth and death for the statistical question at hand. For example, should a spin-off from an existing firm or mergers between firms be considered as new entrants?[5] This MAC paper follows the OECD definitions and does not consider these as new entrants because they are continuing existing business activities.
- the details provided in the administrative data. Dixon *et al.* (2011) pointed out that analysis at firm- or establishment-level data may also lead to different conclusions about a firm as a result of business activities. The ABS Economic Units Model, unlike most countries, does not have an establishment. The statistical unit in ABS is the Type of Activity Unit, which has not yet been incorporated into the prototype GLIDE. This study attempts to identify true firm[6] deaths.

2.3 Different approaches for correcting spurious entries and exits

Different approaches have been used by National Statistical Organisations (NSOs) and researchers to correct for spurious births and deaths.

The U.S. Census Bureau applies an approach that combines administrative and survey data. The maintenance of the Longitudinal Business Database (LBD)[7] involves two steps for minimising spurious births and deaths. First, it uses both deterministic and probabilistic linking methods[8] to correctly link firms and establishments[9] in a particular year to those in the preceding year (Jarmin *et al.*, 2002, Jarmin *et al.*, 2004). Second, it incorporates primary data collections from the Economic Census and other annual surveys to learn about and correct for true establishment entries and exits (Davis *et al.*, 2006).

Another approach is tracking clusters of employee movements across different firm identifiers in the administrative data. If an exiting firm at time t shares a high percentage of employees with a new firm at time t+1, then the two observations are considered to be from the same business (Dixon *et al.*, 2011). Note that these methods are usually applied to a longer time series.

[5] Baldwin *et al.* (2002) highlighted that in the case of Canada, entry rates that account for mergers are higher than ones that do not because of a high level of control changes in firms.

[6] Firms are legal entities in the Australian context.

[7] Since 1972, the Census Bureau maintained the LBD. It contains rich business characteristics information such as name and address information and data on payroll, employment and industrial activity (Jarmin *et al.*, 2004).

[8] Abowd and Vilhuber (2005) discussed the probabilistic matching method to repair longitudinal firm and person records.

[9] An establishment is defined as a physical location where the economic activity occurs. A firm may have one establishment or many establishments (Haltiwanger *et al.*, 2009).

Examples include:

- Statistics Canada's Longitudinal Employment Analysis Program dataset[10] (Dixon *et al.*, 2011);
- Statistics New Zealand's linked employer employee data (LEED) which identifies and tracks movements of common employees between predecessor and successor firms to correct for spurious births and deaths (Kelly, 2003, Fabling *et al.*, 2008).
- Benedetto *et al.* (2007) followed clusters of U.S. workers as they move across administrative entities to reduce spurious business changes when studying insourcing and outsourcing firms; and
- Ibsen *et al.* (2011) followed Denmark's administrative records of persons and firms over time to distinguish true firm births and deaths from spin-offs and mergers.

This paper used a similar approach by tracking where employees go after the firm exits. However, it considers only the connection patterns to distinguish true firm death events between two financial years. For example, if employees of an exit firm disperse widely to other firms than the exit is likely to be a genuine death, but if most go to the same new firm, then this is likely to indicate a takeover, merger or restructure.

2.4 Graphically Linked Information Discovery Environment (GLIDE)

The ABS has developed GLIDE using semantic web techniques. The semantic web provides an alternative approach to data management which can be used to unlock the full potential of interconnected and multi-dimensional datasets (Berners-Lee *et al.*, 2001). Instead of organising datasets as tables with a set of variables, the semantic web models the relationships between the entities the data describes (firms, people etc.).

This allows humans and computers to understand these relationships to:

- create flexible and responsive data structures,
- apply reusable standards for multiple purposes,
- easily explore multiple data dimensions to identify interesting data patterns,
- analyse the network structure of data,
- apply automated machine reasoning and inference.

[10] Covers 1999–2008.

This has immediate advantages for working with linked data. The semantic web stores data in the Resource Description Framework (RDF), a flexible schema-free data model for data interchange and management.[11] RDF facilitates data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without changing the underlying infrastructure to support these changes (W3C, 2014). This makes it easy to incorporate new datasets, as the data sources are not "linked" in the sense of records for the same entity being combined into a single row in a table. They are instead linked virtually through their relationships in a large graph[12] database. For example, Pay-As-You-Go (PAYG[13]) data provides information on the relationships (i.e. jobs) that exist between firms and persons in the labour market.

The ABS already uses conceptual models to describe complex statistical relationships in the data. The semantic web, through the use of ontologies,[14] provides an effective way to do this. Organising data this way enables a flexible data structure which can be used to extract more information from the data. The use of ontologies also enables machines to understand the meaning of these concepts to allow fast retrieving of information (see the Appendix for details).

2.4.1 Data sources in the prototype GLIDE

This prototype GLIDE contains data from the ABS and the ATO. These include the ABS Business Register (ABSBR)[15] and Counts of Australian Businesses, including Entries and Exits (CABEE) data[16], while ATO data[17] include the Business Activity Statement (BAS), Business Income Tax data (BIT), Personal Income Tax data (PIT), and Pay As You Go (PAYG) data.

[11] RDF can be used to solve a common problem of data stored in different format (i.e. ranging from spreadsheets to files in proprietary formats) and provide means of querying and transforming data in standard models.

[12] A graph is a representation of objects that are connected by links. In other words, two things can be related through their relationships that connect them (Dataversity, 2011).

[13] See Glossary.

[14] W3C has developed technology standards e.g. the Web Ontology Language (OWL) is used to express the ontology in RDF (W3C, 2012).

[15] The ABS Business Register is a comprehensive list of organisations that are participating in Australia's formal economy, with structural and classification attributes about each organisation. It includes contact and address details for businesses, and variables such as institutional sector, industry, employment, and turnover. It provides the consistent, coherent and point-in-time survery frames for most ABS economic surveys (ABS, 2014).

[16] Please note that ABS also uses ATO data sources for these products.

[17] The ATO collects this data for compliance purposes, with statistical production not in mind.

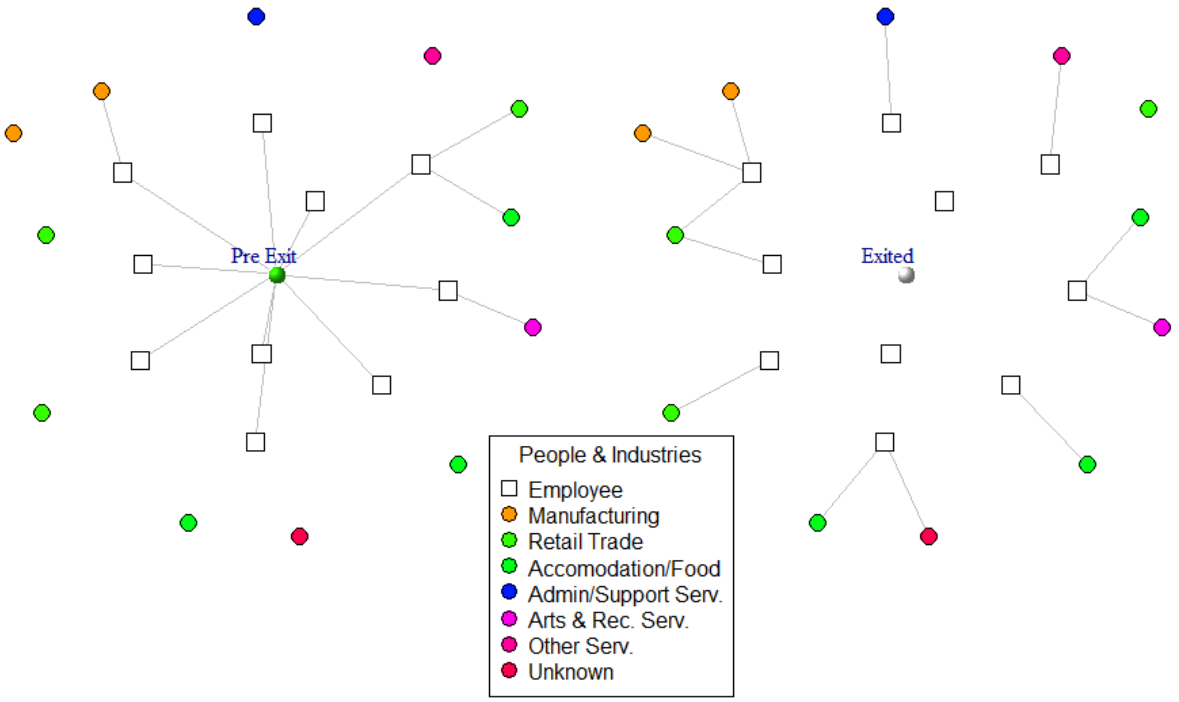
2.4.2 Querying the labour market networks

The main advantage of semantic web techniques for this paper is that it can be used to query and retrieve complex labour market network information. San Martín *et al.* (2009) suggested that using RDF provides excellent data models to process social networks on the Web. Researchers can process large scale data and extract the information underlying the network to perform structural analysis. Querying and transforming network information is a difficult task due to the intrinsic complexity of the networked data. SPARQL Protocol and RDF Query Language (SPARQL) is an integral part of the semantic web, and is an effective tool to simplify this complex task of information exploration and extraction. It can be used to perform basic graph pattern matching against the active graph specified in the query (W3C, 2013), returning the elements of the graph which match the pattern specified in the query.

The labour market network can be represented as a subset of undirected bipartite graphs between employers and employees as $\{i, j, t\}$, where $i \in I$ is the index of workers, $t \in T$ is the time period, and $j \in J$ the employer of i at time t. R_t is the employment network at time t. The vertex set is $V(R_t) = (I, J)$. An element (i, j) of the edge set $E(R_t)$ indicates that a worker i holds a job with an employer j at time t. The labour market network R accumulates information about the employer-employee link over the reference period, $t \in \{1, \ldots, T\}$, by setting $V(R) = (I, J)$ and $E(R) = \bigcup_{t=1}^{T} E(R_t)$[18] (Schmutte, 2014). Figures 2.1 and 2.2 show an example of the bipartite networks derived from the prototype GLIDE. It shows cases for a true firm death and a firm takeover (which is one example of a spurious death). For a given period, a firm can change its characteristics rapidly due to business reasons, e.g. merger or takeover, and appear as a different ABN. In comparison, employee characteristics (e.g. who they work for) do not vary substantially over the period. This provides a way to identify whether a firm is continuing under a different ABN or genuinely dead. For example, if the majority of employees of a deregistered firm are working for the same new firm in the following year, this is likely to reflect continuing economic activity. Conversely, if the employees are dispersed between a number of firms, this is likely to reflect a true firm death (Harwood *et al.*, 2014).

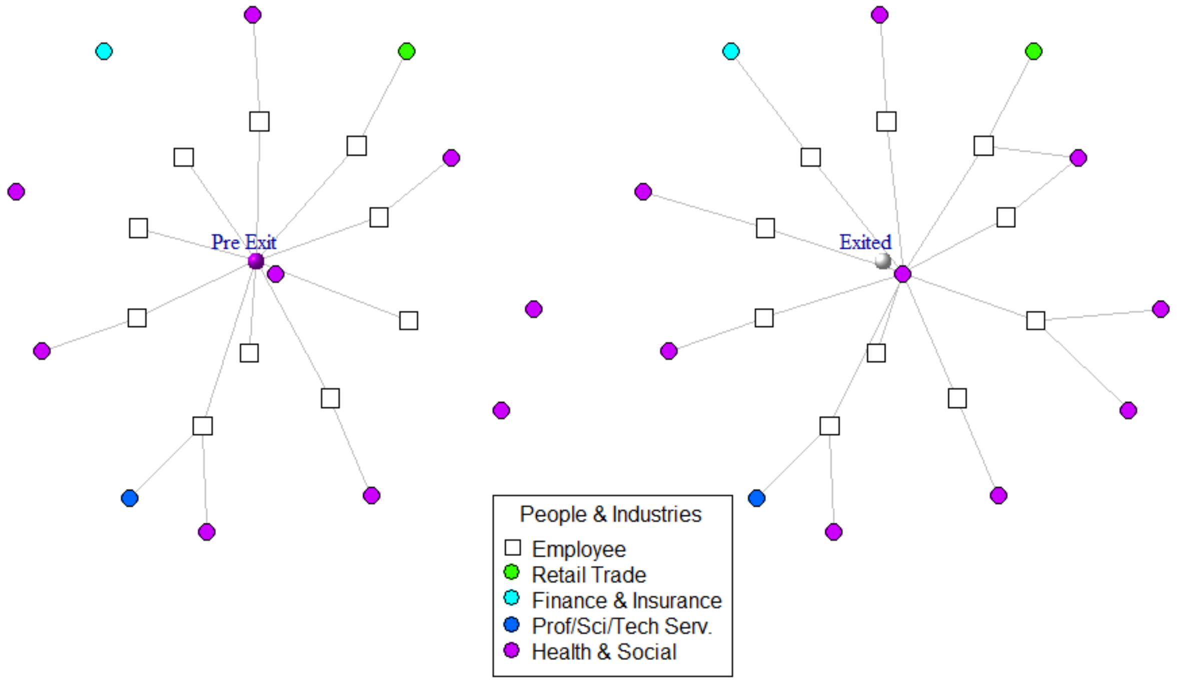
[18] ∪ means the set of those elements which are in period t where $t \in \{1, \ldots, T\}$ in all periods.

2.1 A true firm death



Source: ABS

2.2 A takeover



Source: ABS

2.4.3 Advantages and disadvantages of using GLIDE for analysis

This prototype GLIDE offers several advantages for better tracking firm births and deaths. First, the PAYG data is based on comprehensive administrative records covering a large portion of the Australian workforce. This enables analysis of flows of employees across detailed industry and demographic groups, even for regional economic shocks. Second, PAYG data also provides information on the link between employers and employees, the most valuable aspect to track the flows of employees in the labour market. Finally, the longitudinal nature of the prototype GLIDE allows employers and employees to be followed continuously and longitudinally. These employer-employee links can be used to follow employee movements over time and look for changes of firms and persons characteristics, such as location, for distinguishing between true and spurious births or deaths (Bjelland *et al.*, 2011).

There are also several disadvantages to this prototype GLIDE for this analysis. First, it contains only a short time series.[19] This is different from longitudinal data in other countries such as the United States and New Zealand which have a much longer history in the construction and maintenance of longitudinal administrative datasets. Second, it contains only low frequency data and does not measure employment at a point in time, instead measuring all jobs occupied over a 12 month period. This can create problems distinguishing between the true employer-to-employer flows and multiple job holders. Finally, the timeliness of tax records depends on the administrative processes from which it is obtained, in this case, time needed for the tax returns to be remitted and then processed before being made available to the ABS. These delays can make it difficult to correctly track firm and people links.[20]

2.5 Statistical units

The primary existing ABS publication on firm entries and exits is the *Counts of Australian Businesses, including Entries and Exits* (CABEE) publication (ABS,2015). This is formed by considering each June the set of businesses in the market sector which are actively trading in goods or services, and comparing this to the set from the previous June. Firms which have disappeared since the previous June are considered exits. To be actively trading in goods and services, a firm needs to have an active GST role and have submitted a BAS statement with non-zero dollar amounts within the last 5 quarters (or three years for annual remitters) – firms with a GST role who do not meet this condition are considered 'long term non-remitters' (LTNR) and deemed to

[19] Between 2009–10 and 2011–12.

[20] For example, it can take up to 14 to 16 months processing time (including the waiting period to receive tax return forms) before the ABS receives the PAYG data for the current financial year (Wallace, 2014). The PIT data can also take up to 22 months processing time (Chien *et al.*, 2012).

have exited. Firms which move between the profiled and non-profiled population[21] are also considered to have entered/exited.

This MAC paper uses the same methodology as CABEE (considering if firms are actively trading or not) but considers a wider scope, including businesses which are not in the market sector, and we do not treat moves between the profiled and non-profiled populations as exits. Such moves are purely administrative and will not reflect true firm deaths. Comparing the populations assembled in this way between June 2010 and June 2011 determines the list of possible firm exits which can be assessed to determine if they are deaths.[22] We then further restrict ourselves to only firms which employed at least 10 or more employees in 2010–11, to allow our model to consider relevant firm and employee characteristics such as employee movement across time in detecting true deaths. We also exclude firms which were active again in 2011–12, either considered active by the CABEE definition in June 2012, or who had employees in 2011–12 according to PAYG records. Note that this 'reactivation' information is not necessarily available at the time CABEE is being compiled.

The prototype GLIDE contains only a short time series which means that we cannot use time series methods. The truthing data contains 660 firms with 244 confirmed exits and 416 takeover and merger firms, all of which exited in the 2010–11 financial year. For each of these firms, we include variables which consider the firm alone (its size, and whether several financial variables have increased or decreased substantially from the previous year) and variables which consider the network of employee-firm connections. To generate these network variables, we looked at all the firm's employees in the year it exited, and which firms they worked for in the following year.

For whichever firm had the most employees the following year in common with the original firm, we determined whether the two firms are in the same industry division, the distance between their head office locations, and the percentage of employees from the original firm that now work for the new firm.[23] We then derived dummy variables from this for use in modelling. Table 2.3 contains summary statistics on the variables used in the models.

[21] The ABSBR maintains two populations of businesses – the profiled population, which includes businesses with large and complex business structures, and the non-profiled population, which includes businesses with simple business structures. The non-profiled population contains the majority of businesses.

[22] The exit counts considered in this paper are higher than the published CABEE counts. They exceed CABEE by 500–2000 firms for each time period.

[23] In the case of a tie we broke it arbitrarily, given that in cases where the new firm is genuinely connected to the original firm, we can expect a large number of employees to move; ties in which firm has the most employees in common with the original is only likely for firms that genuinely died and have only a few employees working in common subsequently. It is possible that in rare cases an exiting firm has split into two who have equal number of employees from the original, but as our model only distinguishes true deaths from exits involving other firms (and doesn't distinguish takeovers/mergers from splits) picking either of these should not substantially change the results; and this situation does not occur in the truthing data.

2.3 Summary statistics

	Fewer than 20 employees	*20 or more employees*	*Missing*
No. employees: size1011[24]	65.3 %	34.7 %	0

	Same Industry	*Different Industry*	*Missing*
same_Industry	53.9 %	46.1 %	0

	Within 20 km	*More than 20 km*	*Missing*
distance	66.0 %	34.0 %	37

	Change from previous year			
	Down 25% or more	*No major change*	*Up 25% or more*	*Missing*
turnover1011down[25]	69.2 %	22.3 %	8.5 %	47
OExp1011up[26]	67.9 %	23.1 %	9.0 %	47
PAYG_Instmt1011up[27]	66.2 %	26.1 %	7.7 %	50

	<25%	25–50%	50–75%	75%+
No. common employees: per_moved[28]	35.9 %	20.8 %	24.5 %	18.8 %

2.6 Statistical models

This paper combines both multilevel and Bayesian Networks (BNs) models. Lappenschaar *et al.* (2013) proposed a method to integrate both methods in a single model called multilevel Bayesian networks for analysing hierarchical health care data, but we do not follow that here. Instead, our approach uses the multilevel model to provide us a statistical framework for the BNs model. We explore the use of BNs as one possible approach to distinguish spurious firm deaths.

[24] Multilevel model: 1 in this dummy means a firm with 20 or more employees, otherwise 0. BNs model: use both categories. The cutoff of 20 has been chosen to be in line with cutoffs used in CABEE which inlcudes both medium and large firms (ABS, 2012).

[25] Multilevel model: 1 in this dummy measures Turnover going down 25% or more than the previous year and 0 for the rest of the categories. BNs model: measures all three categories.

[26] Multilevel model: 1 in this dummy measures Operating Expenses going up 25% or more than the previous year and 0 for the rest of the categories. BNs model: measures all three categories.

[27] Multilevel model: 1 in this dummy measures PAYG Installment (i.e. total firm wage bill) going up 25% or more than the previous year and 0 for the rest of the categories. BNs model: measures all three categories.

[28] Note that there are no missing values for this variable.

The advantages of exploring the BNs method in this context include:

- It is computationally efficient by using algorithms to learn about the structure and estimate the parameters based on the training data. For example, the computational time for learning the structure in the BNs model is shorter than the multilevel model with the same number of parameters.
- Both modelling approaches yield similar prediction accuracy but BNs only rely on a subset of variables.
- For prediction purposes, BNs can be used to predict outcome for observations which contain missing variables in the test dataset (Kragt, 2009). This is a useful model feature when we expand the study to incorporate multiple years or firm births. In comparison, multilevel models exclude observations with missing variables in the test dataset.

The main limitation of BNs models is:

- The acyclic property of BNs implies that feedback effects cannot be included in the network (Kragt, 2009).

2.6.1 Multilevel logistic regression

The hierarchical structure of the tax records lends itself to multilevel modelling to distinguish between true firm deaths and spurious deaths (i.e. exiting firms continuing under a different ABN). The main advantages of using this technique, in comparison with other regression techniques e.g. logistic regression, include:

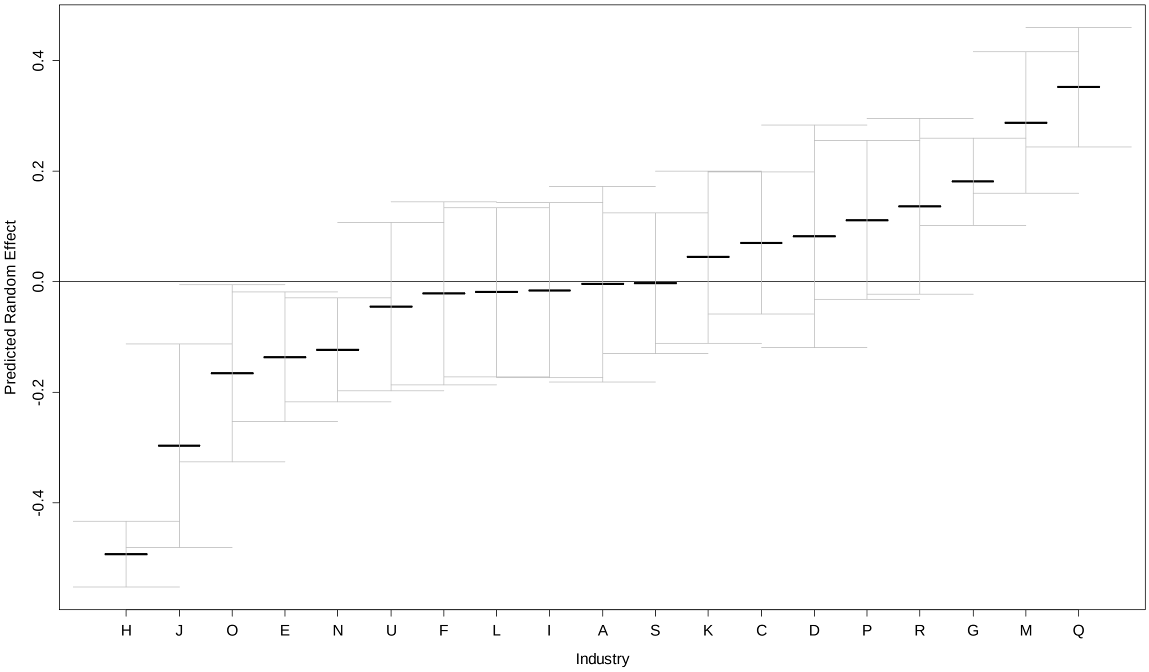
- it captures the cross-level relationship and makes better use of the hierarchical data structure (Snijders *et al.*, 1999, Grassetti *et al.*, 2005). Figure 2.4 is a caterpillar plot showing the industry effects together with 95% confidence intervals. It shows that that there are significant cross industry differences. This suggest that we should consider using multilevel models; and
- it allows for different intercepts. When the data has a nested structure, the observations within groups have similar characteristics because of the selection process and it is not appropriate to use OLS regression (Hox, 2010).

There are some disadvantages of using this technique including:

- The multilevel models assume that the error terms are distributed normally. If there is a violation of this assumption, the asymptotic errors are incorrect which leads to inaccuracy of the confidence intervals. This is particularly important at the higher level (i.e. industry) for the random intercepts (Maas *et al.*, 2004a).

- In addition, the variance component can be underestimated if the sample size at the higher level can be too small;[29] in our case we have only 19 industries (Maas *et al.*, 2004b).

2.4 Differences across selected industries



Source: ABS[30]

The multilevel logistic model can be specified as:

Firm level model: [31]

$$\text{logit}\left(p_{ji}\right) = \log\left(\frac{p_{ji}}{1-p_{ji}}\right) = \beta_{0j} + \sum\nolimits_{k=1}^{K} \beta_{kji} X_{kji} \qquad (1)$$

Industry level model:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (2)$$

[29] Literature recommended the ideal sample size is between 24–30. Browne *et al.* (2000) cited in Maas *et al.* (2004b) suggested that as few as six to twelve is sufficient; in contrast Van Der Leeden *et al.* (1994) cited in Maas *et al.* (2004b) indicated more than 100 is needed.

[30] Note that division 'U' here is 'Unknown' i.e. firms for which industry coding is not available.

[31] $Y_{ji} = \begin{cases} 1 & \text{for a true exit firm.} \\ 0 & \text{for a continuing firm.} \end{cases}$ so $Y_{ji} \mid p_{ji} \sim$ Bernoulli$\left(p_{ji}\right)$, where $p_{ji} = \Pr\left(Y_{ji} = 1\right)$.

Y_{ji} is a binary variable indicating if a firm is truly dead or not.

The random intercept model is thus:

$$\text{logit}\left(p_{ji}\right) = \gamma_{00} + u_{0j} + \sum\nolimits_{k=1}^{K} \beta_{kji} X_{kji} \ .$$

- The logit of p_{ji} for $i =$ firms and $j =$ industries is the sum of a linear function of the explanatory variables and a random industry deviation u_{0j}. A unit difference between the X_{kji} values of two firms in the same industry is associated with a difference of β_{kji} in their log odds.
- $\left\{X_{kji} : k = 1, \ldots, K\right\}$ are the k firm-level binary explanatory variables e.g. at least a 25% fall of turnover between t and $t-1$[32] for firm i in industry j. They also include contextual variables like the firm size. We use semantic technology to derive network statistics for dummies for the relationship between the exiting firm and the firm which had the most employees in common with it the following year:

  – the percentage of shared employees between these two firms;

  – whether the two firms are located within a 20 km or less radius of each other;

  – whether the two firms are in the same industry division.
- β_{0j} is the intercept for industry j;
- $\left\{\beta_{kji} : k = 1, \ldots, K\right\}$ are the corresponding firm-level coefficients that indicate the direction and strength of association between each firm characteristic k and the outcome in industry j;
- γ_{00} is the average intercept across all industries;
- u_{0j} is the industry dependent deviation for the intercept.[33]

Table 2.5 reports the empirical results from the multilevel logistic regressions. Models 1 and 2 use the full data and Models 3 and 4 use the training data[34]. Models 1 and 3 exclude network statistics and Models 2 and 4 include them. We have tested over a couple of thousand combinations of available variables and reported the best results.[35] Note that the difference in the number of observations is the result of different amount of missing data with different sets of variables.

[32] t is the reference year 2010–11, i.e. the firm year the firm 'exits' and $t-1$ is 2009–10.

[33] u_{0j} is assumed to be normally distributed.

[34] For both the models with and without network statistics, we selected a random sample of firms (stratified by industry division) to use as training data for modelling, and kept the rest aside to test the predictive ability of the models. About ¾ of the data is used for training and ¼ for testing, though the testing subset is different for the models with and without network statistics because of different patterns of missingness with the variables included in the final models.

[35] The selection criteria include considering the lowest AIC and BIC and selecting only significant variables which provides the best prediction results.

2.5 Multilevel logistic regressions

	Full data				*Training data*			
	Model 1		*Model 2*		*Model 3*		*Model 4*	
(Intercept)	–0.03	(0.23)	–5.64	(0.75) ***	0.04	(0.26)	–5.41	(0.80) ***
Size1011	0.89	(0.21) ***	1.80	(0.52) ***	0.98	(0.23) ***	1.72	(0.57) **
Turnover1011down	0.67	(0.22) **			0.50	(0.25) *		
OExp1011up	–0.81	(0.40) *			–0.76	(0.45) °		
PAYG_Instmt1011up	–1.11	(0.42) **	–1.71	(0.74) *	–1.25	(0.47) **	–1.41	(0.81) °
Same_Industry			1.76	(0.48) ***			2.00	(0.56) ***
Distance			1.76	(0.49) ***			1.46	(0.53) **
per_moved_25to50			5.42	(0.58) ***			5.37	(0.65) ***
per_moved_50to75			8.47	(1.15) ***			7.98	(1.15) ***
per_moved_g75			24.58	(4061)			24.50	(4905)
AIC	689.0		165.6		546.9		137.8	
BIC	715.1		204.8		571.5		174.7	
Log Likelihood	–338.5		–73.8		–267.50		–59.9	
No. observations ^	573		576		448		446	
No. obs. test data					125		130	
No. groups: A06_DIVISION_CODE	19		20		19		20	
Correct prediction					87		124	
Percent correct					69.6%		95.4%	

^ Different numbers of observations are caused by different missingess in the variables included in each model.
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ° $p < 0.1$

The overall results suggest that large firms (i.e. those with 20 or more employees) are more likely to continue under another ABN (i.e. be spurious deaths) because these firms could restructure in the event of economic downturn. In comparison, firms with a more than 25 per cent increase in the Pay As You Go (PAYG) income tax instalment (i.e. wage bill) are more likely to be true deaths. This is in line with the expectation that firms truly dying will pay off their employees in the event of closure.

Models 1 and 3 show that firms are more likely to continue if their turnover is more than 25 per cent lower and operating expenses are more than 25 per cent lower. Models 2 and 4, containing network statistics, show that being in the same industry and within a similar location as the potential successor firm are also signals for spurious deaths. Similarly, an exiting firm sharing 25 per cent or more employees with another firm also suggests that this firm is continuing. Please note that the insignificant but strongly positive coefficient of the more than 75 per cent of shared employees variable indicates these firms are likely to be spurious deaths.[36]

[36] The insignificance could be the result of few cases in the training data, all of which are continuing firms.

2.5.2 Bayesian networks

This paper considers Bayesian Networks (BNs) as a way to apply a computationally efficient method within a statistical framework. BNs are probabilistic graphical models, used widely for knowledge representation and reasoning under uncertainty (Korb *et al.*, 2010).[37] BNs provide a method to represent and capture the relationships between a set of random variables (or nodes) for discrete or continuous data via a directed acyclic graph (DAG). The dependencies between these nodes are described by conditional probability distributions (Kragt, 2009). Both the structure and the conditional probability distributions can be specified (from theory or prior knowledge) or estimated from the data.[38] The network structure, combined with conditional density functions, specifies a multivariate density function which can be used to predict the value of a variable given the value of other variables. In our case, we use it to predict whether a firm is a genuine death or not.

A general Bayesian Network for a set of variables $X = \{X_k : k = 1, \ldots, K\}$ consists of (1) a network structure S that asserts conditional independence relationships associated with each X and (2) a set of local probability distributions associated with each variable. Consider the general problem of expressing the joint probability distribution for the set of variables X. In general, using the chain rule, this would be:

$$\begin{aligned}\Pr(X_1,\ldots,X_K) &= \Pr(X_1)\times\Pr(X_2|X_1)\ldots,\times\Pr(X_K|X_1,\ldots,X_{K-1}) \\ &= \prod\nolimits_{k=1}^{K}\Pr(X_k|X_1,\ldots,X_{k-1}) \qquad (3)\end{aligned}$$

If we do not specify the dependence structure in a model, we must condition on all other variables. This would result in a model which is quite complex to estimate. By contrast, a Bayesian Network expresses the dependence structure between the variables in a network structure S via a DAG, where each variable is a node in the graph. Under this structure, the value of a node is conditional only on the value(s) of its parent node(s) (Korb *et al.*, 2010). This reduces equation (3) to

$$\Pr\left(X_1,\ldots,X_K\right) = \prod\nolimits_{k=1}^{K}\Pr\left(X_k \mid Pa_k\right) \qquad (4)$$

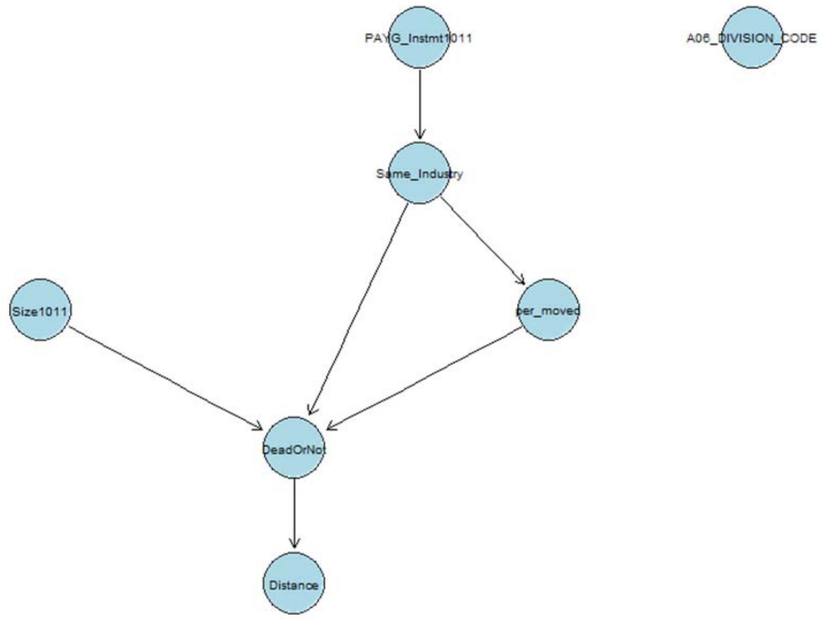
where Pa_k denotes the parent(s) of node X_k. This simplifies the estimation problem when Pa_k is not equal to $X_1,\ldots,X_{k-1}$. We use the significant explanatory variables[39] from the multilevel model and training data to determine the DAG structure for the BNs model.

[37] They are computationally tractable and also used widely in artificial intelligence (Peek *et al.*, 2009)

[38] Learning is a term here borrowed from expert systems theory and artificial intelligence, encompassing both model selection (structure learning) and parameter estimation (parameter learning).

[39] e.g. firm size and shared employees etc.

2.6 Bayesian Network structure



Source. ABS

Figure 2.6 shows the DAG structure of the BNs model. There are three variables – firm size, same industry and shared employees – that are found as parents of whether a firm's death is true or spurious. The joint distribution of all 6 variables, using the ordering (S, P, I, E, C, D)[40], is given by

$$\Pr(S)\,\Pr(P)\,\Pr(I|P)\,\Pr(E|I)\,\Pr(C|S,I,E)\,\Pr(D|C) \qquad (5)$$

This means we only estimate 30 parameters, in comparison with 127, if we were estimating the full conditional distributions without the dependence structure. Please note that the industry division is not found to be relevant in predicting the other variables for the BNs model.

There is a two-step process to use BNs model for prediction:

- First is determining the DAG structure, which is similar to model selections for classic statistical modelling. Pollino *et al.* (2010) suggested that these relationships can be obtained via (1) expert elicitation using modelling methods or (2) estimation from the data using algorithms. We use a combination of the two approaches by starting with those variables which were significant in the multilevel logistic regression, refining the structure through estimation[41], and checking the most

[40] S=Size1011,P=PAYG_Instmt1011, I=Same_Industry, E=per_moved, C=DeadOrNot, D=Distance

[41] This paper uses score based learning, in particular the hill climb algorithm. This applies an application of heuristic optimisation techniques by finding a structure which minimises BIC score. The search space begins from a given network structure (e.g. no connections) and adds, deletes or reverses one arc at a time until the score can no longer be improved (Scutari *et al.*, 2015). See Box 1 for a detailed explanation.

appropriate DAG that provides the best prediction results. For example, the multilevel model results have confirmed that shared employees (per_moved) is an important predictor. However, the hill climb algorithm showed the inverse relationship (making whether a firm is dead or not a parent of the proportion of shared employees) so in the BNs model we assert the relationship in the direction we expected.

Mathematically, Heckerman (1995) suggested that, according to Bayes' theorem, we have: [42]

$$\Pr\left(D^{\text{train}}\middle|S\right)=\frac{\Pr\left(\theta\middle|S\right)\Pr\left(D^{\text{train}}\middle|\theta,S\right)}{\Pr\left(\theta\middle|D^{\text{train}},S\right)} \tag{6}$$

where

– D^{train} is the training dataset in a matrix, which is assumed to be complete.

– S is the configuration of the network which includes both node X_k and their parent node(s) Pa_k.

– θ is the parameters of the structure given dataset D, the posterior distribution can be computed as $\Pr(D^{\text{train}}|\theta,S)$. Parameters θ are assumed to be mutually independent.

- Second is estimating the parameters (i.e the conditional probability tables). These can be estimated using various methods, including Maximum Likelihood estimation[43] and Bayesian estimation. Under Bayesian estimation, parameters are estimated by taking into account both the conditional probabilities and the priors. The posterior probability is derived by the likelihood that conjugates both the conditional probabilities and the priors. The data is then used to update the estimated parameters (Nagarajan *et al.*, 2013). We used the Bayesian estimator with non-informative priors (see the Appendix for an example).

[42] See the glossary for details.

[43] A disadvantage of Maximum Likelihood estimation in this context is that it can produce conditional probabilities of 0 or 1 (if a particular outcome is never/always observed in the training data), and so the model produces errors if a future observation does not fit into the 1 probability catageory; Bayesian estimation by contrast starts from a uniform prior and so smooths the estimated probabilities away from the extremes.

BOX 1: Hill-Climbing Algorithm

1. Select a network structure DAG S .

2. Compute the BIC as

$$\begin{aligned} BIC &= \log \Pr(D^{\text{train}} \big| S) \\ &= \log \Pr(\theta \big| S) + \log \Pr(D^{\text{train}} \big| \theta, S) - \log \Pr(\theta \Big| D^{\text{train}}, S) \\ &\approx \log \Pr(D^{\text{train}} \big| \theta, S) - \frac{d}{2} \log N \end{aligned}$$

   where Heckerman (1995), following Schwarz (1978), suggested that

$$\log \Pr(\theta \big| S) - \log \Pr(\theta \Big| D^{\text{train}}, S)$$

   can be approximated by

$$-\frac{d}{2} \log N$$

   which is the penalising factor for complexity, d is the dimension of θ and N is the number of observations in D^{train} .

3. Repeat the following step so long as BIC increases

   a. for every possible arc addition, deletion or reversal not resulting in a cyclic network, compute BIC'

   b. if $\text{Max}(\text{BIC}') > \text{BIC}$, perform the operation that updates BIC with $\text{Max}(\text{BIC}')$.

4. Return the DAG S .

Scutari *et al.* (2015) suggested that the parameter estimation can be expressed as:[44]

$$CPT(\theta \Big| D^{test}, S) = \prod\nolimits_{k=1}^{K} \Pr(\theta_k \Big| D^{test}, S) \tag{7}$$

where

– D^{test} is the test dataset in a matrix.

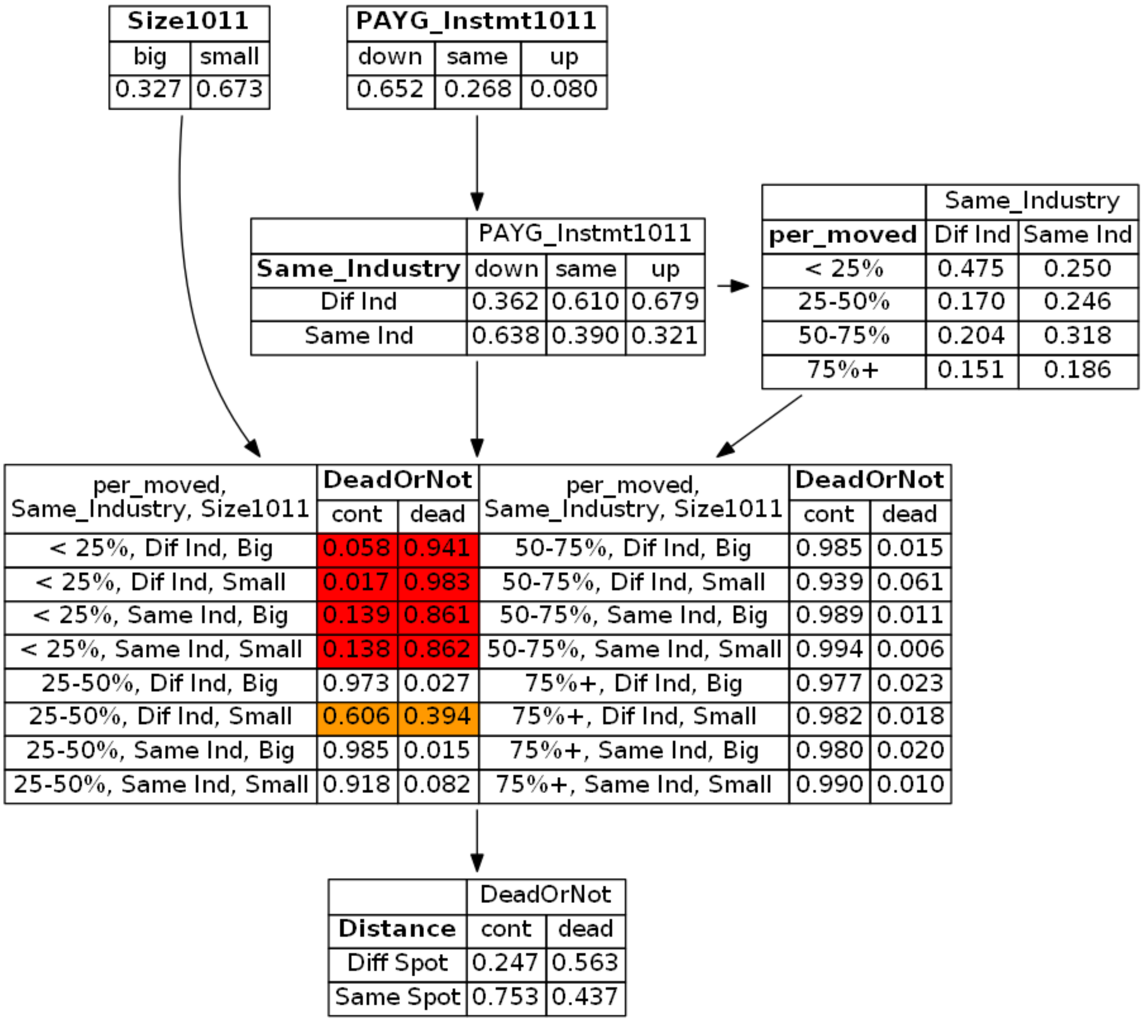
– $\theta = \left\{ \theta_k : k = 1, \ldots, K \right\}$ is the vector of the parameters for the distribution.

– S denotes the joint probability distribution that can be factorised according to the network structure.

[44] Given the dataset D , the posterior distribution can be computed as $\Pr(\theta \big| D, S)$ (Heckerman, 1995, p. 16).

Figure 2.7 shows the BNs structure with the conditional probabilities of the BNs model. We focus on $\Pr(C|S,I,E)$ from equation (5) of the BNs for model prediction. For example, conditional on observing a given proportion of shared employees, whether the two firms have the same industry, and the size of the original firm, what is the probability of that firm being genuinely dead? The results are consistent with the multilevel regressions. As expected, firms with a lower proportion of shared employees are more likely to be a true death (see red cells). Larger firms, with shared employees, are more likely to be continuing. There are mixed results for being in the different industries for predicting if a firm is truly dead or not. The highlighted brown cell indicates high uncertainty for correct prediction in this case because the probabilities are not clear cut.

2.7 Bayesian Network structure with conditional probabilities



Size1011	
big	small
0.327	0.673

PAYG_Instmt1011		
down	same	up
0.652	0.268	0.080

	PAYG_Instmt1011		
Same_Industry	down	same	up
Dif Ind	0.362	0.610	0.679
Same Ind	0.638	0.390	0.321

	Same_Industry	
per_moved	Dif Ind	Same Ind
< 25%	0.475	0.250
25-50%	0.170	0.246
50-75%	0.204	0.318
75%+	0.151	0.186

per_moved, Same_Industry, Size1011	**DeadOrNot**		per_moved, Same_Industry, Size1011	**DeadOrNot**	
	cont	dead		cont	dead
< 25%, Dif Ind, Big	0.058	0.941	50-75%, Dif Ind, Big	0.985	0.015
< 25%, Dif Ind, Small	0.017	0.983	50-75%, Dif Ind, Small	0.939	0.061
< 25%, Same Ind, Big	0.139	0.861	50-75%, Same Ind, Big	0.989	0.011
< 25%, Same Ind, Small	0.138	0.862	50-75%, Same Ind, Small	0.994	0.006
25-50%, Dif Ind, Big	0.973	0.027	75%+, Dif Ind, Big	0.977	0.023
25-50%, Dif Ind, Small	0.606	0.394	75%+, Dif Ind, Small	0.982	0.018
25-50%, Same Ind, Big	0.985	0.015	75%+, Same Ind, Big	0.980	0.020
25-50%, Same Ind, Small	0.918	0.082	75%+, Same Ind, Small	0.990	0.010

	DeadOrNot	
Distance	cont	dead
Diff Spot	0.247	0.563
Same Spot	0.753	0.437

Please note that we bold the parent nodes.

Source. ABS

3. SUMMARIES AND EMPIRICAL RESULTS

This section summarises and discusses the empirical results. This preliminary analysis focuses on distinguishing true and spurious firm deaths. We also compare the distribution of labour productivity before and after adjusting for the spurious deaths.

Having determined the multilevel and Bayesian Network models, using our training data, we test their predictive capability using the subset of firms set aside in the test data[45] (see above for how the data was split). Table 3.1 shows the true number of dead and continuing firms in the test data, divided by industry (Dead indicates a true firm death). Tables 3.2 and 3.3 show the prediction performance of the two different models (TRUE indicates a firm that was correctly predicted). Note that the prediction outcomes are the same here for the models with network statistics (they were different without network statistics), though this may not hold for all datasets.

3.1 Number of true death and continuing firms in the test data by industry

	Industry code																		
	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*UNK*	*Total*
Continuing	1	5	2	4	4	12	11	1		5	4	4	10	3	6	8	4	1	3	88
Dead	1	1		3	1	4	8	1	2	1	1	4	3	1	2	3		4	2	42
Total	2	6	2	7	5	16	19	2	2	6	5	8	13	4	8	11	4	5	5	130

3.2 Prediction success rate for multilevel network model by industry

	Industry code																			
	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*UNK*	*Total*
FALSE						1	3								1				1	6
TRUE	2	6	2	7	5	15	16	2	2	6	5	8	13	4	7	11	4	5	4	124
Total	2	6	2	7	5	16	19	2	2	6	5	8	13	4	8	11	4	5	5	130

3.3 Prediction success rate for Bayesian Networks model by industry

	Industry code																			
	A	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*UNK*	*Total*
FALSE						1	3								1				1	6
TRUE	2	6	2	7	5	15	16	2	2	6	5	8	13	4	7	11	4	5	4	124
Total	2	6	2	7	5	16	19	2	2	6	5	8	13	4	8	11	4	5	5	130

[45] We only show results for the models with network statistics included, as the prediction success rate was substantially higher for these models (up from ~70% to ~95%).

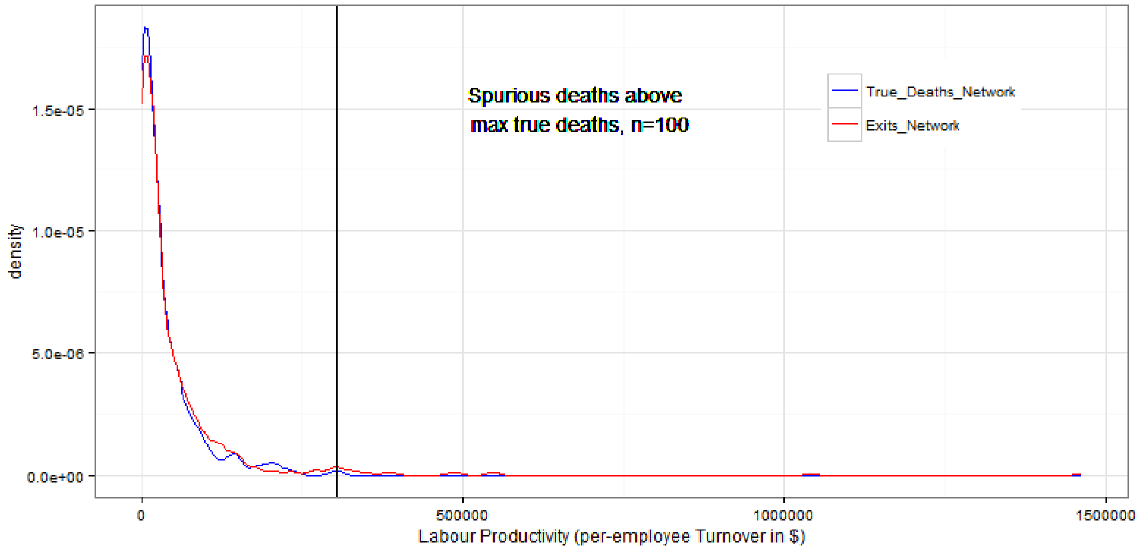
BNs can determine the conditional probability of events even when the values of some covariates are missing in the test data. This makes it possible to predict whether exited firms with missing values are truly dead or not. The results of this are shown in table 3.4, which shows that this model still performs well even with missing values.

3.4 Prediction success rate for BNs model for missing values in test data by industry

	Industry code																	
	A	*C*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*	*Q*	*R*	*S*	*UNK*	*Total*
FALSE			1			2			1					1					5
TRUE	2	4	8	2	13	13	2	1	4	5	5	8	1		4	1	5	1	79
Total	2	4	9	2	13	15	2	1	5	5	5	8	1	1	4	1	5	1	84

Figure 3.5 shows kernel densities of the labour productivity[46] distribution of exiting firms before and after distinguishing true death (for those firms included in the models with network statistics). It clearly shows the difference between the distribution of true deaths compared to all exiting firms (including spurious deaths). For example, there are firms with significant economic contributions (i.e. high labour productivity) which are continuing and should not be considered as deaths – the vertical line shows that there are over 100 firms with spurious deaths which have labour productivity above the maximum of the true deaths. However, the problem of spurious death appears to be less significant at the low end of the distribution. This may be because these smaller firms are less likely to experience restructuring.

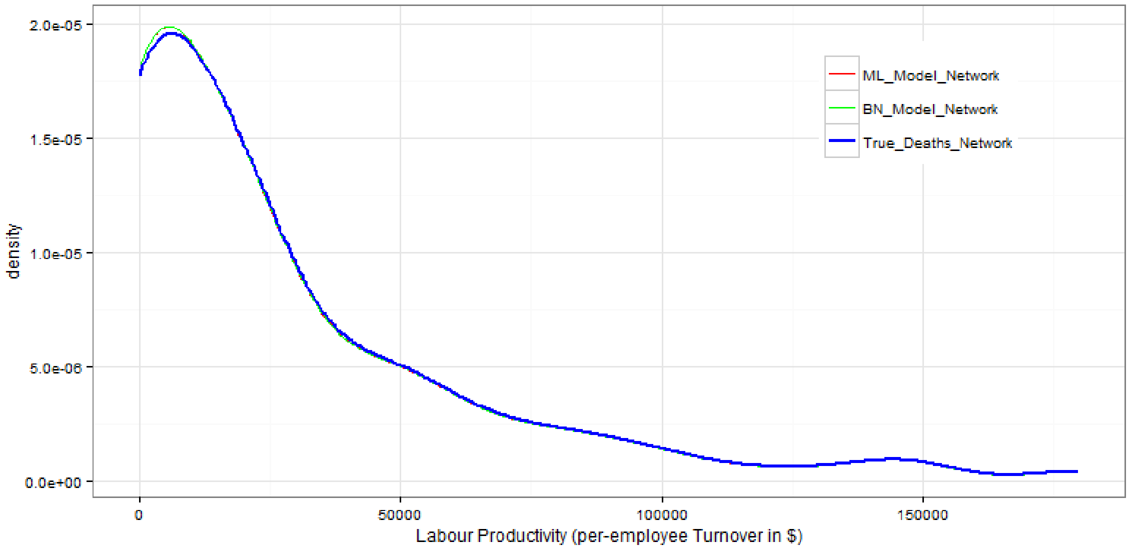
3.5 Kernel of Labour productivity – True Deaths vs all Exits



[46] Measured by per employee turnover.

Figure 3.6 focuses on the lower end of the distribution and compares the predicted results (with models including network statistics) and the true deaths, showing the importance of including network statistics in the models. While the BNs and multilevel models achieve the same prediction results when including network statistics, the BNs model uses only a subset of the explanatory variables, representing efficiency gains above the multilevel model.

3.6 Zoomed kernel of Labour productivity – network models



4. KEY LESSONS

In this MAC paper, we have described a prototype GLIDE and provided an example of using semantic web techniques for deriving network statistics. We used these statistics in the multilevel and Bayesian Networks (BNs) models to detect true firm deaths.

The semantic web is an excellent platform for considering data integration in the context of linked ATO and ABS data to better analyse the complexity of the labour market networks. This is similar to findings from New Zealand and Canada who have used employee networks to detect true firm births and deaths.

We have also shown that it is important to account for spurious death for statistical production. This is because failure to account for spurious firm deaths can result in continuing enterprises being incorrectly classified as firm deaths and as a result it can affect the statistical quality from the perspectives of survey frames and accuracy of the statistics.

We consider both multilevel and BNs models. Our approach applies the BNs method within a statistical framework. We have shown that BNs can handle observations with missing variables in the test data. This paper does not intend to assess both methods on the prediction outcomes. It clearly shows that it is important to incorporate network information for modelling purposes. After including this, the prediction outcomes improved substantially for both models reaching 95% accuracy rate.

This MAC paper demonstrates the statistical values gained from this new data management approach. This preliminary research has many areas for future work:

First, it should be extended to distinguish spurious births. The results can also be cross-validated with the results from this MAC paper which focuses on true deaths, where a continuing firm exits under one ABN and enters as another.

Second, we can split the data to consider different firm sizes. The prediction outcomes may be different if we use different ranges.

Finally, the models should be tested on better truthing data. We are in the process of evaluating another truthing data source.

We conclude that the semantic web is a useful approach for statistical purposes and network analysis can be used to effectively distinguish true firm deaths.

REFERENCES

Abowd, J.M. and Kramarz, F. (1999) "The Analysis of Labor Markets Using Matched Employer-Employee Data", Chapter 40 in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics, Volume 3, Part B*, Elsevier.

Abowd, J.M. and Vilhuber, L. (2005) "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers", *Journal of Business & Economic Statistics*, 23(2), pp. 133–152.

Acid, S.; de Campos, L.M.; Fernández-Luna, J.M.; Rodríguez, S.; Rodríguez, J.M. and Salcedo, J.L. (2004) "A Comparison of Learning Algorithms for Bayesian Networks: A Case Study Based on Data from an Emergency Medical Service", *Artificial Intelligence in Medicine*, 30(3), pp. 215–232.

Australian Bureau of Statistics (2012) *Counts of Australian Businesses, including Entries and Exits: Glossary*, cat. no. 8165.0, ABS, Canberra. <http://www.abs.gov.au/AUSSTATS/abs@.nsf/Previousproducts/8165.0Glossary1Jun 2007 to Jun 2011?opendocument&tabname=Notes&prodno=8165.0&issue=Jun 2007 to Jun 2011&num=&view=>

—— (2013) "The Use of Transactions Data to Compile the Australian Consumer Price Index", Feature Article in *Consumer Price Index, Australia, Sep 2013*, ABS, Canberra. <http://www.abs.gov.au/AUSSTATS/abs@.nsf/Previousproducts/6401.0Main%20Features2Sep%202013?opendocument&tabname=Summary&prodno=6401.0&issue=Sep%202013&num=&view=>

—— (2014) *Australian Bureau of Statistics Business Register*, ABS, Canberra. <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DOSSbytitle/AC79D33ED6045E88CA25706E0074E77A?OpenDocument>

—— (2015) *Counts of Australian Businesses, including Entries and Exits , June 2010 to June 2014*, cat. no. 8165.0, ABS, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/8165.0>

Australian Government Information Management Office (2013) *The Australian Public Service Big Data Strategy*, AGIMO. <http://www.finance.gov.au/sites/default/files/Big-Data-Strategy_0.pdf>

Bachmann, R. and David, P. (2009) "The Importance of Two-Sided Heterogeneity for the Cyclicality of Labour Market Dynamics", *Ruhr Economic Papers*, No. 124, Ruhr Graduate School in Economics.

Baldwin, J.R.; Beckstead, D. and Girard, A. (2002) "The Importance of Entry to Canadian Manufacturing with an Appendix on Measurement Issues", *Analytical Studies Branch Research Paper Series*, Catalogue No. 11F0019MIE – No. 189, Statistics Canada, Ottawa.

Benedetto, G.; Haltiwanger, J.; Lane, J. and McKinney, K. (2007) "Using Worker Flows to Measure Firm Dynamics", *Journal of Business & Economic Statistics*, 25(3), pp. 299–313.

Berners-Lee, T.; Hendler, J. and Lassila, O. (2001) *The Semantic Web*, Scientific American.com.
<http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html>

Bjelland, M.; Fallick, B.; Haltiwanger, J. and McEntarfer, E. (2011) "Employer-to-Employer Flows in the United States: Estimates Using Linked Employer-Employee Data", *Journal of Business & Economic Statistics*, 29(4), pp. 493–505.

Chien, C.-H.; Clarke, C. and Amarasinghe, A. (2012) "Enhancing the ABS's Use of Personal Income Tax Data", *Methodology Advisory Committee Papers*, No. 126, Australian Bureau of Statistics, Canberra.

Chien, C.-H. and Haller, A. (forthcoming) *Statistical Uncertainty and Semantic Web*, Paper prepared for the 14th International Semantic Web Conference, Bethlehem, Pennsylvania, 11–15 October 2015.

Criscuolo, C.; Gal, P.N. and Menon, C. (2014) "The Dynamics of Employment Growth: New Evidence from 18 Countries", *CEP Discussion Papers*, CEPDP1274, London School of Economics and Political Science, London.

Dataversity (2011) *Introduction to: RDF*.
<http://www.dataversity.net/introduction-to-rdf/>

Davis, S.; Haltiwanger, J.; Jarmin, R. and Miranda, J. (2006) "Volatility and Dispersion in Business Growth Rates: Publicly Traded versus Privately Held Firms", *NBER Working Papers*, No. 12354, National Bureau of Economic Research, Inc.

Ding, Z.; Peng, Y. and Pan, R. (2004) "A Bayesian Approach to Uncertainty Modeling in OWL Ontology", in *Proceedings of the International Conference on Advances in Intelligent Systems – Theory and Applications*, Luxembourg, November 2004.
<http://www.csee.umbc.edu/~ypeng/Publications/2004/AISTA2004-137-04Final.pdf>

Ding, Z.; Peng, Y.; Pan, R. and Yu, Y. (2005) "A Bayesian Methodology Towards Automatic Ontology Mapping", in *Proceedings of the AAAI-05 Workshop on Contexts and Ontologies: Theory, Practice and Applications*, Pittsburgh, July 2005.
<http://www.csee.umbc.edu/~ypeng/Publications/2005/C&OFinal.pdf>

Dixon, J. and Rollin, A.-M. (2011) "Firm Dynamics: Employment Growth Rates of Small Versus Large Firms in Canada", *Economic Analysis Research Paper Series*, Catalogue no. 11-622-M – No. 025, Statistics Canada, Ottawa. <http://www.statcan.gc.ca/pub/11-622-m/11-622-m2012025-eng.pdf>

Fabling, R.; Gretton, J. and Powell, C. (2008) *Developing the Prototype Longitudinal Business Database: New Zealand's Experience*, Paper presented at Statistics Research Institute (SRI) Seminar, Daejeon, Korea, 19 May 2008, Statistics New Zealand, Wellington.

Grassetti, L.; Gori, E. and Minotti, S.C. (2005) "Multilevel Flexible Specification of the Production Function in Health Economics", *IMA Journal of Management Mathematics*, 16(4), pp. 383–398.

Gray, M.; Heath, A. and Hunter, B. (2005) "The Labour Force Dynamics of the Marginally Attached", *Australian Economic Papers*, 44(1), pp. 1–14.

Haltiwanger, J.; Jarmin, R. and Miranda, J. (2009) "Business Dynamics Statistics: An Overview", *Ewing Marion Kauffman Foundation BDS Briefs*.

Haltiwanger, J.C.; Jarmin, R.S. and Miranda, J. (2010) "Who Creates Jobs? Small vs. Large vs. Young", *NBER Working Papers*, No. 16300, National Bureau of Economic Research, Inc.

Harwood, A. and Mayer, A. (2014) *Big Data and Semantic Technology: A Future for Data Integration, Exploration and Visualisation*, Paper submitted for 2015 International Association for Official Statistics Prize for Young Statisticians.

Heckerman, D. (1995) *A Tutorial on Learning With Bayesian Networks*, Redmond, Microsoft Corporation. <http://research.microsoft.com/apps/pubs/?id=69588>

Hox, J.J. (2010) "The Basic Two-Level Regression Model", Chapter 2 in *Multilevel Analysis Techniques and Applications*, Lawrence Erlbaum Associates, New Jersey.

Ibsen, R. and Westergård-Nielsen, N.C. (2011) "Job Creation by Firms in Denmark", *IZA Discussion Papers*, No. 5458, Institute for the Study of Labour.

Jarmin, R.; Klimek, S. and Miranda, J. (2004) *Firm Entry and Exit in the U.S. Retail Sector: 1977–1997*, Center for Economic Studies, U.S. Census Bureau.

Jarmin, R.S. and Miranda, J. (2002) *The Longitudinal Business Database*, Center for Economic Studies, U.S. Census Bureau. <https://www.census.gov/ces/pdf/CES-WP-02-17.pdf>

Kelly, N. (2003) *Repairing EMS Employer Longitudinal Links*, Statistics New Zealand, Wellington. <http://www.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/kelly-repairing-ems-links.aspx>

Korb, K.B. and Nicholson, A.E. (2010) "Introducing Bayesian Networks", Chapter 2 in *Bayesian Artificial Intelligence*, Chapman & Hall/CRC Press, Boca Raton.

Kragt, M.E. (2009) *A Beginners Guide to Bayesian Network Modelling for Integrated Catchment Management*, Technical Report No. 9, Landscape Logic. <http://www.landscapelogic.org.au/publications/Technical_Reports/No_9_BNs_for_Integrated_Catchment_Management.pdf>

Leonard, J.S., Mulkay, B. and van Audenrode, M. (1999) "Compensation Policies and Firm Productivity", in J.C. Haltiwanger; J.I Lane; J.R. Spletzer; J.J.M. Theeuwes and K.R. Troske (eds.), *The Creation and Analysis of Employer-Employee Matched Data*, pp. 79–114, Emerald Group Publishing Limited.

Maas, C.J.M. and Hox, J.J. (2004a) "The Influence of Violations of Assumptions on Multilevel Parameter Estimates and their Standard Errors", *Computational Statistics and Data Analysis*, 46(3), pp. 427–440.

Maas, C.J.M. and Hox, J.J. (2004b) "Robustness Issues in Multilevel Regression Analysis", *Statistica Neerlandica*, 58(2), pp. 127–137.

Nagarajan, R.; Scutari, M. and Lèbre, S. (2013) *Bayesian Networks in R – With Applications in Systems Biology*, Springer, New York.

OECD Eurostat (2008) *Eurostat-OECD Manual on Business Demography Statistics*, OECD and Eurostat, Paris.

Peek, N. and Verduijn, M. (2009) "Bayesian Networks", in M. Kattan (ed.), *Encyclopedia of Medical Decision Making*, pp. 64–70, SAGE Publications Inc., Thousand Oaks.

Pollino, C.A. and Henderson, C. (2010) *Bayesian Networks: A Guide for their Application in Natural Resource Management and Policy*, Technical Report No. 14, Landscape Logic. <http://www.utas.edu.au/__data/assets/pdf_file/0009/588474/TR_14_BNs_a_resource_guide.pdf>

Riesenfeld, R.F. (2011) *Bayes' Theorem*, University of Utah. <http://www.eng.utah.edu/~cs5961/Resources/bayes.pdf>

Rijmen, F. (2008) "Bayesian Networks with a Logistic Regression Model for the Conditional Probabilities", *International Journal of Approximate Reasoning*, 48(2), pp. 659–666.

San Martín, M. and Gutierrez, C. (2009) "Representing, Querying and Transforming Social Networks with RDF/SPARQL", in *The Semantic Web: Research and Applications*, pp. 293–307, Springer, Berlin.

Schmutte, I.M. (2014) "Free to Move? A Network Analytic Approach for Learning the Limits to Job Mobility", *Labour Economics*, 29(1), pp. 49–61.

Schwarz, G. (1978) "Estimating the Dimension of a Model", *Annals of Statistics*, 6(2), pp. 461–464.

Scutari, M. and Denis, J.-B. (2015) *Bayesian Networks: With Examples in R*, Chapman & Hall.

Seyb, A. (2003) *The Longitudinal Business Frame*, Statistics New Zealand, Wellington. <http://www.stats.govt.nz/browse_for_stats/income-and-work/employment_and_unemployment/the-longitudinal-business-frame.aspx>

Snijders, T.A.B. and Bosker, R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, SAGE Publications Ltd.

Stoilos, G.; Simou, N.; Stamou, G. and Kollias, S. (2006) "Uncertainty and the Semantic Web", *IEEE Intelligent Systems*, 21(5), pp. 84–87.

Van der Leeden, R.; Busing, F. and Meijer, E. (1997) *Applications of Bootstrap Methods for Two-level Models*, Paper presented at the Multilevel Conference, Amsterdam, 1–2 April 1997.

W3C (2012) *OWL 2 Web Ontology Language Document Overview (Second Edition)* <http://www.w3.org/TR/owl2-overview/>

W3C (2013) *SPARQL 1.1 Query Language* <http://www.w3.org/TR/sparql11-query/>

W3C (2014) *Resource Description Framework (RDF)* <http://www.w3.org/RDF/>

Wallace, D. (2014) *Quality Assessment Employment Size Ranges for ABR Data Extracts*, Australian Business Register, Canberra.

Webb, A.R. and Copsey, K.D. (2011) *Statistical Pattern Recognition*, John Wiley & Sons.

All URLs last viewed on Wednesday 22 July 2015

GLOSSARY

Business Activity Statement

The Business Activity Statement (BAS) is a single form used by businesses to report their taxation obligations and remit their entitlements and obligations for Goods and Services Tax (GST), Pay As You Go (PAYG), Fringe Benefits Tax (FBT), Wine Equalisation Tax (WET), and Luxury Car Tax (LCT). Depending on the business and their reporting requirements, it may be reported in monthly, quarterly or annual statements. This MAC paper follows ABS standard editing methodology,[47] and then summed monthly and quarterly values to get annual BAS values for each ABN.

Pay As You Go

The Pay As You Go (PAYG) data contains information on the wages and salaries paid by companies and businesses to their employees. This includes a scrambled Tax File Number (STFN) and ABN through which the PIT data can be linked to the business data.

Personal Income Tax

The Personal Income Tax (PIT) data comprises all personal income tax records from Australia for that financial year which have been submitted within sixteen months of the end of the given financial year. The file does not contain name and address information but postcodes provide an indication of the individuals' address location. All useful employee-level characteristics for the GLIDE are sourced from the PIT data.

Truthing data

The process is underway to source a possible dataset detailing true deaths or takeovers in the economy. We constructed a preliminary truthing dataset by individually assessing each exiting firm to distinguish between genuine deaths and events involving other firms (whether takeovers, mergers, restructures etc.).

This was done using a SPARQL query to determine which other firms were potentially involved in the exit by looking at which firms employed at least 10% of the exiting firm's 2010–11 employees (minimum 3). For each of these other firms, the number of employees shared, their main and trading names, industry and subdivision code, location, and Enterprise and GST group (where relevant) were compared to those of the exiting firm, to assess if the other firm was involved in the exit. For example, if the

[47] The ABS has developed a standard editing methodology for BAS records, which ensures that values are consistent within a BAS statement (e.g. if Total sales and GST on sales are both 0, but Export sales and/or Other GST-free sales are non-zero, this is an inconsistency and Total sales will be adjusted upwards) as well as consistent across BAS statements (e.g. if an ABN switches from reporting monthly to quarterly, ensuring that no values are double counted across overlapping periods).

other firm took on 50% of the exiting firm's employees, had the same industry code and an identical location, they were deemed to be involved in the exit; whereas if no other firms shared 10% or more employees with the exiting firm, it was deemed to be a true death. In particular, having similar names (e.g. 'Joe's Bakery' and 'Joe's Patisserie' or 'J. Bloggs and J. Smith' and 'J Smith') in combination with shared employees was seen as a strong indication of connection between the exiting and the other firm. Note that truthing in this way has access to further variables than the resulting models (firm names, Enterprise and GST groups, and also looking at shared employees across all years, as the model only looks at shared employees in the year after the firm exits).

We caution interpretation of the results because there were ambiguous cases found using this assessment method. For example, there were several exiting firms employing shearers, where the same employees appear to have worked for half a dozen firms in the following year – it is unclear whether this is a feature of the industry or evidence of connected firms. We have excluded these cases from the truthing data and used clear results to train and test the model.

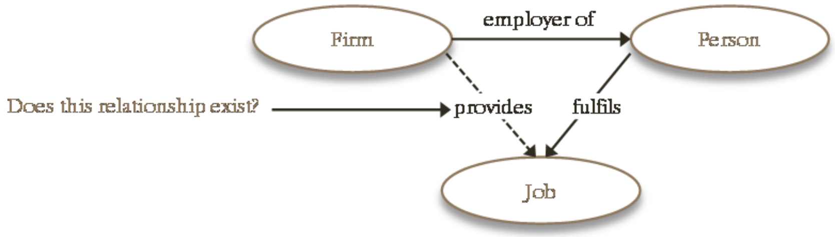
Proof for equation (6)

$$\begin{aligned}\frac{\Pr(\theta|S)\Pr(D^{\text{train}}|\theta,S)}{\Pr(\theta|\boldsymbol{D}^{\text{train}},S)} &= \frac{\dfrac{\Pr(\theta,S)}{\Pr(S)}\times\dfrac{\Pr(D^{\text{train}},\theta,S)}{\Pr(\theta,S)}}{\dfrac{\Pr(\boldsymbol{D}^{\text{train}},\theta,S)}{\Pr(\boldsymbol{D}^{\text{train}},S)}} \\ &= \frac{\Pr(\boldsymbol{D}^{\text{train}},S)}{\Pr(S)} \\ &= \Pr(\boldsymbol{D}^{\text{train}}|S)\end{aligned}$$

APPENDIX

Figure A.1 shows a stylised example of how the ontology describes the simple semantic relationship of a triple. The 'triple' describes two entities (e.g. person and firm) and the semantic relationships between them (e.g. employer/employee of). We could find more information from the data using these relationships. For example, we may not know if the ABS provides a job chief data scientist but we know that Joe X has a job as chief data scientist. We also know that Joe X only works for the ABS. We can deduce that the ABS provides a job chief data scientist. Please note that this paper does not discuss the application of deductive reasoning for data integration. This is an ongoing research area.

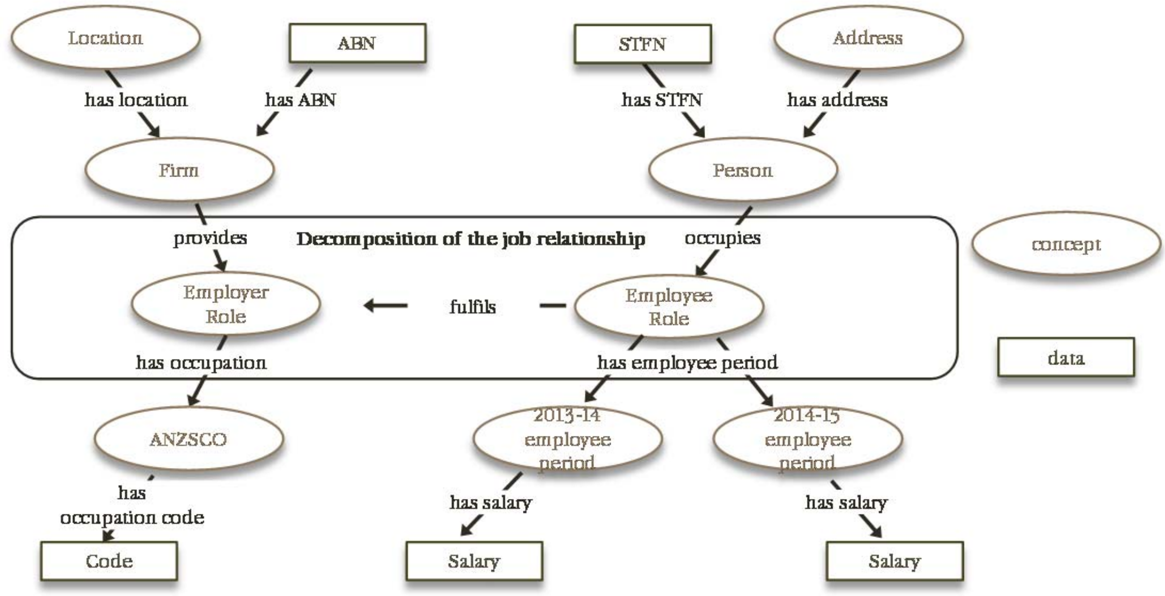
A.1 Some triples for the conceptual relationships



Source: Adapted from (Harwood *et al.*, 2014)

The labour market is complex. The simple example presented above can be easily extended to broader and more complex semantic relationships. New concepts can be added to the ontology as new datasets are integrated, which form new triples to describe the complex relationships. Figure A.2 shows a subset of the ontology specified in the prototype GLIDE. The concept of job, i.e. a relationship between an employee and an employer, is decomposed into two elements in the ontology. This captures the many to many relationships that exist when you consider jobs from the perspective of a firm or a person. For example, consider job sharing, i.e. a firm can provide an employer role which can then be fulfilled by multiple employees. Consider multiple job holders who fulfil different employee roles over multiple periods. By breaking down the basic job relationship into two related concepts (Employee Role and Employer Role), it is possible to ask more detailed questions about these concepts such as, "Who are all the people at the ABS employed as statisticians?" and "Who has been employed in the same occupation at the same company for multiple years?" (Harwood *et al.*, 2014).

A.2 Key elements of the ontology in GLIDE



Source: Authors

ACKNOWLEDGEMENTS

We gratefully acknowledge several people who provided inputs for this paper:
Ric Clarke, Ian Crettenden, Siu-Ming Tam and Peter Rossiter for their comments and guidance at various stage of drafting; Chris Conran, John Machin, Timothy Cosgrove, Michael Possingham and Armin Haller (Statistical Informatics sprint team) for their contributions to the development of GLIDE; David Hansell, David Taylor and Diane Braskic (Innovation and Technology Statistics) and Lachlan McColl, Deepa Wright and Kirsty Leslie (Labour Market Statistics) for their advice and support for the project; Jenny Foster, Pao Franco, Michael Biddington, Paul Romanis and Luisa Ryan (Business Register Unit) for providing data and support for the project; Carmen Kong and David Skutenko (National Tax Data and Business Demography) for providing support and data; Alan Welsh for his advice; Robert Clark (University of Wollongong) for his advice and comments; and members of Methodological Advisory Committee for their comments and advice

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** The ABS website is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics