Australian Bureau of Statistics

Research Paper

# Australian Business Networks

## Australia

## 2021

**1351.0.55.063**

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30AM (CANBERRA TIME) FRI 29 JAN 2021

# AUSTRALIAN BUSINESS NETWORKS

Chien-Hung Chien[1,2], A.H. Welsh[3], and Anton H. Westveld[3]

[1]Australian Bureau of Statistics (ABS)

[2]Australian National University (ANU) Mathematical Science Institute

[3]ANU Research School of Finance, Actuarial Studies & Statistics

**Disclaimer** The results of these studies are based, in part, on tax data supplied by the Australian Taxation Office (ATO) to the ABS 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 or weaknesses is in the context of using the data for statistical purposes, and is not related to the ability of the data to support the ATO's core operational requirements. Views expressed in this paper are those of the authors and do not necessarily represent those of the ABS. Where quoted or used, they should be attributed clearly to the authors.

# ABSTRACT

We demonstrate the value of integrating administrative datasets to study factors that contribute to forming business networks in Australia. We describe how we use a semantic web approach to integrate data, extract business network information and apply statistical network models for analysis. This study uses exponential random graph models (ERGMS) and latent space models (LSMs) to describe the factors contributing to the formation of business networks. We combine different sampling approaches (e.g. stratified sampling, case control sampling and one step snow-ball sampling) to overcome computational problems for the statistical network models. This research shows that it is not appropriate to use a statistical model approach that ignores the endogenous network structure of the data.

We find that larger firms are more likely to form business networks in comparison with small and medium size firms. ERGMs and LSMs suggest that firms have a tendency to form a business network with other firms that have a similar productivity level after GFC. ERGMs suggest that firms are less likely to participate in business networks with other firms that have a similar level of sales after GFC. Firm experience does not affect the probability of forming business networks. This is shown by the insignificant coefficients in most ERGM and LSM results. However, we find that firms with more products are more likely to form business networks.

# Contents

# 1 Introduction

"*A complex business market can be seen as a network where the nodes are business units — manufacturing and service companies and the relationships between them are the threads. Both the threads and the nodes in the business context have their own particular content.*"

Håkansson and Ford (2002, p.133)

As Håkansson and Ford (2002) describe nicely, the market is a business network where firms are nodes and relationships between them are ties. Collaboration, the focus of this study, is one of many different business relationships that lead to the formation of business networks. Ties can also be formed between firms of similar or different sizes, and within or between industries. Firms that do not collaborate become isolates in the network. Nodes and ties exist in the context of every firm's interactions with one another in the market. Firms continuously adapt and change their associations to meet their needs and ensure market success. These firm behaviours affect business outcomes. Firms could gain greater market influence and better market position because of the competitive advantages they develop through participation in business networks (Wilkinson and Young, 2002).

Several studies show that firm participation in business networks (forming ties through collaboration) is positively associated with firm performance (see Belderbos et al. (2004), Miotti and Sachwald (2003) for international studies and Gronum et al. (2012), Soriano et al. (2018), Divisekera and Nguyen (2018) for Australian studies). The Australian Government recognises the benefits that can be gained from firm participation in business networks. It uses a range of initiatives to encourage firm collaboration to enhance business competitiveness and ultimately achieve economic growth. These include the suite of initiatives under the Australian Research Council's Linkage Program, which provide funding for research collaboration between research institutions and industry organisations. The Department of Industry, Innovation and Science's (DIIS) Entrepreneurs' Programme and Cooperative Research Centres provide funding for industry-led collaborations on new technologies, products and services to enhance business competitiveness and productivity (DIIS, 2019b,a).

There are different types of business networks ranging from more structured, e.g. business groups, to less structured, e.g. R&D consortia and commercial associations. These different types of business networks facilitate different degrees of knowledge transfer and create social capital to enhance business performance (Inkpen and Tsang, 2005). We use network analysis to study firms that belong to both R&D and commercial business networks. Focusing on firms belonging to both of these business networks and developing an understanding of how they form business networks will help inform policy that encourages economic growth. Commercialisation of innovation is an area Australia needs to improve in. According to the 2018 Organisation for Economic Co-operation and Development's (OECD) science score board, Australia ranks last in businesses collaborating on innovation with higher education or research institutions (Organisation for Economic Co-operation and Development, 2017). This is despite ranking ninth in research excellence across OECD countries (Organisation for Economic Co-operation and Development, 2017). Commercialising innovation can be an important source of economic growth, especially in the information age when knowledge provides competitive advantages for firms (Jacobs, 2018).

This paper is structured as follows: Section 2 provides the literature review, Section 3 describes scope of the data and our sampling approach to reduce the data to a computationally feasible size for analysis, Section 4 presents the statistical models. We explore approaches that allow statistical inference without independence assumptions. For example, exponential random graph models (ERGMs) allow for inclusion of a set of network statistics derived from the characteristics of vertices or edges (Cranmer and Desmarais, 2011, Cranmer et al., 2017). Section 5 discusses empirical results and the final section gives some conclusions and future directions for further research. Section $C$ in the Appendix discusses estimation methods, Section $D$ in the Appendix provides details for the semantic web data model and queries to extract information on firms belong to both R&D and commercial business networks for analysis.

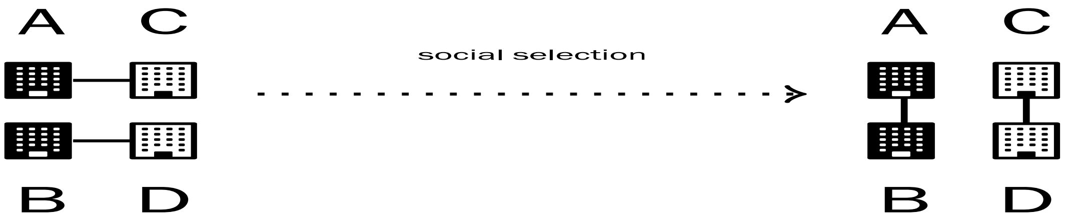
# 2 Literature review

“*Disentangling the effects of selection and influence is one of social science’s greatest unsolved puzzles: Do people befriend others who are similar to them, or do they become more similar to their friends over time?*”

Lewis et al. (2012, p.68)

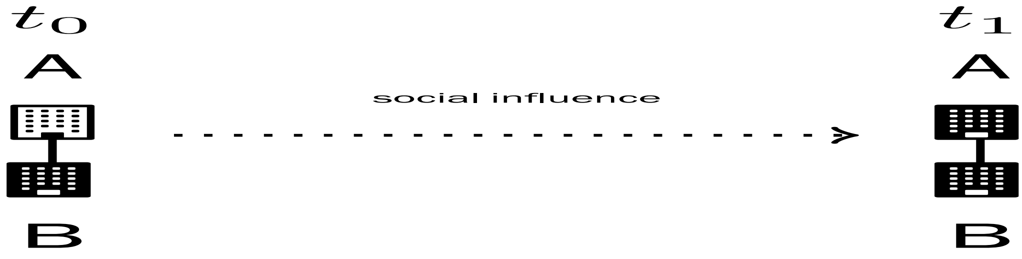
Lewis et al. (2012) succinctly summarise the two key research topics in network analysis: social influence and social selection. The dependent variable in social influence analysis focuses on the nodal attributes of the subject, e.g., productivity, innovation etc. Connections within business networks are included as explanatory variables. In comparison, social selection analysis includes connections within business networks as the dependent variable and nodal and edge attributes of the subjects as explanatory variables. To explain these differences, consider the following example of four different firms (nodes A, B, C and D) (Friemel, 2015). The colour (black or white) indicates the two states of a binary attribute of the firms, e.g., low or high productivity etc. The ties indicate the relations between firms. The relations can be a R&D or a commercial business network. Figure 1 shows the social selection process. Initially, firm A and firm C participate in one business network and firm B and firm D participate in another business network. Over time, the process of homophily, which means similar firms are more likely to participate in a business network with each other, leads to firm A and firm B participating in one business network and firm C and firm D participating in another business network. This is because A and B, as well as C and D are similar (indicated by the colours of the nodes). At the same time we observe zero selections between firm A and C and between firm B and D.

Figure 1: Social selection process



In comparison, Figure 2 shows the social influence process. There exist relations between firm A and B. An influence process affects how firm A adjusts colour to be the same as firm B. For example, if firm B (black) is more productive than firm A, over time firm B exerts influence on firm A so firm A will also become more productive at time $t_1$ .

Figure 2: Social influence process



It is interesting to note that Figures 1 and 2 show that firm A and firm B have the same outcome (indicated by the black colour) despite the underlying selection and influence processes being entirely different.

The literature on the effects of the social influence process on business networks is extensive. Several studies show that firm collaboration in business networks is positively associated with firm performance. Belderbos et al. (2004) and Miotti and Sachwald (2003) analyse Innovation Surveys from the Netherlands and France respectively. They find that firm participation in R&D business networks is an important source of innovation. Firm participation in R&D business networks enables knowledge transfer between firms by sharing technology. Similarly, Australian studies analysing the Australian Bureau of Statistics (ABS) Business Longitudinal Database (BLD), also find that firm participation in business networks plays an important role in enhancing the performance of small & medium size firms. They include Gronum et al. (2012) on innovation and productivity in the manufacturing and services industries, Soriano et al. (2018) on innovation in the food industry and Divisekera and Nguyen (2018) on innovation in the tourism industry.

The literature on the effects of the social selection process on business networks is growing. Kim et al. (2016) use ERGMs to study the formation of board interlock director networks in US publicly listed companies. Mizruchi (1996) describes an interlocking director as a director who is a member of board of directors for at least two firms. Kim et al. (2016) show the importance of including both the nodal attributes of firms and the structural effects of the business network. This is because these structural effects are endogenous and they can influence the formation of business networks when firms share common directors on their boards. Friel et al. (2016) use latent space models (LSMs) to study interlocking directors in Irish publicly listed companies between 2003 and 2013. The LSMs measure the likelihood for two firms to form a tie given the distance between the firms in a latent space (Salter-Townshend and McCormick, 2017). Friel et al. (2016) find the level of interlocking, measured by latent space, increased before and during the Global Financial Crisis (GFC).

Our study adds to the growing literature on the effects of social selection process business networks in Australia. We want to address the following questions: Are similar firms more likely to collaborate when we only consider forms participate in both R&D and commercial business networks? Did the global financial crisis (GFC) have an impact on the process of forming business networks for the sampled firms in our particular setting? This study uses a semantic web approach to integrate open data and ABS data and applies statistical network analysis to address these questions.

# 3 Data

## 3.1 ABS data

The Australian Taxation Office (ATO), Australia Business Register (ABR) and ABS datasets are held in both the Business Longitudinal Analysis Data Environment (BLADE) (ABS and DIIS, 2017) and the prototype Graphically Linked Information Discovery Environment (Chien and Mayer, 2015). The ATO data is provided to the Australian Statistician under the *Taxation Administration Act 1953* and the ABR data is supplied to the Australian Statistician under *A New Tax System (Australian Business Number) Act 1999.* These acts require that these data are only used by the ABS for administering the *Census and Statistics Act 1905.* The ABS is obliged to maintain the confidentiality of individuals and businesses in these ATO and ABR datasets, as well as comply with provisions that govern the use and release of this information, including the *Privacy Act 1988* (ABS, 2015).

This study uses a strict access control protocol. Access to the datasets includes audit trails and is limited to a need to know basis. All ABS officers are legally bound to secrecy under the *Census and Statistics Act 1905.* Officers sign an undertaking of fidelity and secrecy to ensure that they are aware of their responsibilities. The ABS policies and guidelines govern the disclosure of information to maintain the confidentiality of individuals and organisations. This study presents only aggregate results to ensure that they are not likely to enable identification of a worker or a firm. The experimental ABS research dataset contains $10,039,638$ observations containing $2,028,564$ ABNs between 2001–02 and 2012–13.

## 3.2 IP Australia's 2017 Intellectual Property Government Open Data

IPGOD includes over 100 years of IP rights records administered by IP Australia comprising patents, trademarks, designs and plant breeders' rights (IP Australia, 2017, Benjamin et al., 2016). Table 1 describes datasets used to study R&D and commercial business networks. We use the joint patent or trademark applicant information to identify business networks. Patent and trademark applications can be filed by one applicant or multiple applicants. Over the sample period, between 2002–03 and 2012–13 there are $129,306$ applicant–patent application combinations with $21,887$ unique patent applications. The data cleaning step removes $23,530$ applications with no ABN information. In comparison, there are $1,479,276$ applicant–trademark application combinations with $250,378$ unique trademark applications. The data cleaning step removes $40,606$ applications with no ABN information. There are $1,610,202$ applicant–patent or trademark application combinations with $272,480$ unique patent or trademark applications in a combined patents and trademarks dataset. We focus on firms that do and do not participate in both R&D

and commercial business networks. We use a semantic web approach to integrate and extract relevant information on firms (see Appendix $D$ for more discussion). The data cleaning process removes 145 ABNs and 185 applications with no valid records. The sample contains $24,039$ applicant–patent or trademark application combinations with $3,799$ unique ABNs. The number of edges is reduced significantly when we compare raw cleaned IPGOD datasets with the experimental combined ABS—IPGOD datasets.

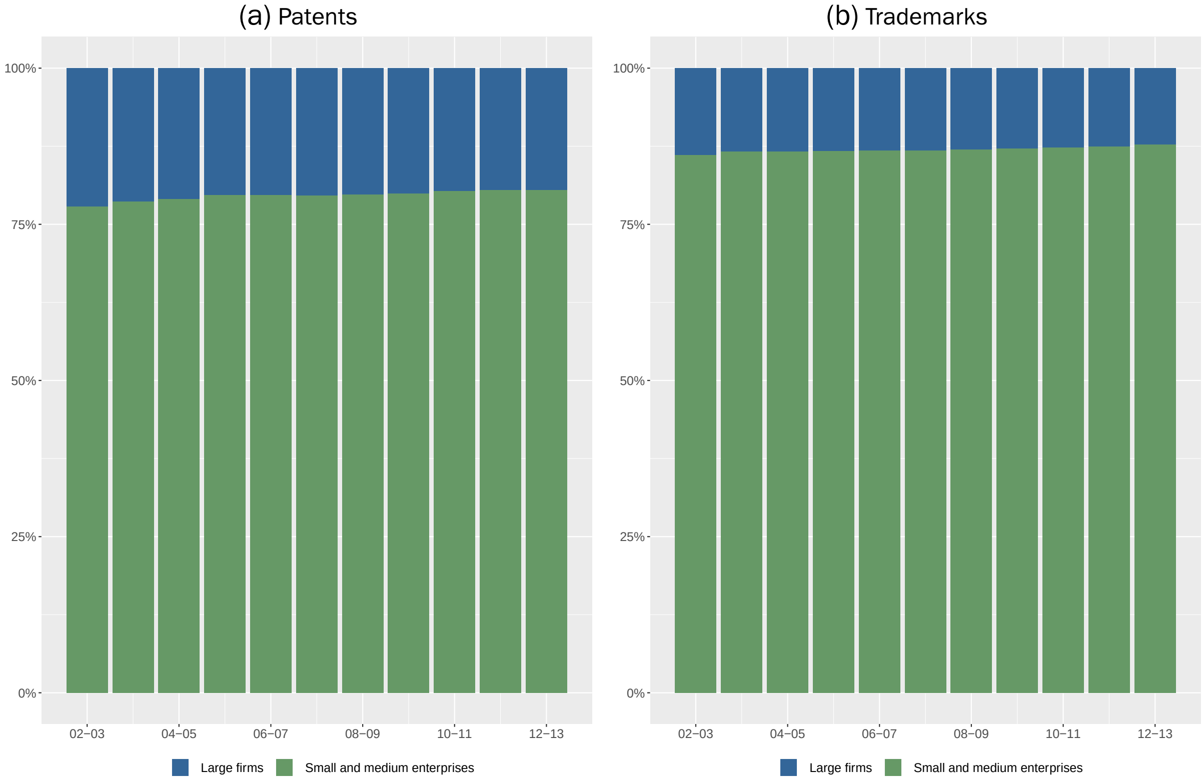
Table 1: Summary of the experimental combined ABS—IPGOD datasets

| **Cleaned IPGOD** | **PAT** | **TMK** | **PATTMK** |
|---|---|---|---|
| Applications firm observations | 129,306 | 1,479,276 | 1,610,202 |
| Number of applications | 21,887 | 250,378 | 272,480 |
| Number of distinct ABN | 7,955 | 82,860 | 86,772 |
| Edges | 17,116 | 45,621 | 59,214 |
| **Experimental combined dataset** | **ABS—PAT** | **ABS—TMK** | **ABS—PATTMK** |
| Application firm observations | 36,291 | 381,305 | 24,039 |
| Number of distinct ABN | 6,228 | 67,686 | 3,799 |
| Edges | 3,826 | 17,867 | 1,306 |

Note. PAT = Patents, TMK = Trademarks, PATTMK = Patents and Trademarks

The five applicant types are international, small or medium-sized enterprises, large firms, private applicant and unknown. We focus only on small and medium-sized enterprises and large firms because these are the only firms we can link to the ABS datasets using Australian Business Numbers (ABNs) or Australian Company Numbers (ACNs). The 100% stacked bar charts in Figure 3 show the proportion of applicant types for patents and trademarks over the sample period before combining datasets. The majority of patent and trademark applicants are from small and medium-sized enterprises.

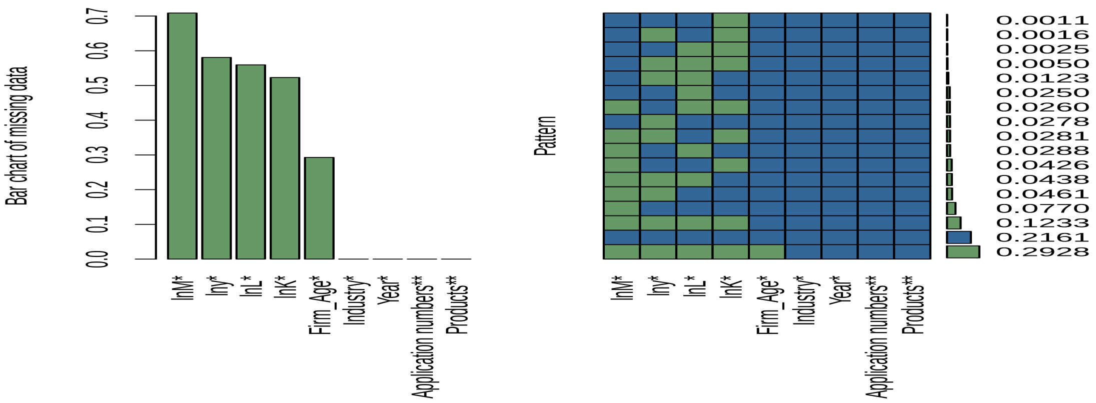
Figure 3: Proportion of applicant types for patents and trademarks between 2002--03 and 2012--13 before data integration



## 3.3 Combining administrative data sources

The benefits and challenges associated with using administrative data for statistical purposes are well documented in Tam and Clarke (2015). Administrative data sources contain deterministic linking keys such as ABNs or ACNs which enable high quality linked datasets. Missing data can still arise, even when quality linking keys are available. Missing data can be caused by the timing of processing or the scope of firms included in the data sources. There are many different approaches to handle missing data, ranging from complete case analysis and mean data imputation to more complex multiple imputation approaches. See Graham (2009) for a detailed discussion. There is no single correct approach to handle missing data. Figure 4 shows that we would lose a significant amount of information if we perform complete case analysis.

Figure 4: Missing data patterns in integrated datasets



Note. The green tiles indicate missing data. The blue tiles indicate non missing data. The left panel is a bar chart showing the propostion of missing data for each variable. The right panel shows the missing data patterns in the data and the proportion of each pattern. These proportions are scaled to increase the readability of the plot (Templ et al., 2012). We remove the subscripts to simplify the notation. The variables ln$L$, ln$K$, ln$M$ and ln$Firm_Age$ are the logarithms of labour for firms, capital for firms, materials used for production and firm age, respectively. The trademark or patent registration numbers is $Application\,numbers$. The number of patent or trademark applications for each firm is $Products$. See Appendix $E$ for more information.
1. * ABS data ** IPA data

We prefer to minimise the loss of information when we integrate datasets for our analysis. Therefore, we assume data are missing at random and impute the missing values using sequential regression in SAS `proc mi` procedure. We create 10 imputed datasets and select the one that maximises the likelihood for model (21) below from the 10 datasets in Appendix $E$. A detailed discussion of the method used in our study can be found in Appendix $B$ and Chien et al. (2018b). All subsequent analyses use the imputed dataset.

## 3.4 A semantic web approach for multiple business networks

The scope of the analysis includes firms that participate (i.e. $Group^{BN}$) and do not participate (i.e. $Group^{\notin BN}$) in R&D and commercial (or multiple) business networks. We define these two groups as

$$\begin{cases} Group^{BN} = & \textbf{Firms that participate in a business network} \\ & \text{Firms file at least one patent application or at least one trademark} \\ & \text{application with a firm that also files } \textit{at least one patent and one} \\ & \textit{trademark} \text{ application.} \\ Group^{\notin BN} = & \textbf{Firms that do not participate in a business network} \\ & \text{Firms file } \textit{at least one patent and one trademark} \\ & \text{on their own and no application with another firm.} \end{cases} \tag{1}$$

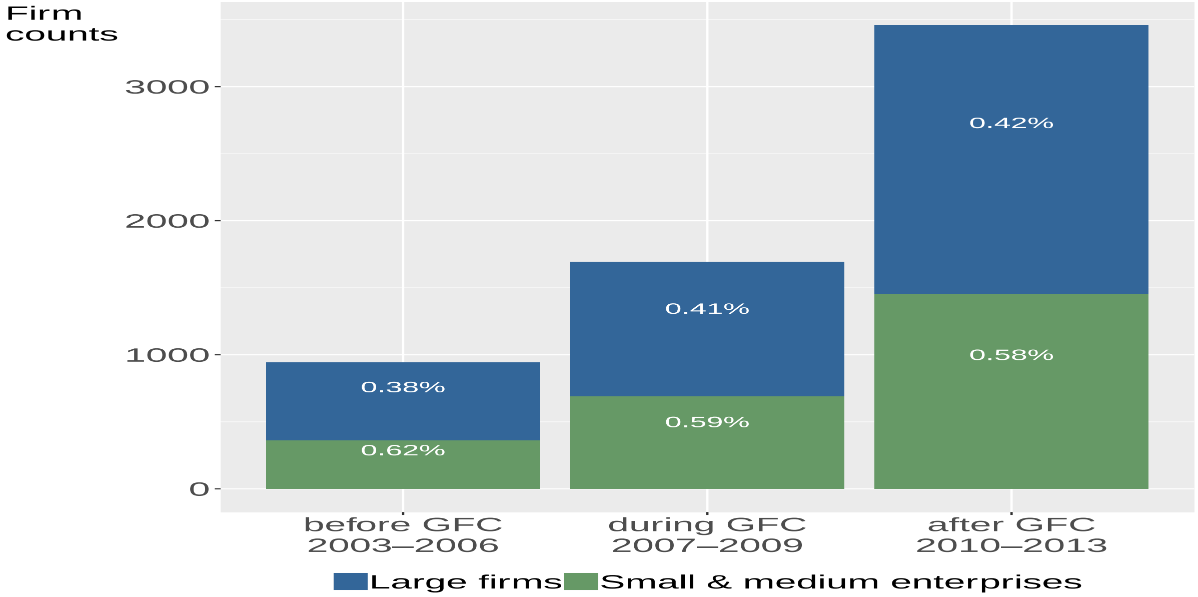
We use the semantic web to integrate datasets from ABS and IPGOD. This approach is well suited to integrate data from multiple sources and to extract information on firms participate in and do not participate in multiple business networks (see Appendix *D*). The semantic web approach is often used to link information online. The research on using this approach to link administrative data is growing (Chien and Mayer, 2015, Clarke and Chien, 2015, 2017).

The semantic web approach allows us to organise administrative data sources by relationships in a graph rather than in a table with different columns (Harwood and Mayer, 2016). The entity of interest is firm which has an unique firm identifier, e.g., ABNs or ACNs. The relationship of interest is whether or not firms belong to multiple business networks. Firms that belong to multiple business networks are connected by joint patent and trademark applications. Conversely, firms that file one patent application and one trademark application on their own are considered isolates. The approach extract network information on firms belong to multiple business networks (see details on the semantic web approach in Chien et al. (2018a)). The semantic web approach is not the only way to organise data in a graph format for network analysis (see Csardi et al. (2006) for an introduction to `igraph` software). However, it is not as easy to capture data provenance when creating graph objects in `igraph`.

Data provenance here refers to the database which contains the administrative records. The semantic web approach also provides a platform for integrating data from multiple sources in a machine interpretable way to enable future extensions and data visualisation (Clarke and Chien, 2017, Berners-Lee et al., 2001).

Figure 5 compares the proportion of large firms to small & medium enterprises in these multiple business networks. The number of business networks has grown over time in the sample. Overall, there are no significant differences between the proportion of large firms and small & medium enterprises over different periods. However, we observe a larger increase in the number of firms after the GFC (2010 to 2013) period.

Figure 5: Proportion of applicants types between 2003 and 2013 in business networks

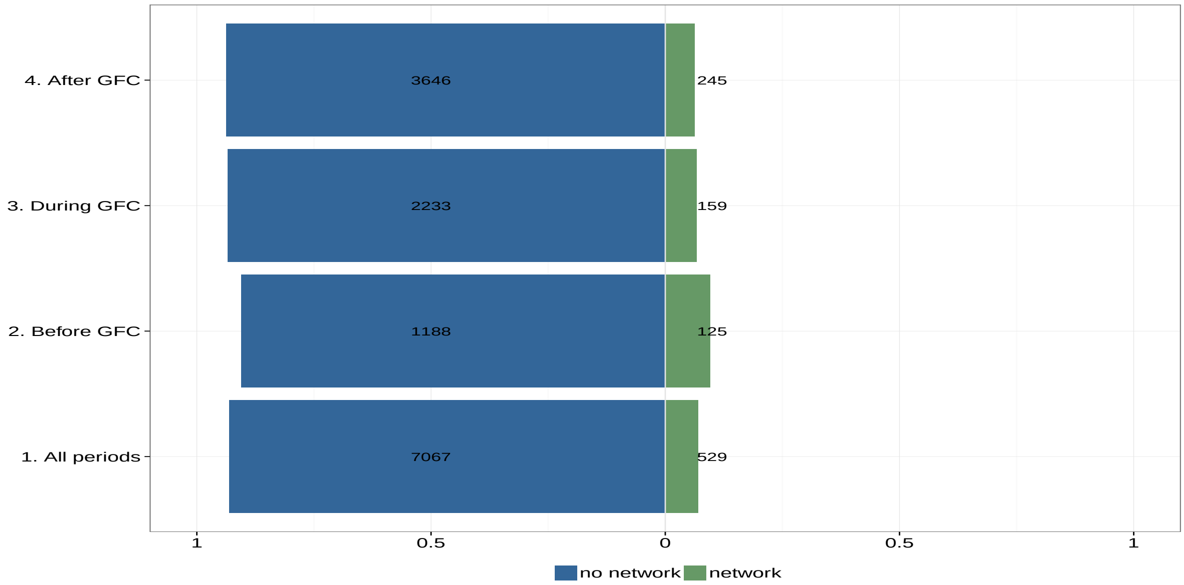


## 3.5 Sampling

We explore both ERGMs and LSMs in this analysis. We discuss both approaches in detail in the next section. Both methods are computationally expensive because Markov chain Monte Carlo (MCMC) process are not well suited to handle large datasets. For example, ERGMs evaluate the probability of an observed connection given all the possible ties in an observed network (Hunter et al., 2012). We have more than $17,000$ firms in our sample which means that MCMC needs to evaluate the probability of an observed tie over more than 289 million possible ties. Salter-Townshend and Murphy (2013) also discuss in detail the computational difficulties of LSMs with MCMC methodology for a network consisting of 604 nodes with 4640 ties.

Therefore we need to reduce the data to a computationally feasible size by taking a sample of the data for analysis. However, we cannot use a simple random sampling technique for our analysis due to the large number of firms that have no connection to any other firms (isolates). Figure 6 shows there is a larger number of firms in $Group^{\notin BN}$ than firms in $Group^{BN}$ in our sample.

Figure 6: Number of firms in $Group^{\notin BN}$ and $Group^{BN}$



Note. GFC = global financial crisis.

Our sample consists of two groups of firms (see the definitions in Section 3.4). We first separate the time series data into three periods: before, during and after the GFC. We want to avoid selecting the same edge or the same firm more than once in the same period. For firms in $Group^{BN}$, if they share the same application number more than once, we randomly select one edge between two firms. Similarly, for firms in $Group^{\notin BN}$, if a firm has more than one observation, we randomly select one observation for that firm. For example, if Firm A has observations in 2003, 2004 and 2005 in the before-GFC period, we randomly choose one to represent that firm in that period. Note that Firm A can be included in a subsequent period.

In Step 1 of the sampling step, we use stratified sampling by industry to select 30% of firms in $Group^{BN}$ from each industry from a total of 18 industries. In Step 2, we use one-step snowball sampling to select all firms that participate in a business network with a firm from Step 1. Finally in Step 3, for firms in $Group^{\notin BN}$, we use stratified sampling by industry to select an equal number of firms in each industry from $Group^{BN}$ (see Algorithm 1 for details). Our sampling approach is similar to case–control sampling. Case–control sampling is frequently used to study factors that contribute to rare diseases. The approach compares subjects who have a disease (treatment group) with similar subjects who do not have the disease (control group; Breslow, 1982). Our 'treatment group', $Group^{BN}$, contains three types of firms in business networks: firms that file at least one patent application, firms that file at least one trademark application and firms that file at least one patent and one trademark application.

Consider the following example for a particular period and industry for our sampling scheme. Table 2 has an adjacency matrix with six firms: $firm\_\alpha$ and $firm\_\beta$ file a joint patent application and $firm\_\beta$, $firm\_\gamma$ and $firm\_\delta$ file a joint trademark application. $Firm\_\epsilon$, $firm\_\zeta$, $firm\_\eta$ and $firm\_\theta$ file both patent and trademark applications on their own. If we sample firm $firm\_\beta$ from Step 1 (as it participates in both patent and trademark business networks), we consider the adjacency is made up of four firms ($firm\_\alpha, firm\_\beta, firm\_\gamma, firm\_\delta$; filling in all connections).

Our control group, $Group^{\notin BN}$, contains only one type of firm: firms that file at least one patent and one trademark application on their own. For example, we sample $firm\_\epsilon$ and $firm\_\eta$ from four isolate firms in Table 2[2].

Table 2: A small adjacency matrix example

| | $firm\_\alpha$ | $firm\_\beta$ | $firm\_\gamma$ | $firm\_\delta$ | $firm\_\epsilon$ | $firm\_\eta$ |
|---|---|---|---|---|---|---|
| $firm\_\alpha$ | 0 | 1 | 0 | 0 | 0 | 0 |
| $firm\_\beta$ | 1 | 0 | 1 | 1 | 0 | 0 |
| $firm\_\gamma$ | 0 | 1 | 0 | 1 | 0 | 0 |
| $firm\_\delta$ | 0 | 1 | 1 | 0 | 0 | 0 |
| $firm\_\epsilon$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $firm\_\eta$ | 0 | 0 | 0 | 0 | 0 | 0 |

It is important to note that we use stratified sampling to select firms that participate in both R&D and commercial business networks in Step 1. We then use snow-ball sampling to select firms that connect with firms in Step 1. This means in our sample we do not have firms that participate in business networks because they connect with firms that only participate in R&D or only participate in commercial business networks.

[2] There are different interpretations to our approach. Some may argue that it is case–control sampling because we are still comparing firms that do or do not participate in business networks. Others may argue that it is not, because if the treatment firms participate in R&D business networks only, then the control group should be sampled from firms that do not participate in R&D business networks. Likewise, if treatment firms participate in commercial business networks only, then the control group should be sampled from firms that do not participate in commercial business networks. Finally, if treatment firms participate in both R&D and commercial business networks, then the control group should be sampled from firms that do not participate in both R&D and commercial business networks.

Algorithm 1 provides information on our sampling approach.

**Algorithm 1** Data Reduction Algorithm

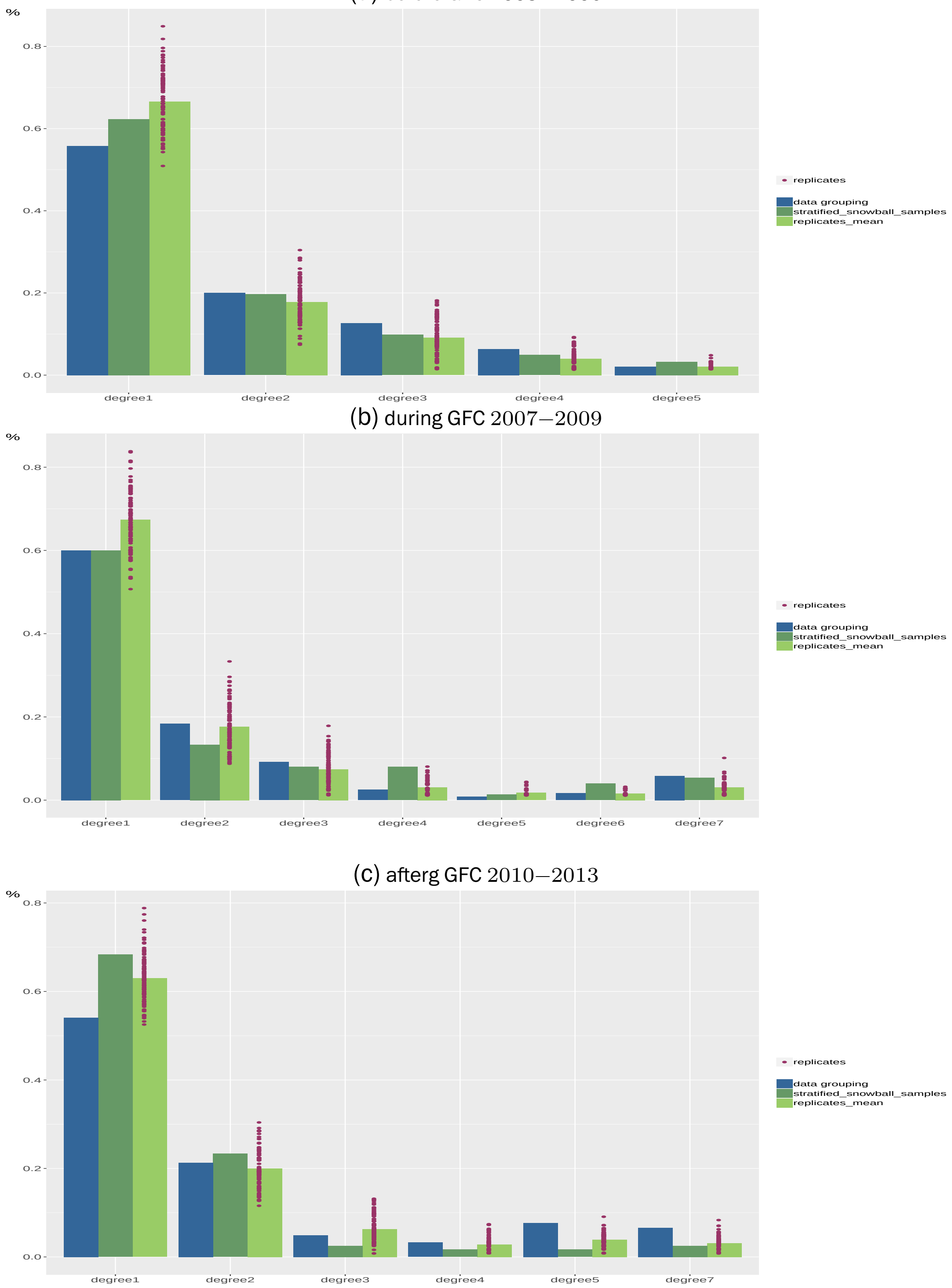
```
procedure
    1. data grouping step groups data into before GFC, during GFC and post GFC.
    for each period $\in$ before GFC, during GFC and post GFC do
        for each firm $i \in Group^{BN}$ do
            randomly select an edge if two firms share the same application number $> 1$.
        for each firm $j \in Group^{\notin BN}$ do
            randomly select one observation if a firm has more than one observation.
    2. sampling step
    for each period $\in$ before GFC, during GFC and post GFC do
        for each firm $i \in Group^{BN}$ do
            Step 1: use stratified sampling by industry to select 30% of firms that participate
            in both patent and trademark business networks.
            Step 2: use one-step snowball sampling to select firms that participate in a business
            network (either patent or trademark) with firms in Step 1 (see Table 2).
        for each firm $j \in Group^{\notin BN}$ do
            Step 3: use stratified sampling and select equal number of firms in each industry
            from the previous step (see Table 2).
```

Figures 10, 11 and 12 in Appendix $F$ show business networks and summary statistics for our sample. They show that more large firms participate in both R&D and commercial business networks than small & medium enterprises. There are more firms participating in business networks during and after the GFC than before the GFC.

Figure 7 compares the degree distributions between the data grouping step and the sampling steps. The blue bars represent the degree distributions from the data grouping step to reduce the data size. The dark green bars are the degree distributions from the sampling step. The light green bars are the degree distributions from the mean values from the 100 replicates from the sampling step. The brown points are the degree distributions from the 100 replicates. Figures $7(a)$, $7(b)$ and $7(c)$ show that the sampling step provides a representative sample because the degree distributions are similar between the data grouping sample and the stratified and snow-ball sample. The degree distributions of the stratified and snow-ball sample are also within the 100 replicates.

Figure 7: Degree distributions of full and case control samples

(a) before GFC 2003—2006



(b) during GFC 2007—2009

(c) afterg GFC 2010—2013

Note. Points are the proportion for 100 replicates of our sampling approach.

# 4 Statistical models

"*The applied statistician should avoid models that are contradicted by observed data in relevant ways-frequency calculations for hypothetical replications can monitor a model's adequacy and help to suggest more appropriate models.*"

Rubin (1984, p.1171)

Rubin (1984) highlights the importance of using appropriate methods to draw valid inference from the observed data. Hoff et al. (2002) argue against using statistical models that assume independence of observations, i.e., actors do not affect each other's outcomes, to study network data. These models cannot capture the intricate relationships that exist in the business network data. There are many interdependent social processes in the data that drive the formation of business networks. The formation of business networks can be influenced by the presence (or absence) of other ties in the network. This complexity is shown by the business network formed between Verizon Wireless and Google in 2009. Verizon Wireless, one of the key wireless telecommunications providers in the United States, wanted to become less reliant on Apple and iPhone to deliver its service to high-paying customers (Svensson, 2009). This was mainly because AT&T, one of Verizon Wireless's main competitors, had already established a close working relationship with Apple (Cohan, 2013). The successful relationship between Verizon Wireless and Google has led to forming business networks with other Android phone manufacturers like Samsung (Tibken, 2009).

Networks are inherently relational so the occurrence of a particular relationship, or tie, could depend on the occurrence of other ties (Koskinen and Daraganova, 2013). Firms consider factors beyond the simple evaluation of the suitable characteristics of prospective partners. Verizon Wireless's decision to form ties with Google and Samsung involves a strategic response to compete against AT&T. An observed business network can result from a combination of simultaneous processes with interdependent endogenous factors (Kim et al., 2016). We follow Cranmer et al. (2017) approach, which compares results from ERGMs and LSMs with results from logistic regression models that assume independence of observations, to show the importance of using models which take into account network structural effects to study Australian business networks.

## 4.1 Exponential random graph models

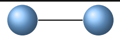
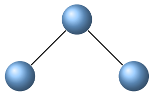
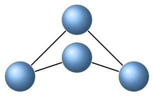
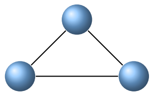
ERGMs take into account the underlying network structure, characteristics of firms and the characteristics of the ties in the inference. ERGMs have two main functions: first, to describe if a given network structure, e.g. edge or stars observed in a network, occurs more than expected by chance; and second, to determine whether there is an association between network ties and firm characteristics, between network ties and tie characteristics or between network ties and both firm and tie characteristics (Valente, 2010).

We build on the work of Desmarais and Cranmer (2012), Broekel and Hartog (2013) and Balland et al. (2016) and use (7) in Appendix $C$.1 to define ERGMs as

$$Pr(\mathbf{Y} = \mathbf{y} \mid \mathbf{X}, \theta) = \frac{1}{k(\theta)}\, exp\,[\theta^{\top} g(\mathbf{y}, \mathbf{X})], \tag{2}$$

where $\mathbf{y}$ is the observed business network and the components of $\mathbf{y}$ take value 1 if there is a tie between firm $i$ and firm $j$ and 0 otherwise. The symbol $\theta$ represents unknown parameters of interest and determines the effects of the network statistics. We use $g(\mathbf{y}, \mathbf{X})$ to represent the network statistics in the model. The term $k(\theta)$ is the normalising constant. We follow Broekel and Hartog (2013) and explore three network statistics to study R&D collaborations. Table 3 shows examples of different network structures in ERGMs (Robins et al., 2007a).

Table 3: Non-directed network structures

| network structures | graphic configurations |
|---|---|
| edge |  |
| two-star |  |
| two-paths |  |
| triangle |  |

It is convenient to write the term $\mathbf{X}$ which describes exogenous explanatory variables including network statistics in Wilkinson and Rogers (1973) notation as $intercept + gwdegree + gwdsp + gwesp + |Mfp_i - Mfp_j| + |Sales_i - Sales_j| + Largefirm + SME + Firm\_Age + Products + Industry + State$ (See Appendix $E$ for information on the variables). The term *intercept* measures the effects of edge on forming business networks. The term *gwdegree* is the geometrically weighted degree statistic. The *gwdegree* statistic measures the existence of preferential attachment where there is a tendency for nodes in a growing network to form connections with nodes with high numbers of connections (de Blasio et al., 2007). The *gwdegree* term can also be considered as an equivalent to the more traditional k-star statistic (Hunter, 2007). The geometrically weighted dyadwise share partner *gwdsp* statistic measures shared partners for firms, whether or not these firms participate in a business network or not (Harris, 2014). For example, regardless whether or not firm $i$ and firm $j$ participate in a business network, if they have a partner in common, are they more likely than expected by chance to have a second partner firm in common? In comparison, the geometrically weighted edgewise share partner *gwesp* statistic measures shared partners only for firms participate in a business network (Hunter, 2007, Harris, 2014). For example, given that firm $i$ and firm $j$ participate in a business network, are they more likely than expected to have multiple shared partners? The firm pair-specific characteristics, $|Mfp_i - Mfp_j|$, is the pair-specific absolute difference in the level of productivity between firm $i$ and firm $j$. The variable $|Sales_i - Sales_j|$ is the pair-specific absolute difference in the level of sales between firm $i$ and firm $j$. We measure homophily, which means similar firms are more likely to participate in a business network, by using indicator variables $LargeFirm$, $SME$ and $LargeFirmSME$. The $LargeFirm$ indicates a tie is formed between two large firms. The $SME$ indicates a tie between two small & medium enterprises. The reference group $LargeFirmSME$ is a tie form between a large firm and a small & medium enterprise. The firm-specific characteristics are $Products$, the number of registered patents or trademarks for each firm, $Firm\_Age$, the firm age derived by year $t$ minus the year of incorporation of firm $j$, $Industry$, the industry dummy indicator variables, $State$, the state dummy indicator variable.

### 4.1.1 Exponential random graph model term

ERGMs capture the change of the probability distribution of the network configurations when we introduce a new tie. ERGMs become unstable if an additional tie significantly increases the number of network configurations (e.g., by adding an edge, a two-star transforms into a triangle at the same time, introducing two new two-stars; a two-star becomes a three-star etc.). The cascading effects cause the probability density to have high weights on selected configurations (Snijders et al., 2006). This causes the probability distribution to degenerate in the model. This problem is referred to as ‘model degeneracy’, where large regions of the parameter space are concentrated on a small number of network configurations (Li, 2015, Hunter et al., 2012). A common symptom is that the MCMC algorithms do not converge (Handcock et al., 2003).

Robins et al. (2007b) describe the problem of model degeneracy when ERGMs are used to capture two-star network configurations, commonly observed in business network data. This problem becomes prevalent when the number of nodes increases because it exacerbates the cascading effects (van der Pol, 2018). We use a network specification proposed by Hunter and Handcock (2006) and Hunter (2007) to handle model degeneracy. The method introduces a weighted degree distribution term in the model. This term gives a higher weight to low density while decreasing the weight as the observed occurrence of a network configuration increases (see Appendix $C$.1.1).

## 4.2 Latent space model

LSMs use a latent space to capture dependency of the network structure without the need to specify the network configurations (Shortreed et al., 2006). This approach can avoid the model degeneracy commonly found in ERGMs. However, the theoretical framework is still being developed, particularly the number of latent space dimensions. Increasing the number of dimensions generally captures the dependency structure better, but at the same time makes the model specification more complex and difficult to interpret (Cranmer et al., 2017).

We follow Hoff et al. (2002), Westveld and Hoff (2011) and build on the work of Friel et al. (2016) and use (13) in Appendix $C.2$ to define LSMs as

$$Pr(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z}, \mathbf{X}, \theta^{*}) = \prod_{i \neq j} Pr(y_{i,j} \mid z_i, z_j, \mathbf{X}, \theta^{*}), \tag{3}$$

where $\mathbf{y}$ is the observed business network and it takes value 1 if there is a tie between firm $i$ and firm $j$ and 0 otherwise. The symbols $\theta^{*}$ and $\mathbf{Z}$ are the unknown parameters of interest and the latent positions to be estimated. We include the same explanatory variables as ERGMs in the design matrix. The estimated parameters $\theta^{*}$ include an intercept term $\alpha$ but exclude all the network structure terms from ERGMs.

## 4.3 Logistic regression model

We follow Cranmer et al. (2017) to show the importance of using ERGMs and LSMs that capture network structure from the data to study Australian business networks. We compare the results from ERGMs and LSMs with the results from logistic regression models that assume independence of observations. We use Appendix $C.4$ to specify the formula for the logistic regression models as

$$Pr(\mathbf{Y} = \mathbf{y} \mid \mathbf{X}, \theta^{**}) = exp\,(\mathbf{X}^{\top}\theta^{**}), \tag{4}$$

where $\mathbf{y}$ is the observed business network and the components of $\mathbf{y}$ take value 1 if firm $i$ participates in business networks. We include the same explanatory variables as ERGMs in the design matrix including network structure terms. The estimated parameters $\theta^{**}$ also include an intercept term $\alpha^{**}$.

# 5 Empirical results

We use ergm (Handcock et al., 2018) and latentnet (Handcock and Krivitsky, 2008) to estimate our models. Some useful references and tutorials can be found in Hunter (2007), Krivitsky (2014), Levy (2016). The results are in log-odds, as discussed in Appendix $C.1$ and C.2. We calculate $exp(\hat{\theta})/[1 + exp(\hat{\theta})]$ to interpret the estimated coefficients as probability (Mood, 2009). For ERGMs and LSMs, we specify the decay parameters by relying on a manual iterative trial-and-error process of estimating model specifications (Broekel and Hartog, 2013). The process ends when the models are converged, i.e., the trace plots are horizontal and density plots look normal. The trace and density plots are in Appendix $H$.

Table 4, 5 and 6 in Appendix $G$ show the estimated results for three models for three different periods. While most results are broadly consistent, some coefficients are different in magnitude and significance and have different signs. This shows that different modelling approaches can lead to different conclusions. Similar to Cranmer et al. (2017), we have found that coefficients have different signs in the logistic regression models. While there are more significant coefficients in both ERGMs and LSMs results, it is interesting to note that there are more in the ERGMs. These differences are likely due to LSMs capturing more network dependency structure in the estimation. We have included different network configurations, but adding more terms led to the MCMC algorithms for fitting the ERGMs not converging.

Our preliminary results suggest that the network's degree distribution *gwdegree* does not help in explaining firms that participate in the business networks before and during GFC. There is evidence to support preferential attachment processes after GFC because the estimated coefficient is positive and significant. There are mixed results for the geometrically weighted dyadwise share partner *gwdsp* statistics. The *gwdsp* statistic is negative and insignificant before GFC, negative and significant during GFC, but positive and significant after GFC. This means that firms are less likely to be indirectly connected regardless of whether they participate in a business network or not before GFC and during GFC. However, they are more likely to be indirectly connected after the GFC period. In comparison, we find positive and significant coefficients of the *gwesp* statistic over the three periods. This implies that triangles are a common feature of the network in this sample, which also corresponds to the visual inspection of the network (Snijders et al., 2006, Broekel and Hartog, 2013) (see Appendix $F$).

We find mixed evidence on the absolute difference in the level of productivity between two firms affects the probability of firms that participate in business networks. The magnitude of coefficients are similar in both ERGMs and LSMs, but they have different signs in logistic regression models. The coefficients are negative and insignificant in ERGMs and

LSMs before GFC, positive and insignificant in ERGMs and LSMs during GFC, and positive and significant in ERGMs and positive and insignificant in LSMs after GFC. ERGMs ($\approx 0.52$ significant at 5% level) and LSMs ($\approx 0.56$ significant at 10% level) suggest that firms have a tendency to form a business network with other firms that have a similar productivity level after GFC. Similarly, we find the absolute difference in the level of sales between two firms contributes differently to the probability of forming business networks over the three periods. The coefficients have consistent signs over the three periods but they are mostly insignificant. ERGMs suggest that firms are less likely ($\approx 0.51$ significant at 5% level) to participate in business networks with other firms that have a similar level of sales after GFC. LSMs suggest that firms are more likely ($\approx 0.58$ significant at 10% level) to participate in business networks with other firms that have a similar level of sales before GFC.

We have mixed evidence for homophily: it is true for $LargeFirm$, but not for $SME$. The coefficients for $LargeFirm$ are all positive and significant, while the coefficients for $SME$ are all negative but insignificant in LSMs compared to the reference group (a tie between $LargeFirm$ and $SME$) during and after GFC. There is evidence in our study to support the conclusion that large firms are more likely to participate in business networks with other large firms because the estimated probability is at least 0.72 significant at 5% level before GFC and higher in other periods. Small firms are less likely to participate in business networks with other small firms because ERGMs suggest that the estimated probability is at least 0.62 significant at 5% level post GFC and the probabilities are higher in other periods in this sample. Our results are similar to Kim et al. (2016), who find that larger firms with more resources are more likely to form business networks.

The coefficients for the firm-specific characteristics $Firm\_Age$ and $Products$ are positive. However, firm experience does not affect the probability of forming business networks. This is shown by the insignificant coefficients in both the ERGMs and LSMs results. However, we find that firms with more products are more likely to form business networks. The estimated probabilities are higher from LSMs than ERGMs. The probability is 0.7 significant at 5% level for LSMs in comparison with 0.57 significant at 5% level in ERGMs for before GFC, 0.75 significant at 5% level for LSMs in comparison with 0.55 significant at 5% level in ERGMs for during GFC and 0.67 significant at 5% level for LSMs in comparison with 0.56 significant at 5% level in ERGMs for after GFC.

## 6 Conclusions, limitations and future directions

We demonstrate the possibility of using administrative data to study firms participating in multiple business networks. We use a semantic web approach to integrate administrative data from different sources and extract business network information so we can fit statistical network models to study factors contributing to forming these business networks.

We show that it is not appropriate to use a statistical model approach that ignores the endogenous network structure of the data. As we expected, using models that incorporate the network dependence, which are more suitable for the data, lead to different results than a model which assumes independence.

We find broadly consistent with some differences in results from two statistical network modelling approaches. In our analysis using 30% of firms in the sample, we find that large firms are more likely to form multiple business networks. This may suggest that they have more resources to commercialise their innovations. As we expect, firms with more registered products are also more likely to form multiple business networks.

It is important to note that we do not observe all types of business networks in the administrative data. This is because not all firm collaborations will lead to joint patent or trademark applications. This highlights the opportunities of combining administrative data with survey data because information on other types of business networks is generally not collected by administrative agencies.

Our research could be extended in several areas. One possibility is to compare our results with more computationally efficient variational Bayesian approaches (e.g., Salter-Townshend and Murphy (2013)). These approaches would allow us to analyse larger datasets. It would be useful to compare the results of this study with the results from using these approaches. It would be interesting to compare our results with results using the approach of Tranmer et al. (2014) which captures both the hierarchical and network structures in the data for the analysis.

## 7 Acknowledgements

We would like to express our gratitude to the following ABS colleagues - Lisette Aaron, Michael Beahan, Diane Braskic, Bruce Hockman, Dr Anders Holmberg, Grace Kim, Sybille McKeown, Katrina Richardson, Paul Roberts, Dr Siu-Ming Tam, David Taylor, Liza Tiy and Carter Wong for their helpful comments and support for this research. We acknowledge Laurent Lefort and Chris Conran, who provided advice. We also thank those who provided comments at the 2018 SemStats workshop in the 17th International Semantic Web Conference and The Third Annual Australian Social Network Analysis Conference 2018. We remain solely responsible for the views expressed in this paper.

# 8 References

Abowd, J. M., Creecy, R. H., and Kramarz, F. (2002), "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," Report, US Census Bureau, `ftp://ftp2.census.gov/ces/tp/tp-2002-06.pdf`[Accessed: 12022016].

ABS (2015), "Information Paper: Construction of Experimental Statistics on Employee Earnings and Jobs from Administrative Data, 2011-12," `https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/6311.0main+features12011-12`[Accessed: 01032018].

— (2018), "8158.0 - Innovation in Australian Business, 2016-17," `https://www.abs.gov.au/ausstats/abs@.nsf/mf/8158.0`[Accessed: 01022019].

ABS and DIIS (2017), "Business Longitudinal Analysis Data Environment," `https://industry.gov.au/Office-of-the-Chief-Economist/Data/Pages/Business-Longitudinal-Analysis-Data-Environment.aspx`[Accessed: 06062017].

Agresti, A. (2007), *An introduction to categorical data analysis*, New Jersey: John Wiley & Sons, Inc.

Balland, P.-A., Belso-Martínez, J. A., and Morrison, A. (2016), "The Dynamics of Technical and Business Knowledge Networks in Industrial Clusters: Embeddedness, Status, or Proximity?" *Economic Geography*, 92, 35–60.

Belderbos, R., Carree, M., and Lokshin, B. (2004), "Cooperative R&D and firm performance," *Research Policy*, 33, 1477–1492.

Benjamin, M., Matthew, J., Bradley, M., and Luke, M. (2016), "Intellectual Property Government Open Data: Australian Business Number Links to All Intellectual Property Data in Australia," *Australian Economic Review*, 49, 96–104, `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8462.12147`[Accessed: 01022017].

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001), "The semantic web," *Scientific American*, 284, 28–37.

Breslow, N. (1982), "Design and analysis of case-control studies," *Annual review of public health*, 3, 29–54.

Breunig, R. and Wong, M.-H. (2008), "A Richer Understanding of Australia's Productivity Performance in the 1990s: Improved Estimates Based Upon Firm-Level Panel Data," *Economic Record*, 84, 157–176.

Broekel, T. and Hartog, M. (2013), *Determinants of cross-regional R&D collaboration networks: an application of exponential random graph models*, Springer, pp. 49–70.

Chib, S. and Greenberg, E. (1995), "Understanding the metropolis-hastings algorithm," *The american statistician*, 49, 327–335.

Chien, C.-H., Haller, A., and Westveld, A. H. (2018a), "Firm Business Networks," in *SemStats@ISWC*, http://ceur-ws.org/Vol-2317/article-13.pdf[Accessed: 20122018].

Chien, C.-H. and Mayer, A. (2015), "A New Analytical Platform to Explore Linked Data," Report, Australian Bureau of Statistics, http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.151[Accessed: 05082016].

Chien, C.-H., Welsh, A. H., and Breunig, R. V. (2019), "Approaches to Analysing Micro-Drivers of Aggregate Productivity," Report, Australian Bureau of Statistics, https://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.164[Accessed: 20062019].

Chien, C.-H., Welsh, A. H., and Moore, J. D. (2018b), "Synthetic Microdata - A Possible Dissemination Tool," Report, Australian Bureau of Statistics, http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.163[Accessed: 16112018].

Clarke, F. and Chien, C.-H. (2015), "Connectedness and Meaning: New Analytical Directions for Official Statistics," in *SemStats@ISWC*, http://ceur-ws.org/Vol-1551/article-02.pdf[Accessed: 15122018].

— (2017), *Visualising Big Data for Official Statistics*, IGI Global.

Cohan, P. (2013), "Project Vogue: Inside Apple's iPhone Deal With AT&T," https://www.forbes.com/sites/petercohan/2013/09/10/project-vogue-inside-apples-iphone-deal-with-att/#6122230d4d3c[Accessed: 03122018].

Cranmer, S. J. and Desmarais, B. A. (2011), "Inferential network analysis with exponential random graph models," *Political Analysis*, 19, 66–86.

Cranmer, S. J., Leifeld, P., McClurg, S. D., and Rolfe, M. (2017), "Navigating the range of statistical tools for inferential network analysis," *American Journal of Political Science*, 61, 237–251.

Csardi, G., Nepusz, T., et al. (2006), "The igraph software package for complex network research," *InterJournal, Complex Systems*, 1695, 1–9.

Czepiel, S. A. (2002), “Maximum likelihood estimation of logistic regression models: theory and implementation,” `http://ww.saedsayad.com/docs/mlelr.pdf`[Accessed: 01032019].

Davison, A. C. (2003), *Statistical models*, vol. 11, New York, UNITED STATES: Cambridge University Press.

de Blasio, B. F., Svensson, k., and Liljeros, F. (2007), “Preferential attachment in sexual networks,” *Proceedings of the National Academy of Sciences*, 104, 10762.

Desmarais, B. and Cranmer, S. (2012), “Statistical mechanics of networks: Estimation and uncertainty,” *Physica A: Statistical Mechanics and its Applications*, 391, 1865 – 1876.

DIIS (2019a), “Cooperative Research Centres Projects,” `https://www.adelaide.edu.au/research-services/funding/initiatives/crc-projects/`[Accessed: 24032019].

— (2019b), “Entrepreneurs’ Programme,” `https://www.business.gov.au/assistance/entrepreneurs-programme`[Accessed: 24032019].

Divisekera, S. and Nguyen, V. K. (2018), “Drivers of innovation in tourism: An econometric study,” *Tourism Economics*, 24, 998–1014.

Drechsler, J. (2011), *Synthetic Datasets for Statistical Disclosure Control Theory and Implementation*, Lecture Notes in Statistics, New York: Springer.

Friel, N., Rastelli, R., Wyse, J., and Raftery, A. E. (2016), “Interlocking directorates in Irish companies using a latent space model for bipartite networks,” *Proceedings of the National Academy of Sciences*, 113, 6629.

Friemel, T. N. (2015), *Opinion Leadership| Influence Versus Selection: A Network Perspective on Opinion Leadership*, vol. 9 of *2015*, `https://ijoc.org/index.php/ijoc/article/view/2806`[Accessed: 23032019].

Goodreau, S. M., Kitts, J. A., and Morris, M. (2009), “Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks,” *Demography*, 46, 103–125.

Graham, J. W. (2009), “Missing data analysis: Making it work in the real world,” *Annual review of psychology*, 60, 549–576.

Gronum, S., Verreynne, M.-L., and Kastelle, T. (2012), “The Role of Networks in Small and Medium-Sized Enterprise Innovation and Firm Performance,” *Journal of Small Business Management*, 50, 257–282, `http://dx.doi.org/10.1111/j.1540-627X.2012.00353.x`[Accessed: 20012017].

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. (2018), *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*, The Statnet Project (`http://www.statnet.org`), r package version 3.9.4.

Handcock, M. S. and Krivitsky, P. N. (2008), “Fitting Latent Cluster Models for Networks with latentnet,” *Journal of Statistical Software*, 24.

Handcock, M. S., Robins, G., Snijders, T., Moody, J., and Besag, J. (2003), “Assessing degeneracy in statistical models of social networks,” Tech. rep., Citeseer, access at .

Harris, J. K. (2014), *Building a Useful Exponential Random Graph Model*, Thousand Oaks, California: SAGE Publications, Inc.

Harwood, A. and Mayer, A. (2016), “Big data and semantic technology: A future for data integration, exploration and visualisation,” *Statistical Journal of the IAOS*, 32, 613–626.

Håkansson, H. and Ford, D. (2002), “How should companies interact in business networks?” *Journal of Business Research*, 55, 133–139.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2001), “Latent Space Approaches to Social Network Analysis,” Report, University of Washington, `https://www.stat.washington.edu/index.php/article/tech-report/latent-space-approaches-social-network-analysis`[Accessed: 01072018].

— (2002), “Latent space approaches to social network analysis,” *Journal of the american Statistical association*, 97, 1090–1098.

Hunter, D. R. (2007), “Curved Exponential Family Models for Social Networks,” *Social networks*, 29, 216–230.

Hunter, D. R. and Handcock, M. S. (2006), “Inference in curved exponential family models for networks,” *Journal of Computational and Graphical Statistics*, 15, 565–583.

Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), “Computational statistical methods for social network models,” *Journal of Computational and Graphical Statistics*, 21, 856–882.

Inkpen, A. C. and Tsang, E. W. K. (2005), “Social Capital, Networks, and Knowledge Transfer,” *The Academy of Management Review*, 30, 146–165.

IP Australia (2017), “Intellectual Property Government Open Data 2017,” http://data.gov.au/dataset/intellectual-property-government-open-data-2017[Accessed: 03082017].

Jacobs, I. (2018), “Australia can do a better job of commercialising research—here’s how,” , 1–4http://theconversation.com/australia-can-do-a-better-job-of-commercialising-research-heres-how-95526[Accessed: 03032019].

Kim, J. Y., Howard, M., Cox Pahnke, E., and Boeker, W. (2016), “Understanding network formation in strategy research: Exponential random graph models,” *Strategic management journal*, 37, 22–44.

Koskinen, J. and Daraganova, G. (2013), “Exponential random graph model fundamentals,” in *Exponential random graph models for social networks: Theory, methods, and applications*, eds. Lusher, D., Koskinen, J., and Robins, G., Cambridge University Press, chap. 6, pp. 49–76.

Krivitsky, P. N. (2014), “Latent space models with latentnet,” http://statnet.csde.washington.edu/workshops/SUNBELT/current/latentnet/latentnet.pdf[Accessed: 23032019].

Levy, M. (2016), “ERGM Tutorial,” http://michaellevy.name/blog/ERGM-tutorial/[Accessed: 22092018].

Lewis, K., Gonzalez, M., and Kaufman, J. (2012), “Social selection and peer influence in an online social network,” *Proceedings of the National Academy of Sciences*, 109, 68–72.

Li, K. (2015), “Degeneracy, Duration, and Co-evolution: Extending Exponential Random Graph Models (ERGM) for Social Network Analysis,” Ph.D. thesis, https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/34190/Li_washington_0250E_14640.pdf?sequence=1&isAllowed=y[Accessed: 23032019].

Mare, D. C., Hyslop, D. R., and Fabling, R. (2017), “Firm productivity growth and skill,” *New Zealand Economic Papers*, 302–326.

Martina Morris, Steven M. Goodreau, S. M. J. (2018), “Network Modeling for Epidemics,” http://statnet.github.io/nme/index.html[Accessed: 01032019].

Miotti, L. and Sachwald, F. (2003), “Co-operative R&D why and with whom?” *Research Policy*, 32, 1481–1499.

Mizruchi, M. S. (1996), “What Do Interlocks Do? An Analysis, Critique, and Assessment of Research on Interlocking Directorates,” *Annual Review of Sociology*, 22, 271–298.

Mood, C. (2009), “Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It,” *European Sociological Review*, 26, 67–82.

Morris, M., Handcock, M. S., Butts, C. T., Hunter, D. R., Goodreau, S. M., de Moll, S. B., and , Krivitsky, P. N. (2016), “Exponential Random Graph Models (ERGMs) using statnet tutorial,” `https://statnet.org/trac/raw-attachment/wiki/Sunbelt2016/ergm_tutorial.html`[Accessed: 01032019].

Morris, M., Handcock, M. S., and Hunter, D. R. (2008), “Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects,” *Journal of statistical software*, 24, 1548–7660.

Nguyen, T. and Hansell, D. (2014), “Firm dynamics and productivity growth in Australian manufacturing and business services Oct 2014,” Report, ABS, `https://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.052`[Accessed: 20052017].

Organisation for Economic Co-operation and Development (2017), *OECD Science, Technology and Industry Scoreboard 2017: The digital transformation*, OECD, `https://www.oecd-ilibrary.org/science-and-technology/oecd-science-technology-and-industry-scoreboard-2017/research-excellence-and-specialisation_sti_scoreboard-2017-14-en`[Accessed: 03032019].

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), “A multivariate technique for multiply imputing missing values using a sequence of regression models,” *Survey methodology*, 27, 85–96.

Reiter, J. P. (2005), “Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 185–205.

Robert, C. P. (2015), *The Metropolis–Hastings Algorithm*, American Cancer Society.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a), “An introduction to exponential random graph (p*) models for social networks,” *Social Networks*, 29, 173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b), “Recent developments in exponential random graph (p*) models for social networks,” *Social Networks*, 29, 192–215.

Rubin, D. B. (1984), “Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician,” *The Annals of Statistics*, 12, 1151–1172.

Salter-Townshend, M. and McCormick, T. H. (2017), “Latent space models for multiview network data,” *The annals of applied statistics*, 11, 1217.

Salter-Townshend, M. and Murphy, T. B. (2013), “Variational Bayesian inference for the latent position cluster model for network data,” *Computational Statistics & Data Analysis*, 57, 661–671.

Schomaker, M. and Heumann, C. (2014), “Model selection and model averaging after multiple imputation,” *Computational Statistics and Data Analysis*, 71, 758–770.

Shortreed, S., Handcock, M. S., and Hoff, P. (2006), “Positional estimation within a latent space model for networks,” *Methodology*, 2, 24–33.

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), “New specifications for exponential random graph models,” *Sociological methodology*, 36, 99–153.

Soriano, F. A., Villano, R. A., Fleming, E. M., and Battese, G. E. (2018), “What’s driving innovation in small businesses in Australia? The case of the food industry,” *Australian Journal of Agricultural and Resource Economics*.

Svensson, P. (2009), “Verizon, Google in Android partnership,” `http://www.nbcnews.com/id/33192558/ns/technology_and_science-tech_and_gadgets/t/verizon-google-android-partnership/#.XEKWjlxzSUl`[Accessed: 03122018].

Tam, S.-M. and Clarke, F. (2015), “Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics,” *International Statistical Review*, 83, 436–448.

Templ, M., Alfons, A., and Filzmoser, P. (2012), “Exploring incomplete data using visualization techniques,” *Advances in Data Analysis and Classification*, 6, 29–47.

Tibken, S. (2009), “Samsung, Verizon will partner on 5G smartphone in first half of 2019,” `https://www.cnet.com/news/samsung-verizon-will-partner-on-5g-smartphone-in-first-half-of-2019`[Accessed: 03122018].

Tranmer, M., Steel, D., and Browne, W. J. (2014), “Multiple-membership multiple-classification models for social network and group dependences,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177, 439–455.

Valente, T. W. (2010), *Social networks and health: Models, methods, and applications*, vol. 1, Oxford University Press New York.

van der Pol, J. (2018), “Introduction to Network Modeling Using Exponential Random Graph Models (ERGM): Theory and an Application Using R-Project,” *Computational Economics*, `https://doi.org/10.1007/s10614-018-9853-2`[Accessed: 01062019].

Westveld, A. H. and Hoff, P. D. (2011), “A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict,” *The Annals of Applied Statistics*, 5, 843–872.

Wilkinson, G. N. and Rogers, C. E. (1973), “Symbolic Description of Factorial Models for Analysis of Variance,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22, 392–399.

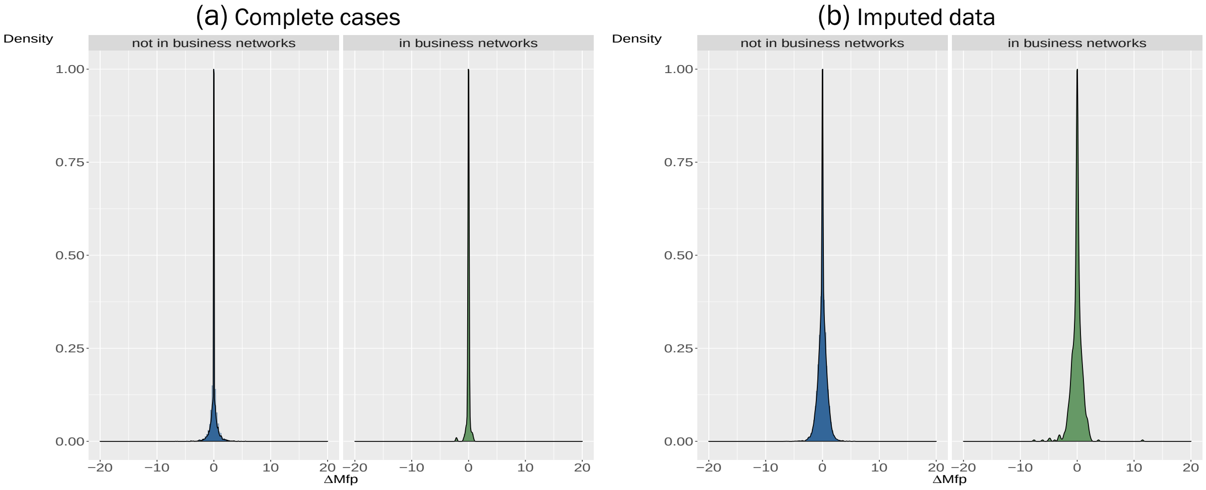
Wilkinson, I. and Young, L. (2002), “On cooperating: firms, relations and networks,” *Journal of Business Research*, 55, 123–132, `http://www.sciencedirect.com/science/article/pii/S0148296300001478`[Accessed: 01122016].

Zellner, A., Kmenta, J., Dr, xe, and ze, J. (1966), “Specification and Estimation of Cobb-Douglas Production Function Models,” *Econometrica*, 34, 784–795.

# A COMPLETE CASES ANALYSIS

The simplest way to handle missing data is using complete cases analysis, which means removing cases with missing data. Figure 8($a$) shows distribution for the changes in Mfp when we use complete cases analysis from the experimental datasets. These distributions look narrow and most changes are closed to 0. In comparison, if we use imputed data we see normal distributions for the changes in multifactor productivity. Figure 8($b$) show that the distributions of the changes in multifactor productivity are closer to what we expect to see.

Figure 8: Histogram of changes in multifactor productivity in experiment ABS, Patents and Trademarks



# B MISSING DATA IMPUTATION

## B.1 Imputation methods for categorical data

We first use the information from IPGOD to allocate firm $j$ belonging to an unknown industry $U$ into different industries. The font—$\boldsymbol{\mathcal{X}}$—represents observed dataset in the notation. The formula to allocate firms into different industries is

$$\begin{aligned} Pr(j = k \mid \boldsymbol{\mathcal{X}}_{jkt}) &= \frac{exp\,(\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k)}{1+\sum_{k=1}^{K-1} exp\,(\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k)}, k = 1, \cdots, K-1 \\ \vdots \quad &= \quad \vdots \\ Pr(j = K \mid \boldsymbol{\mathcal{X}}_{jkt}) &= \frac{1}{1+\sum_{k=1}^{K-1} exp\,(\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k)}. \end{aligned} \tag{5}$$

The one terms in the denominator and in the numerator of the $Pr(j = K \mid \boldsymbol{\mathcal{X}}_{jkt})$ ensure probabilities over the response categories sums to 1 (Czepiel, 2002, Agresti, 2007). It is convenient to write the term $\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k$ in Wilkinson and Rogers (1973) notation as $Products{+}BusinessNetwork{+}FirmAge{+}Time{+}State$. Here $Products$ is the number of products firm $j$ register at time $t$. The indicator $BusinessNetwork = 1$ for scenario 3 if firm $j$ is in a business network and 0 otherwise. The variable $FirmAge$ is the age of firm $j$ at time $t$. The variable $Time$ is represented by 10 time indicator variables, one for each year with 2002–03 as baseline. The variable $State$ is represented by 8 indicator variables, one for each state with Northern Territory as a reference group. This makes each $\boldsymbol{\mathcal{X}}_{jkt}^{\top}\mathbf{a}_k$ a sum of 18 terms. The formula is applied to the complete cases to obtain the industry coefficients $\mathbf{a}_k$ with $k = 1, \cdots, 17$ industries. We combine these estimated coefficients with firm characteristics data $\boldsymbol{\mathcal{X}}_{jkt}$ for firms with the missing industry. We allocate firm $j$ to an industry with the highest predictive probability.

## B.2 Imputation methods for continuous data

Next, we assume MAR and impute missing values in the combined ABS and IPGOD datasets by imputed industry. We use sequential regression in SAS `proc mi` procedure for the imputation. We adapt a similar notation to Reiter (2005). The experimental dataset consists of $[\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{y}$ is an $N \times 1$ vector that includes the dependent variable, and $\boldsymbol{\mathcal{X}}$ is an $N \times 15$ matrix that includes all the independent variables from (21). This gives 15 unknown regression parameters in (21). We impute missing variables ln$y$, ln$K$ and ln$M$. The observed dataset consists of two $N \times 16$ matrices, $\boldsymbol{\mathcal{D}} = [\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{\mathcal{X}}$ includes all the independent variables from (21); and the response indicator matrix $\boldsymbol{\mathcal{R}}$ which we use to partition $\boldsymbol{\mathcal{D}}$ into the observed $\boldsymbol{\mathcal{D}}^{obs}$ and the missing $\boldsymbol{\mathcal{D}}^{mis}$. We use $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{(K)}$ and $\boldsymbol{\mathcal{X}}^{(M)}$ to denote the design matrix for imputing missing data in ln$y$, ln$K$ and ln$M$, respectively.

We impute the missing values in ln$y$, ln$K$ and ln$M$ separately, using sequential regression

(SR). The SR method uses appropriate regression models for different variable types. For example, continuous variables are imputed using a normal model and binary variables using a logit model. The SR method generates a continuous vector $\boldsymbol{y}^{seq}$ from the parameters directly estimated from the fitted regression following Raghunathan et al. (2001). The SR formula for generating missing data for $\boldsymbol{y}$ is:

$$\boldsymbol{y} = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}. \tag{6}$$

We apply (6) three times, with $\boldsymbol{y}$ denoting each of the three variables ln$y$, ln$K$ and ln$M$. We use $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{(K)}$ and $\boldsymbol{\mathcal{X}}^{(M)}$ to denote the design matrix for creating missing data in ln$y$, ln$K$ and ln$M$, respectively. If the missing data variable is ln$y$, then $\boldsymbol{\mathcal{X}}$ includes all the independent variables from (21). In comparison, if the missing data variable is ln$K$, then $\boldsymbol{\mathcal{X}}^{(K)}$ includes all the independent variables and ln$y$ but excludes ln$K$. Similarly, if the missing data variable is ln$M$, then $\boldsymbol{\mathcal{X}}^{(M)}$ includes all the independent variables and ln$y$ but excludes ln$M$. Algorithm 2 describes the basic concept of the algorithm (Drechsler, 2011).

**Algorithm 2** Sequential regression algorithm

```
procedure
    Step 1: draw a new value $\theta = (\sigma^2, \boldsymbol{\beta})$ from $Pr(\theta \mid \boldsymbol{y}_{obs})$
        draw variance from $\sigma^2 \mid \boldsymbol{\mathcal{X}}_{obs} \sim (\boldsymbol{y}_{obs} - \boldsymbol{\mathcal{X}}_{obs}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y}_{obs} - \boldsymbol{\mathcal{X}}_{obs}\widehat{\boldsymbol{\beta}})\chi^{-2}_{n-k}$, where $n$ is the total number of observations and $k$ is the number of parameters
        draw coefficients from $\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{\mathcal{X}}_{obs} \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, (\boldsymbol{\mathcal{X}}'_{obs}\boldsymbol{\mathcal{X}}_{obs})^{-1}\sigma^2)$
    Step 2: draw an imputed value $\boldsymbol{y}^{seq}$ from $Pr(\boldsymbol{y}^{seq} \mid \boldsymbol{y}_{obs}, \theta)$
        draw from fitted regression $\boldsymbol{y}^{seq} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\mathcal{X}}_{obs} \sim \mathcal{N}(\boldsymbol{\mathcal{X}}_{obs}\boldsymbol{\beta}, \sigma^2)$
    repeat Step 1 and Step 2 to impute each variable sequentially
```

We create 10 imputed datasets in each imputed industry and we select the best imputed dataset which maximises the likelihood for equation (21) in Appendix ($E$) from the 10 datasets in each industry (Schomaker and Heumann, 2014, Chien et al., 2018b).

# C Estimation methods

We start by defining the indicator of a relationship between firm $i$ and $j$ as

$$y_{i,j} = \begin{cases} 1, & \text{if a firm } i \text{ is in a business network with firm } j \\ 0, & \text{otherwise,} \end{cases}$$

for $i, j = 1, \cdots, N$ total number for firms within the network. The network is described by an $N \times N$ socio-matrix $\mathbf{Y}$, and possibly additional characteristics $\mathbf{X}$ which can include one or both of nodal attributes $X$ or pair-specific attributes $x_{i,j}$.

## C.1 Exponential Random Graph Models

The exponential random graph models (ERGMs) use the probability of the observed networks over the networks with the same number of vertices that could have been observed to estimate parameters. We follow Morris et al. (2016) and specify the general form for an ERGM as:

$$P(\mathbf{Y} = \mathbf{y}) = \frac{1}{k(\theta)} exp[\theta^{\top} g(\mathbf{y}, \mathbf{X})], \tag{7}$$

where $\mathbf{Y}$ is the vector for the state of the network and $\mathbf{y}$ is the observed networks and $g(\mathbf{y}, \mathbf{X})$ are the summary statistics from the observed networks. We use $\mathbf{X}$ to denote the observed firm characteristics. The symbol $\theta$ represents unknown parameters of interest and determines the effects of the network statistics. The symbol $k(\theta)$ is the normalising constant, it represents the quantity in the numerator summed over all possible networks (typically constrained to be all networks with the same number of node set as $\mathbf{y}$). The formula (7) can be re-expressed in terms of the conditional log-odds of a single tie between two firms as

$$logit\left[Pr(Y_{i,j} = 1 \mid \mathbf{y}_{i,j}^{\mathsf{C}})\right] = \text{log odds}\left[\frac{(Y_{i,j} = 1 \mid \mathbf{y}_{i,j}^{\mathsf{C}})}{(Y_{i,j} = 0 \mid \mathbf{y}_{i,j}^{\mathsf{C}})}\right] \tag{8}$$

$$= \theta^{\top} \delta(\mathbf{y}_{i,j}) \tag{9}$$

where $Y_{i,j}$ is the random variable for the state of the firm pair i,j (with realisation $\mathbf{y}_{i,j}$). We use $\mathbf{y}_{i,j}^{\mathsf{C}}$ to denote the complement of $\mathbf{y}_{i,j}$, i.e. all connections in the network except $y_{i,j}$. The vector $\delta(\mathbf{y}_{i,j})$ contains the change statistic for each model term. The change statistic records how the $g(\mathbf{y}, \mathbf{X})$ term changes when $y_{i,j}$ is toggled from 0 to 1 (Goodreau et al., 2009).

This means that the coefficients $\theta$ are interpreted as the log-odds of an individual tie conditional on all other ties (Martina Morris, 2018). This is the major departure from the logit or probit model. The inclusion is necessary because $Pr(Y_{i,j})$ is dependent on the tie-wise outcome of all other ties (Koskinen and Daraganova, 2013).

### C.1.1 Geometrically weighted degree and shared partners statistics

The term *gwdegree* is the geometrically weighted degree statistic and the terms *gwdsp* and *gwesp* are the shared edgewise partner (EP) and shared dyadic partner (DP) statistics used in the ERGMs. Snijders et al. (2006) propose a new approach by multiplying decreasing weights on the higher observed network configurations using degree counts. Hunter (2007) reformulates the equation by multiplying the frequency for each value of degree by a weighting parameter and summing the values. The *gwdegree* statistics is defined as:

$$u(\mathbf{y}, \lambda^{(gwdegree)}) = e^{\lambda^{(gwdegree)}} \sum_{i}^{N-1} [1 - (1 - e^{-\lambda^{(gwdegree)}})^{i}] D_i(\mathbf{y}) \tag{10}$$

and the shared edgewise partner statistic *gwesp* is formulated as:

$$v(\mathbf{y}, \lambda^{(gwesp)}) = e^{\lambda^{(gwesp)}} \sum_{i}^{N-2} [1 - (1 - e^{-\lambda^{(gwesp)}})^{i}] EP_i(\mathbf{y}) \tag{11}$$

and the shared dyadic partner statistic *gwdsp* equals:

$$w(\mathbf{y}, \lambda^{(gwdsp)}) = e^{\lambda^{(gwdsp)}} \sum_{i}^{N-2} [1 - (1 - e^{-\lambda^{(gwdsp)}})^{i}] DP_i(\mathbf{y}), \tag{12}$$

where $\mathbf{y}$ is the observed network, $\lambda^{(gwdegree)}$, $\lambda^{(gwesp)}$ and $\lambda^{(gwdsp)}$ are the decay parameters determining the geometric rate of decay of the log-odds (the higher the value of the parameter, the slower the decay). The geometrically weighted degree distribution statistics model the observed network's frequency distribution for nodal degrees (Morris et al., 2008). The term $D_i(\mathbf{y})$ represents the number of nodes in $\mathbf{y}$ with degree $i$. This statistic is based on the geometric sequence $(1-e^{\lambda})^k$. The edgewise shared partner statistics measure a set of distinct k-triangles that share a common edge and are denoted as $EP_0(\mathbf{y}), \cdots, EP_{N-2}(\mathbf{y})$, with $EP_i(\mathbf{y})$ representing the pairs that have exactly $k$ common neighbours regardless of whether $y_{ij} = 1$ or $y_{ij} = 0$. The dyadic shared partner statistics measure the number of distinct k-twopaths joining the same pair of nodes and are denoted as $DP_0(\mathbf{y}), \cdots, DP_{N-2}(\mathbf{y})$, with $DP_i(\mathbf{y})$ representing the number of pairs that have exactly $k$ common neighbours (Hunter, 2007).

## C.2 Latent Space Model

The latent position models, proposed by Hoff et al. (2001, 2002), Westveld and Hoff (2011), assume conditional independence so the presence or absence of a tie between two firms is independent of all other ties in the network, given the latent positions of the two firms. Consequently,

$$Pr(\mathbf{Y} \mid \mathbf{Z}, \mathbf{X}, \theta^*) = \prod_{i \neq j} Pr(y_{i,j} \mid z_i, z_j, x_{i,j}, \theta^*), \tag{13}$$

where $\mathbf{X}$ may include pair-specific or firm specific values, and $\mathbf{Z}$ and $\theta^*$ are the latent positions and parameters to be estimated. One way to parametrise $Pr(y_{i,j} \mid z_i, z_j, x_{i,j}, \theta^*)$ is using a logistic regression model where the probability of forming a tie depends on the Euclidean distance between the latent positions $z_i$ and $z_j$ of firms $i$ and $j$ as well as covariate $x_{i,j}$ which measure the characteristics of the tie:

$$\eta_{i,j} = \log \text{odds}(y_{i,j} = 1 \mid z_i, z_j, x_{i,j}, \alpha, \beta) = \alpha + x_{i,j}^\top \beta - |z_i - z_j|. \tag{14}$$

We interpret equation (14) as the distance from firm $i$ to firm $j$ and from firm $i$ to firm $k$, i.e. the log odds ratio of $i \to j$ versus $i \to k$, is $(x_{i,j} - x_{i,k})^\top \beta$. The distance between $i$ and $j$ i.e. $|z_i - z_j|$ can be replaced by an arbitrary set of distances $d_{i,j}$ e.g., $\frac{z_i' z_j}{|z_j|}$ as long as they satisfies the triangle inequality, $d_{i,j} \leq d_{i,k} + d_{k,j} \forall i, j, k$ (Hoff et al., 2002).

The latent position approach is inherently reciprocal and transitive. For example, if $i \to j$ and $j \to k$, then the distances $d_{i,j}$ and $d_{j,k}$ are not too large, that makes the events $j \to i$ (reciprocity) and $i \to k$ (transitivity) more likely. This feature makes the latent space model well suited to study undirected relations where the parameter space has a lower dimension than the data. We simplify (14) by excluding covariate information with undirected relation $y_{i,j} = y_{j,i}$ as

$$\log \text{odds}(y_{i,j} = 1 \mid d_{i,j}, \alpha) = \alpha(1 - d_{i,j}). \tag{15}$$

A set of distances $d_{i,j}$ represents the network $\mathbf{Y}$ if

$$\begin{aligned} d_{i,j} &> 1 \forall i, j : y_{i,j} = 0, \text{ and} \\ d_{i,j} &< 1 \forall i, j : y_{i,j} = 1. \end{aligned} \tag{16}$$

The probability of the data under parametrisation of (15) will converge to unity as $\alpha \to \infty$ for such set od distances. The approach model the distances as being Euclidean distances in some $k-$ dimensional space, we say a network is $d_k$ representable if there are points $z_i \in \mathbb{R}_k$ such that the distances $d_{i,j} = |z_i - z_j|$ satisfy (16) (Hoff et al., 2001). There are many example of social networks which are $d_k-$ representable for $k < n$. Examples include $k-$ star networks or $k-$chain networks (Hoff et al., 2002).

## C.3 Calculations of latent distance

We demonstrate how to calculate the pair-wise latent distance by considering a simple cross-sectional latent space model without covariates for firm networks in both patents and trademarks in year 2003. We follow Hoff et al. (2002) and specify the model as

$$Pr(\mathbf{Y} \mid \alpha, \mathbf{Z}) = \prod_{i \neq j}^{N} Pr(y_{i,j} \mid z_i, z_j) \tag{17}$$

$$logit(y_{i,j} = 1 \mid z_i, z_j, \alpha) = \alpha - |z_i - z_j|, \tag{18}$$

where the $z_i$ lies in $\mathbb{R}_2$. The probability of the data depends only on the distances which are invariant under reflection, rotation and location shift. Therefore for each $2 \times N$ matrix of latent positions $Z$. It is important to note that there are an infinite number of other positions that give the same log-likelihood i.e. $log\, Pr(\mathbf{Y} \mid \mathbf{Z}, \alpha) = log\, Pr(\mathbf{Y} \mid \mathbf{Z}^*, \alpha)$ for any $\mathbf{Z}^*$ which is equal to $\mathbf{Z}$ under the operations of reflection, rotation, or translation.

This problem can be solved by making inference on equivalence classes of latent positions. We let $[\mathbf{Z}]$ be the class of positions equivalent to $\mathbf{Z}$ under rotation, reflection, and translation. For each $[\mathbf{Z}]$, there is one set of distances between nodes which is called the *configuration*. We make inference on particular elements of configurations that are comparable across configurations. So for a given $[\mathbf{Z}]$, we choose $\mathbf{Z}^* = argmin_{T\mathbf{Z}} \mathrm{tr}(\mathbf{Z}_0 - T\mathbf{Z})'(\mathbf{Z}_0 - T\mathbf{Z})$, where $\mathbf{Z}_0$ is a fixed set of positions and $T$ ranges over the set of rotations, reflections and translations. $\mathbf{Z}^*$ is a *procrustes transformation* of $\mathbf{Z}$, being the element of $[\mathbf{Z}]$ closest to $\mathbf{Z}_0$ in terms of the sum of squared positional differences and is unique if $\mathbf{Z}_0\mathbf{Z}'$ is non-singular.

We assume $\mathbf{Z}$ and $\mathbf{Z}_0$ are centred at the origin and compute $\mathbf{Z}^* = \mathbf{Z}_0\mathbf{Z}'(\mathbf{Z}\mathbf{Z}_0'\mathbf{Z}_0\mathbf{Z}')^{-\frac{1}{2}}\mathbf{Z}$. We typically take $\mathbf{Z}_0 = \widehat{\mathbf{Z}}_{MLE}$, maximum likelihood estimates of the latent positions centred at the origin. Given prior information on $\alpha$ and $\mathbf{Z}$. The procedure of sampling from the posterior distribution of the latent space is given in Algorithm 1.

**Algorithm 3** Pseudocode for the Gibbs sampler

```
procedure
    Using $\mathbf{Z}_0 = \widehat{\mathbf{Z}}$ as starting points,
    Constructing a Markov Chain over model parameters as follows:
        sample a proposal proposal $\widetilde{\mathbf{Z}}$ from $J(\mathbf{Z} \mid \mathbf{Z}_k)$ a symmetric proposal distribution;
        accept $\widetilde{\mathbf{Z}}$ as $\mathbf{Z}_{k+1}$ with the probability $\frac{Pr(\mathbf{Y} \mid \widetilde{\mathbf{Z}}\alpha)\pi(\widetilde{\mathbf{Z}})}{Pr(\mathbf{Y} \mid \mathbf{Z}_k\alpha)\pi(\mathbf{Z}_k)}$ ($\pi$ represents Gaussian distribution), otherwise set $\mathbf{Z}_{k+1} = \mathbf{Z}_k$.
        store $\widehat{\mathbf{Z}}_{k+1} = argmin_{T\mathbf{Z}_{k+1}} \mathrm{tr}(\widehat{\mathbf{Z}} - T\mathbf{Z}_{k+1})'(\widehat{\mathbf{Z}} - T\mathbf{Z}_{k+1})$
    Update $\alpha$ with a Metropolis-Hastings algorithm[3].
```

Each configuration can be represented by its unique procrustean statistic, the posterior distribution of the configuration around $\widehat{\mathbf{Z}}$ is represented by samples of $\widetilde{\mathbf{Z}}$ from the Markov chain (Hoff et al., 2002, Shortreed et al., 2006).

[3] Metropolis-Hastings algorithm is a MCMC method to produce a sequence of samples from a complex probability distribution from which direct sampling may prove near impossible or quite costly (Robert, 2015). The sequence of sample can be used to approximate the complex distribution. The algorithm uses on Markov chain theory to validate the convergence of the chain to the distribution of interest (Chib and Greenberg, 1995).

## C.4 Logistic Regression Models

Davison (2003) shows the formula for the logistic regression can be simplified as

$$logit\left[\frac{Pr(Y_{i,j}=1}{Pr(Y_{i,j}=0}\right] = \log\text{ odds}\left[\frac{exp\,(\mathbf{X}^{\top}\theta^{**}) \times (1+exp\,(\mathbf{X}^{\top}\theta^{**}))^{-1}}{1 \times (1+exp\,(\mathbf{X}^{\top}\theta^{**}))^{-1}}\right] \tag{19}$$

$$= exp\,(\mathbf{X}^{\top}\theta^{**}) \tag{20}$$

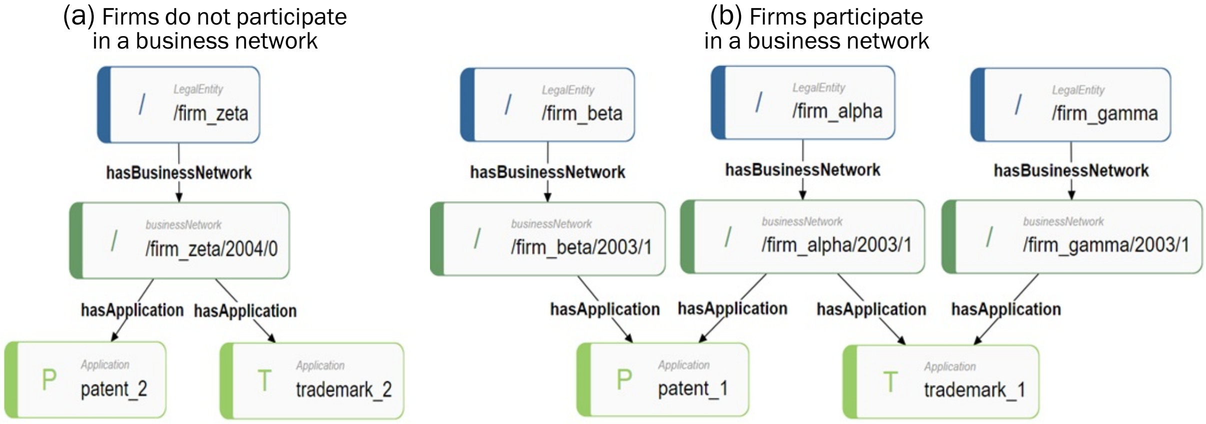
## D Semantic web for data integration and analysis

We use the semantic web to integrate two datasets - patents and trademarks from 2017 Intellectual Property Government Open Data (IPGOD) (Chien et al., 2018a) . The semantic web approach is well suited to integrate data from multiple sources and to extract information on firms in multiple business networks. We assign an unique uniform resource identifier (URI) for each firm using the unique firm identifier ABN or Australian Company Number (ACN). We attach different firm attributes e.g. states and number of applications from different data sources to each firm.

Our analysis focuses on firms with both patent and trademark applications. We then use SPARQL, a query language to retrieve data stored in Resource Description Framework (RDF) format, to extract information on firms have applications with at least one other firm (in business network) or by itself (not in business network). In addition, we separate our sample into three periods: before global financial crisis (GFC) from 2003 to 2006, during the GFC from 2007 to 2009 and after the GFC from 2010 to 2013 in our sample.

For our analysis, it is important to know the data provenance to correctly compare firms with and without multiple business networks. Data provenance here refers to the database which contains the administrative records. We use named graphs in the semantic web architecture to distinguish different data sources by adding a prefix in the URIs. These prefixes are then used in the SPARQL queries to retrieve correct information. Figure 9 shows the ontology for our data model. We use Ontodia - an OWL and RDF diagramming tool to visualise our data model. The Business Network node qualifies how firms can be connected through joint patent or trademark applications. For example, *firm_alpha*, *firm_beta* and *firm_gamma* participate in a business network because *firm_alpha* shares at least one patent with *firm_beta* and it also shares at least one trademark application with *firm_gamma*. We also consider *firm_zeta* does not participate in a business network because it files at least one patent and one trademark alone.

Figure 9: Ontology



We have 518,870 triples of firms in the patent named graph `http://patents`, 5,638,915 triples of firms in the trademark named graph `http://trademarks` with a total of 6,157,785 triples in the integrated database. We use legal entities to represent firms. The IPGOD and ASX datasets contain unique firm identification numbers (ABNs or ACNs) for firms. We use patent and trademark applications (application number) to identify firms in business networks. We use unique ABNs and ACNs to form the URI for the legal entities. These URIs serve as unique linking keys to correctly retrieve firm information from different sources using SPARQL queries. An example below shows how we construct a SPARQL query to retrieve firms belonging to both patent and trademark networks in the period before the GFC from 2003 to 2006.

Listing 1: intersection SPARQL query

```
prefix pat: <http://patents>
prefix tmk: <http://trademarks>
SELECT ?ABN
FROM NAMED pat:
FROM NAMED tmk:
WHERE {
values (?BN) {("2003_BN") ("2004_BN") ("2005_BN") ("2006_BN")}
{GRAPH pat:{
?LegalEntity fnet:hasAustralianBusinessNumber ?ABN;
fnet:hasBusinessNetwork ?businessNetwork.}}
FILTER EXISTS
{GRAPH tmk: {
?LegalEntity fnet:hasAustralianBusinessNumber ?ABN;
fnet:hasBusinessNetwork ?businessNetwork.}}}
```

# E FIRM PERFORMANCE AND BUSINESS NETWORK MEASURES

## E.1 Firm productivity

This study will use the productivity measure from Chien et al. (2019). Following Zellner et al. (1966), Breunig and Wong (2008), Nguyen and Hansell (2014), Mare et al. (2017), the statistical model for the firm production function is specified as

$$\ln y_{jkt} = \beta_k + \beta_1 \ln L_{jkt} + \beta_2 \ln K_{jkt} + \beta_3 \ln M_{jkt} + \beta_4 Firm\_Age_{jkt} + \tau_{kt} + \varepsilon_{jkt}, \qquad (21)$$

where the formula for firm value added $\ln y_{jkt}$ is

$$log \left[ \frac{\text{(total sales - the repurchase of stocks)}}{\text{gross value added implicit price deflators by industry}} \right]$$

for firm $j$ in industry $k$ at time $t$ (ABS, 2018). We use the method proposed by Abowd et al. (2002) to derive the logarithm of estimated firm average labour components, $\ln L_{jkt}$ for firm $j$ in industry $k$ at time $t$. The formula for the logarithm of capital cost per employee $\ln K_{jkt}$ is

$$log \left[ \frac{\text{(equipment depreciation} + \text{business rental expenses} + \text{capital investment deductions)}}{\text{consumption of fixed capital deflators by industry}} \right].$$

We calculate the per employee logarithm of material inputs $\ln M_{jkt}$ as

$$log \left[ \frac{\text{materials used in the production process}}{\text{Producer Price Index for intermediate goods}} \right]$$

for firm $j$ in industry $k$ at time $t$. The logarithm of age for firm $j$ in industry $k$ at time $t$ is $Firm\_Age_{jkt}$. The estimated time fixed effect for firm $j$ in industry $k$ at time $t$ is denoted as $\tau_{jkt}$. The term $\varepsilon_{jkt}$ are assumed to satisfy $\varepsilon_{jkt} \overset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$ to estimate unbiased coefficients for the Cobb Douglas production function.

We follow Mare et al. (2017) and define productivity measure as $Mfp_{jkt} = \hat{\tau}_{kt} + \hat{\varepsilon}_{jkt}$, which includes both time fixed effects $\tau_{jkt}$ and the estimated multi-factor productivity $\hat{\varepsilon}_{jkt}$ for firm $j$ in industry $k$ at time $t$.

## E.2 Products

We measure products by counting the number of unique patent and/or trademark applications a firm has in each year over the period between 2003 and 2013.

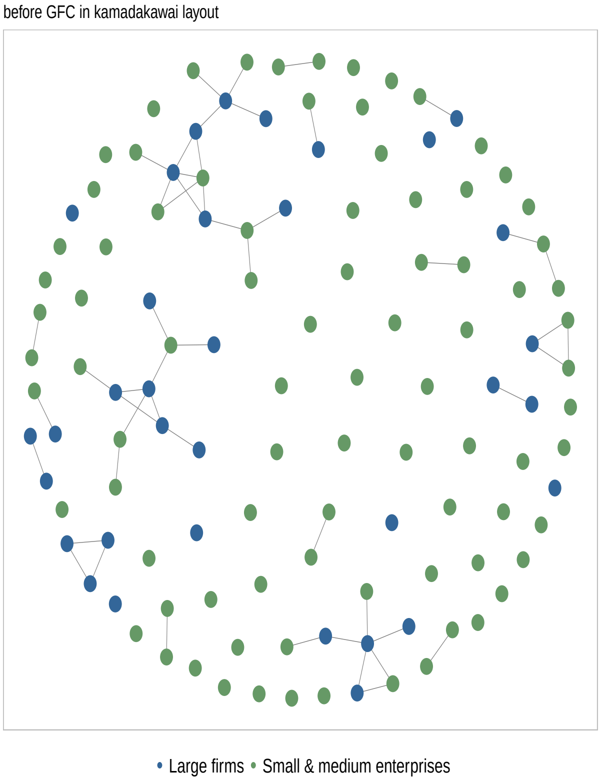
## E.3 Firm sales

The formula for firm sales, $\ln Sales$, is:

$$log \left[ \frac{\text{(total sales - the repurchase of stocks)}}{\text{gross value added implicit price deflators by industry}} \right]$$

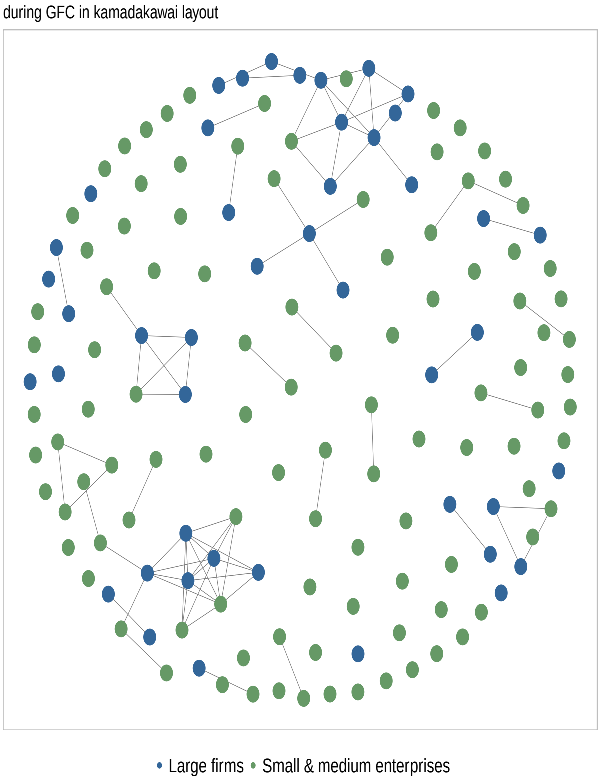
## F Summary Statistics

Figure 10: Patents and trademarks business networks before the Global Financial Crisis


before GFC in kamadakawai layout
● Large firms ● Small & medium enterprises

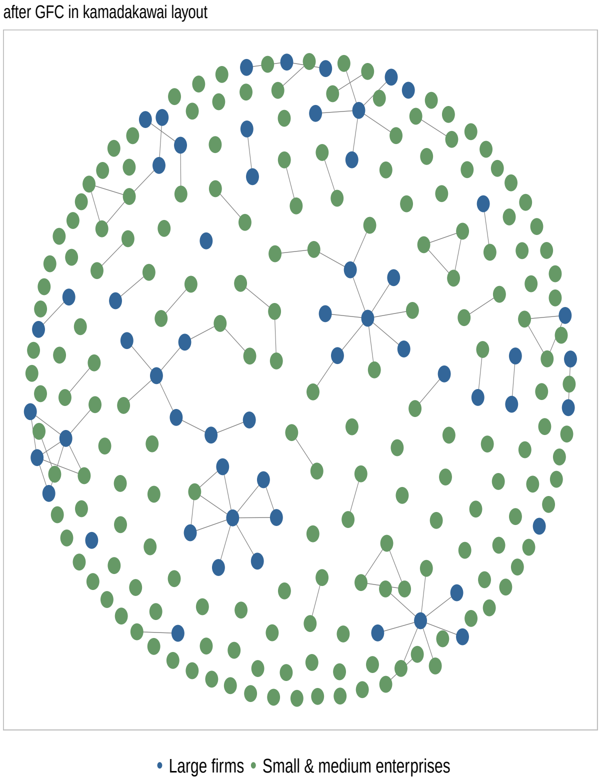
| Statistic | N | $P_{1st}$ | $P_{50th}$ | $P_{99th}$ | St. Dev. |
|---|---|---|---|---|---|
| **Data grouping step** | | | | | |
| $\lvert Sales_i - Sales_j \rvert$ | 93 | 5.74 | 14.75 | 22.86 | 3.18 |
| $\lvert Mfp_i - Mfp_j \rvert$ | 93 | -2.86 | 2.4 | 7.6 | 2.01 |
| FIRM_AGE | 2,036 | 0 | 1.39 | 2.56 | 0.71 |
| products | 2,036 | 0 | 1.1 | 3.96 | 0.9 |
| **Sampling step** | | | | | |
| $\lvert Sales_i - Sales_j \rvert$ | 51 | 5.74 | 14.83 | 20.29 | 3.06 |
| $\lvert Mfp_i - Mfp_j \rvert$ | 51 | -0.33 | 2.13 | 5.8 | 1.65 |
| FIRM_AGE | 122 | 0 | 1.39 | 2.48 | 0.6 |
| products | 122 | 0 | 1.39 | 5.06 | 1.24 |

Figure 11: Patents and trademarks business networks during the Global Financial Crisis


during GFC in kamadakawai layout
● Large firms ● Small & medium enterprises

| Statistic | N | $P_{1st}$ | $P_{50th}$ | $P_{99th}$ | St. Dev. |
|---|---|---|---|---|---|
| **Data grouping step** | | | | | |
| $\|Sales_i - Sales_j\|$ | 125 | 2.99 | 13.39 | 32.53 | 5.295 |
| $\|Mfp_i - Mfp_j\|$ | 125 | -3.56 | 0.47 | 7.02 | 1.94 |
| FIRM_AGE | 4,492 | 0 | 1.95 | 2.77 | 0.672 |
| products | 4,492 | 0 | 1.39 | 4.37 | 0.917 |
| **Sampling step** | | | | | |
| $\|Sales_i - Sales_j\|$ | 79 | 1.98 | 13.43 | 31.24 | 5.43 |
| $\|Mfp_i - Mfp_j\|$ | 79 | -3.61 | 0.47 | 6.05 | 1.98 |
| FIRM_AGE | 150 | 0 | 1.95 | 2.77 | 0.54 |
| products | 150 | 0 | 1.79 | 5.91 | 1.36 |

Figure 12: Patents and trademarks business networks after the Global Financial Crisis


after GFC in kamadakawai layout
● Large firms ● Small & medium enterprises

| Statistic | N | $P_{1st}$ | $P_{50th}$ | $P_{99th}$ | St. Dev. |
|---|---|---|---|---|---|
| **Data grouping step** | | | | | |
| $\|Sales_i - Sales_j\|$ | 205 | 6.26 | 14.76 | 39.94 | 7.42 |
| $\|Mfp_i - Mfp_j\|$ | 205 | -3.57 | 0.89 | 9.6 | 2.66 |
| FIRM_AGE | 7,303 | 0 | 2.2 | 3 | 0.71 |
| products | 7,303 | 0 | 1.37 | 4.48 | 0.95 |
| **Sampling step** | | | | | |
| $\|Sales_i - Sales_j\|$ | 93 | 7.85 | 14.76 | 40.56 | 6.41 |
| $\|Mfp_i - Mfp_j\|$ | 93 | -3.07 | 0.89 | 9.89 | 2.73 |
| FIRM_AGE | 240 | 0 | 2.3 | 2.89 | 0.5 |
| products | 240 | 0 | 1.95 | 6.18 | 1.48 |

# G Empirical results

Table 4: Patents and trademarks business networks before the Global Financial Crisis

| | logit | | ergm | | latent | |
|---|---|---|---|---|---|---|
| Intercept | −6.06 | [−9.61; −2.51]* | −6.00 | [−9.95; −2.06]* | 0.21 | [−4.79; 5.32] |
| gwdegree $\lambda = 0.25$ | 1.26 | [0.26; 2.25]* | 0.72 | [−0.62; 2.07] | | |
| gwdsp $\lambda = 0.2$ | −0.11 | [−0.47; 0.25] | −0.27 | [−0.70; 0.16] | | |
| gwesp $\lambda = 0.05$ | 1.78 | [1.42; 2.14]* | 1.78 | [1.19; 2.38]* | | |
| $\lvert Mfp_i - Mfp_j \rvert$ | 0.07 | [−0.14; 0.28] | −0.07 | [−0.26; 0.13] | −0.06 | [−0.86; 0.85] |
| $\lvert Sales_i - Sales_j \rvert$ | 0.03 | [−0.09; 0.16] | 0.10 | [−0.02; 0.22] | 0.32 | [−0.12; 0.89]† |
| Largefirm | 0.98 | [0.10; 1.86]* | 0.95 | [0.19; 1.70]* | 4.35 | [0.86; 9.68]* |
| SME | −1.04 | [−1.92; −0.16]* | −0.97 | [−1.86; −0.08]* | −2.85 | [−6.45; −0.20]* |
| products | 0.31 | [0.07; 0.55]* | 0.29 | [0.06; 0.52]* | 0.84 | [0.11; 1.80]* |
| FIRM_AGE | 0.15 | [−0.28; 0.59] | 0.14 | [−0.31; 0.59] | 0.71 | [−0.61; 2.12] |
| AIC | | 484.44 | | 551.85 | | |
| BIC (Likelihood) | | 655.86 | | 723.26 | | 1964.82 |
| BIC (Latent Positions) | | | | | | 1374.01 |
| BIC (Overall) | | | | | | 3651.70 |

* significant at 5% level; † significant at 10% level

Table 5: Patents and trademarks business networks during the Global Financial Crisis

| | logit | | ergm | | latent | |
|---|---|---|---|---|---|---|
| Intercept | −9.51 | [−14.27; −4.75]* | −6.34 | [−9.83; −2.84]* | 0.46 | [−4.87; 5.82] |
| gwdegree $\lambda = 0.25$ | 0.31 | [−0.50; 1.12] | 0.03 | [−0.93; 1.00] | | |
| gwdsp $\lambda = 0.7$ | −0.80 | [−1.13; −0.47]* | −0.57 | [−0.76; −0.38]* | | |
| gwesp $\lambda = 0.35$ | 2.68 | [2.23; 3.13]* | 2.41 | [1.87; 2.94]* | | |
| $\lvert Mfp_i - Mfp_j \rvert$ | 0.02 | [−0.17; 0.21] | 0.10 | [−0.05; 0.26] | 0.15 | [−0.64; 0.94] |
| $\lvert Sales_i - Sales_j \rvert$ | 0.10 | [0.01; 0.18]* | 0.01 | [−0.05; 0.07] | 0.20 | [−0.19; 0.64] |
| Largefirm | 2.25 | [1.39; 3.12]* | 1.46 | [0.90; 2.02]* | 12.66 | [6.36; 21.01]* |
| SME | −0.98 | [−1.80; −0.16]* | −0.81 | [−1.50; −0.12]* | −2.25 | [−5.79; 0.94] |
| products | 0.34 | [0.13; 0.56]* | 0.19 | [0.01; 0.36]* | 1.02 | [0.22; 1.90]* |
| FIRM_AGE | 0.21 | [−0.38; 0.80] | 0.25 | [−0.20; 0.71] | 0.59 | [−0.69; 1.89] |
| AIC | | 521.38 | | 678.41 | | |
| BIC (Likelihood) | | 726.38 | | 883.41 | | 157.88 |
| BIC (Latent Positions) | | | | | | 2736.21 |
| BIC (Overall) | | | | | | 3374.66 |

* significant at 5% level; † significant at 10% level

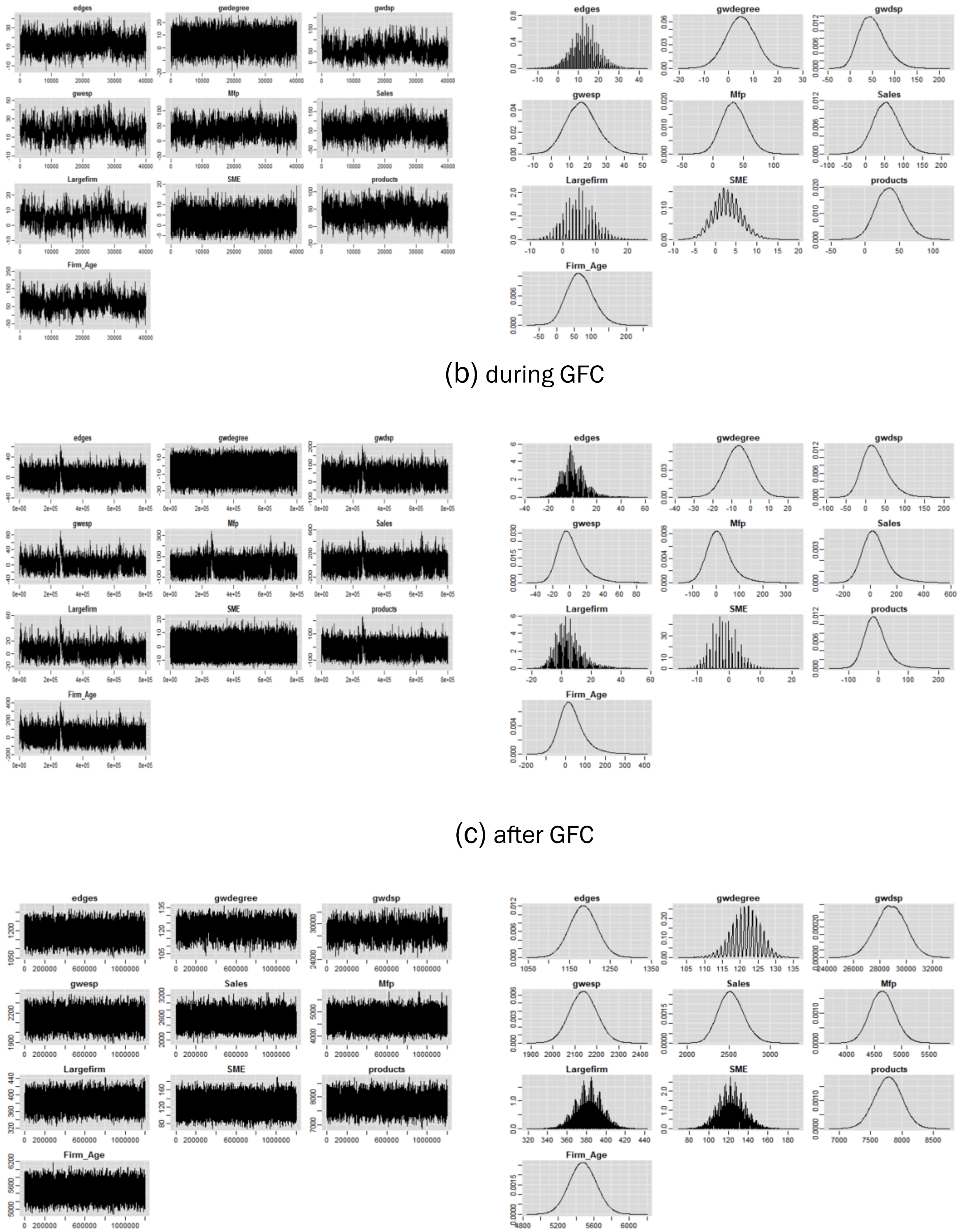
Table 6

| | logit | | ergm | | latent | |
|---|---|---|---|---|---|---|
| Intercept | $-42.40$ | $[-1616.35;\ 1531.55]$ | $-40.06$ | $[-40.94;\ -39.17]^*$ | $-1.38$ | $[-6.67;\ 4.02]$ |
| gwdegree $\lambda = 0.1$ | $2.73$ | $[2.10;\ 3.35]^*$ | $1.33$ | $[0.07;\ 2.58]^*$ | | |
| gwdsp $\lambda = 0.4$ | $0.21$ | $[0.08;\ 0.35]^*$ | $0.02$ | $[0.02;\ 0.03]^*$ | | |
| gwesp $\lambda = 0.6$ | $1.85$ | $[1.57;\ 2.13]^*$ | $2.78$ | $[2.35;\ 3.20]^*$ | | |
| $\lvert Mfp_i - Mfp_j \rvert$ | $0.13$ | $[0.04;\ 0.22]^*$ | $0.09$ | $[0.07;\ 0.12]^*$ | $0.27$ | $[-0.12;\ 0.81]^\dagger$ |
| $\lvert Sales_i - Sales_j \rvert$ | $-0.01$ | $[-0.06;\ 0.05]$ | $-0.06$ | $[-0.08;\ -0.05]^*$ | $-0.03$ | $[-0.27;\ 0.25]$ |
| Largefirm | $2.41$ | $[1.74;\ 3.07]^*$ | $1.40$ | $[1.20;\ 1.60]^*$ | $11.01$ | $[4.24;\ 22.07]^*$ |
| SME | $-1.02$ | $[-1.70;\ -0.33]^*$ | $-0.43$ | $[-0.64;\ -0.23]^*$ | $-0.56$ | $[-3.37;\ 1.79]$ |
| products | $0.22$ | $[0.07;\ 0.38]^*$ | $0.21$ | $[0.17;\ 0.25]^*$ | $0.69$ | $[0.13;\ 1.44]^*$ |
| FIRM_AGE | $1.04$ | $[0.50;\ 1.59]^*$ | $0.58$ | $[0.43;\ 0.73]^*$ | $0.53$ | $[-0.49;\ 1.53]$ |
| AIC | 945.11 | | 2214.44 | | | |
| BIC (Likelihood) | 1176.51 | | 2445.83 | | 4324.90 | |
| BIC (Latent Positions) | | | | | 3483.44 | |
| BIC (Overall) | | | | | 8410.27 | |

$^*$ significant at 5% level; $^\dagger$ significant at 10% level

# H Diagnostics

Figure 13: ERGMs MCMC Convergence Diagnostics

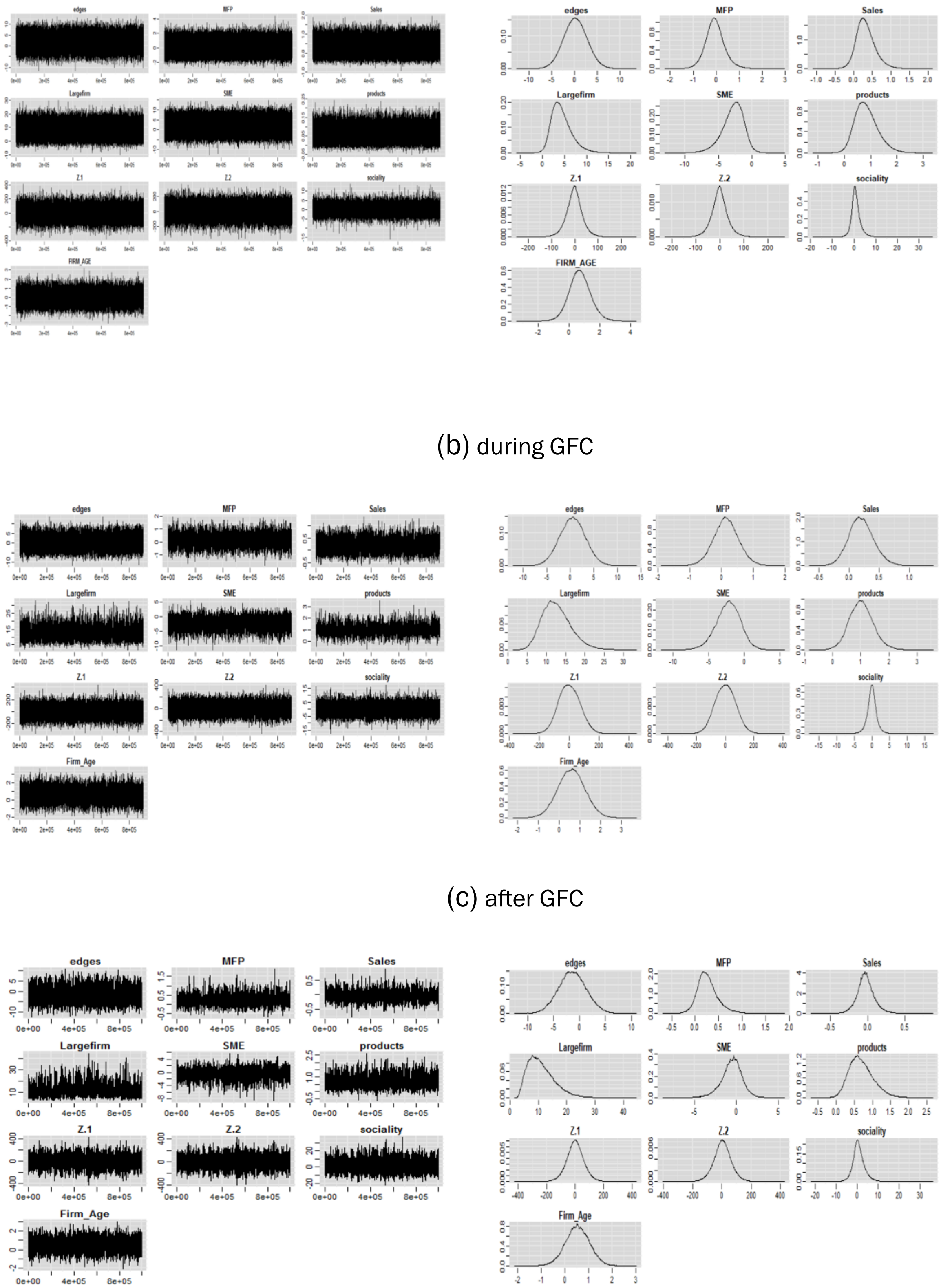


(a) before GFC

(b) during GFC

(c) after GFC

Figure 14: LSMs MCMC Convergence Diagnostics


(a) before GFC

(b) during GFC

(c) after GFC

ABS Catalogue No. 1351.0.55.063

Produced by the Australian Bureau of Statistics.

## INQUIRIES

For further information about these and related statistics, contact the National Information and Referral Service on 1300 135 070.

Reissue

# Research Paper

# Australian Business Networks

## Methodology Transformation Branch

Methodology Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 29 JAN 2021

## FOR MORE INFORMATION . . .

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

### 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. 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.

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FAX 1300 135 211

EMAIL client.services@abs.gov.au

PHONE 1300 135 070

## FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

Produced by the Australian Bureau of Statistics