Connectedness and Meaning: A New Analytical Platform

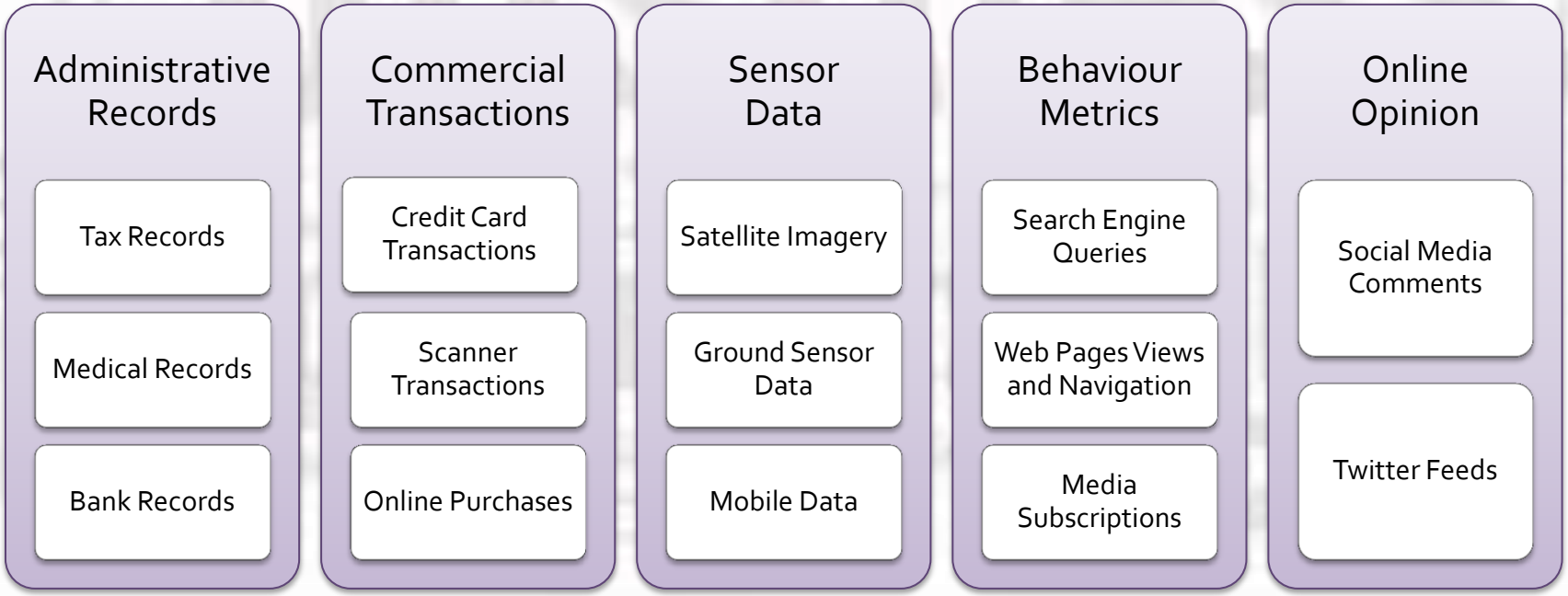
Joseph (Chien-Hung) Chien
Australian Bureau of Statistics

Disclaimer:

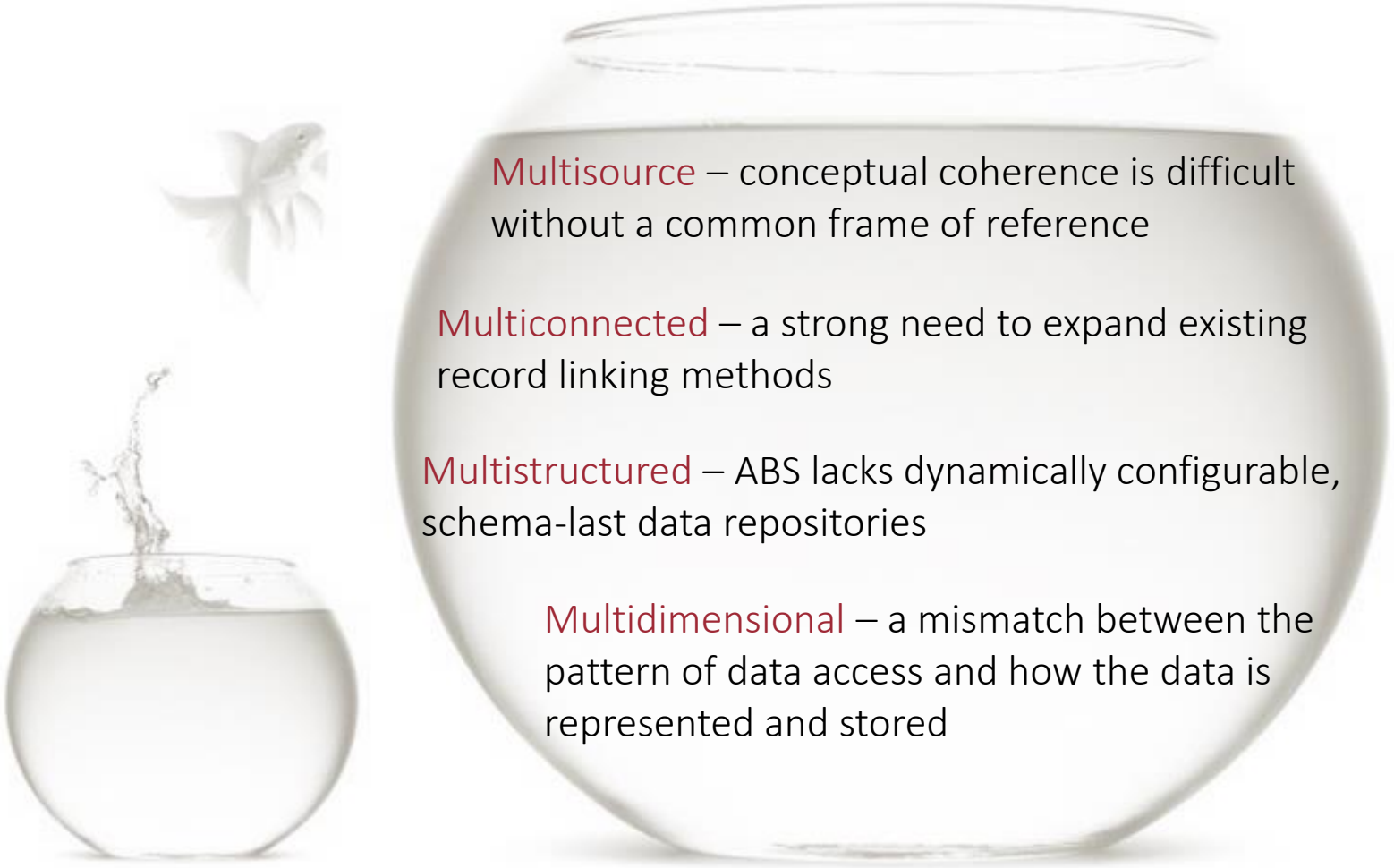
The information presented is not for official statistics and only synthetic data is used for demonstration purposes. The opinions expressed in this presentation are those of the presenter, not the ABS

Big Data = Big Data sources

- Administrative records and Scanner Data are not new.
- Sensor Data has potential – but has significant challenge
- Behaviour metrics and online opinion – potentially large inherent statistical biases.



Analytical challenges of Big Data



Multisource – conceptual coherence is difficult without a common frame of reference

Multiconnected – a strong need to expand existing record linking methods

Multistructured – ABS lacks dynamically configurable, schema-last data repositories

Multidimensional – a mismatch between the pattern of data access and how the data is represented and stored

From structured data sets …

Business Activities Statements (ATO)

ABN	PERIODICITY	REPORTING_PERIOD	STATE	INDUSTRY_CODE	TURNOVER	WAGES
1234546789	…	…	…	…	…	…
	ANNUAL	2010-11	NSW	1	$200,000	$100,000
…	…	…	…	…	…	…

Business Income Tax (ATO)

ABN	FINANCIAL_YEAR	STATE	PROFIT_AND_LOSS	INDUSTRY_CODE	ASSETS	SALES
…	…	…	…	…	…	…
123454678910	2010-11	NSW	$20,000	1	$500,000	$200,000
…	…	…	…	…	…	…

Pay As You Go Statements (ATO)

SCRAMBLED TFN	ABN	GROSS_SALARY	FINANCIAL_YEAR
…	…	…	…
567784321	123454678910	$50,000	2010-11
…	…	…	…

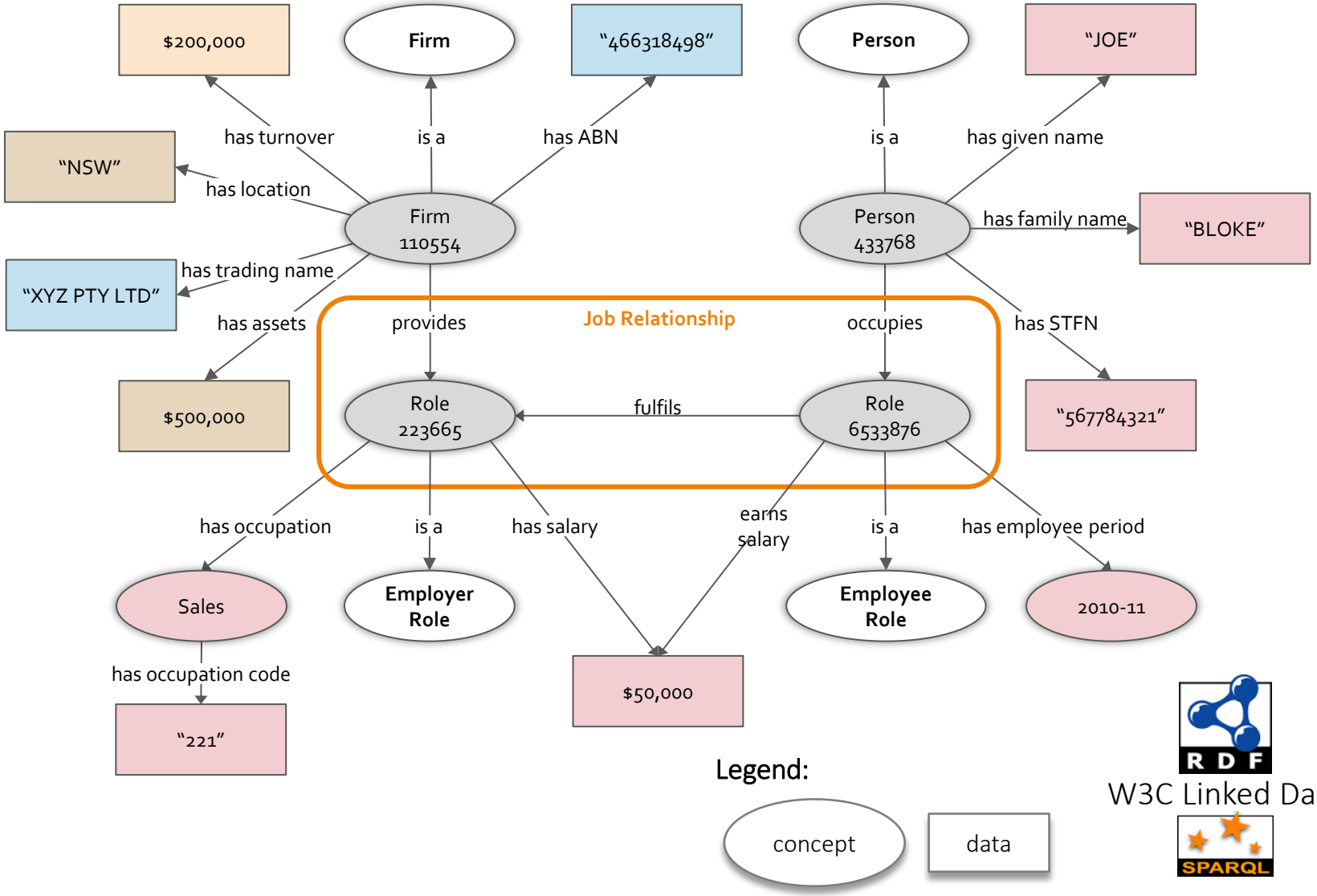
Personal Income Tax (ATO)

SCRAMBLED_TFN	OCCUPATION	FINANCIAL_YEAR	FAMILY_NAME	GIVEN_NAME	OCC_CODE	SALARY
…	…	…	…	…	…	…
567784321	Sales	2010-11	BLOKE	JOE	221	$50,000
…	…	…	…	…	…	…

ABS Business Register (ABS)

ABN	STATE	TRADING_NAME	ECONOMIC_ACTIVITY	POST_CODE	SECTOR	SURVEY_ID
…	…	…	…	…	…	…
123454678910	NA	XYZ PTY LTD	RETAIL	2000	PRIVATE	Eco1234
…	…	…	…	…	…	…

… to a network of entities and relationships



Application 1: Research question and Methodology

Research question: Can we distinguish true and spurious firm death events from analysing the network connections in the prototype semantic LEED?

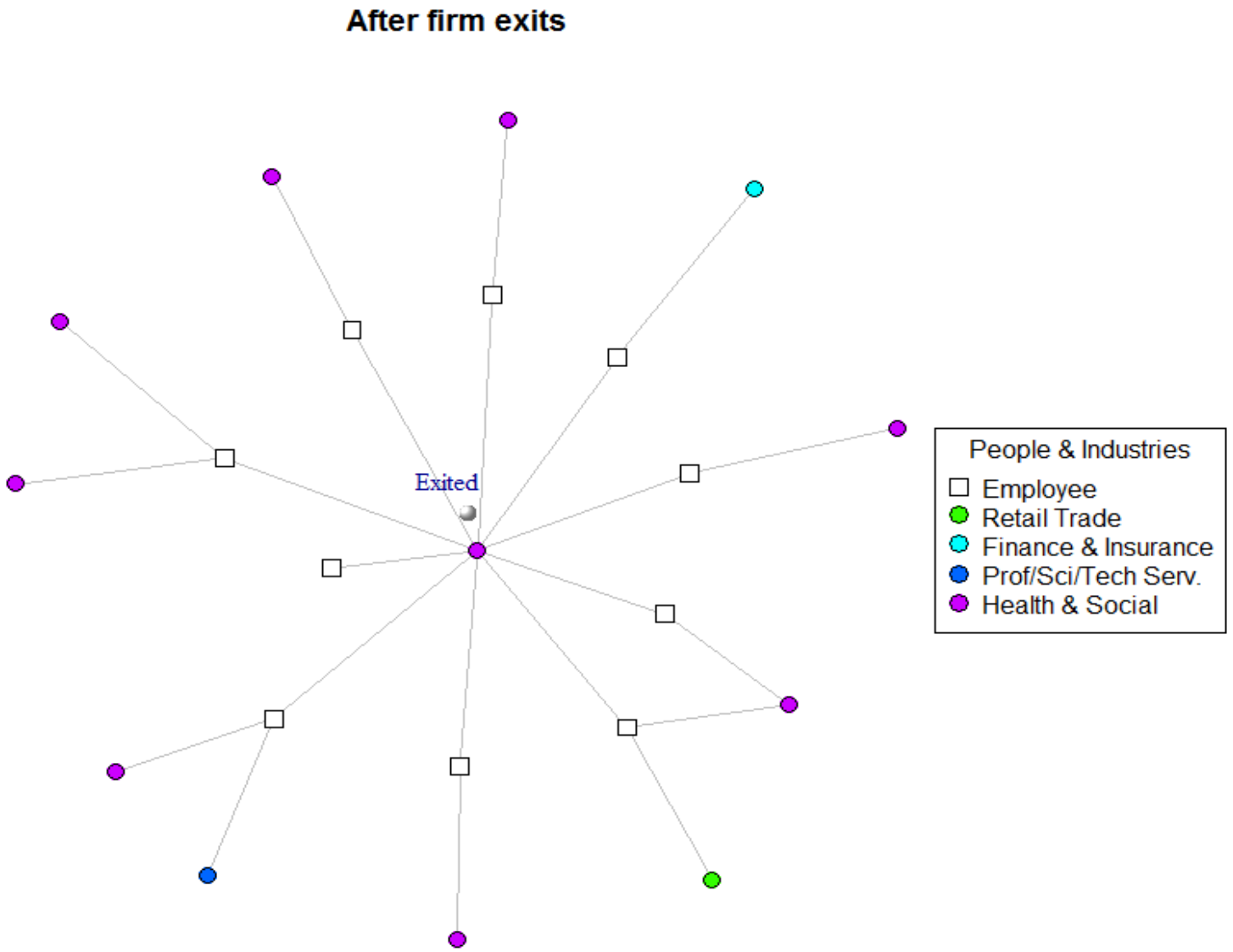
The statistical importance of accounting for true firm deaths.

- Exits ≠ Deaths (OECD)
- Statistical bias if we don't correct them.

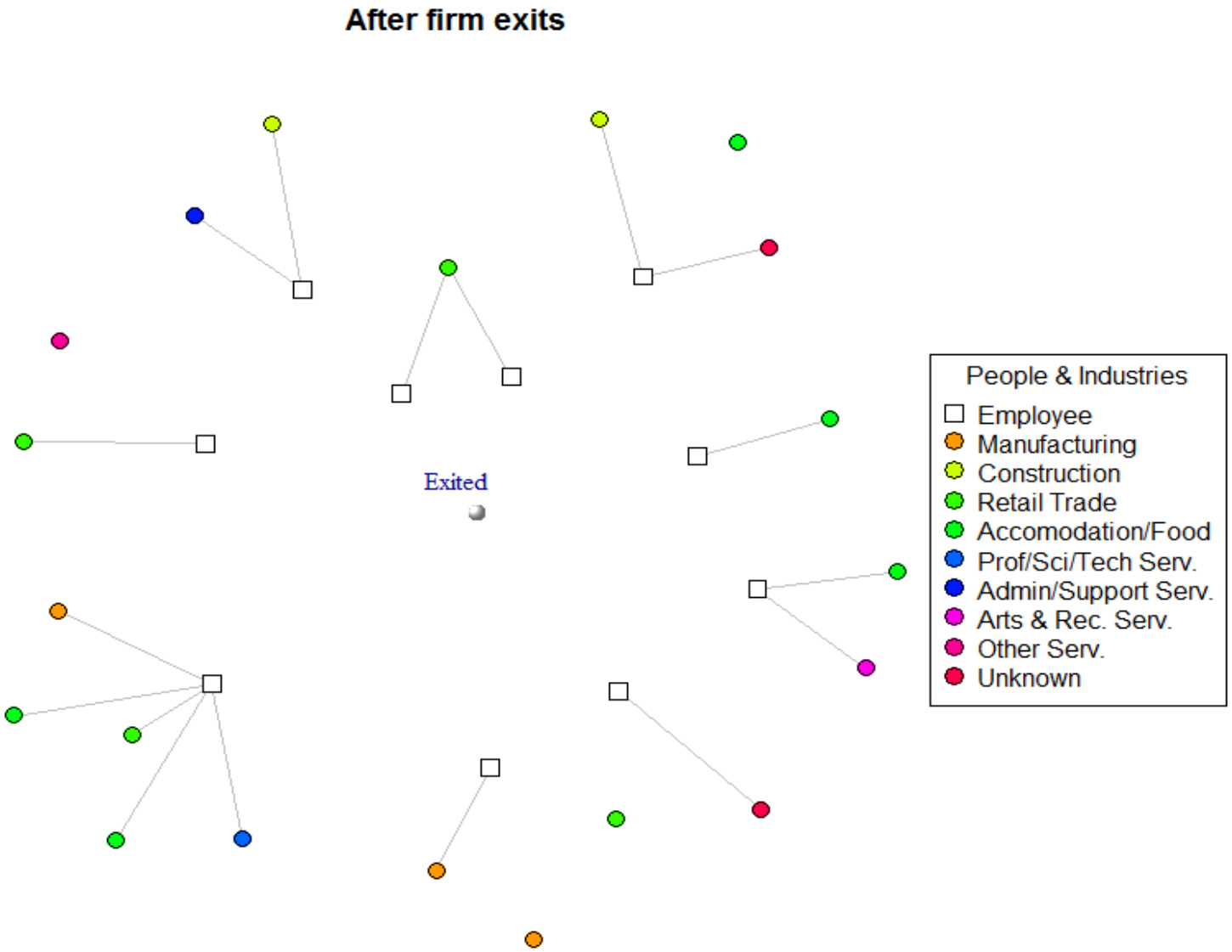
Methodology

- Derive network statistics using sematic LEED (Important).
- Combining Multilevel modelling and Bayesian network.

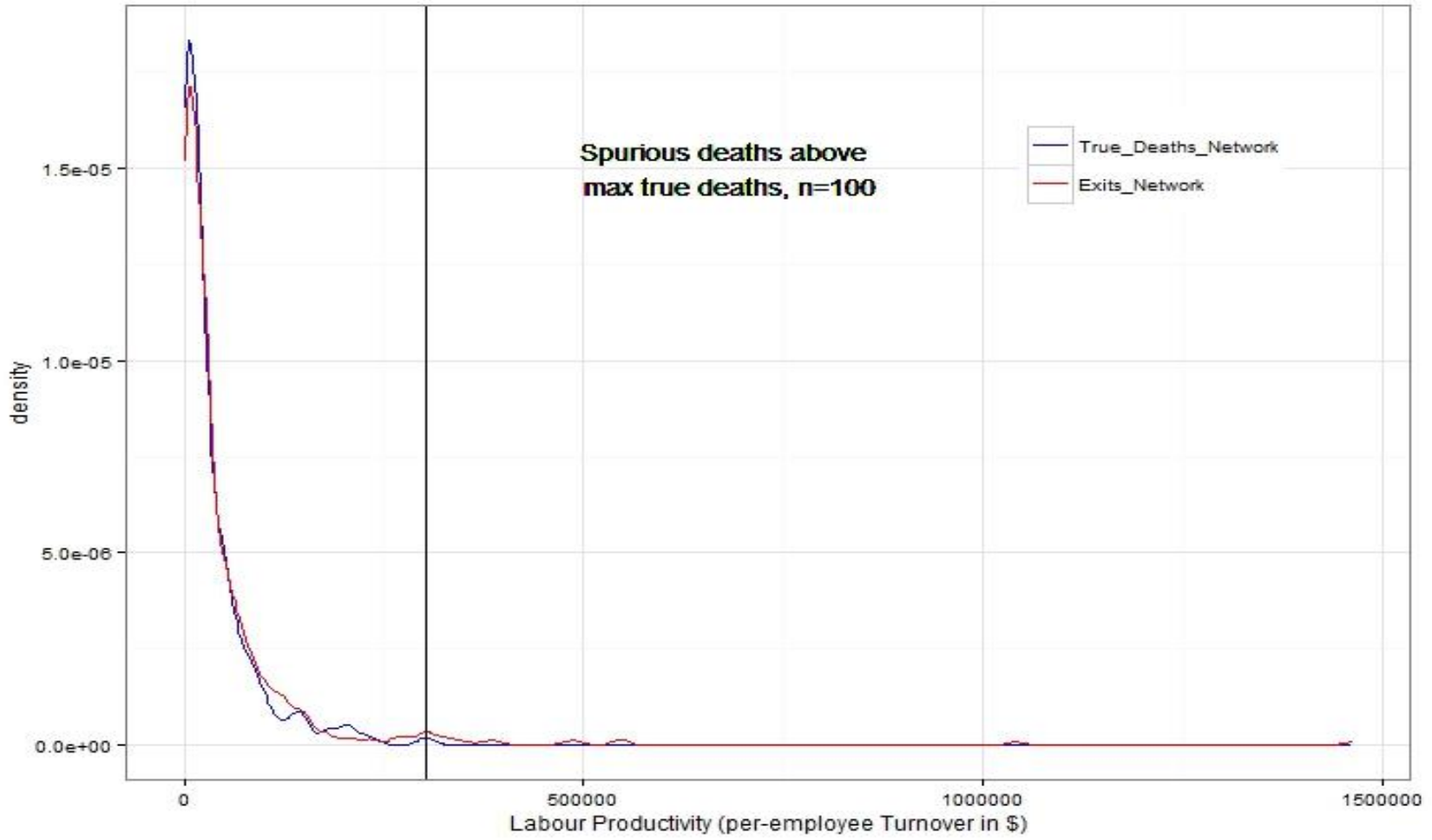
Firm exit: A continuing example



Firm exit: a death example



Summaries and results – bias without correction

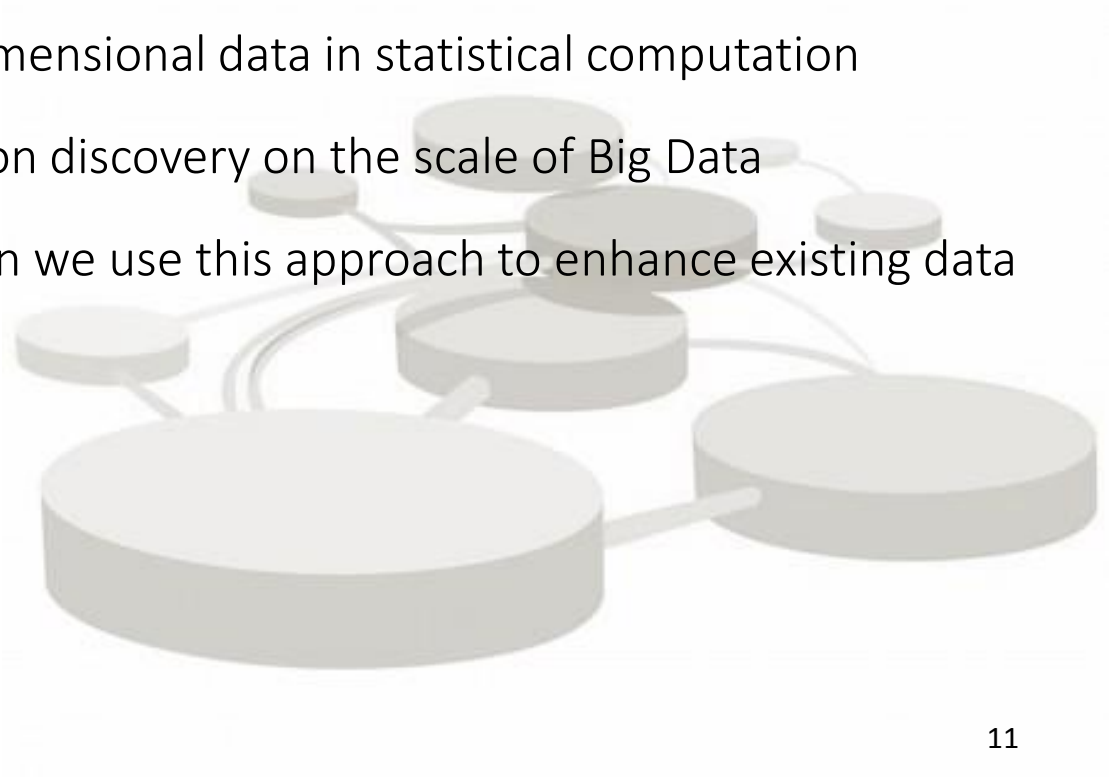


Application 2: DEMO

Australian Bureau of Statistics

Advantages for Official Statistics

- ✓ Consistency in the use of statistical concepts across data collections;
- ✓ Linking multiple disparate datasets for different analytical perspectives
- ✓ Integrating dynamic structured and unstructured content
- ✓ Manipulating highly multidimensional data in statistical computation
- ✓ Fast and adaptive information discovery on the scale of Big Data
- ? **Research direction** – how can we use this approach to enhance existing data linking methods in the ABS?



Questions?

Email: joseph.chien@abs.gov.au
chien-hung.chien@anu.edu.au