



Australian Government

IP Australia

Practice and challenges

Name harmonisation & entity resolution in Australia



Contents

- Setting the scene
- Australian data sources
- External sources
- Problem
- Overview of name harmonisation approaches
- Canopy/block, score & threshold
- Network science approaches
- Machine learning assistance in ER process
- Australia's ER products



Setting the scene

- Over 120 years of IP rights filings
 - Multiple use cases for name data
 - Applicants/ Inventor names for granting rights
 - Name data for analytics purposes
 - Legacy systems
 - Name identification/ matching not done up front
 - Duplicate records
-
- How do you solve this problem?



IP Australia data sources

Bibliographic

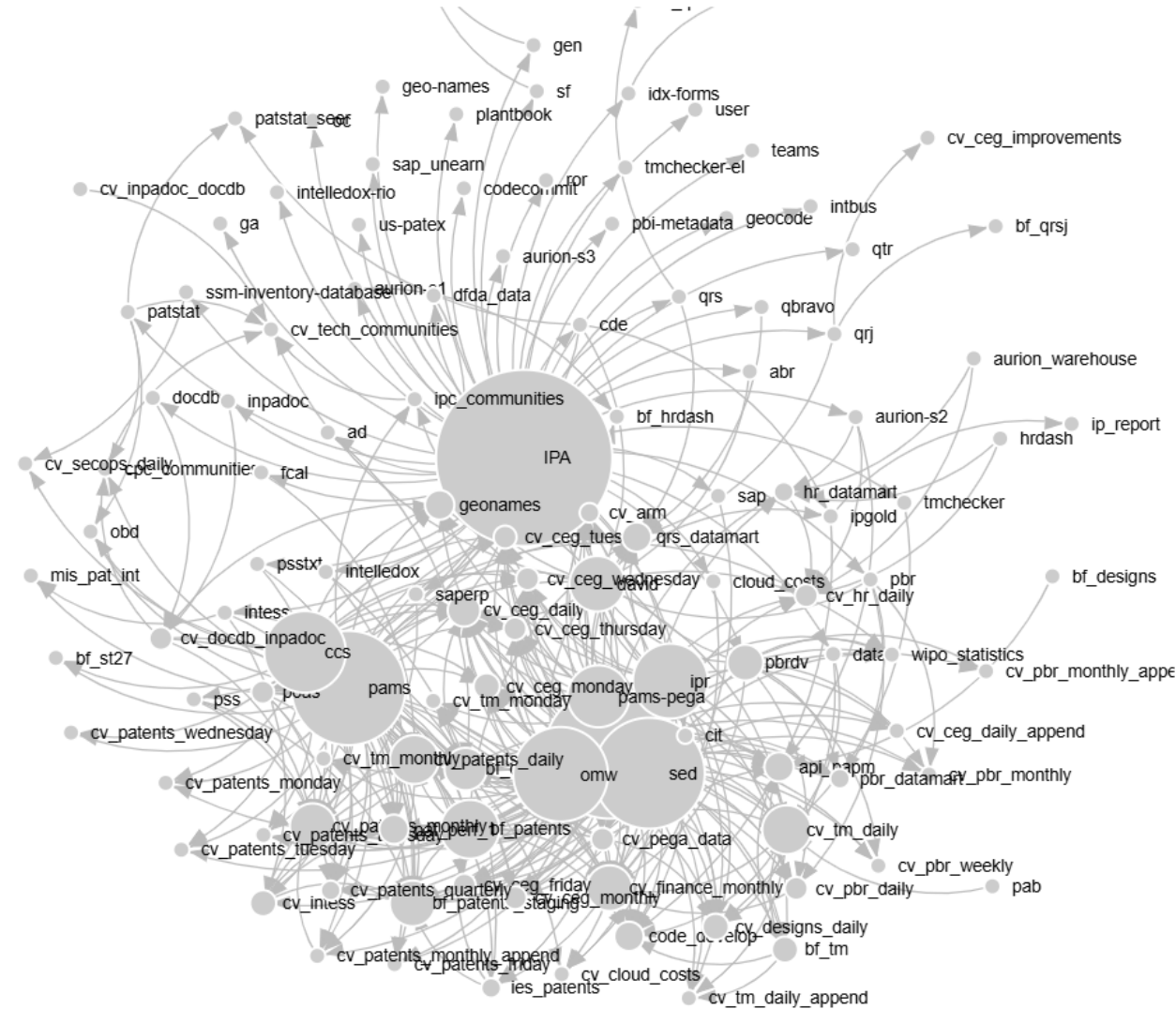
- Designs and trade marks
- Patents
- PBRDV – plant breeder's rights

Customer

- Across multiple sources depending on IP right

Financial

- Duplicated
- Scores of reporting datasets



External data sources

International IP right data

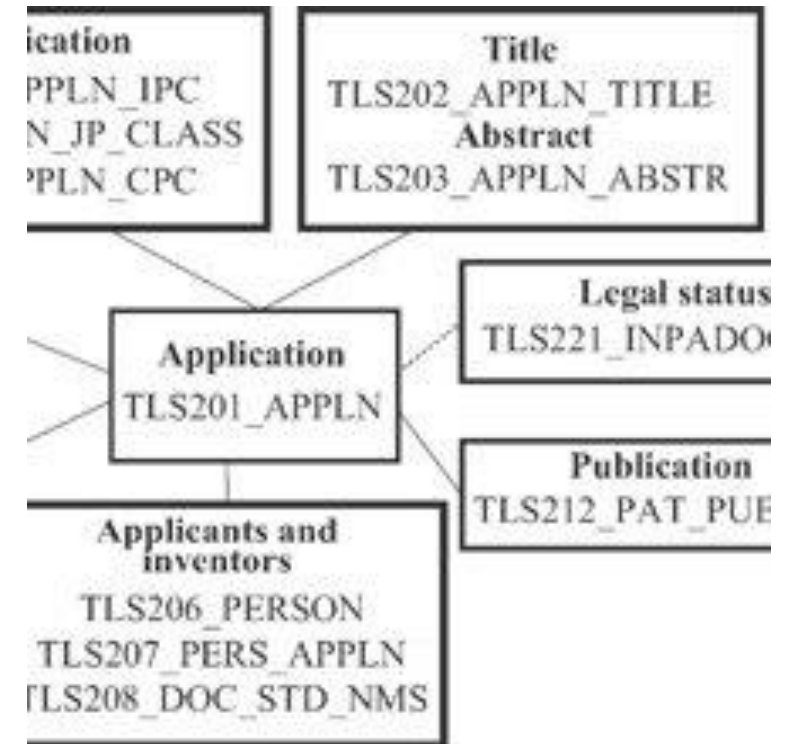
- PATSTAT
- INPADOC/DOCDB

Entity references

- Research organisation registry
- Australian business register
- ATO (open) data
- DISR data

Geospatial Information

- Geonames
- Australian GNAF



Australian Government
Australian Business Register

Problem

- **Simple question – complex context**
 - Now vs then?
 - Identity is non-stationary.
 - People and companies change who they are across time.
 - Who's asking? and why?
 - Different reckonings of identity are relevant to different contexts
 - Authentication vs analytics
 - Assumption of intentionality?
- **As a result ground truth is frequently absent or mutable**
- **Computationally complex without good heuristics**

1918 – 1955

MATSUSHITA ELECTRIC

1955 – 1965



1965 - 1968



1968 - 1971



1971 - now

Panasonic



Entity resolution in Australia — an overview

- Some hand tooling
- ... but data driven approaches more and more common
- Canopy/block, score and (ML assisted) thresholding methods
- Network analysis methods
- Machine learning assistance



Australian Government

IP Australia

Canopy/block, score & threshold

Canopy/block

- Define a heuristic to eliminate bad comparisons and bring it within computational power.

Score

- Score comparisons based upon similarity metrics
 - Vectorise & cosine similarity
 - String/set similarity (jaccard, jaro-winkler)
 - Others?

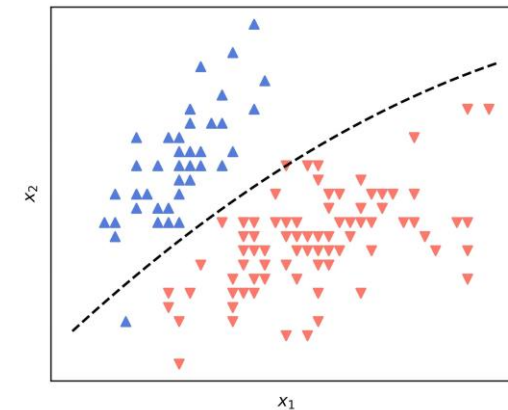
Threshold

- Set a minimum required standard for pairings to be considered matches
- Or make a computer do it for you
 - ML

Canopy/block

Colour	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

Score



Threshold

$O(n^2)$

	Cat	Catfish	Horse	Horse fly
Cat				
Catfish				
Horse				
horsefly				

$O(n \log n)?$

	Cat	Catfish	Horse	Horse fly
Cat				
Catfish				
Horse				
horsefly				

Network approaches

Treat entity resolution as a network problem

- Treat name variations as a nodes in a network
- Treat similarities between nodes as network edges
- Treat identity as a clustering (or community detection) problem

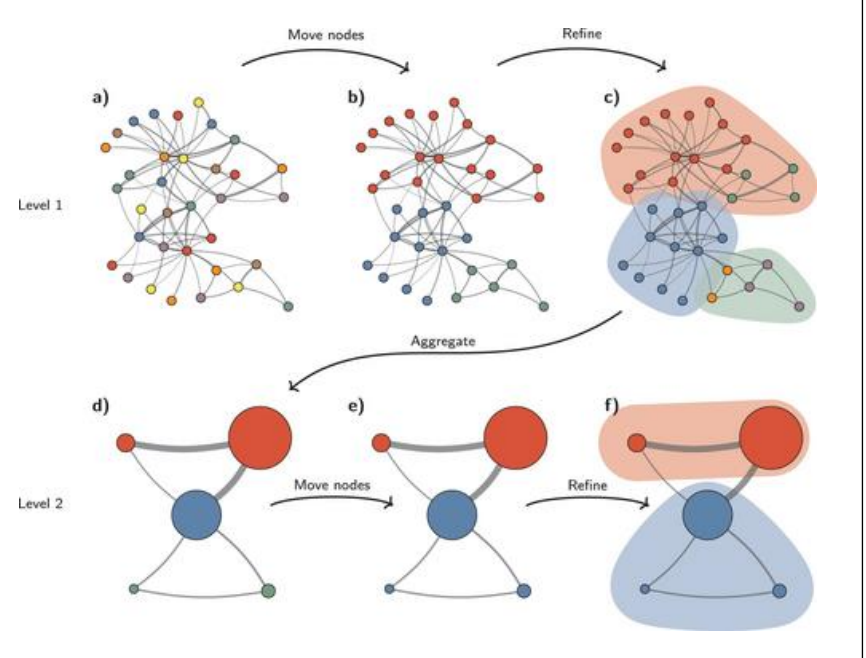
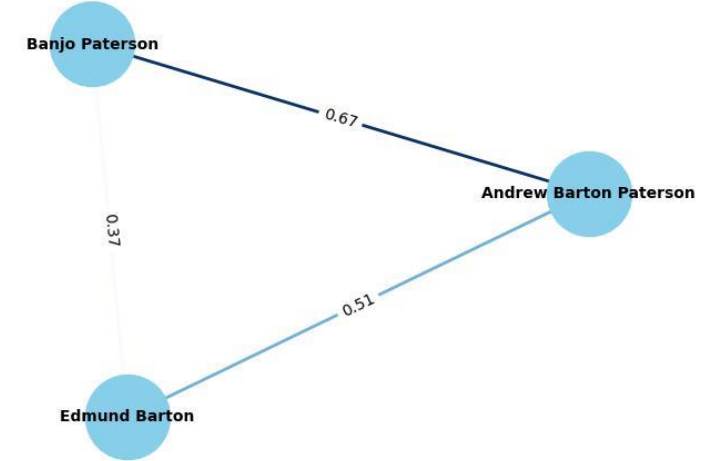
Many conventional clustering approaches require a priori declaration of the number of clusters

- Impossible to know
- Computationally intractable

Community detection methodologies

- Girvan-Newman
 - Intractable at scale
- Modularity optimisation methods
 - Leiden & Louvain algorithms.
 - Select “optimal” resolution parameter to approximate stochastic block model

Network diagram with name similarity



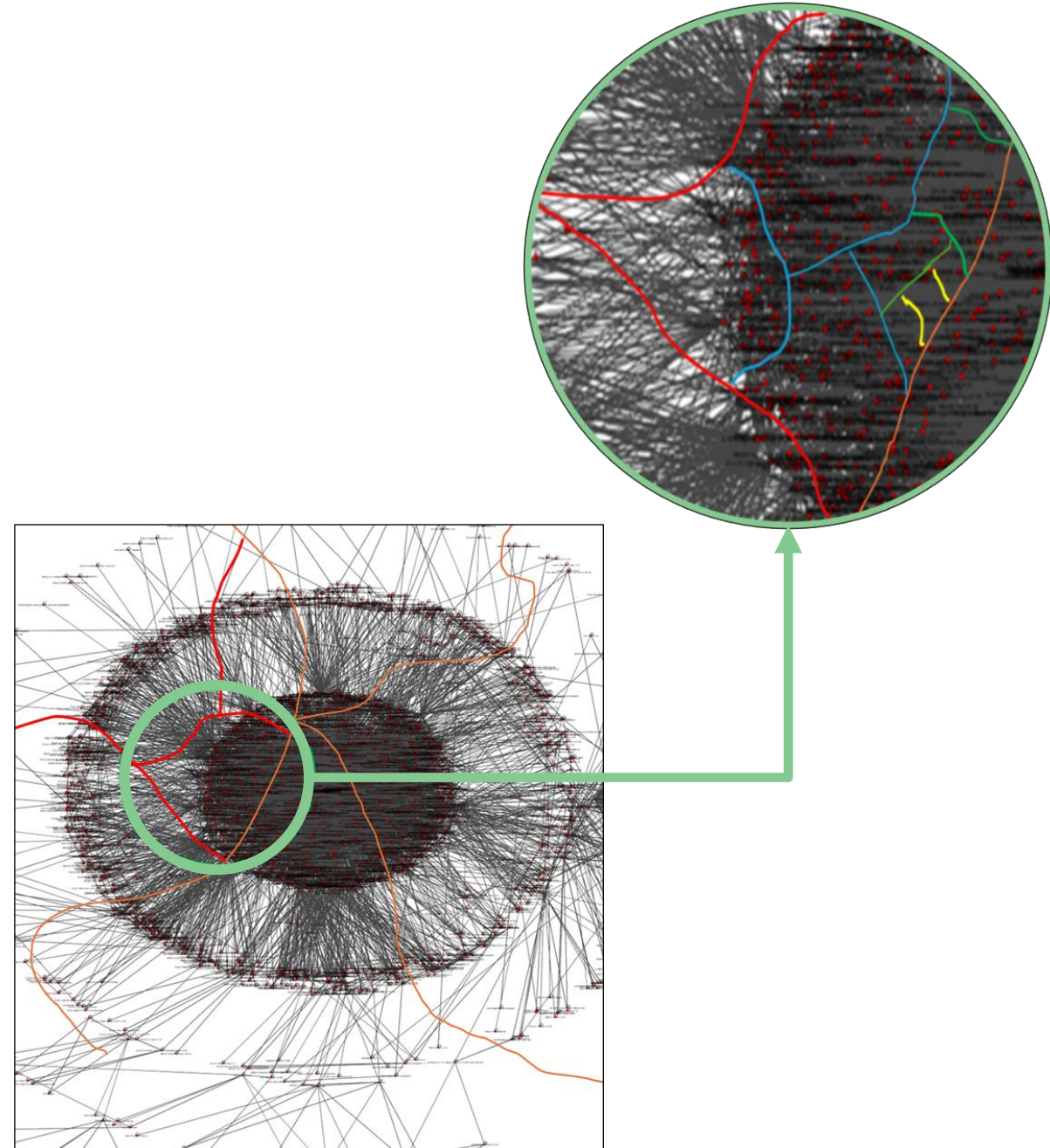
Network approaches – iterative recurrent leiden

Address is fractally decomposable – Identity should therefore be addressable

Iterative recurrent leiden recurrently subdivides communities into smaller and smaller components by reapplying the leiden algorithm and ramping up the resolution parameter

Process:

1. Initialise network
2. For level in required resolution levels:
 1. Apply leiden algorithm at resolution level
 2. Append to each node's address the community number at the current level of resolution
 3. $\text{Resolution} = \text{resolution} * \text{increase_factor}$
3. Report each node's position using resolution address
 1. 1-125-0-0-0-1-0-1-1-2-0-0-0
 2. Report identity using above addresses at specified depth



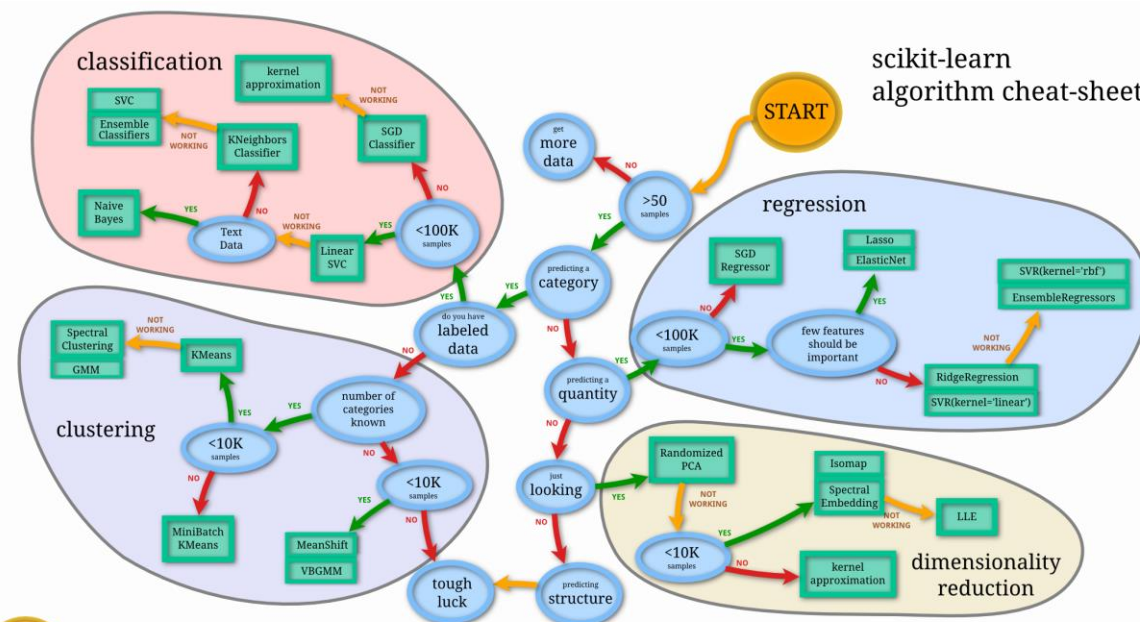
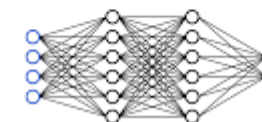
Machine learning assisted entity resolution

Most common use is to match entities in the scoring space

- Sklearn
- Keras
- Whatever you like

Limited use of LLMs at the present time

- Names (usually) lack semantic content
 - LLMs are very good at “understanding” semantic content
 - But names don’t really “mean” anything so this ability is less helpful.



Australia's entity resolution products

DAVID & IPRapid

- Internal and open copies of Australia's bibliographic data
 - First line reporting datasets for internal & external customers
 - Uses an (old) block, score & threshold approach with machine learning assistance
 - Harmonized ID across internal Australian bibliographic & customer data.

PATSTAT SEER

- Creation of an extra name harmonization step over PATSTAT (and soon inpadoc/docdb data).
- Uses iterative recurrent leiden to assign a network based ID
- Also utilises blocking, scoring and thresholding with machine learning assistance
- Internal dataset



ER adjacent products

Geonames

- Australian privacy law interaction with data security and sovereignty.
- Conventional geocoding is undesirable under some of these constraints.
 - Expensive & unnecessary
 - You don't need precise geospatial coordinates – but you need to know where Sydney is.
- Geonames is an open data product which contains place names worldwide.
- Partial geocoding using address string matching in a similar way to an ER workflow.
- Integrated into Internal data using a block, score & threshold approach.



The GeoNames geographical database covers all countries and contains over eleven million placenames that are available for download free of charge.



Australian Government

IP Australia

Contact us

☎ 1300 65 1010 (9am-5pm)

🌐 ipaustralia.gov.au

✉ data@ipaustralia.gov.au

