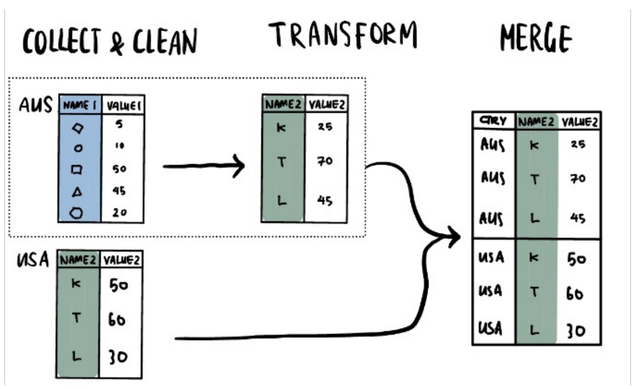
Using ***Graphs, Matrices and Edge Lists*** to investigate, illuminate and improve ***Ex-Post Data Harmonisation***

Crossmaps: A principled approach to ex-post data harmonisation and dataset integration

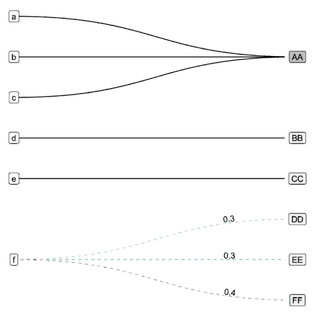
Harmonising and merging data collected under different statistical classifications, taxonomies or nomenclatures is often required to analyse and compare social, political and economic phenomena across time or countries. Procedures used to achieve comparability are broadly known as **Ex-Post Harmonisation,** and include the transformation of data collected under a **source** taxonomy into harmonised data classified according to a **target** taxonomy. We refer to this sub-task as a **Cross-Taxonomy Transformation,** and encapsulate the transformation logic in a new information structure: the **Crossmap.**

Cross-Taxonomy Transformation is an imputation ***from source to target*** data

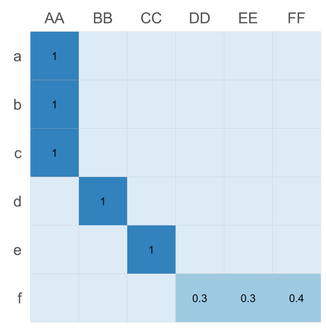


Crossmaps is a unified framework for ***specifying***, *validating, implementing, documenting and analysing* cross-taxonomy transformations

Weighted Bi-Partite Graph



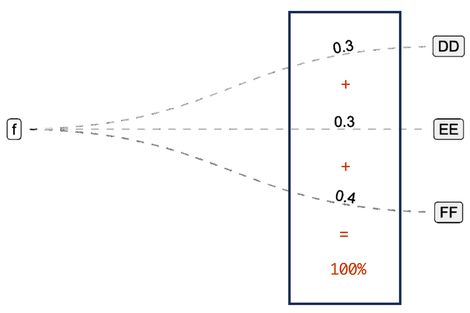
Transition / Bi-Adjacency Matrix



Edge List Table / Adjacency List

from	to	weights
a	AA	1.0
b	AA	1.0
c	AA	1.0
d	BB	1.0
e	CC	1.0
f	DD	0.3
f	EE	0.3
f	FF	0.4

Transformation logic can be *validated via graph properties* instead of ad-hoc assertions or line-by-line code review



Condition for preserving numeric totals

Data transformation can be *implemented* using validated crossmaps ***via matrix multiplication*** [1] performed ***as database operations*** on the edge list [2]

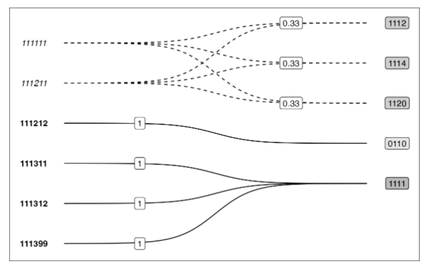
```
anzsco_xmap
# A tibble: 10 × 3
   anzsco22 isco8 weights
   <chr>    <chr>   <dbl>
 1 111111   1112    0.333
 2 111111   1114    0.333
 3 111111   1120    0.333
 4 111211   1112    0.333
 5 111211   1114    0.333
 6 111211   1120    0.333
 7 111212   0110    1
 8 111311   1111    1
 9 111312   1111    1
10 111399   1111    1

anzsco22_stats
# A tibble: 6 × 2
  anzsco22 count
  <chr>    <dbl>
1 111111    1000
2 111211     500
3 111212      40
4 111311     300
5 111312     150
6 111399      10

apply_xmap(.data = anzsco22_stats,
           .xmap = anzsco_xmap)
# A tibble: 5 × 2
  isco8 new_count
  <chr>     <dbl>
1 0110         40
2 1111        460
3 1112        500
4 1114        500
5 1120        500

# mock-up of apply_xmap() function
apply_xmap <- function(.data, .xmap) {
  left_join(
    x = .data,
    y = .xmap,
    by = "anzsco22") |>
  mutate(part_count = count * weights) |>
  group_by(isco8) |>
  summarise(new_count = sum(part_count))
}
```

Bi-graph visualisation and summary techniques can be used to ***design*** data provenance ***documentation*** [3]

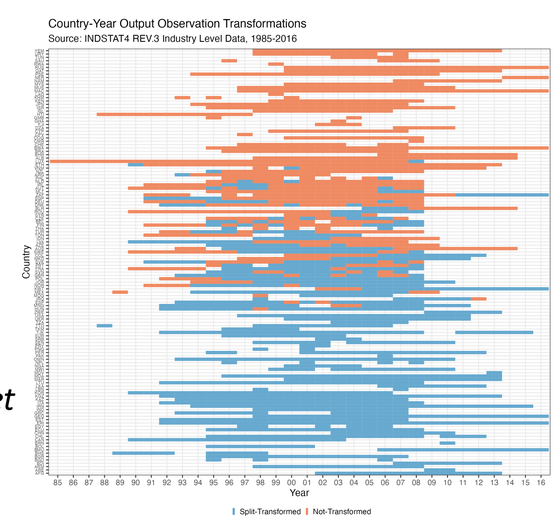


anzsco22	anzsco22_descr
111111	Chief Executive or Managing Director
111211	Corporate General Manager
111212	Defence Force Senior Officer
111311	Local Government Legislator
111312	Member of Parliament
111399	Legislators nec

isco8	isco8_descr
1112	Senior government officials
1114	Senior officials of special-interest organizations
1120	Managing directors and chief executives
0110	Commissioned armed forces officers
1111	Legislators

Metrics based on *crossmap properties* can be used to *quantify and compare:*

- How does the degree and extent of imputation differ between crossmaps?
- How robust are downstream results to alternative harmonisation designs?
- How much imputation has been performed on a given dataset with a given crossmap?
- *Which observations in a harmonised dataset have undergone the most (or least) transformation?*



Crossmaps integrates multiple complementary perspectives from ***graph theory, matrix algebra and relational databases*** to **explore properties** of ex-post harmonised datasets **and unify related** cross-taxonomy transformation **workflows.**

References

[1] Hulliger, Beat. 1998. "Linking of Classifications by Linear Mappings." *Journal of Official Statistics* 14 (January): 255–66.

[2] Zhou, Xiantian, and Carlos Ordonez. 2020. "Matrix Multiplication with SQL Queries for Graph Analytics." In *2020 IEEE International Conference on Big Data (Big Data)*, 5872–73. Atlanta, GA, USA IEEE https://doi.org/10.1109/BigData50022.2020.9378275.

[3] Huang, Cynthia A. 2023. "Visualising Category Recoding and Numeric Redistributions." August 12, 2023 http://arxiv.org/abs/2308.06535.

Acknowledgements

Thank you to Laura Puzzello for her ongoing support and funding of earlier iterations of this work. Many thanks also to Rob Hyndman, Sarah Goodwin, Simon Angus, Emi Tanaka, Patrick Li, Mitch O'Hara-Wild and my other colleagues at Monash EBS and Monash SoDa Labs for their helpful guidance, feedback and suggestions. The author is supported in part by top-up scholarships from Monash Data Futures Institute and the Statistical Society of Australia.

Cynthia A. Huang

Department of Econometrics and Business Statistics, Monash University

Supervised by Prof. Rob J Hyndman, Dr. Sarah Goodwin and Assoc. Prof. Simon Angus