Using Personal Names to Explore Cultural, Ethnic and Linguistic Structure in Populations

David O'Sullivan
Department of Geography

Berkeley Demography
September 24, 2014

Diversity and segregation in urban communities

Long standing interest in urban neighbourhoods, urban neighbourhood change, diversity and segregation of urban populations

Some previous publications

Hong S-Y and O’Sullivan, D. 2012. Detecting ethnic residential clusters using an optimisation clustering method. *International Journal of Geographical Information Science*, 26, 1257–1277.

Xue J, Friesen W and O’Sullivan D. 2012. Diversity in Chinese Auckland: Hypothesising multiple ethnoburbs. *Population, Space and Place*, 18, 579–595.

O’Sullivan D. 2009. Changing neighborhoods - neighborhoods changing: a framework for spatially explicit agent-based models of social systems. *Sociological Methods and Research*, 37, 498–530.

Reardon SF, Farrell CR., Matthews SA, O’Sullivan D, Bischoff K and Firebaugh G. 2009. Race and space in the 1990s: changes in the geographic scale of racial residential segregation, 1990-2000. *Social Science Research*, 38, 55–70.

Lee BA, Reardon SF, Firebaugh G, Farrell CR, Matthews SA and O’Sullivan D. 2008. Beyond the census tract: patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73, 766–791.

Reardon SF, Matthews SA, O’Sullivan SA, Lee BA, Firebaugh G, Farrell CR and Bischoff K. 2008. The geographic scale of metropolitan segregation. Demography, 45, 489-514.

O’Sullivan D and Wong DW. 2007. A surface-based approach to measuring spatial segregation. Geographical Analysis, 39, 147–68.

Reardon SF and O’Sullivan D. 2004. Measures of Spatial Segregation. *Sociological Methodology*, 34, 121–62.

O’Sullivan D. 2002. Toward micro-scale spatial modelling of gentrification. *Journal of Geographical Systems*, 4, 251–74.

Some personal background

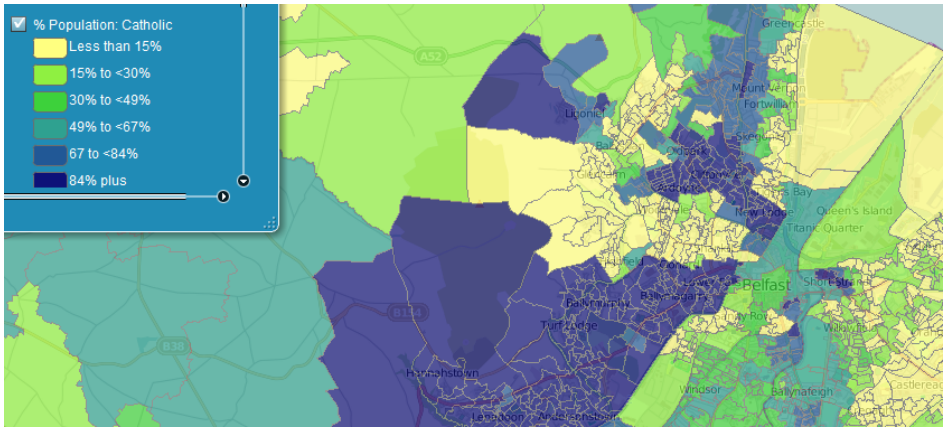
- Belfast 'Catholic'
- Meet Fintan and Malachy. . .
- Still matters today

Some personal background

- Belfast ‘Catholic’
- Meet Fintan and Malachy. . .
- Still matters today

Some personal background

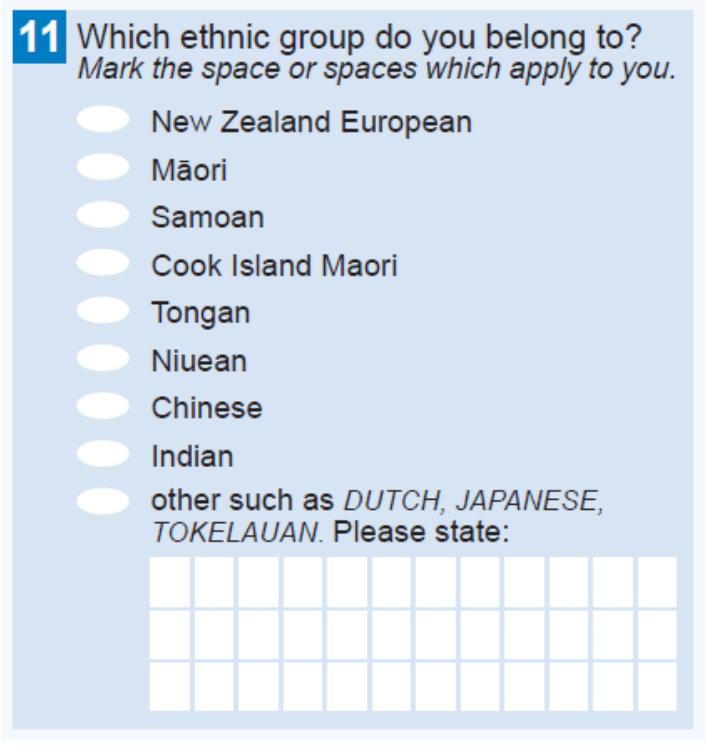
- Belfast ‘Catholic’
- Meet Fintan and Malachy. . .
- Still matters today



Source: airomaps.nuim.ie

The limits of conventional ‘identity’ data

- ‘Identity’ is a tricky, multi-dimensional concept
- Standards for how to collect data don’t match
- Conventional instruments don’t capture it well
  - Census ‘tick boxes’
  - Census ‘write-ins’



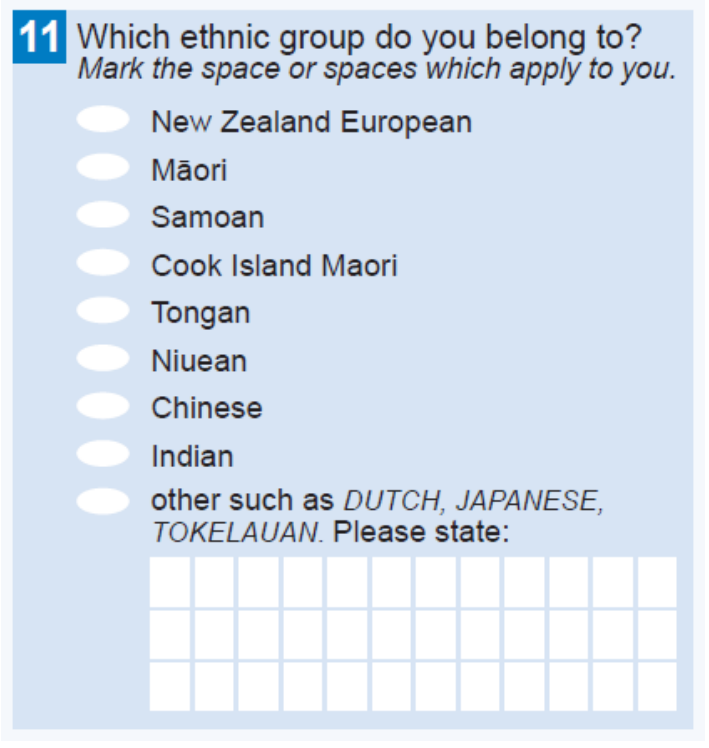
11 Which ethnic group do you belong to?
Mark the space or spaces which apply to you.

- New Zealand European
- Māori
- Samoan
- Cook Island Maori
- Tongan
- Niuean
- Chinese
- Indian
- other such as *DUTCH, JAPANESE, TOKELAUAN.* Please state:

Source: Statistics New Zealand, www.stats.govt.nz

The limits of conventional ‘identity’ data

- ‘Identity’ is a tricky, multi-dimensional concept
- Standards for how to collect data don’t match
- Conventional instruments don’t capture it well
  - Census ‘tick boxes’
  - Census ‘write-ins’



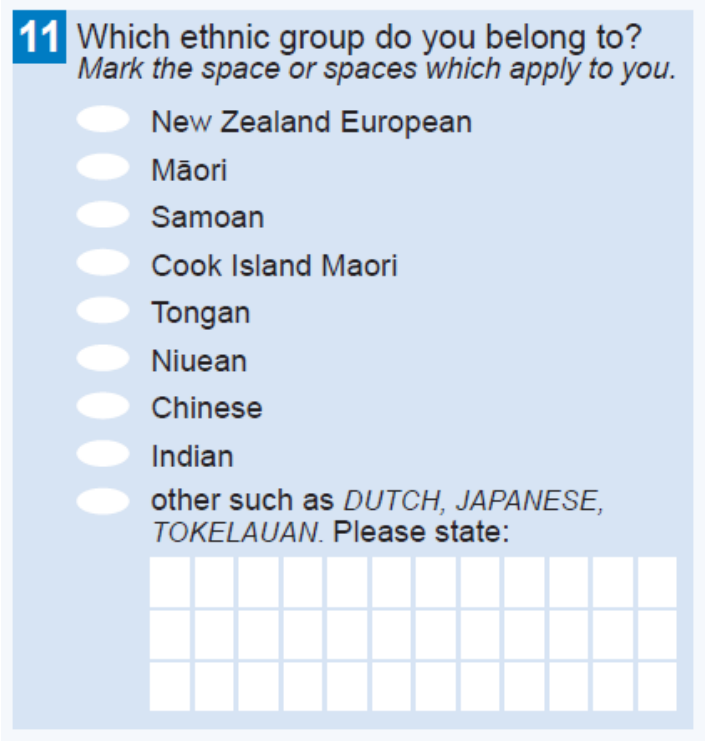
11 Which ethnic group do you belong to?
Mark the space or spaces which apply to you.

- New Zealand European
- Māori
- Samoan
- Cook Island Maori
- Tongan
- Niuean
- Chinese
- Indian
- other such as *DUTCH, JAPANESE, TOKELAUAN.* Please state:

Source: Statistics New Zealand, www.stats.govt.nz

The limits of conventional ‘identity’ data

- ‘Identity’ is a tricky, multi-dimensional concept
- Standards for how to collect data don’t match
- Conventional instruments don’t capture it well
  - Census ‘tick boxes’
  - Census ‘write-ins’



11 Which ethnic group do you belong to?
Mark the space or spaces which apply to you.

- New Zealand European
- Māori
- Samoan
- Cook Island Maori
- Tongan
- Niuean
- Chinese
- Indian
- other such as *DUTCH, JAPANESE, TOKELAUAN.* Please state:

Source: Statistics New Zealand, www.stats.govt.nz

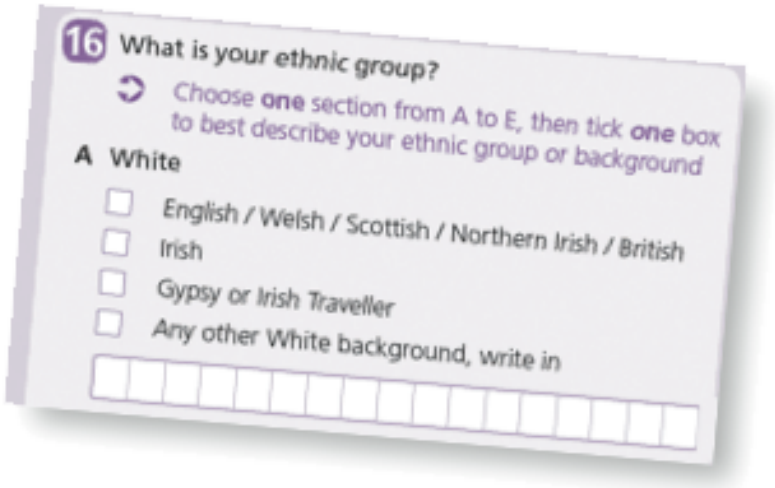
The limits of conventional 'identity' data

- 'Identity' is a tricky, multi-dimensional concept
- Standards for how to collect data don't match
- Conventional instruments don't capture it well
  - Census 'tick boxes'
  - Census 'write-ins'

Source: www.irishinbritain.org/

The limits of conventional ‘identity’ data

- ‘Identity’ is a tricky, multi-dimensional concept
- Standards for how to collect data don’t match
- Conventional instruments don’t capture it well
  - Census ‘tick boxes’
  - Census ‘write-ins’



Source: www.irishinbritain.org/

The limits of conventional ‘identity’ data

- ‘Identity’ is a tricky, multi-dimensional concept
- Standards for how to collect data don’t match
- Conventional instruments don’t capture it well
  - Census ‘tick boxes’
  - Census ‘write-ins’

The New Zealand Herald

nzherald.co.nz WEDNESDAY OCTOBER 8, 2008 12:54AM NZT

News Business Election 08 Sport Technology Entertainment Life &

National World Weather Politics Crime Health Environment Science

National ShareThis Print Email RSS

Email urges 'New Zealander' for Census

By Julie Middleton

A fast-spreading email appeal is urging people to state their ethnicity as "New Zealander" in next Tuesday's Census.

Census

- Make your vote count this year, Koreans urged
- One in three working more than 50 hours a week - Census

The email, which came to the Herald from several sources, reads: "Maybe we can get the powers-that-be to sit up and recognise that we are proud of who we are and that we want to be recognised as such, not divided into sub-categories and all treated as foreigners in our own country.

Names as markers of identity

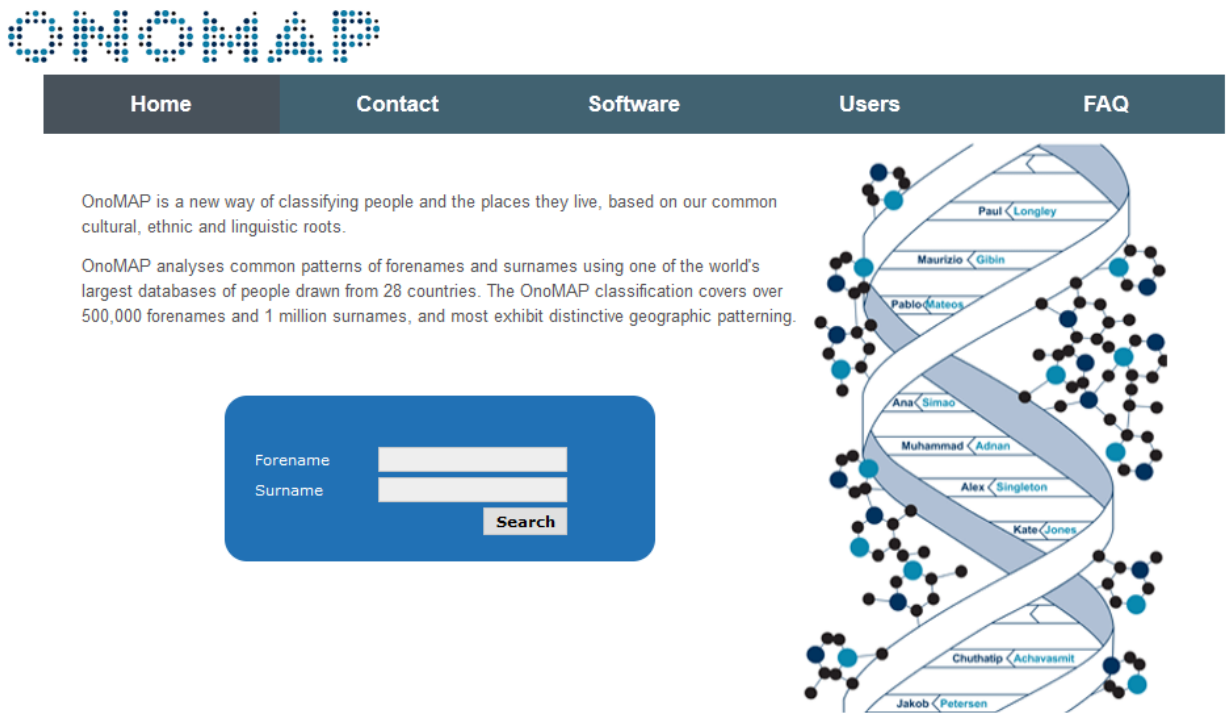
- Readily available
- Minimally invasive
- May carry complex (if limited) information

Previous work

- Mateos P. 2007. A review of name-based ethnicity classification methods and their potential in population studies, *Population Space and Place*, 13, 243–63.
- Mateos P, Webber R and Longley PA. 2007. *The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names*. CASA Working Paper 116, University College London.
- Mateos P. 2007. *An Ontology of Ethnicity Based upon Personal Names: With Implications for Neighbourhood Profiling*. PhD Thesis, University College London

Onomap and the names project at UCL

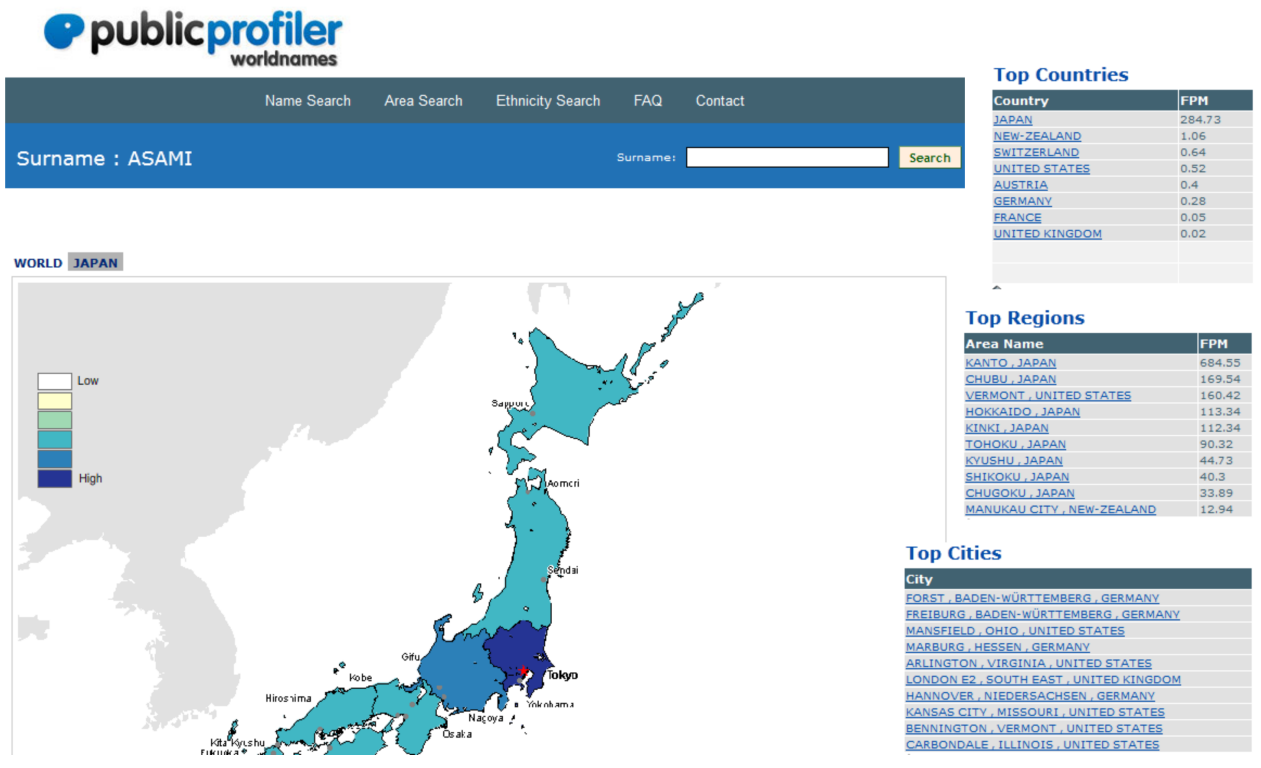
Mateos’s work helped to develop the *Onomap* classification of names into cultural-ethnic-linguistic (CEL) groups, which has been made spatial by the World Names Mapper project



www.onomap.org

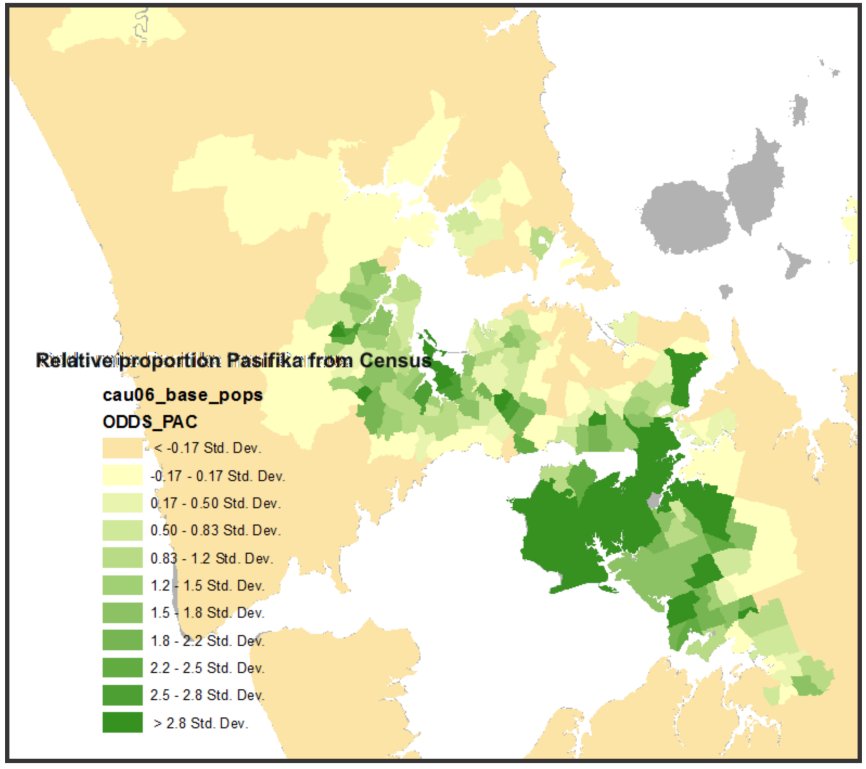
Onomap and the names project at UCL

Mateos’s work helped to develop the *Onomap* classification of names into cultural-ethnic-linguistic (CEL) groups, which has been made spatial by the World Names Mapper project

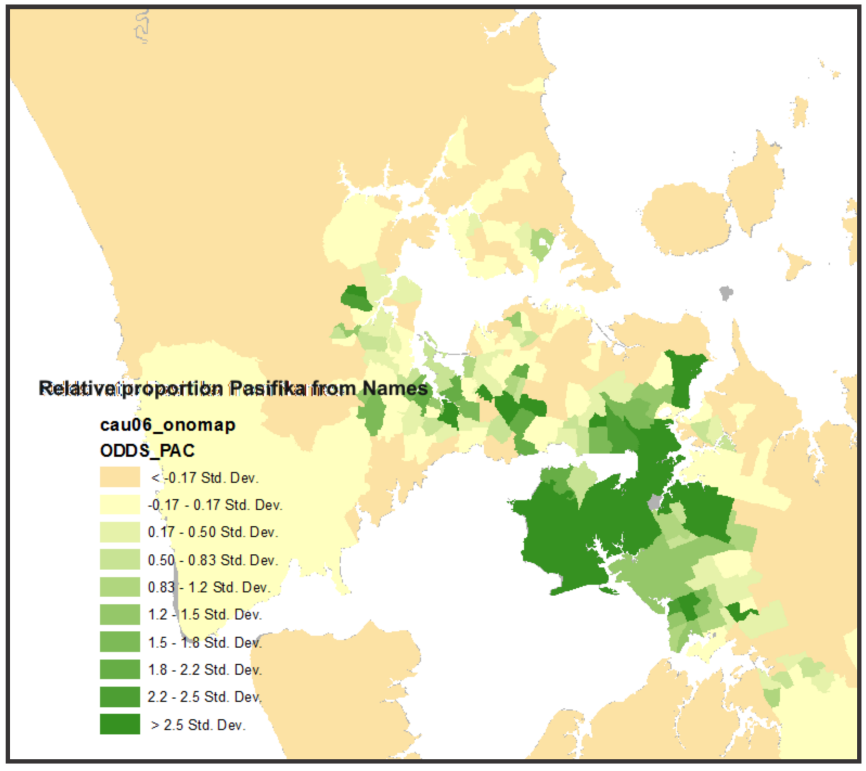


worldnames.publicprofiler.org/Default.aspx

Example names maps from New Zealand

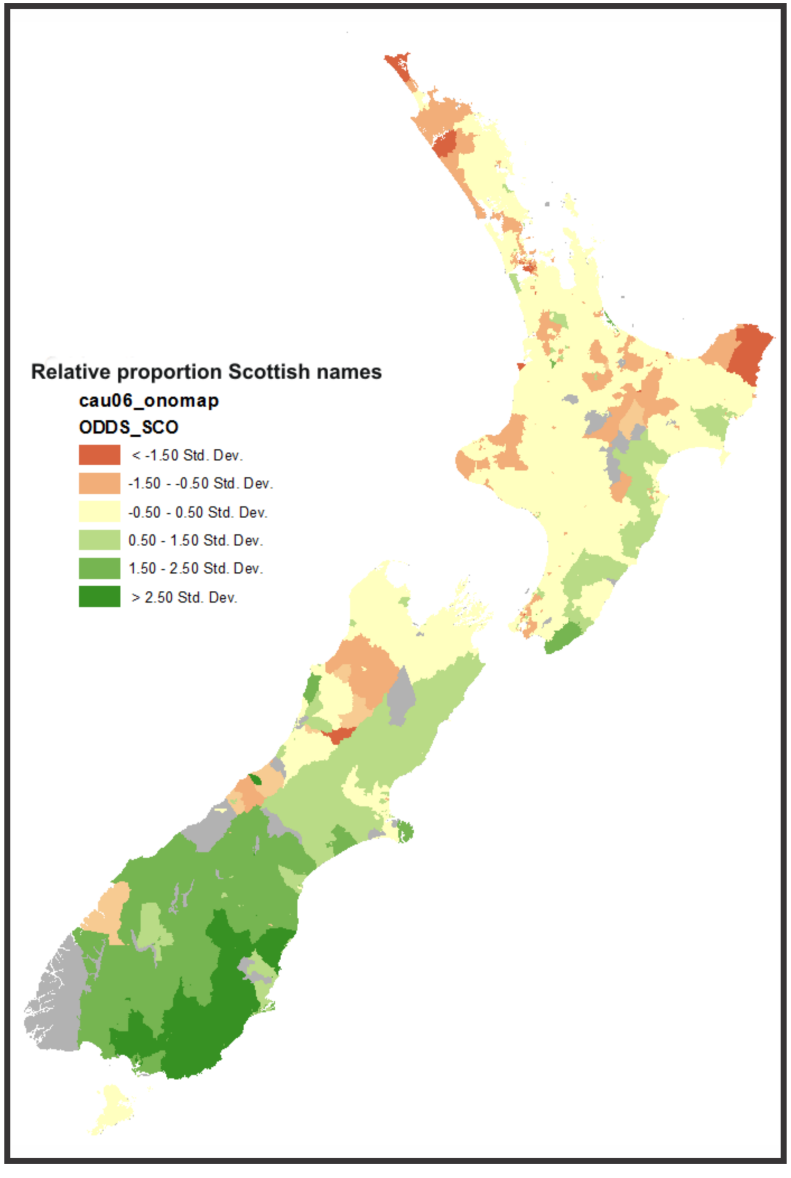


Pacific people south Auckland (census)



Pacific names

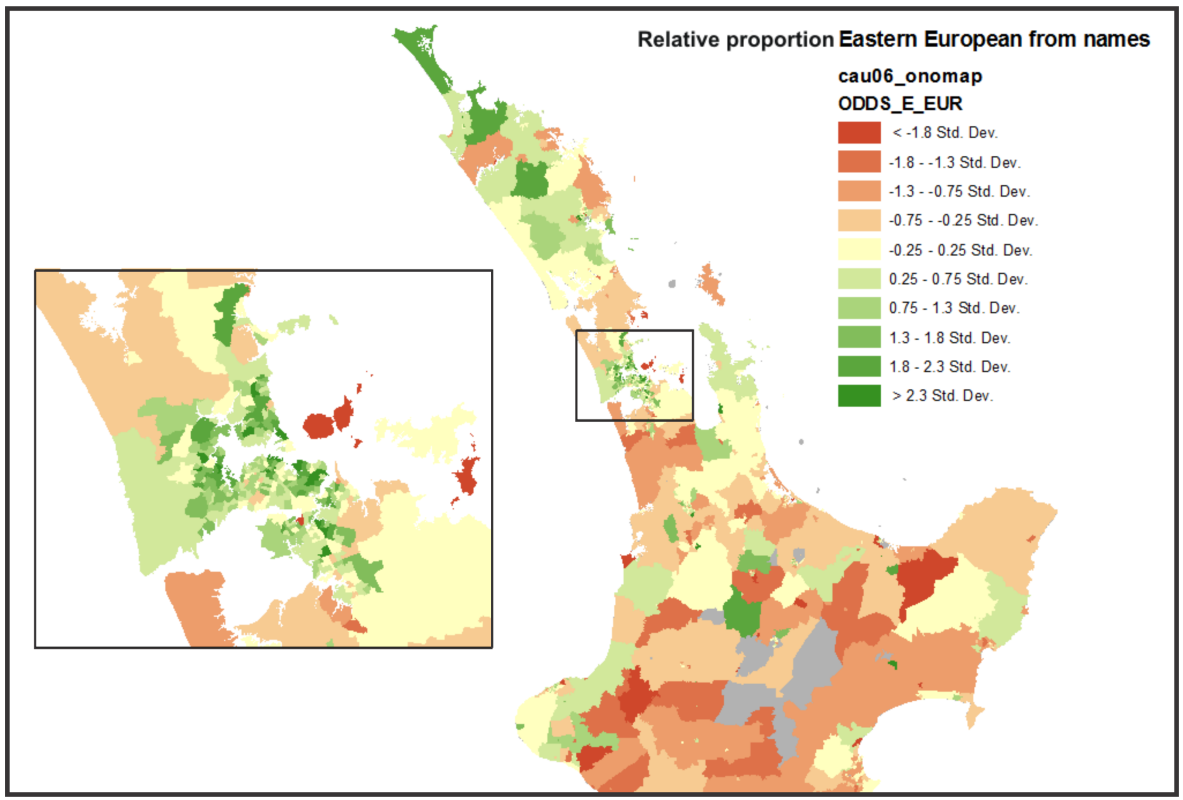
Example names maps from New Zealand



Scottish names countrywide

- No ’Scottish’ category in the New Zealand census
- Counts derived from Electoral Roll
- Map reflects the disproportionately Scottish origins of early European migration to the South Island (as well as more recent non-European immigration in the North Island)

Example names maps from New Zealand

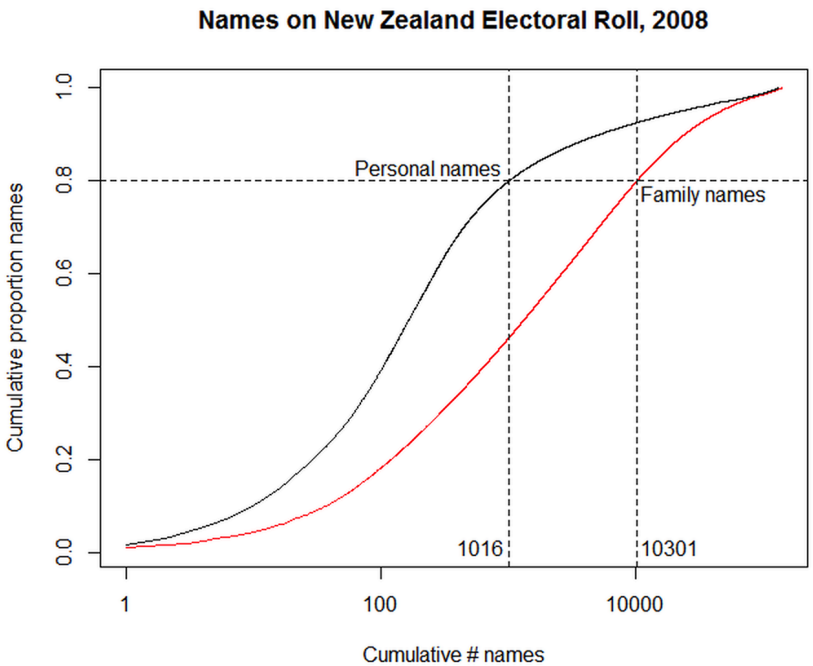


Eastern European names, west Auckland

Babich, Fistonich (Villa Maria), Mazuran, Nobilo, Sapich, Soljan, Yukich (Montana)...

Difficulties with names as data

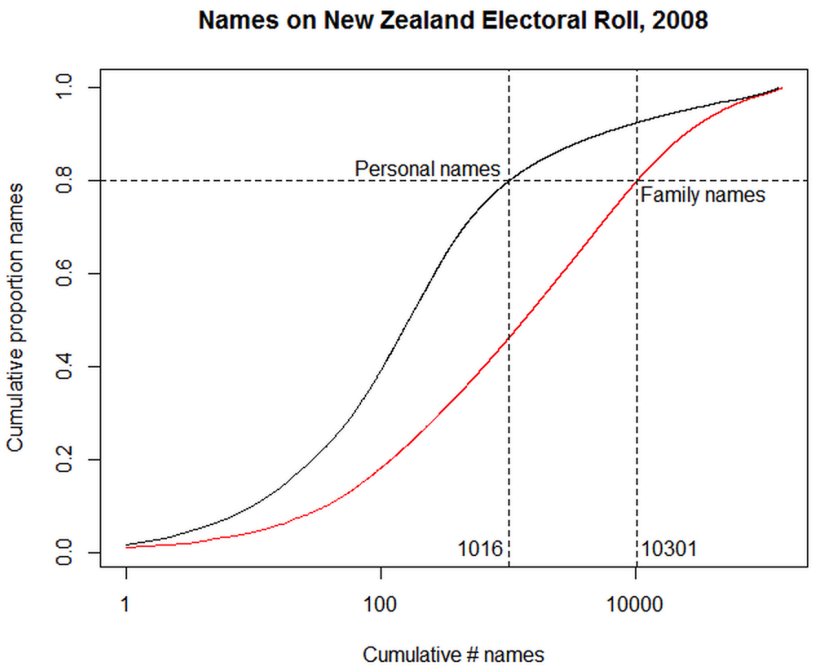
- *Very* unevenly distributed



- Also, how surnames are passed via marriage. . .
- Prone to misspellings, with no easy fixes, e.g.,
  - Is MICHEAL a real name?
  - Should ABBAGAIL - ABBEGAIL - ABBIGAIL - ABBYGAIL be 'corrected'?
- Interpretation labour-intensive, dependent on expert knowledge

Difficulties with names as data

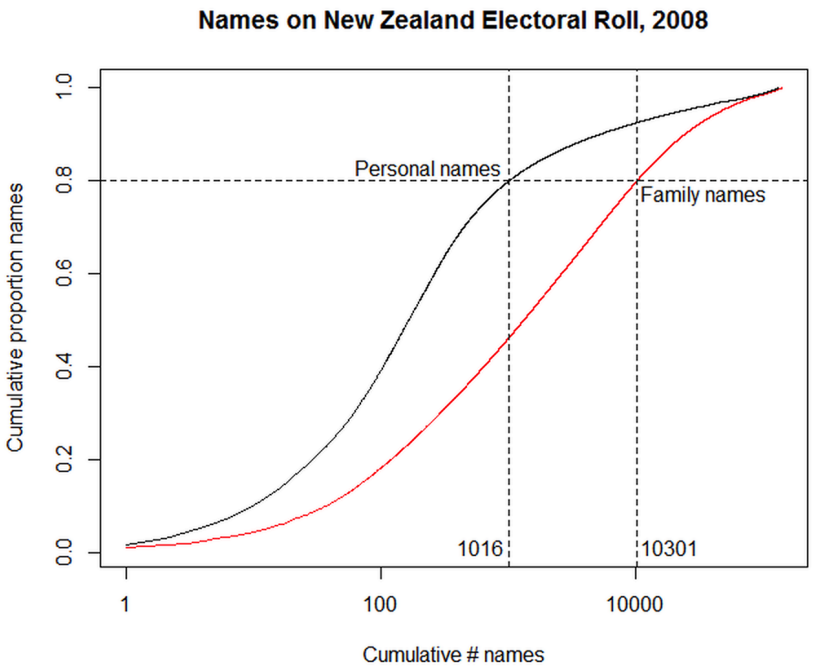
- *Very* unevenly distributed



- Also, how surnames are passed via marriage. . .
- Prone to misspellings, with no easy fixes, e.g.,
  - Is MICHEAL a real name?
  - Should ABBAGAIL - ABBEGAIL - ABBIGAIL - ABBYGAIL be 'corrected'?
- Interpretation labour-intensive, dependent on expert knowledge

Difficulties with names as data

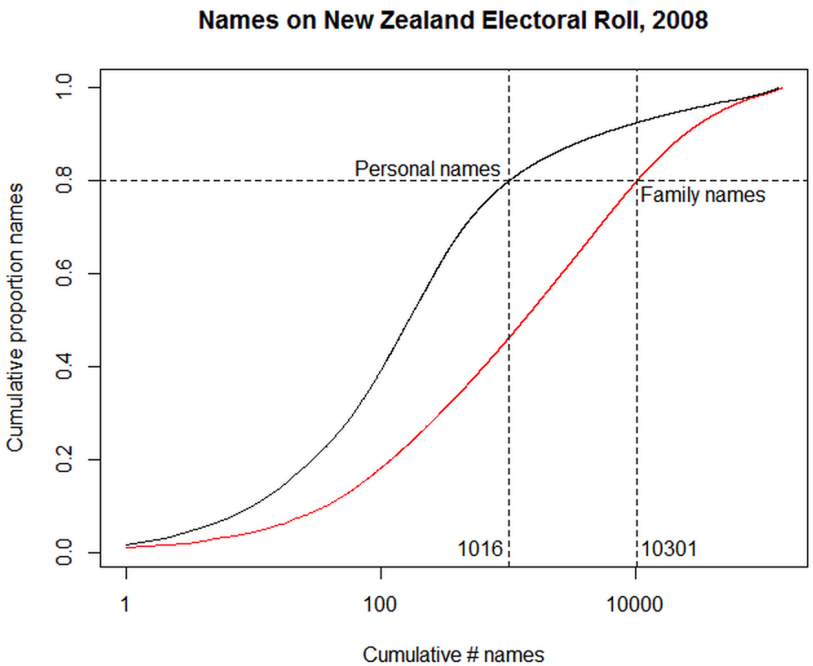
- *Very* unevenly distributed



- Also, how surnames are passed via marriage. . .
- Prone to misspellings, with no easy fixes, e.g.,
  - Is MICHEAL a real name?
  - Should ABBAGAIL - ABBEGAIL - ABBIGAIL - ABBYGAIL be 'corrected'?
- Interpretation labour-intensive, dependent on expert knowledge

Difficulties with names as data

- *Very* unevenly distributed



- Also, how surnames are passed via marriage. . .
- Prone to misspellings, with no easy fixes, e.g.,
  - Is MICHEAL a real name?
  - Should ABBAGAIL - ABBEGAIL - ABBIGAIL - ABBYGAIL be 'corrected'?
- Interpretation labour-intensive, dependent on expert knowledge

What about when we meet a name we don't know?

The New Zealand case presented some specific challenges

- A major problem dealing with names unknown to Onomap from previous experience, esp. Māori, Pacific names
- We had limited expert knowledge
- ⇒ Is there a way to infer 'new' groups from the relationships between names?
- ⇒ concept of a network of names

What about when we meet a name we don't know?

The New Zealand case presented some specific challenges

- A major problem dealing with names unknown to Onomap from previous experience, esp. Māori, Pacific names
- We had limited expert knowledge
- $\Rightarrow$ Is there a way to infer 'new' groups from the relationships between names?
- $\Rightarrow$ concept of a network of names

What about when we meet a name we don't know?

The New Zealand case presented some specific challenges

- A major problem dealing with names unknown to Onomap from previous experience, esp. Māori, Pacific names
- We had limited expert knowledge
- $\Rightarrow$ Is there a way to infer 'new' groups from the relationships between names?
- $\Rightarrow$ concept of a network of names

What about when we meet a name we don't know?

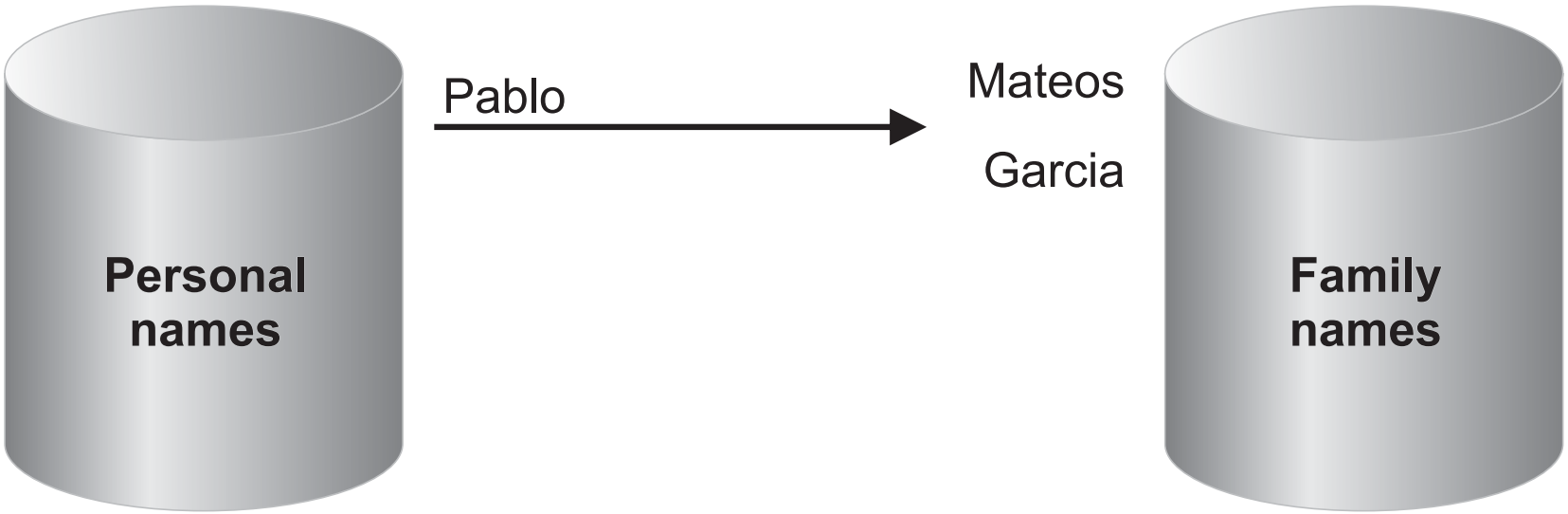
The New Zealand case presented some specific challenges

- A major problem dealing with names unknown to Onomap from previous experience, esp. Māori, Pacific names
- We had limited expert knowledge
- $\Rightarrow$ Is there a way to infer 'new' groups from the relationships between names?
- $\Rightarrow$ concept of a network of names

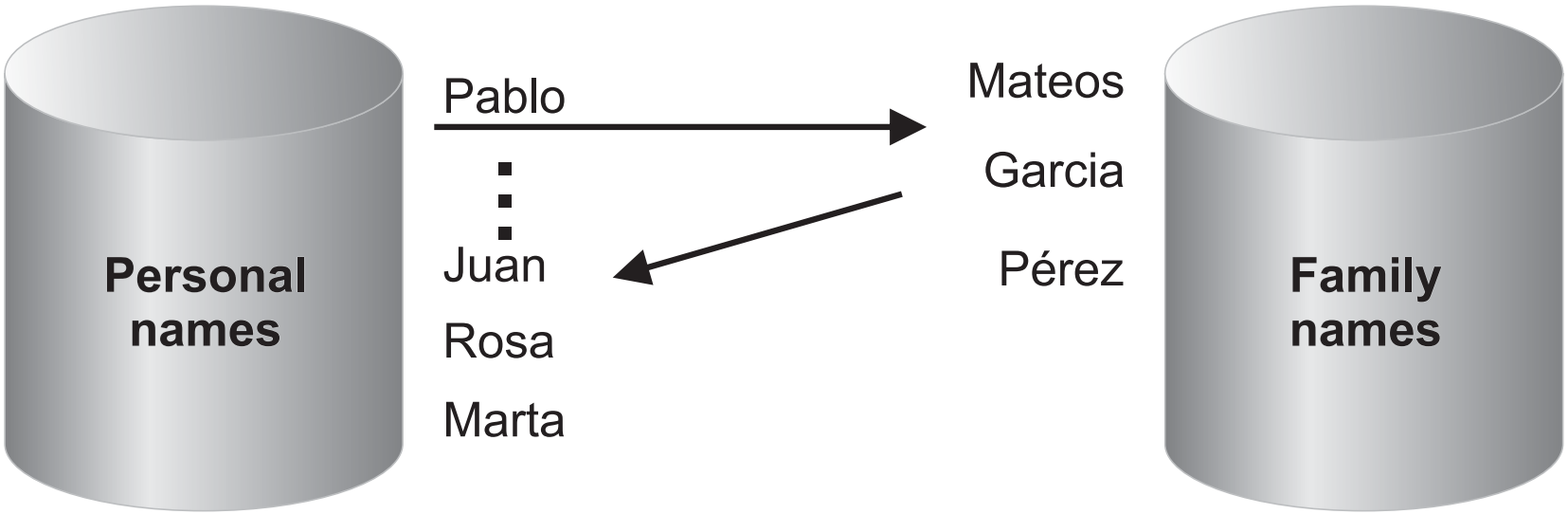
Relationships between names: from lists to networks

Pablo

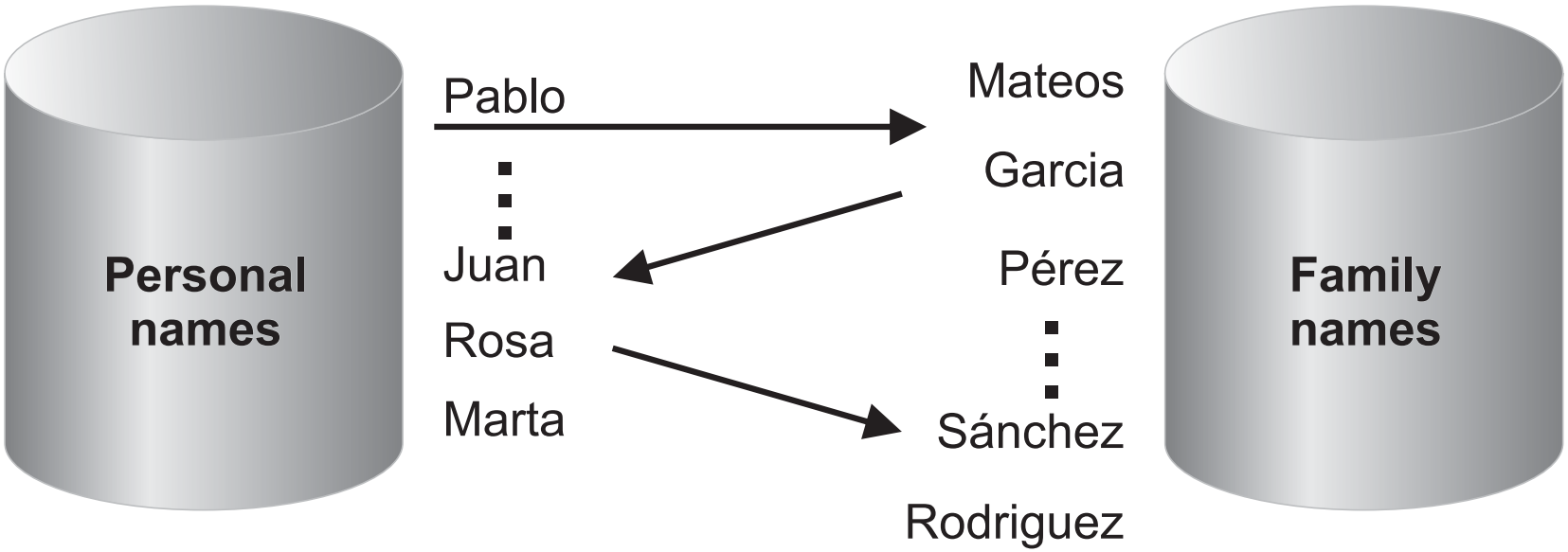
Relationships between names: from lists to networks



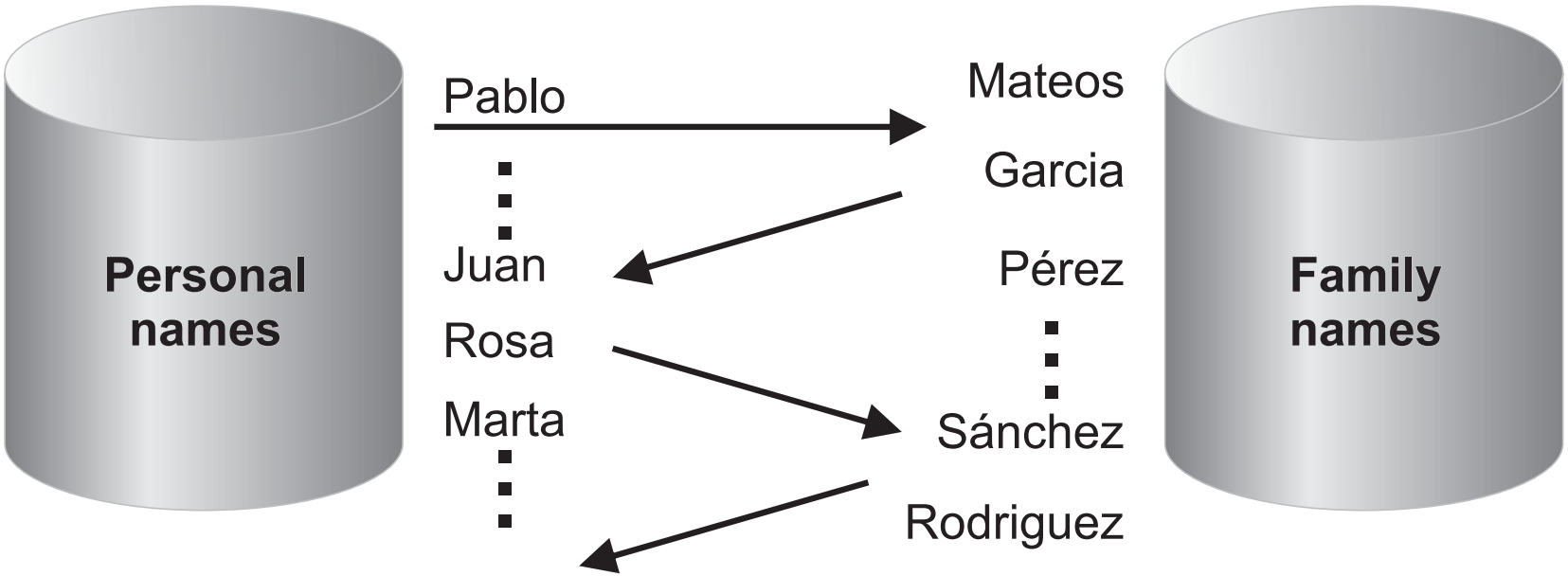
Relationships between names: from lists to networks



Relationships between names: from lists to networks

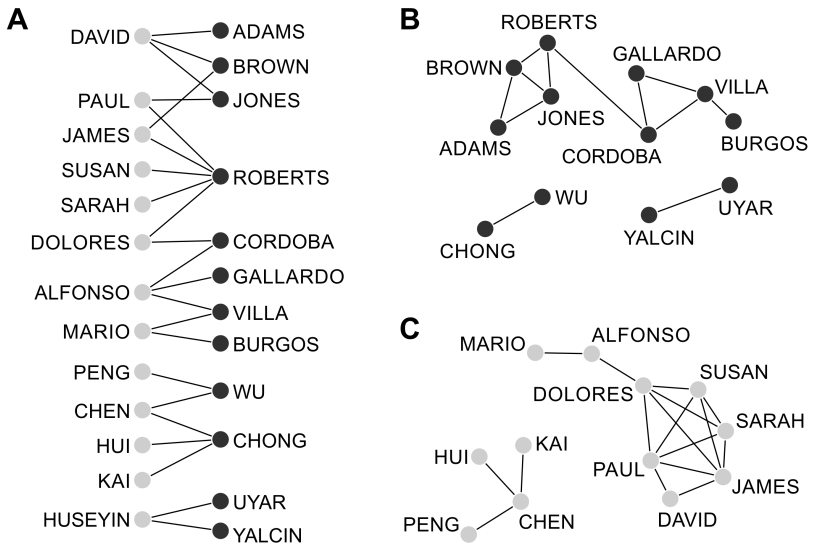


Relationships between names: from lists to networks



From two-mode to one-mode networks

We can replace all the back and forth database querying with some matrix arithmetic

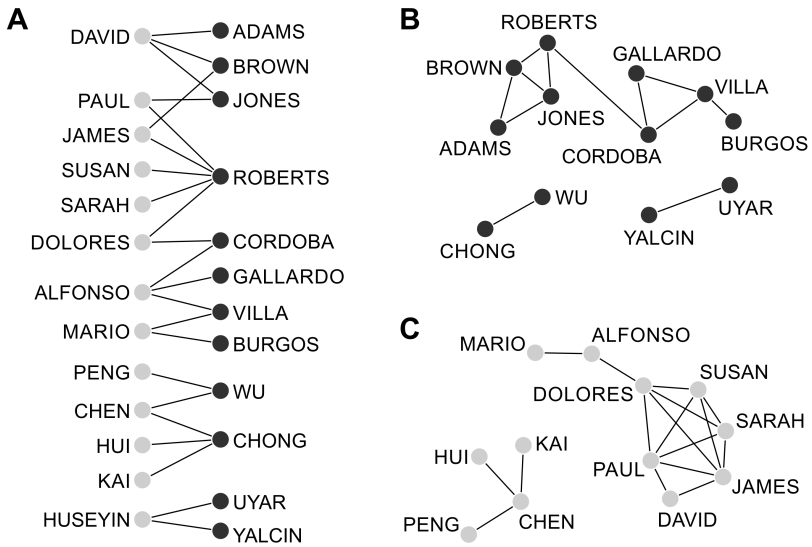


Graph adjacency matrix maths

Form a given name-family name coincidence matrix, $\mathbf{A} = [w_{gf}]$, where w_{gf} is strength of association between g-name g and f-name f

From two-mode to one-mode networks

We can replace all the back and forth database querying with some matrix arithmetic



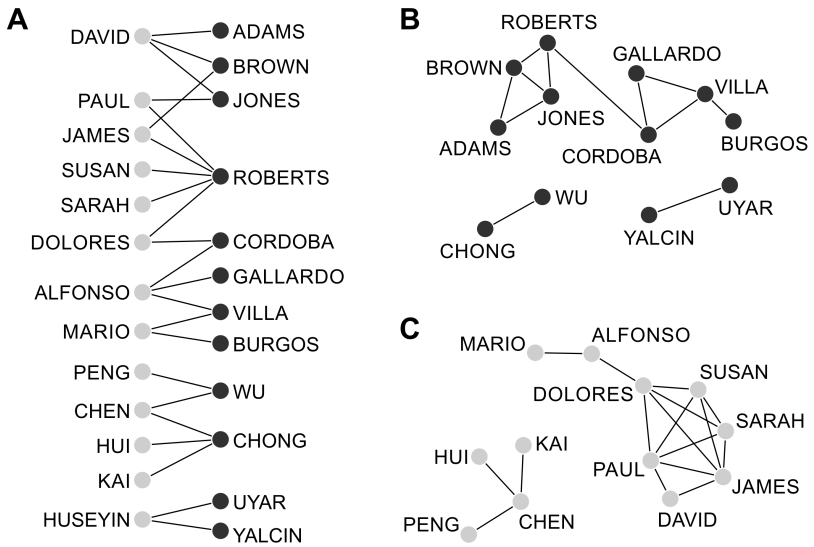
Graph adjacency matrix maths

Form a given name-family name coincidence matrix, $\mathbf{A} = [w_{gf}]$, where w_{gf} is strength of association between g-name g and f-name f

Then, $\mathbf{B} = \mathbf{A}^{\mathbf{T}}\mathbf{A}$ is the adjacency matrix of a family name-only network, and

From two-mode to one-mode networks

We can replace all the back and forth database querying with some matrix arithmetic



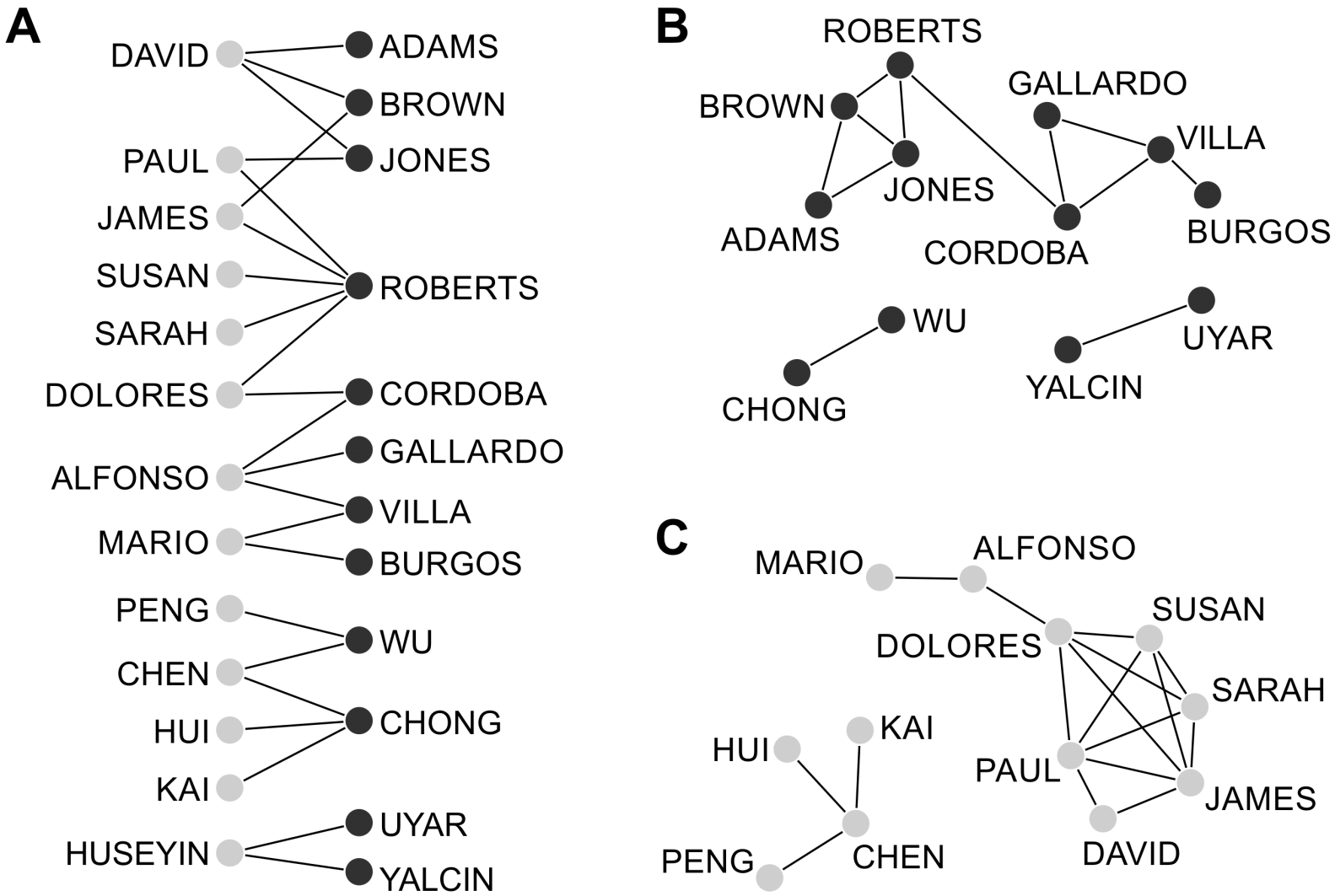
Graph adjacency matrix maths

Form a given name-family name coincidence matrix, $\mathbf{A} = [w_{gf}]$, where w_{gf} is strength of association between g-name g and f-name f

Then, $\mathbf{B} = \mathbf{A}^{\mathbf{T}}\mathbf{A}$ is the adjacency matrix of a family name-only network, and

$\mathbf{C} = \mathbf{A}\mathbf{A}^{\mathbf{T}}$ is the adjacency matrix of a given name-only network

From two-mode to one-mode networks



Salience matters more than frequency

A key issue is how to determine the w_{gf} strengths (weights) of association

- Common names form majority of given name $\Leftrightarrow$ family name links
- But *interesting* given name links are those that are common with respect to a family name relative to overall prevalence
- After some experimentation, we arrived at

$$w_{gf} \propto \frac{n_{gf}}{\sqrt{n_g \left(n_g - 1\right)}}$$

where n_{gf} is the number of times a particular combination occurs, and n_g is the total number of occurrences (across all family names) of that given name

Salience matters more than frequency

A key issue is how to determine the w_{gf} strengths (weights) of association

- Common names form majority of given name $\Leftrightarrow$ family name links
- But *interesting* given name links are those that are common with respect to a family name relative to overall prevalence
- After some experimentation, we arrived at

$$w_{gf} \propto \frac{n_{gf}}{\sqrt{n_g \left(n_g - 1\right)}}$$

where n_{gf} is the number of times a particular combination occurs, and n_g is the total number of occurrences (across all family names) of that given name

Salience matters more than frequency

A key issue is how to determine the w_{gf} strengths (weights) of association

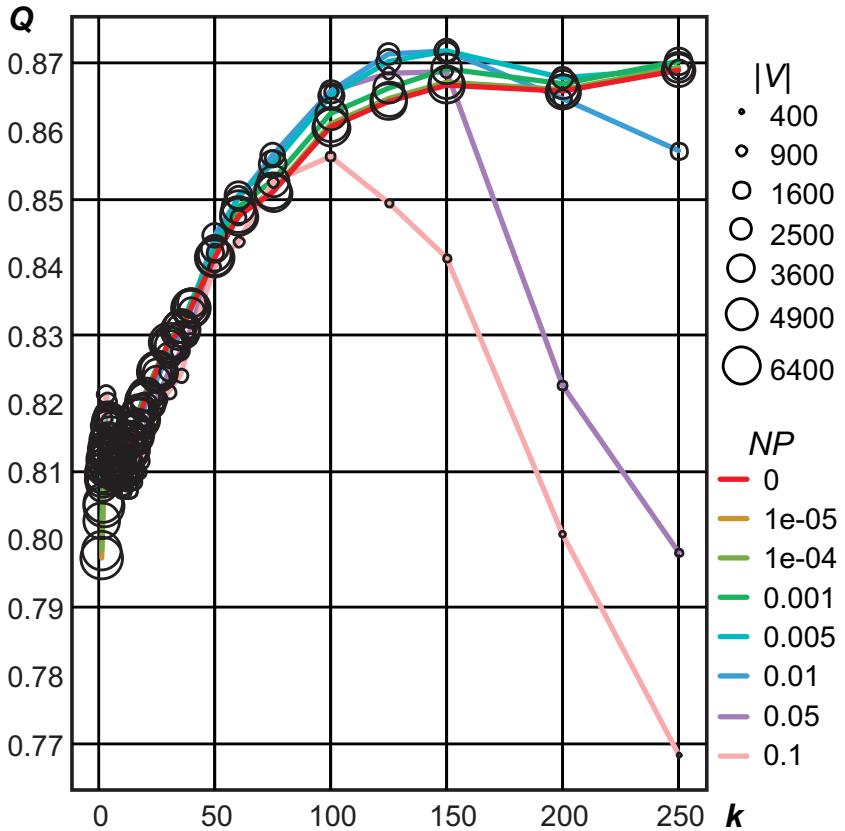
- Common names form majority of given name $\Leftrightarrow$ family name links
- But *interesting* given name links are those that are common with respect to a family name relative to overall prevalence
- After some experimentation, we arrived at

$$w_{gf} \propto \frac{n_{gf}}{\sqrt{n_g \left(n_g - 1\right)}}$$

where n_{gf} is the number of times a particular combination occurs, and n_g is the total number of occurrences (across all family names) of that given name

Salience matters more than frequency

- 'Raw' networks extremely densely connected
- Filter by removing links between family names weaker than some NP threshold
- Also where w_{gf} is less than some factor k above expectation
- In practice, the quality (modularity Q) of subsequent graph community detection depends on these filters



k is relative prevalence of a given-name vs expectation
Q is modularity, a measure of how 'clean' the clusters are

Datasets

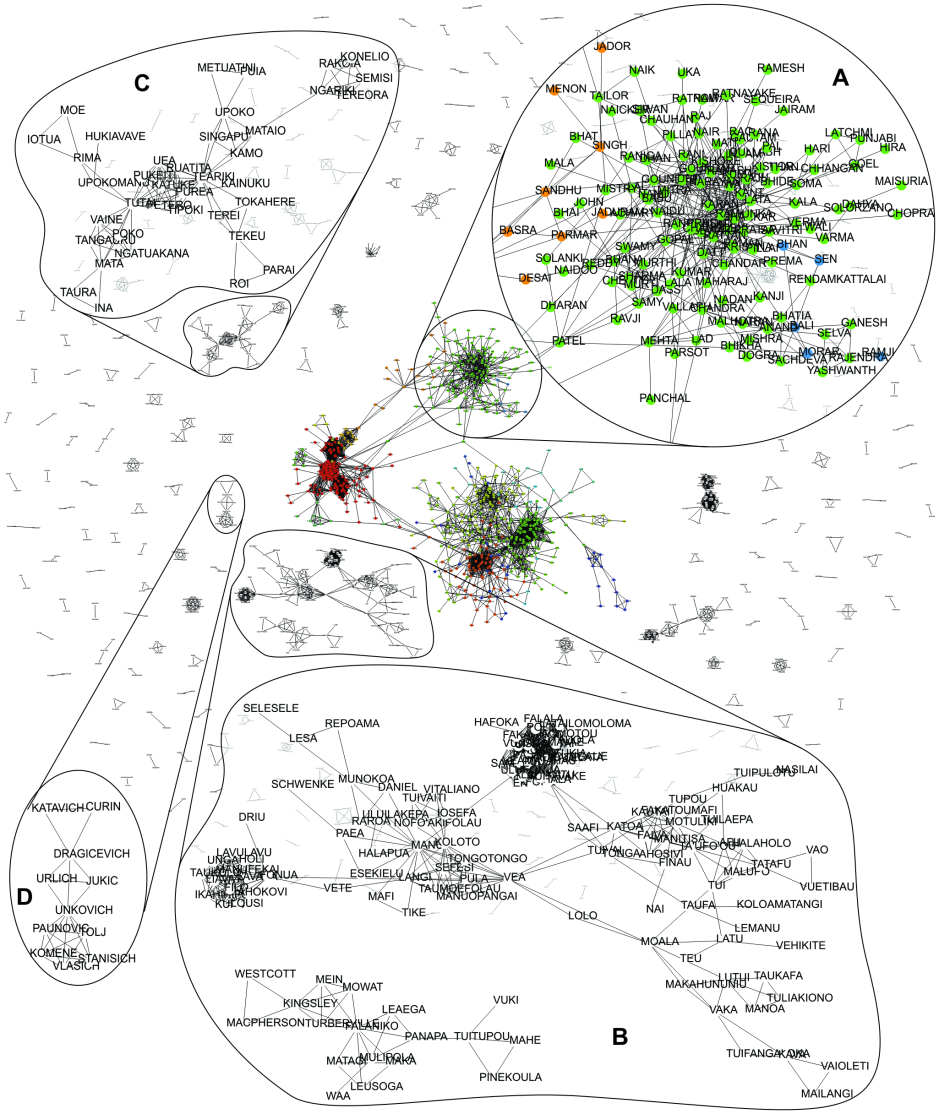
Auckland, New Zealand

- 912,858 registered voters from the 2008 Electoral Roll
- 79,855 unique f-names
- 88,760 unique g-names
- 711,807 unique p-f pairs
- Small yet ethnically and culturally diverse
- Many names somewhat unique to New Zealand
- A good testbed for methods

'Diagnostic' synthetic dataset

- 30,479 surnames tagged as one of 40 cultural, ethnic and linguistic (CEL) groups
- tagged names used to retrieve a subset of names from a 300 million person database
- 118.3 million persons from 17 countries
- 4.6 million unique f-names
- 1.8 million unique g-names
- 46.3 million unique g-f pairs

The Auckland names network

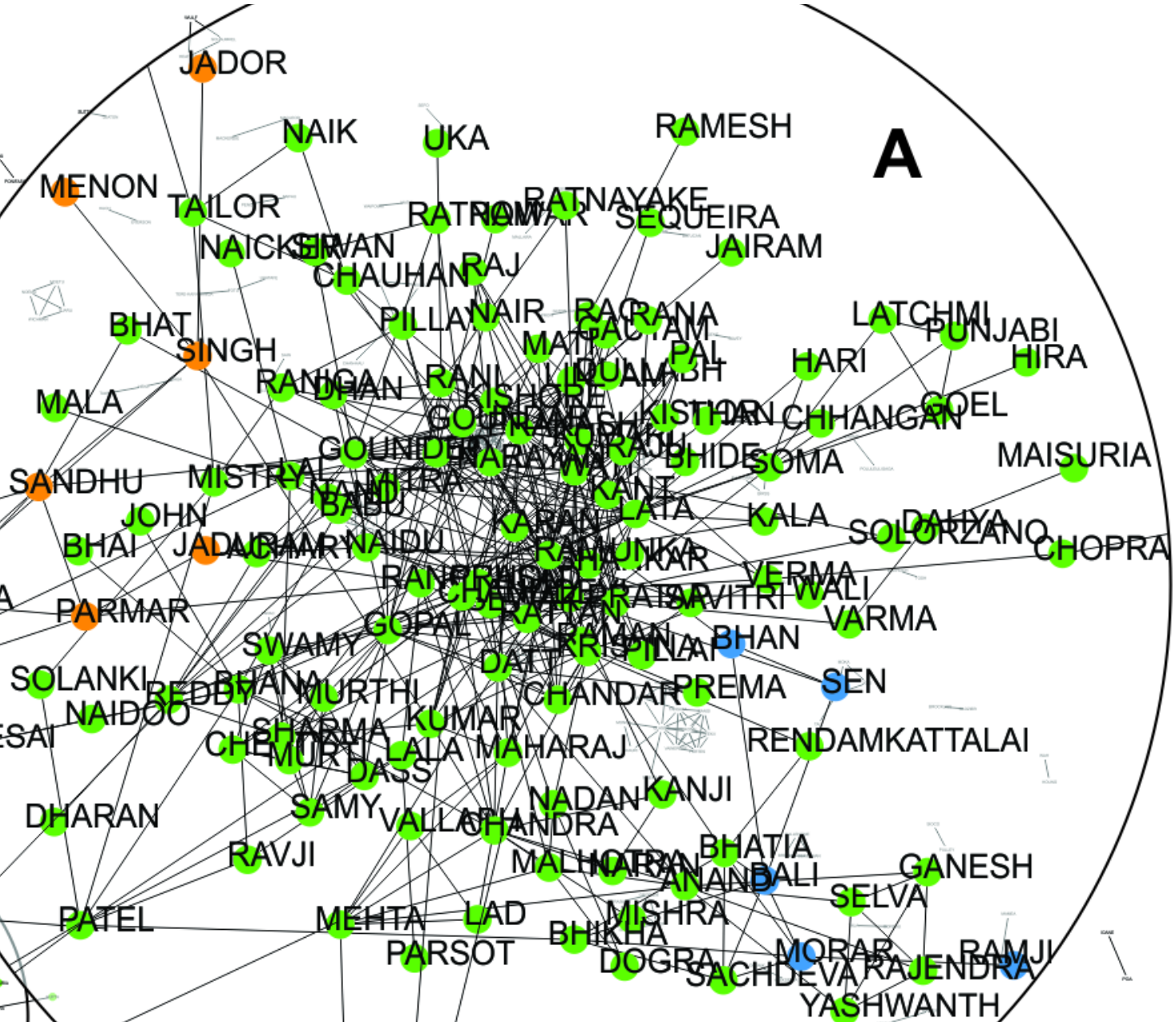


The filtered Auckland network

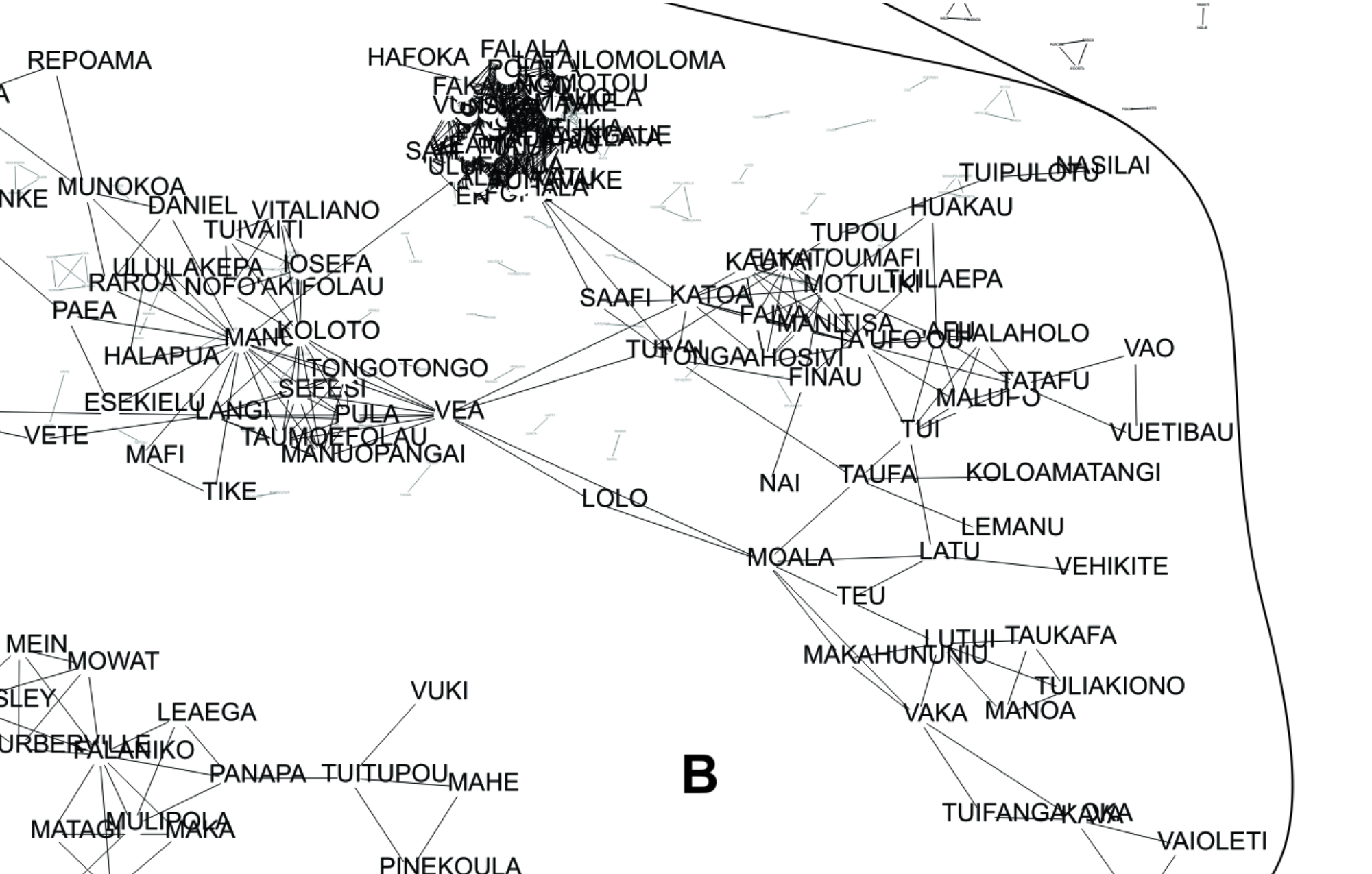
- A - South Asian names in 3 or 4 clusters
- B - Samoan / Tongan / Pacific names
- C - Māori names
- D - Eastern European names

Previously untagged names grouped separately from those already tagged. Colours are clusters found by community detection method of Clauset, Newman and Moore

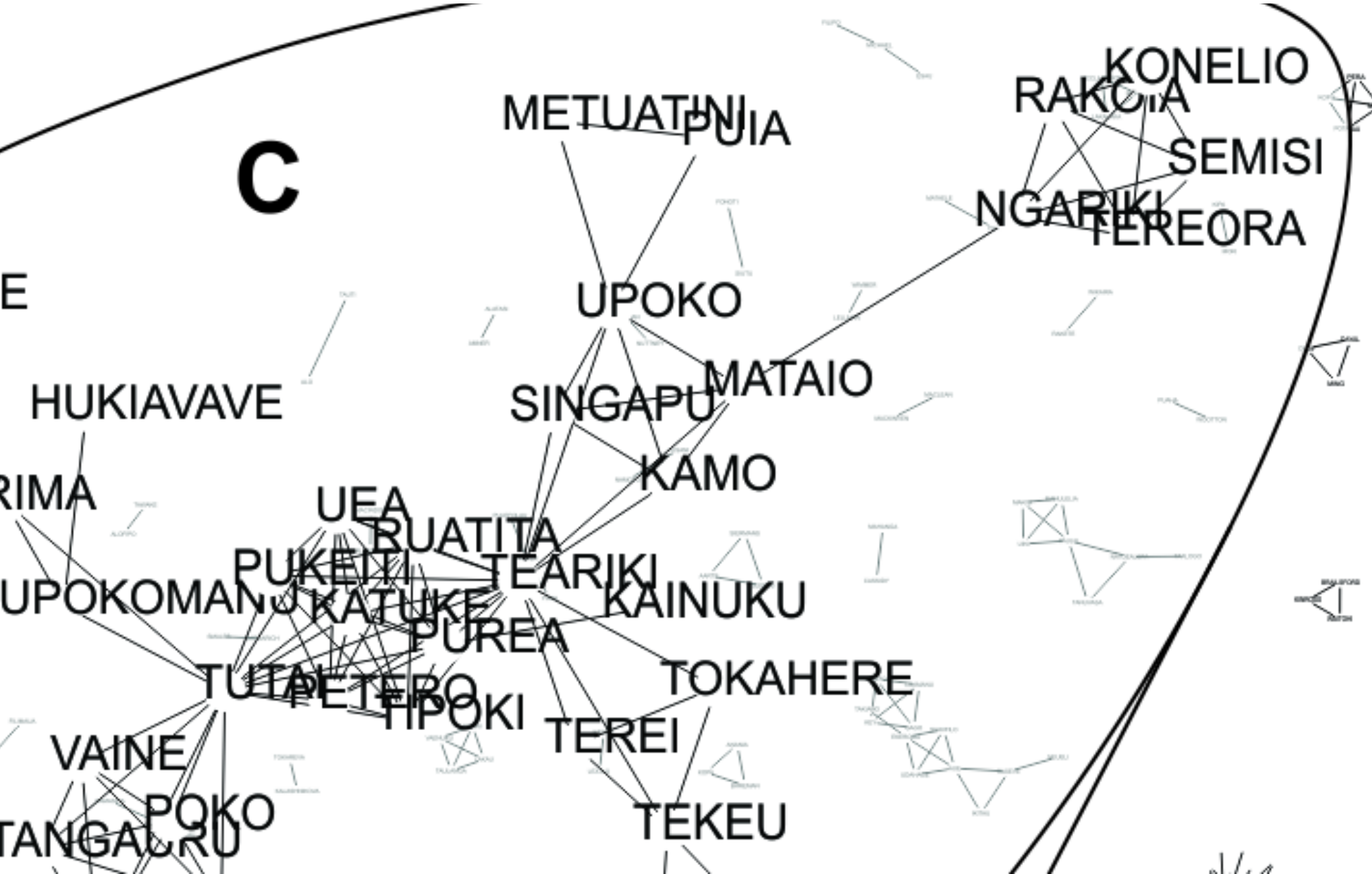
The Auckland names network



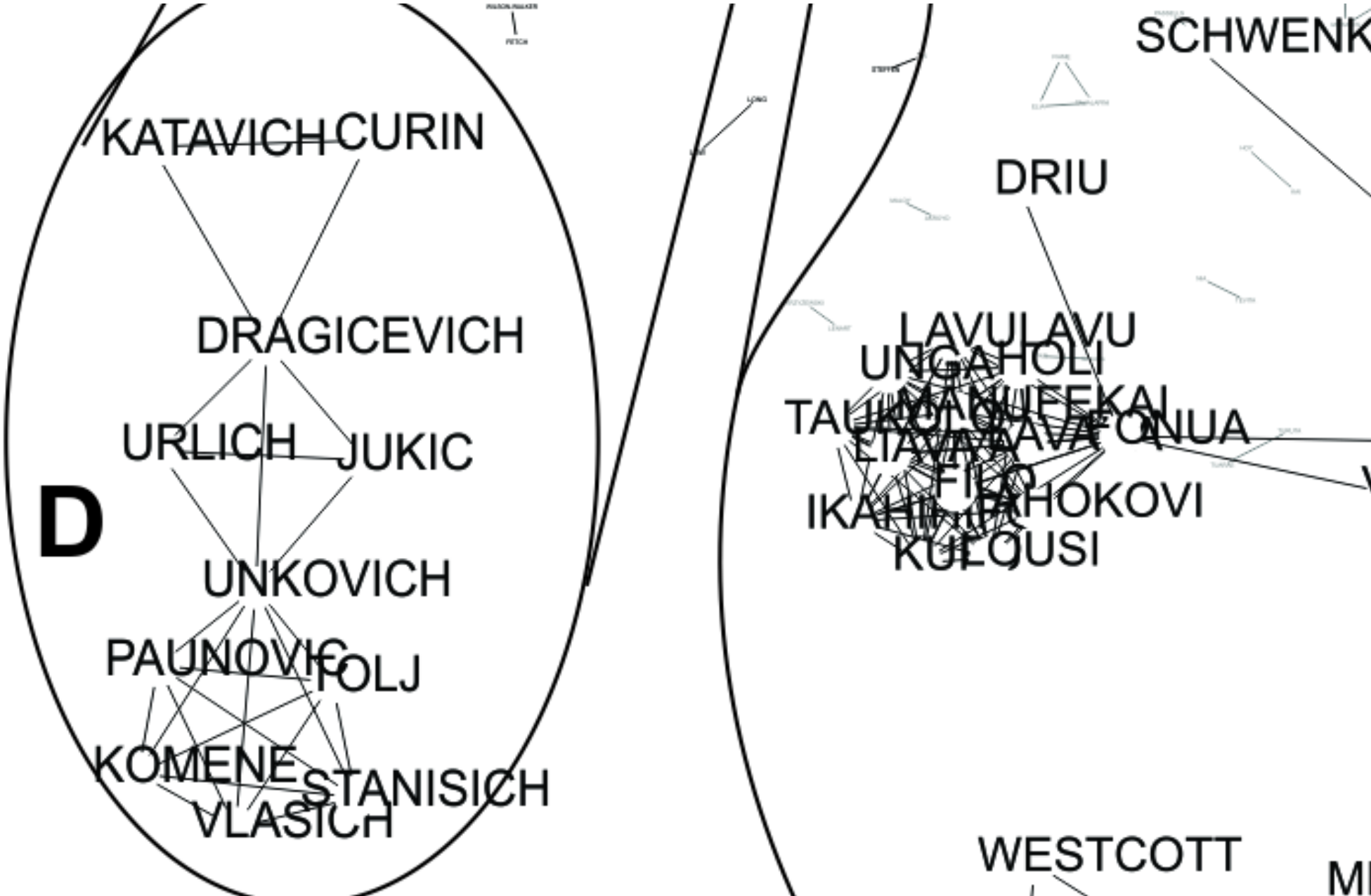
The Auckland names network



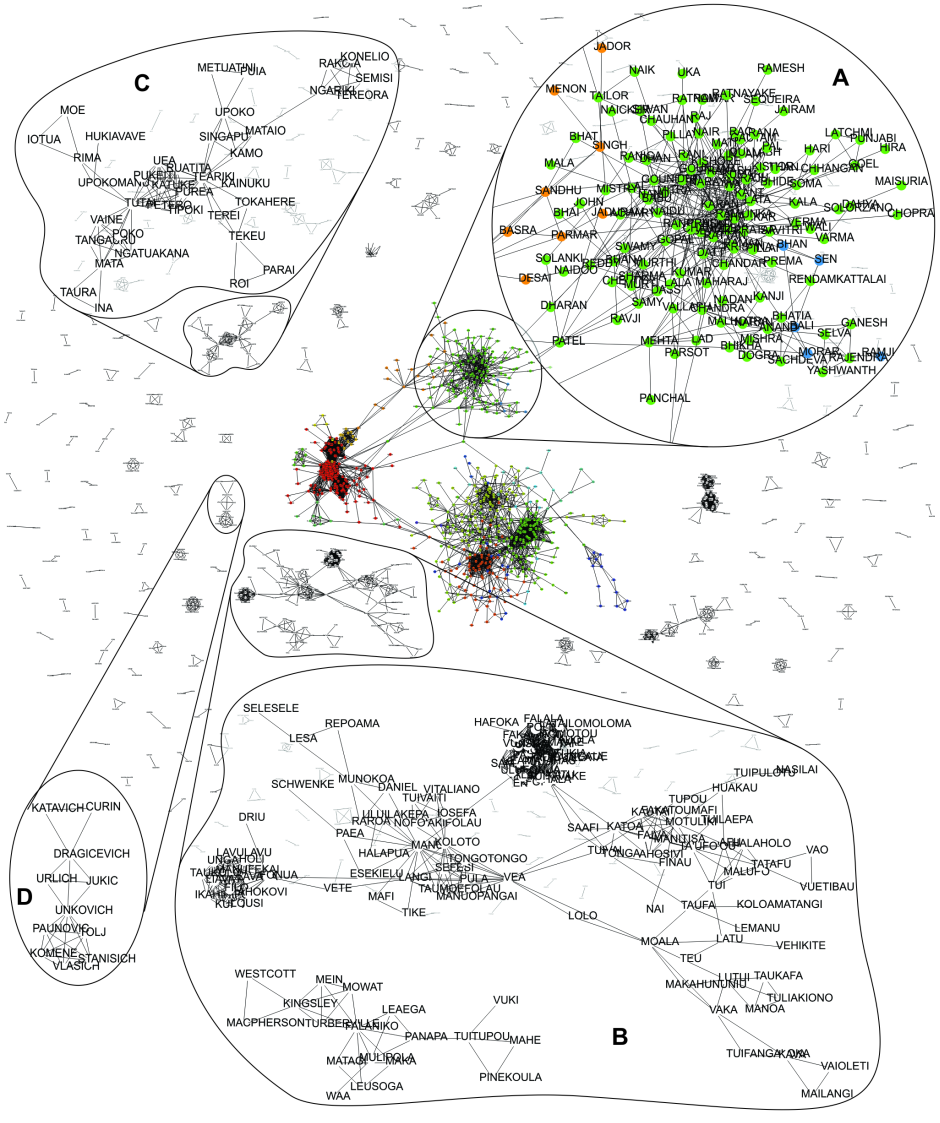
The Auckland names network



The Auckland names network



The Auckland names network

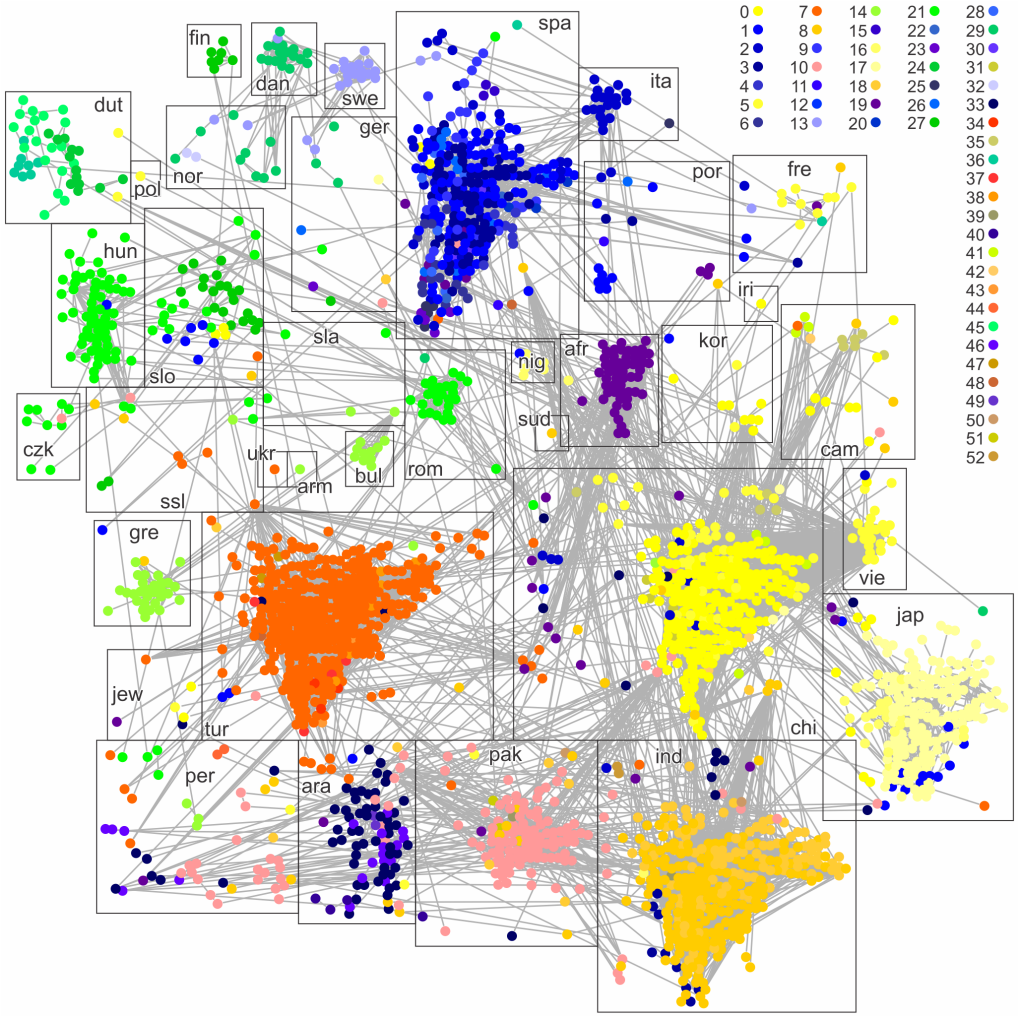


The filtered Auckland network

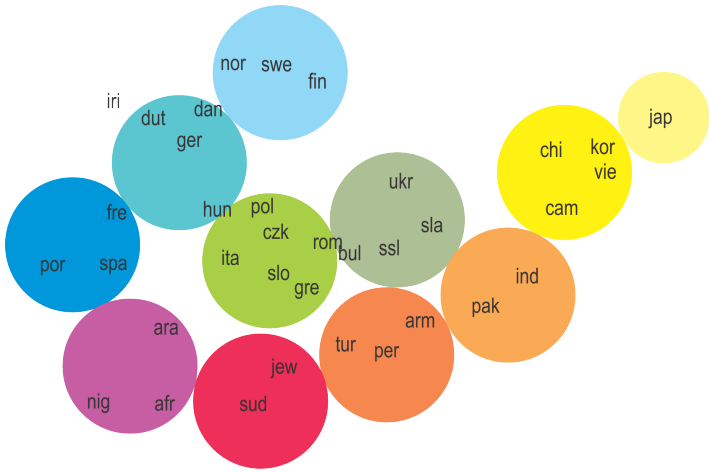
- A - South Asian names in 3 or 4 clusters
- B - Samoan / Tongan / Pacific names
- C - Māori names
- D - Eastern European names

Previously untagged names grouped separately from those already tagged. Colours are clusters found by community detection method of Clauset, Newman and Moore

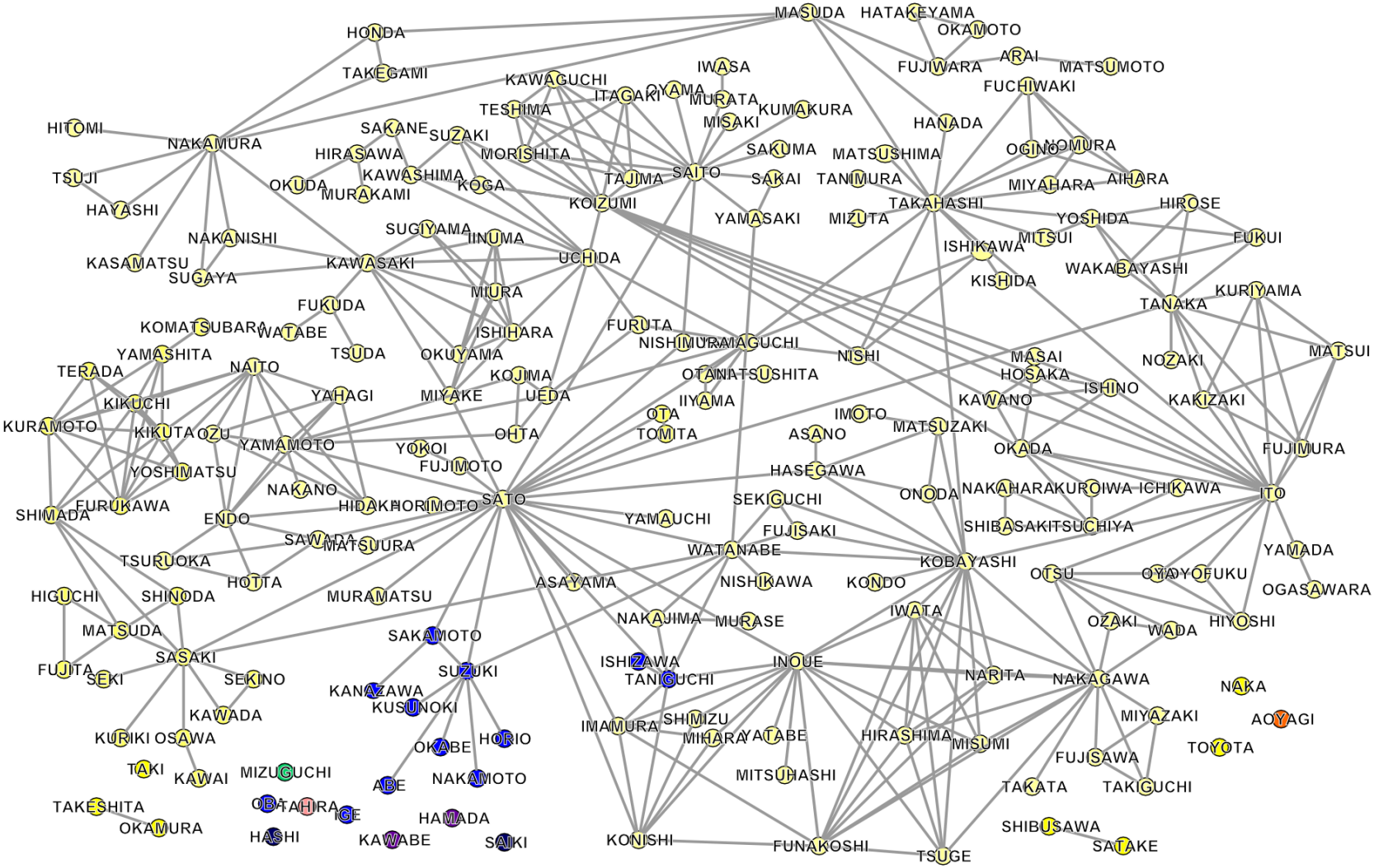
‘Diagnostic names’ network



- Colours reflect assignment to different ‘communities’ by the community detection algorithm
- Boxed groupings are expert-determined CEL associations



'Japanese' names network



Much remains to be done...

Technically:

- Investigate methods for automatically 'cleaning' the names
- Perhaps automate determination of parameters to partition networks cleanly...
- ...or better, find a way to visualize all the detail
- Look into 'label propagation' methods that might allow inference of group associations for new or unknown names

Much remains to be done...

Technically:

- Investigate methods for automatically 'cleaning' the names
- Perhaps automate determination of parameters to partition networks cleanly...
- ...or better, find a way to visualize all the detail
- Look into 'label propagation' methods that might allow inference of group associations for new or unknown names

Much remains to be done...

Technically:

- Investigate methods for automatically 'cleaning' the names
- Perhaps automate determination of parameters to partition networks cleanly...
- ...or better, find a way to visualize all the detail
- Look into 'label propagation' methods that might allow inference of group associations for new or unknown names

Much remains to be done...

Technically:

- Investigate methods for automatically 'cleaning' the names
- Perhaps automate determination of parameters to partition networks cleanly...
- ...or better, find a way to visualize all the detail
- Look into 'label propagation' methods that might allow inference of group associations for new or unknown names

Much remains to be done...

More substantively:

- Explore degree to which apparent structure within groups reflects real intra-group relationships
- Consider how to spatially embed the network
  - Links between places based on names
  - Links between names based on places
- Consider how historical evolution of the network is spatially structured—perhaps a way to unpack details of migration patterns

Much remains to be done...

More substantively:

- Explore degree to which apparent structure within groups reflects real intra-group relationships
- Consider how to spatially embed the network
  - Links between places based on names
  - Links between names based on places
- Consider how historical evolution of the network is spatially structured—perhaps a way to unpack details of migration patterns

Much remains to be done...

More substantively:

- Explore degree to which apparent structure within groups reflects real intra-group relationships
- Consider how to spatially embed the network
  - Links between places based on names
  - Links between names based on places
- Consider how historical evolution of the network is spatially structured—perhaps a way to unpack details of migration patterns

Much remains to be done...

More substantively:

- Explore degree to which apparent structure within groups reflects real intra-group relationships
- Consider how to spatially embed the network
  - Links between places based on names
  - Links between names based on places
- Consider how historical evolution of the network is spatially structured—perhaps a way to unpack details of migration patterns

Much remains to be done...

More substantively:

- Explore degree to which apparent structure within groups reflects real intra-group relationships
- Consider how to spatially embed the network
  - Links between places based on names
  - Links between names based on places
- Consider how historical evolution of the network is spatially structured—perhaps a way to unpack details of migration patterns

Conclusions

- Names appear a promising approach for looking at community structures and inter-relationships
- They perhaps enable a more nuanced view of CEL affiliation than pre-defined 'tick-box' affiliations
- How naming networks have changed (or not) over time may prove a particularly interesting direction to look

Conclusions

- Names appear a promising approach for looking at community structures and inter-relationships
- They perhaps enable a more nuanced view of CEL affiliation than pre-defined 'tick-box' affiliations
- How naming networks have changed (or not) over time may prove a particularly interesting direction to look

Conclusions

- Names appear a promising approach for looking at community structures and inter-relationships
- They perhaps enable a more nuanced view of CEL affiliation than pre-defined 'tick-box' affiliations
- How naming networks have changed (or not) over time may prove a particularly interesting direction to look

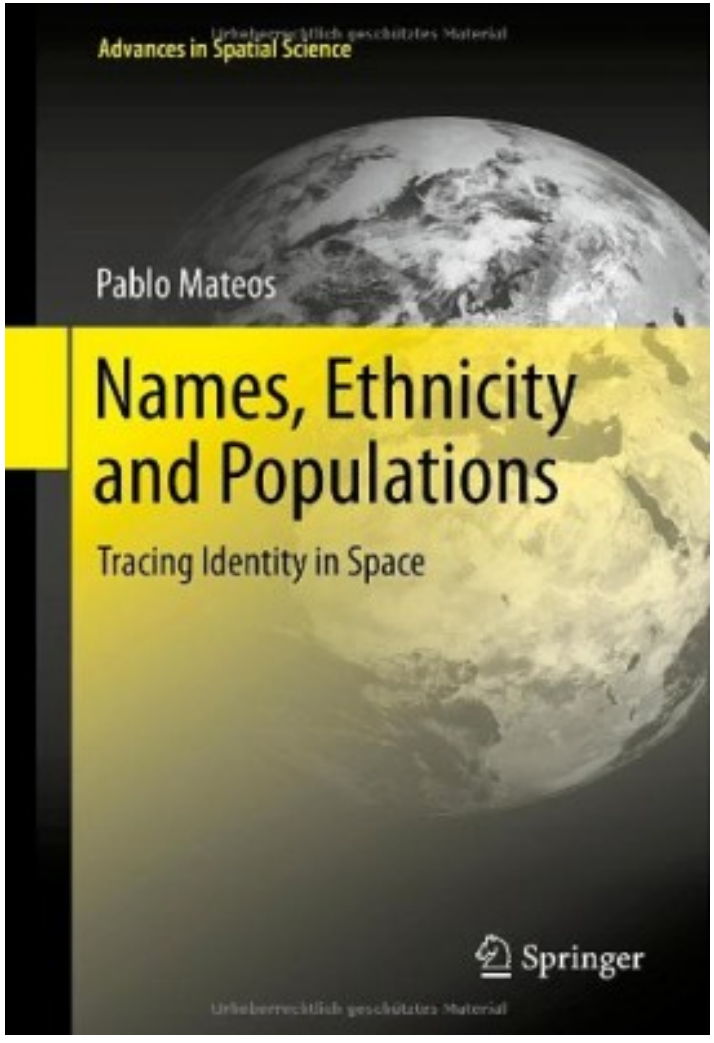
More information

Published paper

Mateos, P., P. A. Longley, and D. O'Sullivan. 2011. Ethnicity and Population Structure in Personal Naming Networks. *PLoS ONE*, **6**(9) e22943. DOI: 10.1371/journal.pone.0022943

Websites

http://www.onomap.org
http://worldnames.publicprofiler.org



Acknowledgments

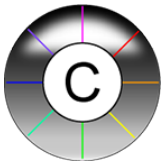
Pablo and Paul
UK Economic and Social Research Council (ESRC) grants RES-000-22-0400 Surnames as a quantitative evidence resource for the social sciences
RES-172-25-0019 Web-based dissemination of the geography of genealogy
RES-149-25-1078 The Genesis Project: GENerative E-Social Science
PTA-026-27-1521 The Geography and Ethnicity of People's Names

Pablo
Royal Society International Travel Grant TG092248
New Zealand Performance Based Research Fund 2009 (University of Auckland)

David
UK HEFCE Spatial Literacy in Teaching (SpLinT) Fellowship 2008
University of Auckland Study Leave grant-in-aid 2008
University of Tokyo Visiting Position 2012

Free software and tools
Python `numpy`, `scipy`, `igraph`

Graph visualization tools
Cytoscape (from biosciences)



yEd (for graph drawing)