Naming networks and population structure

David O'Sullivan[1] Pablo Mateos[2]

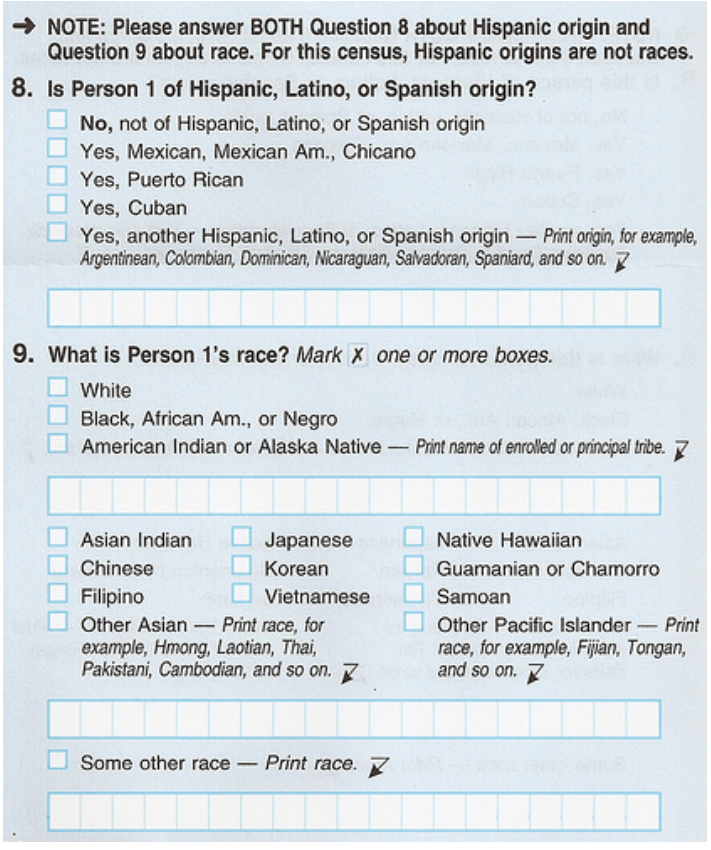
[1]School of Environment
University of Auckland
Te Whare Wānanga o Tāmaki Makaurau

[2]Department of Geography
University College London

Annual Meeting
Association of American Geographers
2011

Limits to more conventional 'identity' data

- Ethnicity (or culture or identity) is a tricky and 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'



➜ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ **No,** not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark ✗ one or more boxes.*

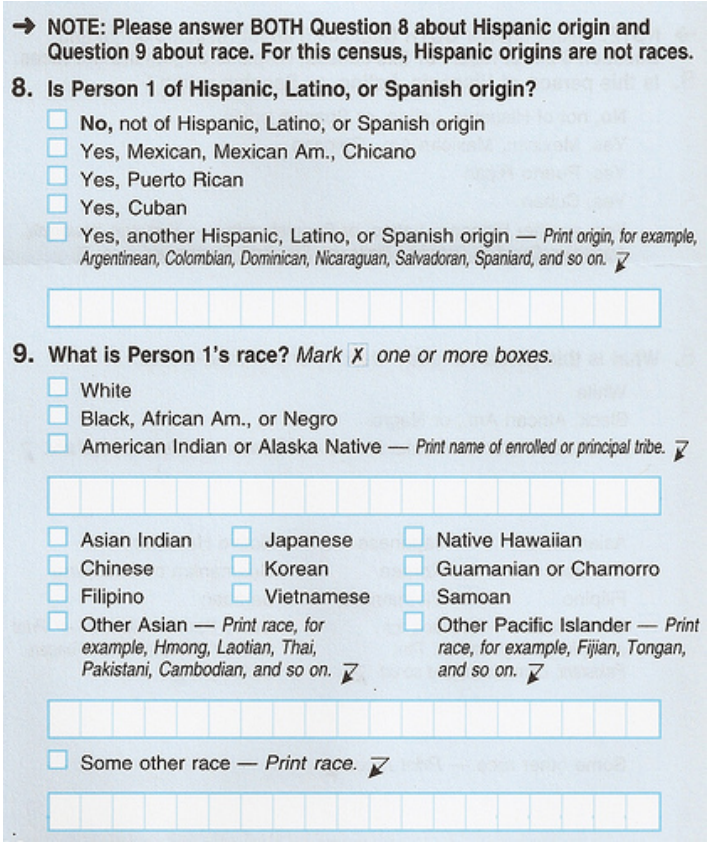
- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

- ☐ Some other race — *Print race.*

Limits to more conventional 'identity' data

- Ethnicity (or culture or identity) is a tricky and 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'



➜ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ **No,** not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark ✗ one or more boxes.*

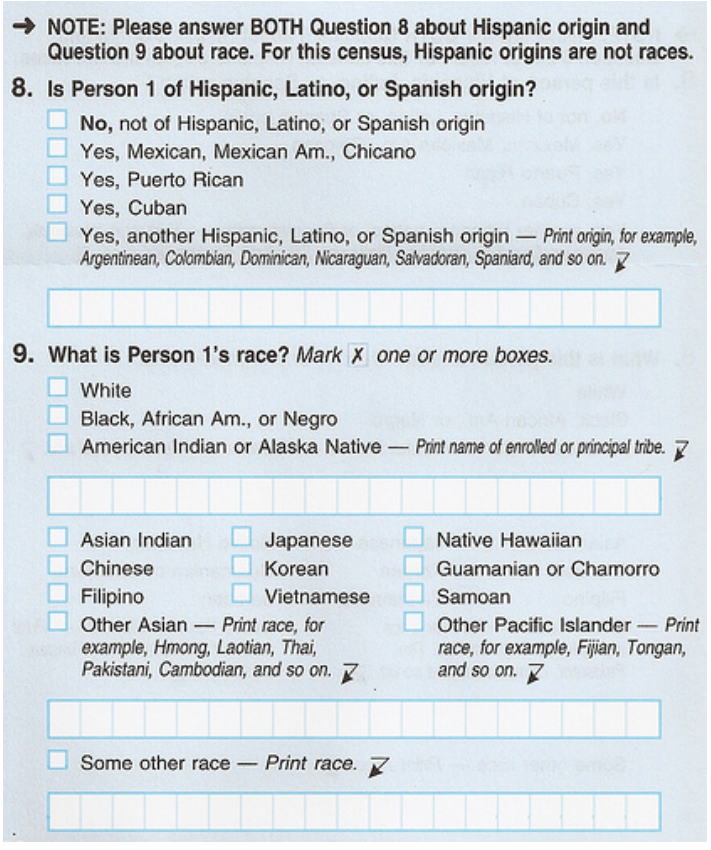
- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

- ☐ Some other race — *Print race.*

Limits to more conventional 'identity' data

- Ethnicity (or culture or identity) is a tricky and 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'



➜ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ **No,** not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark ✗ one or more boxes.*

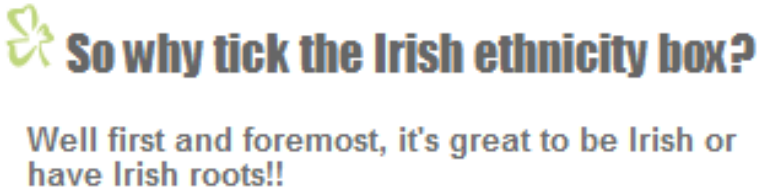
- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

- ☐ Some other race — *Print race.*

Limits to more conventional 'identity' data

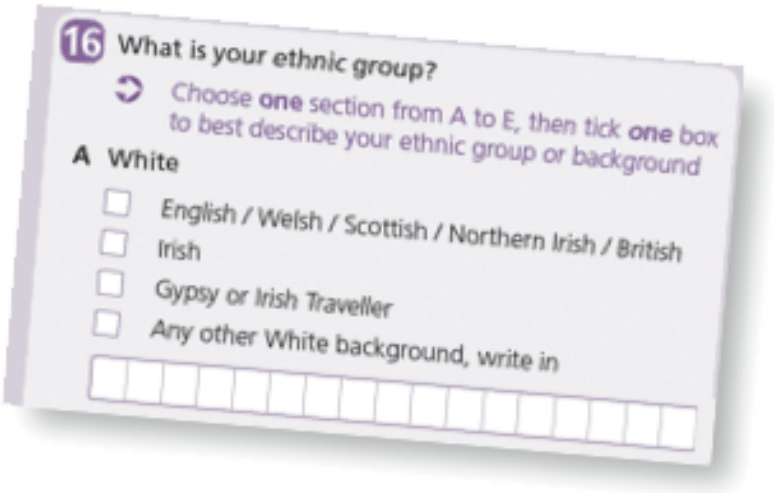
- Ethnicity (or culture or identity) is a tricky and multi-dimensional concept
- Standards for how to collect data don't match
- Conventional instruments don't capture it well
  - Census 'tick boxes'
  - Census 'write-ins'



Source: howirishareyou.com

Limits to more conventional 'identity' data

- Ethnicity (or culture or identity) is a tricky and 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: howirishareyou.com

Limits to more conventional 'identity' data

- Ethnicity (or culture or identity) is a tricky and 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.

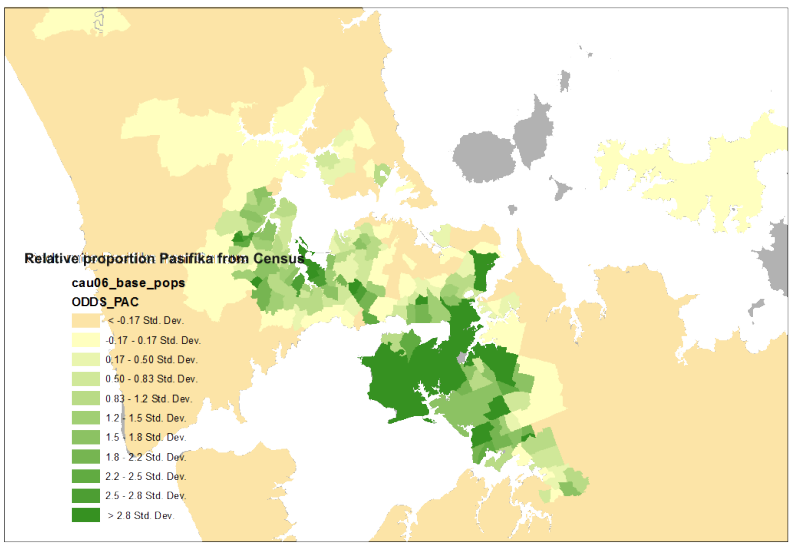
Names as markers of identity

- Minimally invasive
- Readily available
- Carry complex (perhaps limited) information

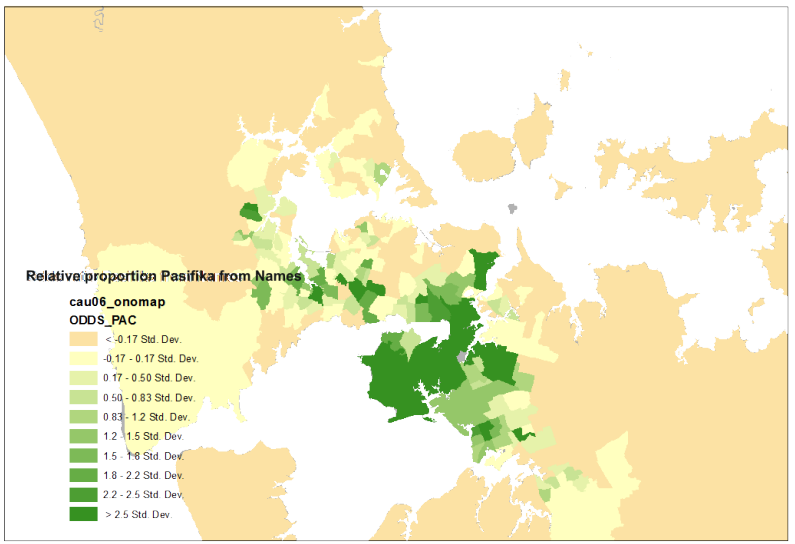
Previous work

- Mateos, P. (2007) A review of name-based ethnicity classification methods and their potential in population studies, *Population Space and Place*, 13(4): 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

Example names maps from New Zealand

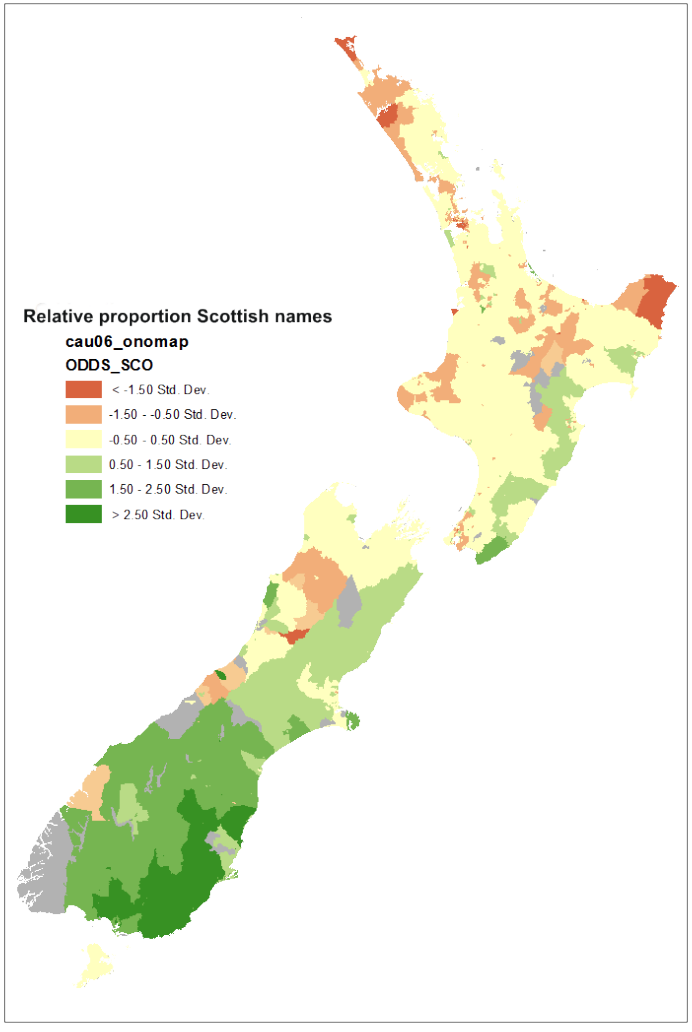


Pacific people south Auckland (census)



Pacific names

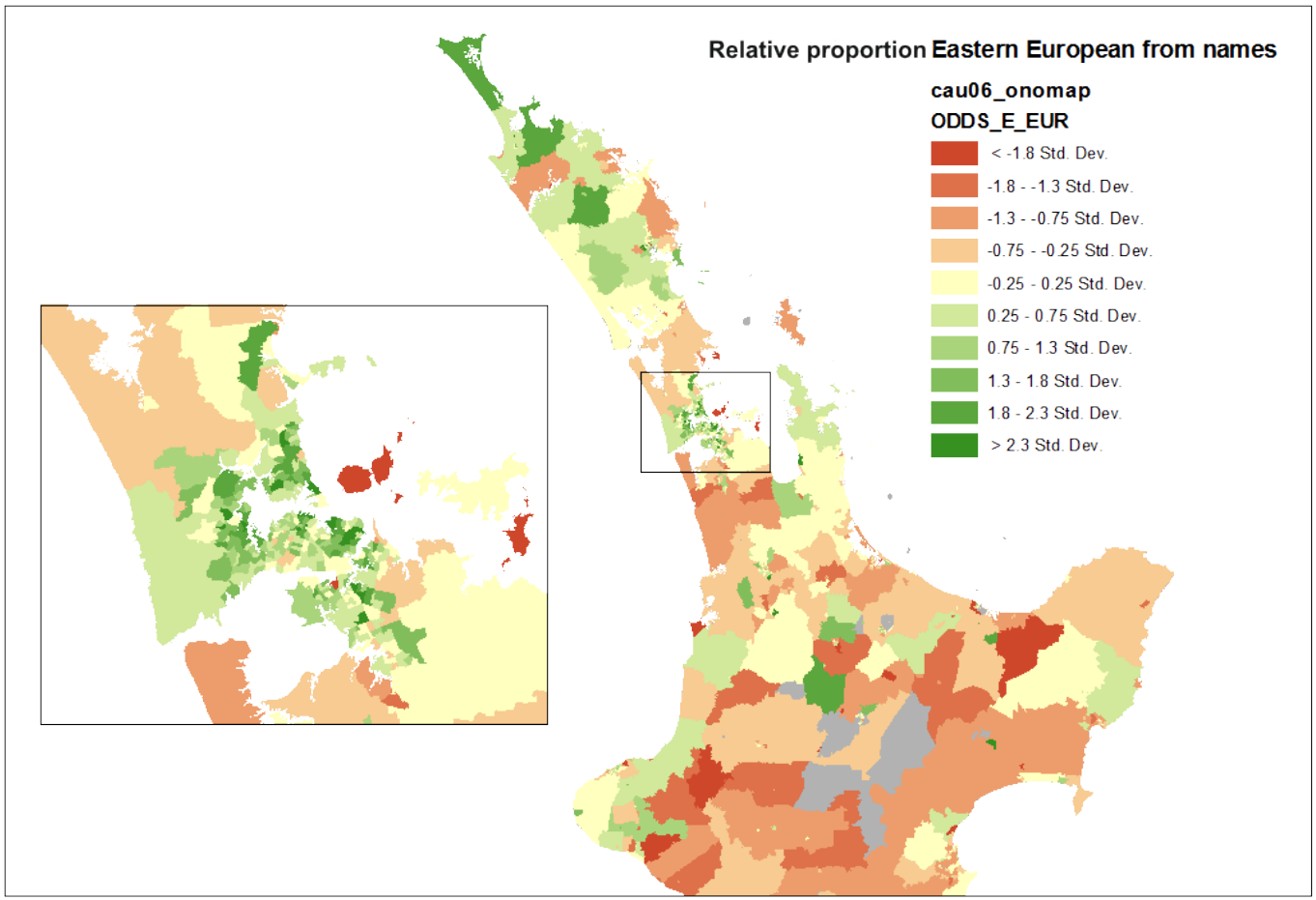
Example names maps from New Zealand



Scottish names countrywide

- No 'Scottish' New Zealand category in the 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)

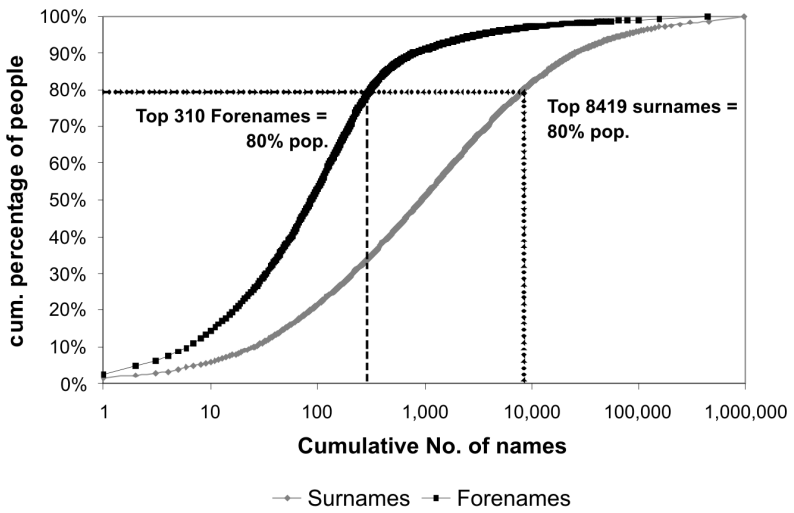
Example names maps from New Zealand



Eastern European names, west Auckland

Difficulties with names as data

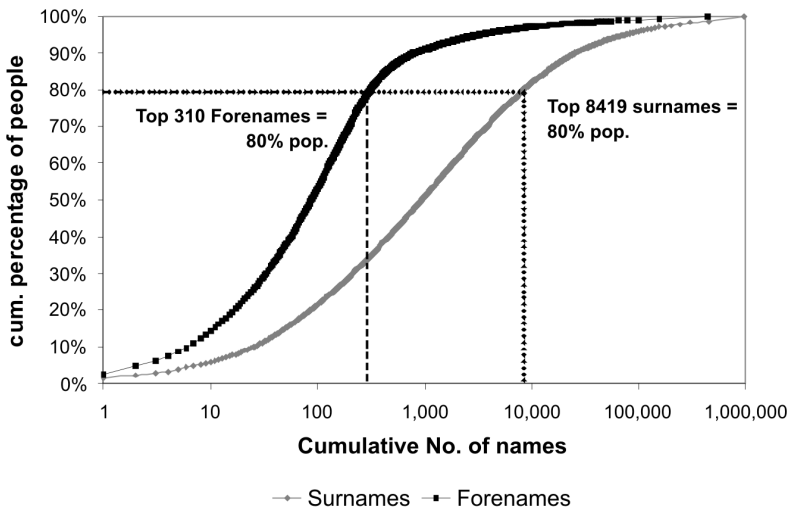
- *Very* unevenly distributed



- 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

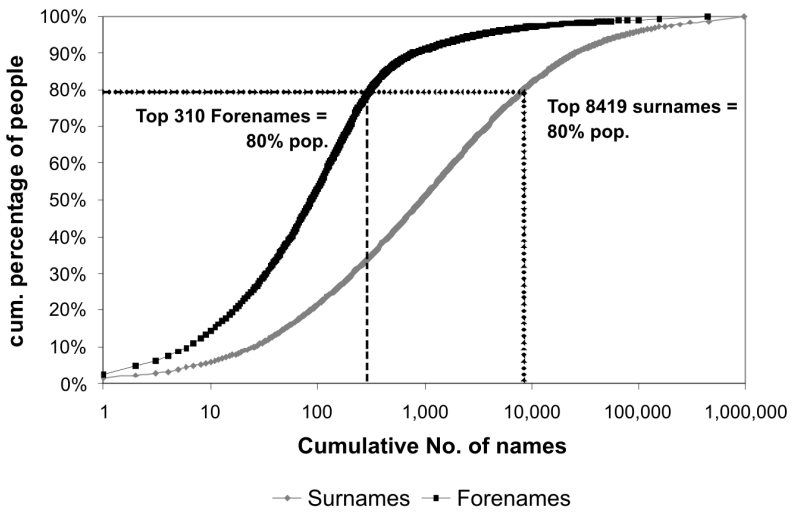
- *Very* unevenly distributed



- 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

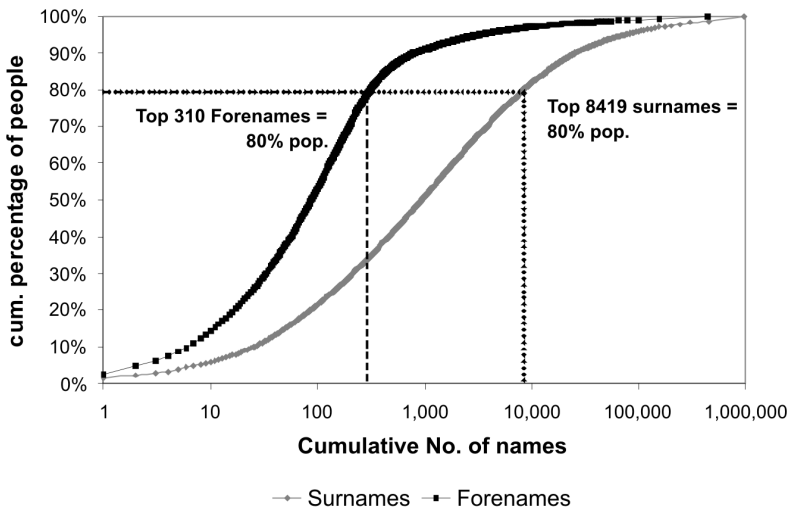
- *Very* unevenly distributed



- 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 from previous experience, esp. Māori, Pacific names
- Limited expert knowledge
- 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 from previous experience, esp. Māori, Pacific names
- Limited expert knowledge
- 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

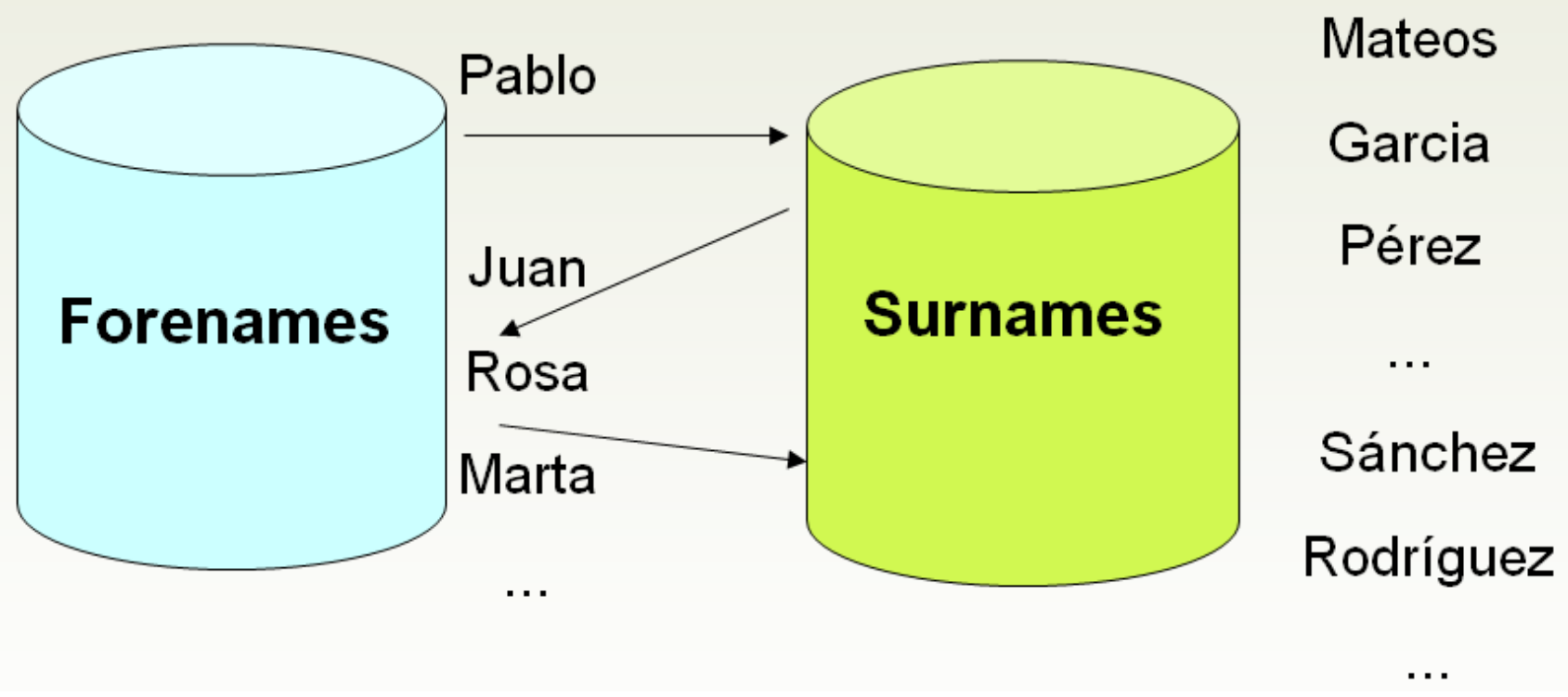
- A major problem dealing with names unknown from previous experience, esp. Māori, Pacific names
- Limited expert knowledge
- Is there a way to infer 'new' groups from the relationships between names?
- $\Rightarrow$ concept of a network of names

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

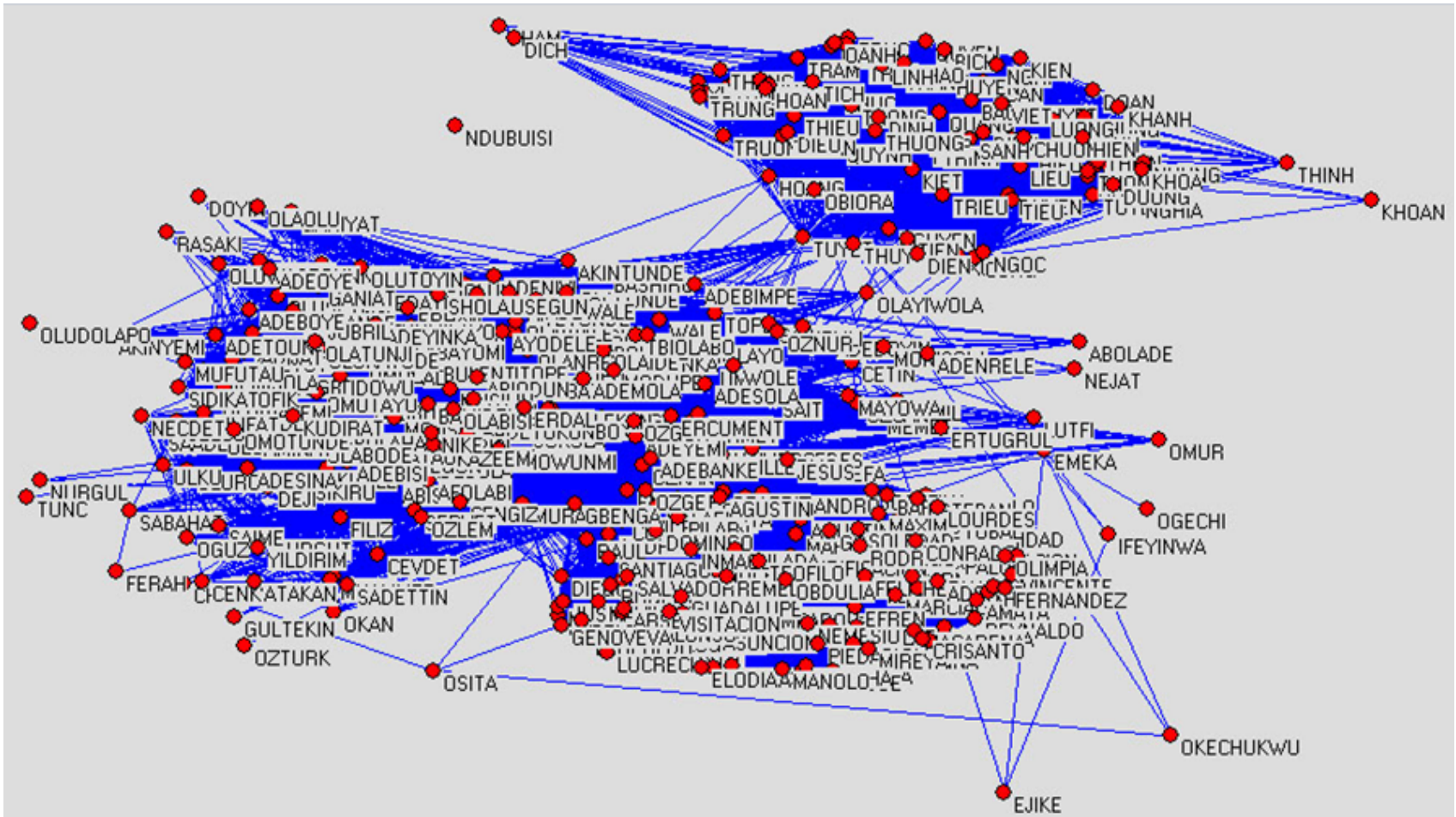
The New Zealand case presented some specific challenges

- A major problem dealing with names unknown from previous experience, esp. Māori, Pacific names
- Limited expert knowledge
- 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

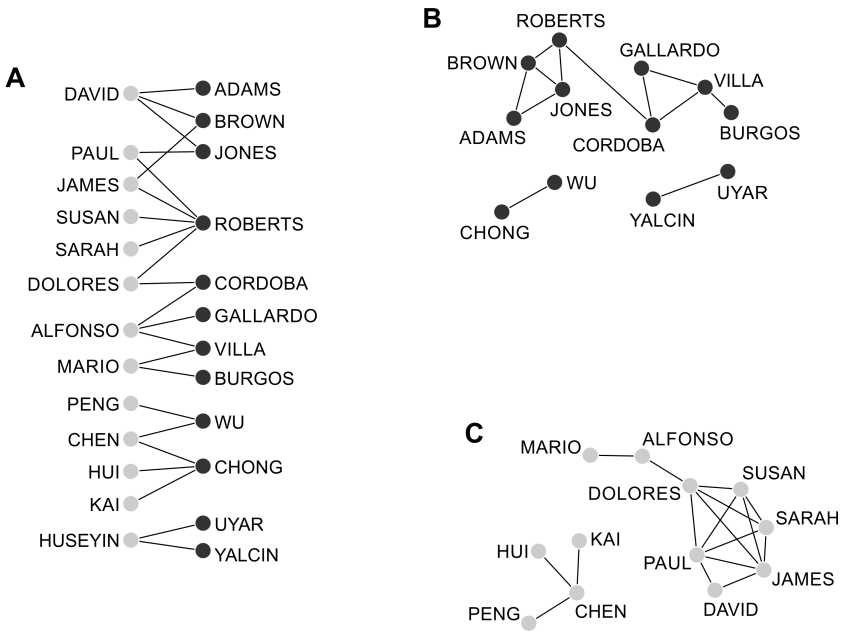


Relationships between names: from lists to networks



From two-mode to one-mode networks

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

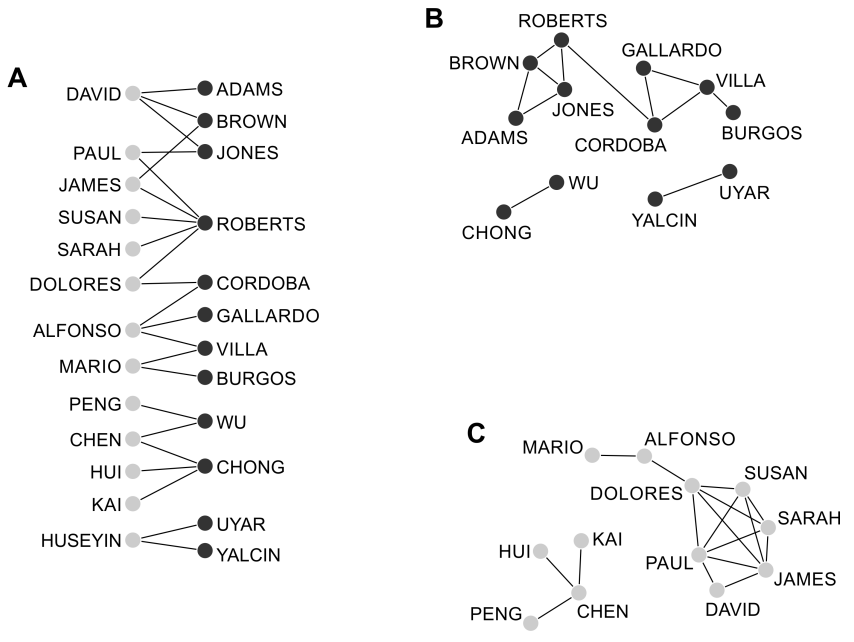


Graph adjacency matrix maths

Form a forename-surname coincidence matrix, $\mathbf{A} = [w_{fs}]$, where w_{fs} is the weight of association between forename f and surname s

From two-mode to one-mode networks

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



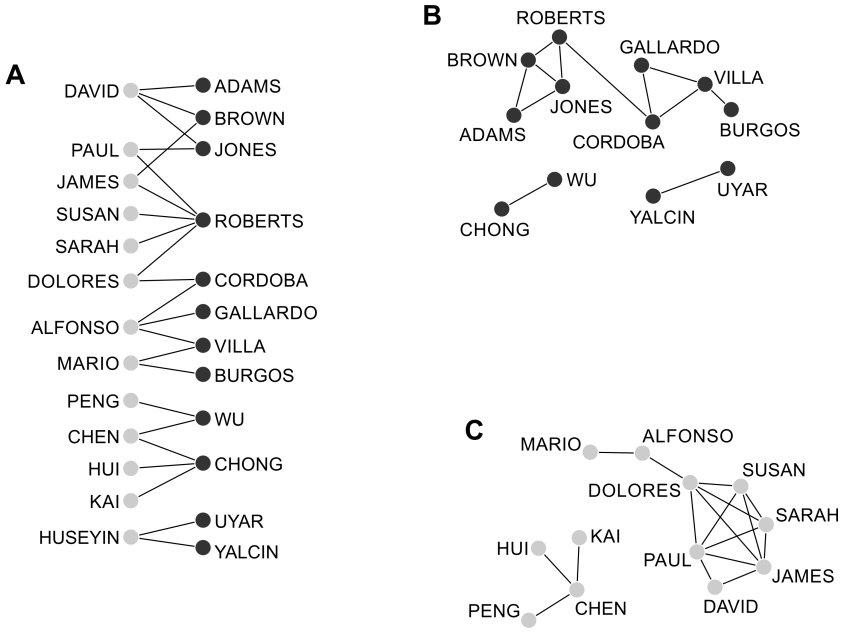
Graph adjacency matrix maths

Form a forename-surname coincidence matrix, $\mathbf{A} = [w_{fs}]$, where w_{fs} is the weight of association between forename f and surname s

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

From two-mode to one-mode networks

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



Graph adjacency matrix maths

Form a forename-surname coincidence matrix, $\mathbf{A} = [w_{fs}]$, where w_{fs} is the weight of association between forename f and surname s

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

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

Salience matters more than frequency

A key issue is how to determine the w_{fs} weights

- Common names form the majority of forename↔surname links
- But interesting forename-links are those that are common with respect to a surname relative to overall prevalence
- After some experimentation, we arrived at

$$w_{fs} \propto \frac{n_{fs}}{\sqrt{n_f \left(n_f - 1\right)}}$$

where n_{fs} is the number of times a particular forename-surname combination occurs, and n_f is the total number of occurrences (across all surnames) of that forename

- The 'raw' networks produced by this process are extremely densely connected, so it is necessary to filter them by removing links between surnames that are weaker than some chosen threshold.

Salience matters more than frequency

A key issue is how to determine the w_{fs} weights

- Common names form the majority of forename↔surname links
- But interesting forename-links are those that are common with respect to a surname relative to overall prevalence
- After some experimentation, we arrived at

$$w_{fs} \propto \frac{n_{fs}}{\sqrt{n_f\left(n_f - 1\right)}}$$

where n_{fs} is the number of times a particular forename-surname combination occurs, and n_f is the total number of occurrences (across all surnames) of that forename

- The 'raw' networks produced by this process are extremely densely connected, so it is necessary to filter them by removing links between surnames that are weaker than some chosen threshold.

Salience matters more than frequency

A key issue is how to determine the w_{fs} weights

- Common names form the majority of forename↔surname links
- But interesting forename-links are those that are common with respect to a surname relative to overall prevalence
- After some experimentation, we arrived at

$$w_{fs} \propto \frac{n_{fs}}{\sqrt{n_f \left(n_f - 1\right)}}$$

where n_{fs} is the number of times a particular forename-surname combination occurs, and n_f is the total number of occurrences (across all surnames) of that forename

- The 'raw' networks produced by this process are extremely densely connected, so it is necessary to filter them by removing links between surnames that are weaker than some chosen threshold.

Salience matters more than frequency

A key issue is how to determine the w_{fs} weights

- Common names form the majority of forename↔surname links
- But interesting forename-links are those that are common with respect to a surname relative to overall prevalence
- After some experimentation, we arrived at

$$w_{fs} \propto \frac{n_{fs}}{\sqrt{n_f \left(n_f - 1\right)}}$$

where n_{fs} is the number of times a particular forename-surname combination occurs, and n_f is the total number of occurrences (across all surnames) of that forename

- The 'raw' networks produced by this process are extremely densely connected, so it is necessary to filter them by removing links between surnames that are weaker than some chosen threshold.

Tools of the trade

Recent developments in free software and tools for handling large networks have enormously benefitted this research:

- Python libraries scientific computing, `numpy` and `scipy`
- Python libraries for network analysis, `igraph` particularly network 'community detection'
- Graph visualization tools, Cytoscape (from biosciences) and yEd (for graph drawing)

Tools of the trade

Recent developments in free software and tools for handling large networks have enormously benefitted this research:

- Python libraries scientific computing, `numpy` and `scipy`
- Python libraries for network analysis, `igraph` particularly network 'community detection'
- Graph visualization tools, Cytoscape (from biosciences) and yEd (for graph drawing)

Tools of the trade

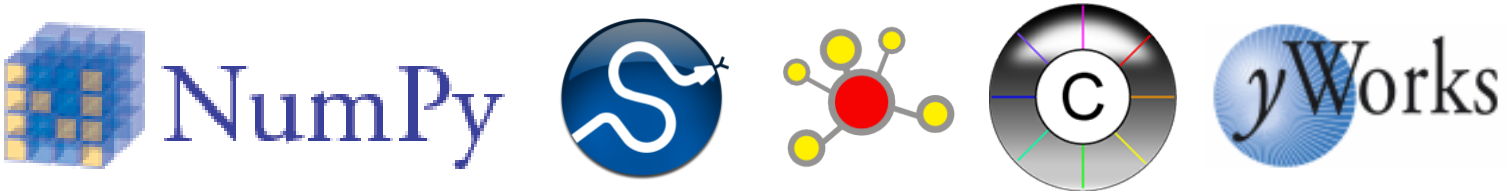
Recent developments in free software and tools for handling large networks have enormously benefitted this research:

- Python libraries scientific computing, `numpy` and `scipy`
- Python libraries for network analysis, `igraph` particularly network 'community detection'
- Graph visualization tools, Cytoscape (from biosciences) and yEd (for graph drawing)

Tools of the trade

Recent developments in free software and tools for handling large networks have enormously benefitted this research:

- Python libraries scientific computing, `numpy` and `scipy`
- Python libraries for network analysis, `igraph` particularly network 'community detection'
- Graph visualization tools, Cytoscape (from biosciences) and yEd (for graph drawing)



Datasets

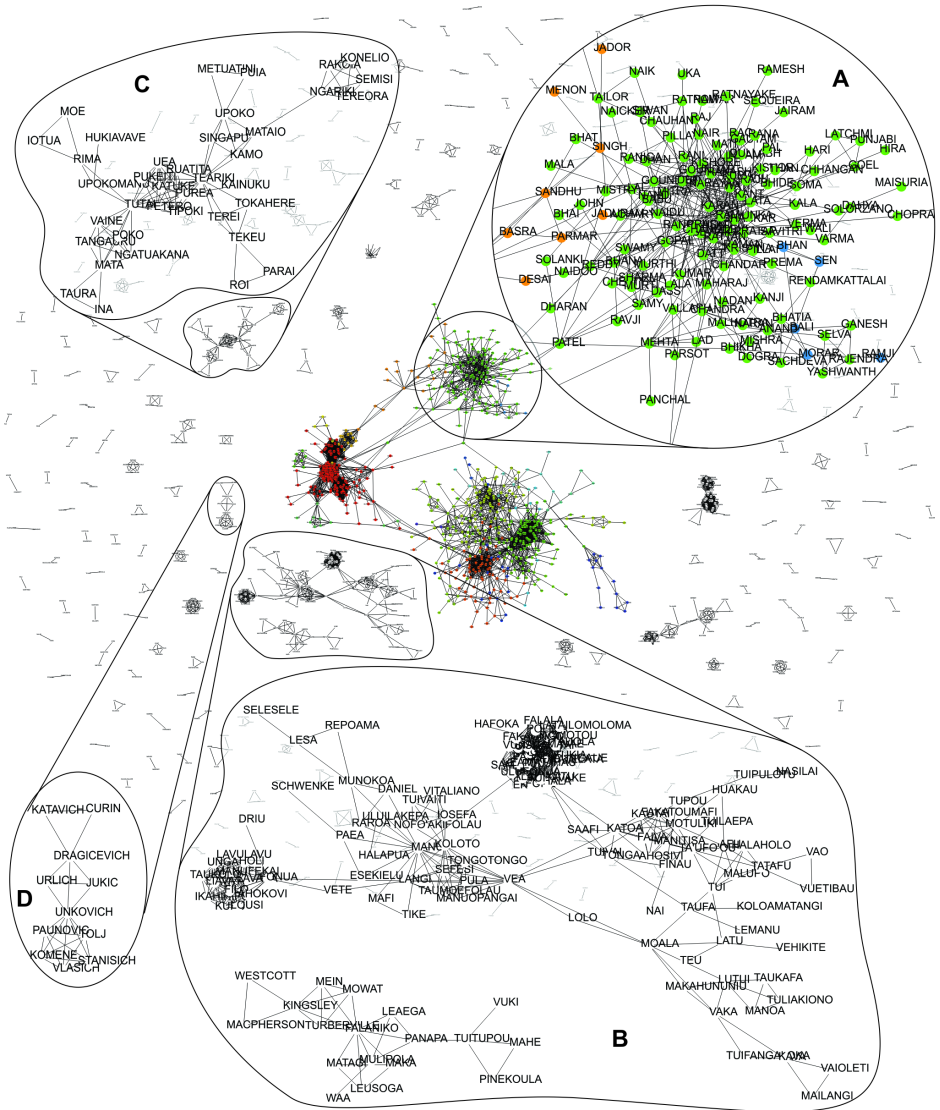
Auckland, New Zealand

- 912,858 registered voters from 2008 Electoral Roll
- 79,855 unique surnames
- 88,760 unique forenames
- 711,807 unique f-s pairs
- Small yet ethnically, 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 surnames
- 1.8 million unique forenames
- 46.3 million unique f-s pairs

The Auckland names network

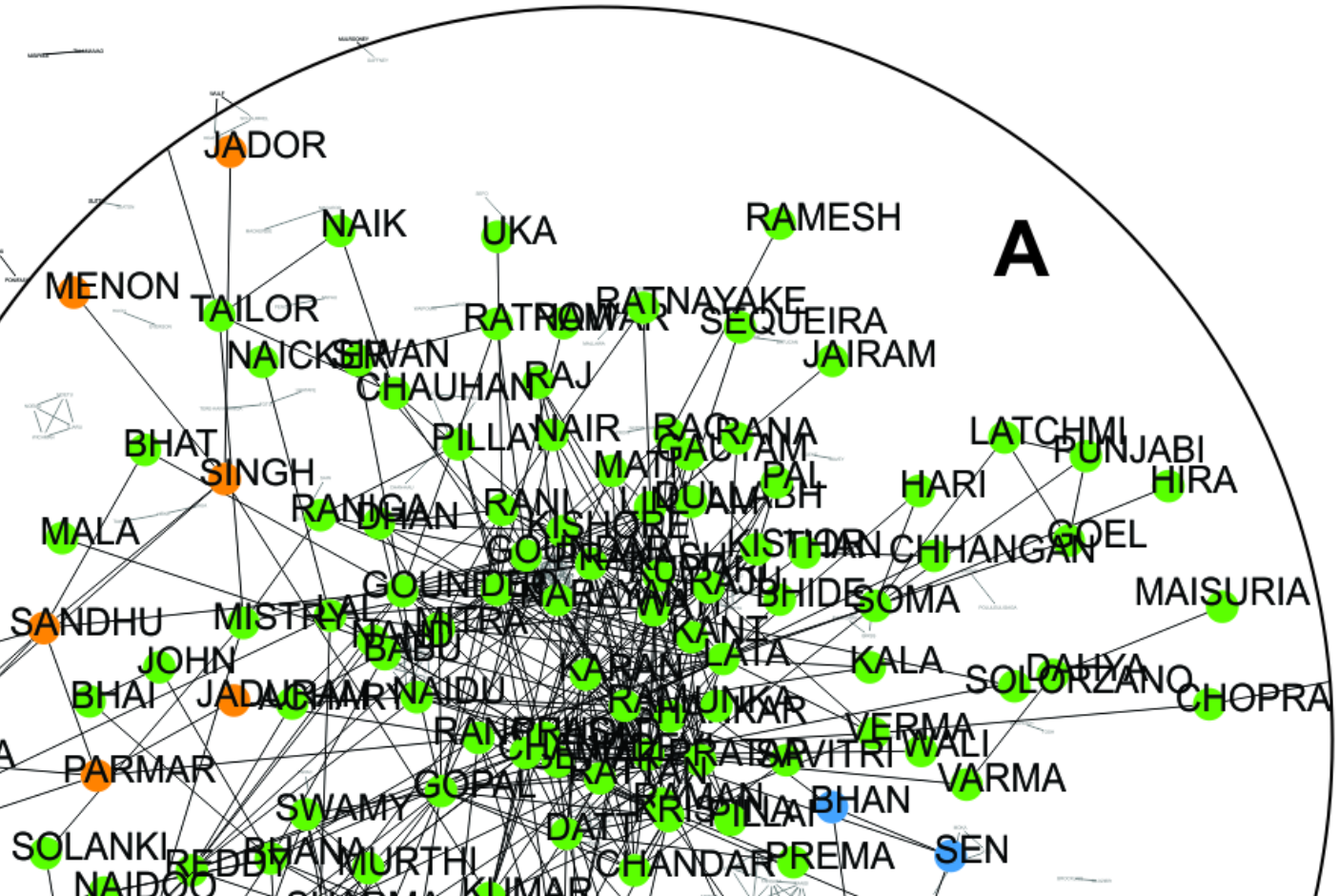


A filtered version of the Auckland network:

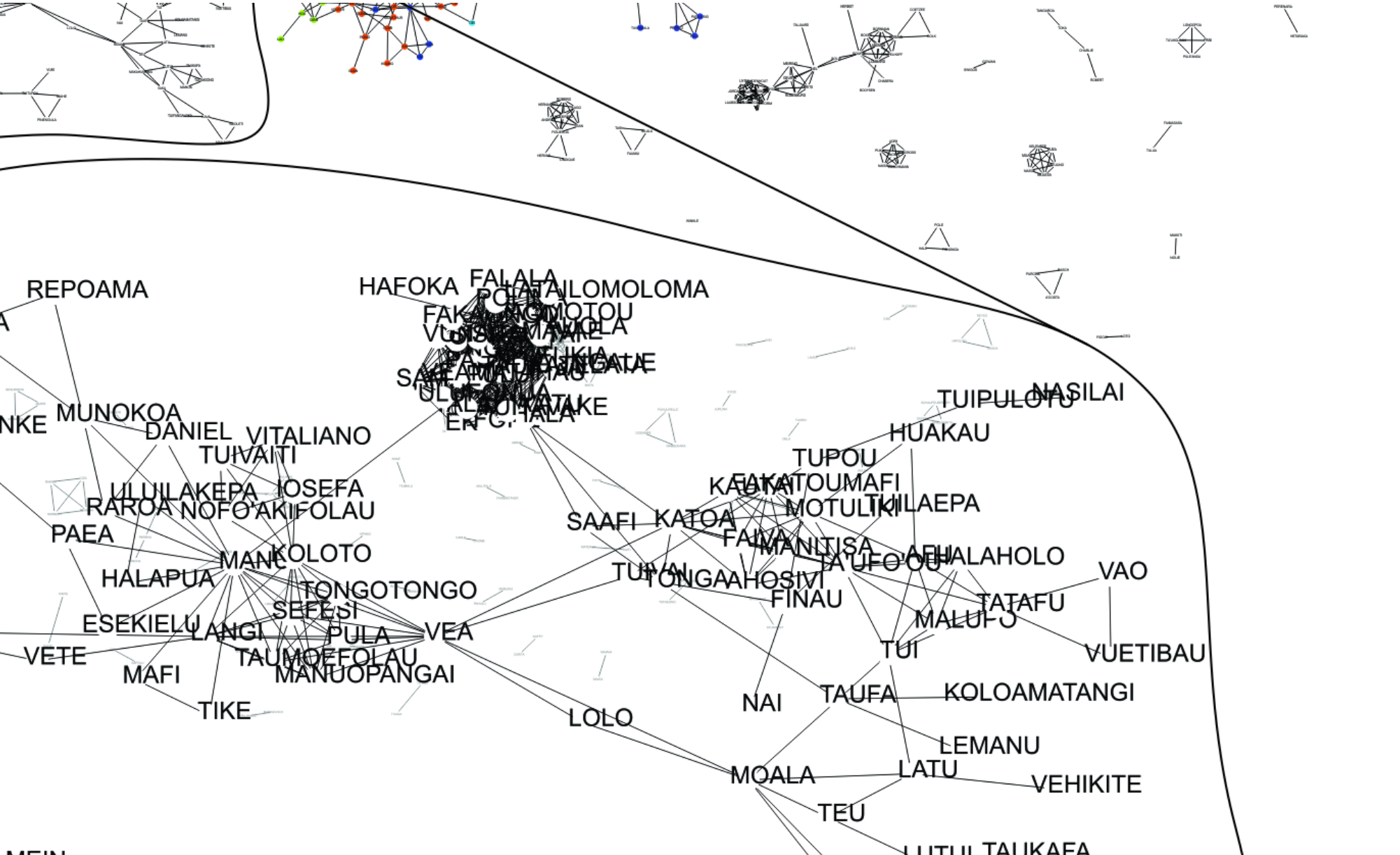
- A - Subcontinent names in 3 or 4 clusters
- B - Samoan / Tongan / Pacific names
- C - Māori names
- D - Eastern European names

Encouraging that previously untagged names are grouped separately from those already tagged. Note that colours are clusters determined by a community detection

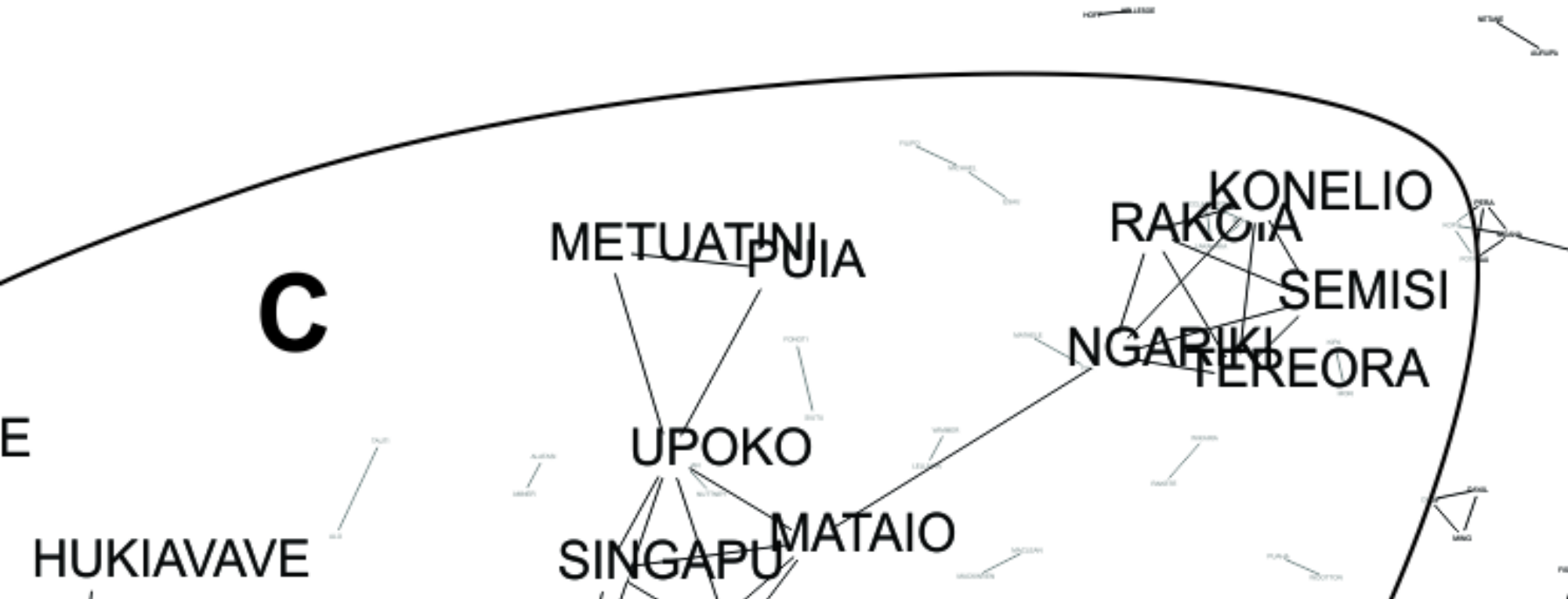
The Auckland names network



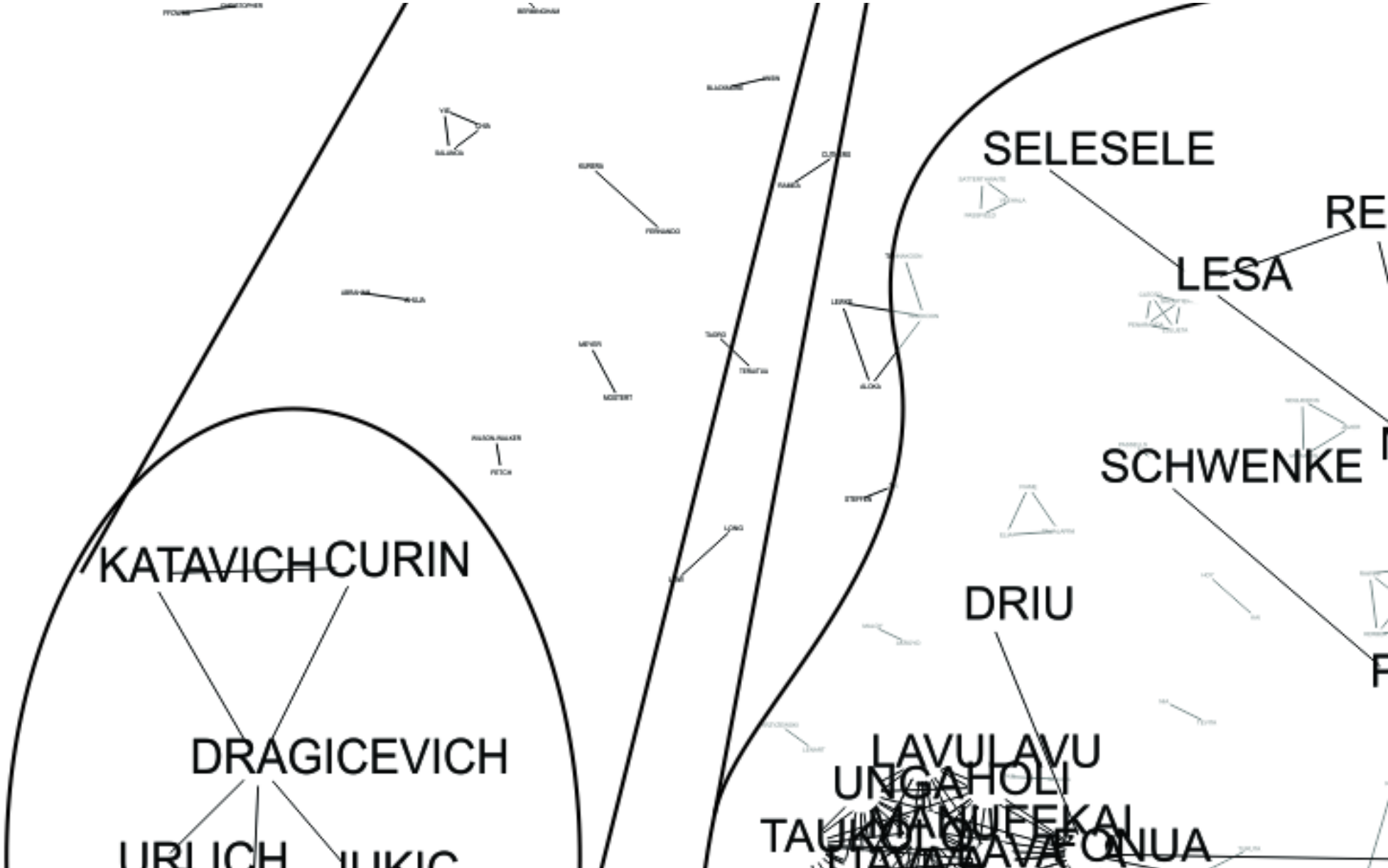
The Auckland names network



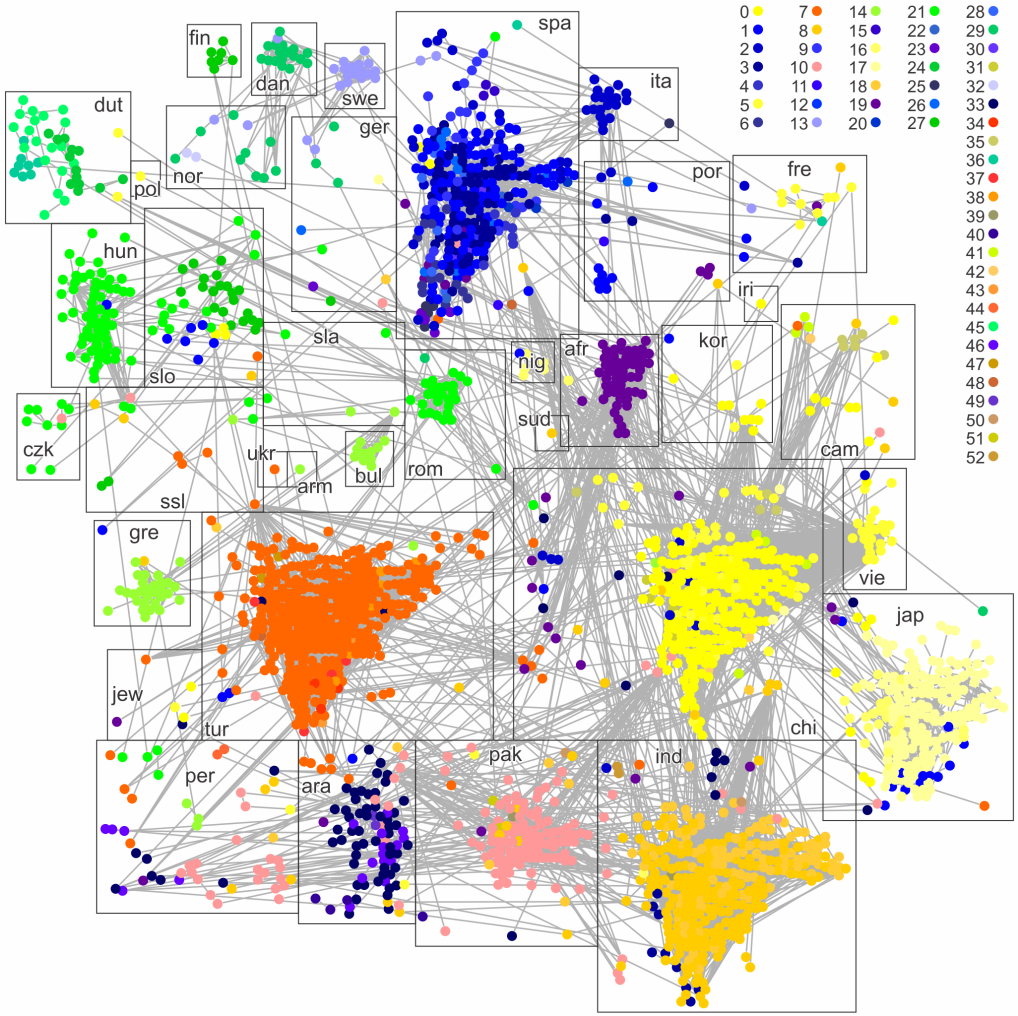
The Auckland names network



The Auckland names network



‘Diagnostic names’ network



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

Further work

Much remains to be done. . .

- Investigate methods for automatically cleaning up the names themselves (not particularly hopeful)
- Develop an automated way to determine best combinations of parameters for networks that partition cleanly
- Examine the degree to which (if at all) apparent structure in some groups reflects real intra-group divisions and relationships
- Consider possibilities offered by co-clustering methods
- Explore how 'label propagation' methods from graph theory might enable automated inference of CEL associations for unknown names

Further work

Much remains to be done. . .

- Investigate methods for automatically cleaning up the names themselves (not particularly hopeful)
- Develop an automated way to determine best combinations of parameters for networks that partition cleanly
- Examine the degree to which (if at all) apparent structure in some groups reflects real intra-group divisions and relationships
- Consider possibilities offered by co-clustering methods
- Explore how 'label propagation' methods from graph theory might enable automated inference of CEL associations for unknown names

Further work

Much remains to be done. . .

- Investigate methods for automatically cleaning up the names themselves (not particularly hopeful)
- Develop an automated way to determine best combinations of parameters for networks that partition cleanly
- Examine the degree to which (if at all) apparent structure in some groups reflects real intra-group divisions and relationships
- Consider possibilities offered by co-clustering methods
- Explore how 'label propagation' methods from graph theory might enable automated inference of CEL associations for unknown names

Further work

Much remains to be done. . .

- Investigate methods for automatically cleaning up the names themselves (not particularly hopeful)
- Develop an automated way to determine best combinations of parameters for networks that partition cleanly
- Examine the degree to which (if at all) apparent structure in some groups reflects real intra-group divisions and relationships
- Consider possibilities offered by co-clustering methods
- Explore how 'label propagation' methods from graph theory might enable automated inference of CEL associations for unknown names

Further work

Much remains to be done. . .

- Investigate methods for automatically cleaning up the names themselves (not particularly hopeful)
- Develop an automated way to determine best combinations of parameters for networks that partition cleanly
- Examine the degree to which (if at all) apparent structure in some groups reflects real intra-group divisions and relationships
- Consider possibilities offered by co-clustering methods
- Explore how 'label propagation' methods from graph theory might enable automated inference of CEL associations for unknown names

Conclusions

- Names appear a promising approach for looking at community structures and inter-relationships
- 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
- 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
- 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

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

Paper under-review: *Ethnicity and Population Structure in Personal Naming Networks*