Ethnic Structure in Global Naming Networks

K. K. Kowalska[1], P. A. Longley[1], M. Musolesi[2]

[1]Department of Geography, University College London, London, WC1E 6BT, UK
Email: {kira.kowalska.13; p.longley}@ucl.ac.uk

[2]School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK
Email: m.musolesi@cs.bham.ac.uk

1. Introduction

In today's multicultural society, there is a growing need to understand the detailed composition of different ethnic groups and the interactions between them. These large-scale group dynamics emerge from the aggregation of millions of ethnic self-identifications of individuals. Mateos et al. (2011) and others have begun to demonstrate the extent to which cultural, ethnic and linguistic (CEL) assignments can be inferred using 'naming networks' of forename-surname pairs of any population.

In further developing this work, we construct a global personal naming network from over 300 million name records from 23 countries. We use this to detect distinct social and ethno-cultural clusters in the network using Louvain community detection algorithm, and examine the interactions between them by inspecting the network structure. The results reveal the degree of isolation, integration or overlap between different human populations, and hence provide new insights into studies on migration, identity, integration or social interaction around the world.

2. Methods

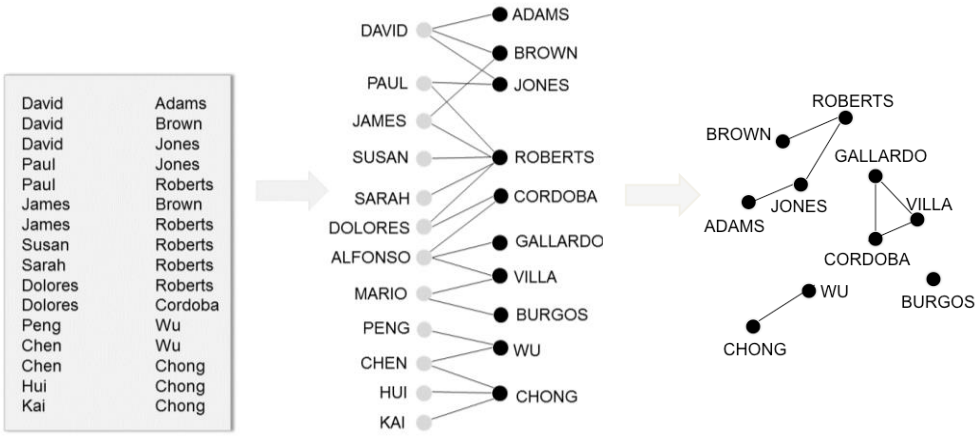
The central rationale to our analysis is that CEL classifications manifest themselves as topological features of networks in which unique surnames are represented as nodes. The subsections below outline how these networks are constructed from raw names data, and how community detection as well as other network statistics can be applied to the 'naming networks' in order to provide insight into the composition and interactions between different ethnic groups.

2.1 Names as a network

The first step in our analysis is to visualise names on a network. Given a dataset of people's names, a naming network is formed by treating unique forenames and surnames as network nodes and by placing links between the nodes if a person is identified by a particular forename-surname combination (see Figure 1 (a)).

Having represented names as a network, network links are weighted according to a technique outlined in Mateos et al. (2011). The weighing step adjusts link weights to ensure that very common forenames and surnames do not obscure the network topology, i.e. that the strength of links reflect CEL similarity of names instead of their overall popularity.

Finally, the forename-surname (two mode) network is converted into a one-mode network of surnames only. An example of such a transformation is shown in Figure 1 (b). The weights of the surname network are a result of a simple matrix multiplication of the weights in the forename-surname network (Mateos et al. 2011).



(a) (b)

Figure 1. Converting (a) people's names into a forename-surname network, (b) a forname-surname (two-mode) network to a network of surnames only.

2.2 Network community detection

Once names are represented in a surname network format, their ethnic origins are detected using the Louvain network community detection algorithm (Blondel et al. 2008). The algorithm inspects the network structure to unveil clusters of interconnected surname nodes, which can be interpreted as distinct ethnic groups. The Louvain method is chosen for this project because of its ability to deal with very large networks (up to tens of millions of nodes) with *weighted* links. It is iterative and hence enables investigation of ethnic communities at different levels of resolution. The 'best' level of resolution, i.e. leading to the most distinct communities, is characterized by the highest modularity score (close to 1), where modularity is defined as the fraction of links that fall within the obtained clusters minus the expected value of the fraction if links were distributed at random (Newman and Girvan 2004).

2.3 Semi-automated community labelling

Ethnic groups detected using the Louvain method can be automatically assigned their nationality making use of the richness of the naming data used in the project. The data come from 23 countries, enabling the calculation of percentage distribution of each surname across the 23 countries. The average distribution of surnames in one ethnic group could indicate their most probable country of origin. Since not all world countries are present in the data, some statistics are needed to decide whether a community comes from one of the represented countries, making automated labelling possible, or not, hence leaving the labelling to human expertise. The statistics investigated in the project, based on average percentage distributions of communities (or surnames they contain) across countries, are:

1. Percentage in dominant country (*Geographic Dominance*)
2. Standard deviation across countries (*Geographic Spread*)
3. Mean cosine similarity of surnames assigned (*Geographic Integrity*)

The assumption is that communities with *high* geographic dominance, *high* geographic integrity and *low* geographic spread could be automatically assigned nationality of their dominant country. The remaining communities could represent nationalities missing from the data or

ethnic groups that do not belong to any country (e.g. Romani people), and hence would require further investigation.

2.4 Network measures of interaction

Node properties of degree, betweenness and farness (Newman, 2010) are used to quantify interactions between surnames (nodes) in the naming networks. Surnames with high *degree* share similar naming practices with a large pool of other surnames, and hence might belong to a large ethnic group or one with unusually large surname heterogeneity. Surnames with high *betweenness* play an important role in connecting different parts of the global naming network and hence could be called 'cultural connectors' as they are most likely to interact with people from different ethnic backgrounds. Finally, surnames with high *farness* are least integrated within the global community; average farness of a community could be used as a measure of CEL isolation.

2.5 Data

Data used in this analysis come from a very extensive database of over 300 million people's names from 23 countries in four continents, collected from telephone directories and electoral registers and analysed as part of the 'Uncertainty of Identity' project at University College London (http://www.uncertaintyofidentity.com/). The data represent each country at a varying level of accuracy (i.e. the percentage of total population captured for various countries ranges from 0.3% to 79%). Therefore, before constructing the global naming network, name frequencies from each country are proportionally weighted to represent their total country's population.

3. Results

The world names data were converted into a naming network according to the steps outlined in Section 2.1. Firstly, a two-mode network was created with unique forenames and surnames as nodes (1,497,327 forename, 1,128,970 surname nodes). The two-mode network was then converted to a one-mode network of surnames only, which was subsequently used for the analysis of ethnic population structure around the world.

Ethnic communities in the global naming network were detected using the Louvain method. The algorithm started by assigning each surname node to a separate community, and then iteratively merged highly-connected communities until it arrived at the maximum modularity score of 0.628303. The resulting partition consisted of 7,947 ethnic communities of sizes varying from 2 to 157,889 surnames.

The geographic properties of *dominance*, *spread* and *integrity* of the ethnic communities were quantified using the statistics introduced in Section 2.3 (see Figure 2) in order to select communities suitable for automated nationality labelling. Varying thresholds on the statistics had impact on the number of surnames labelled, as shown in Figure 3.

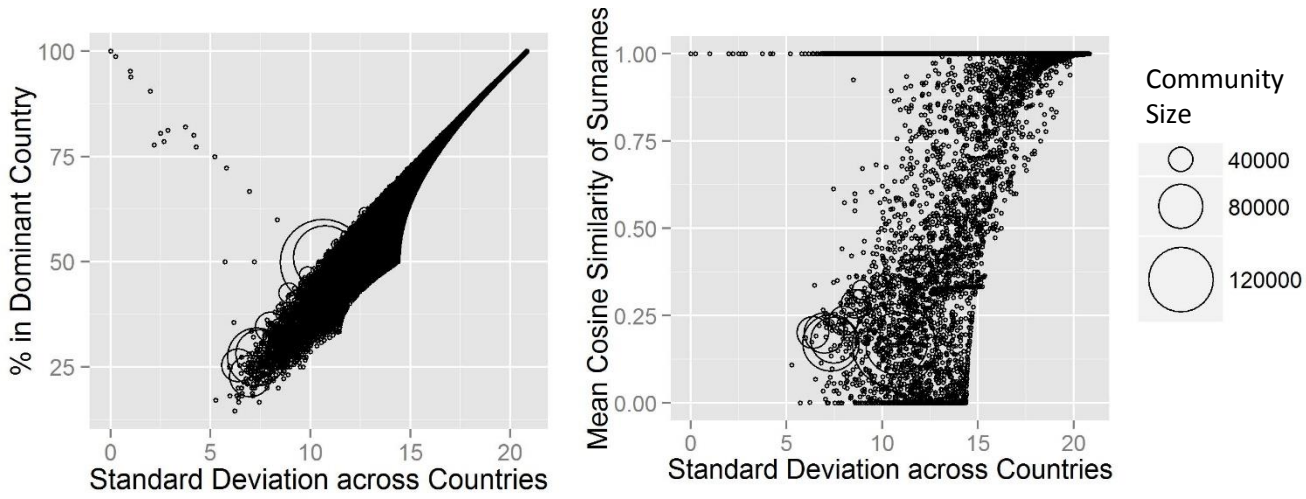


Figure 2. Ethnic communities scattered according to their geographic *dominance* (% in dominant country), *spread* (standard deviation) and *integrity* (mean cosine similarity of their surnames).

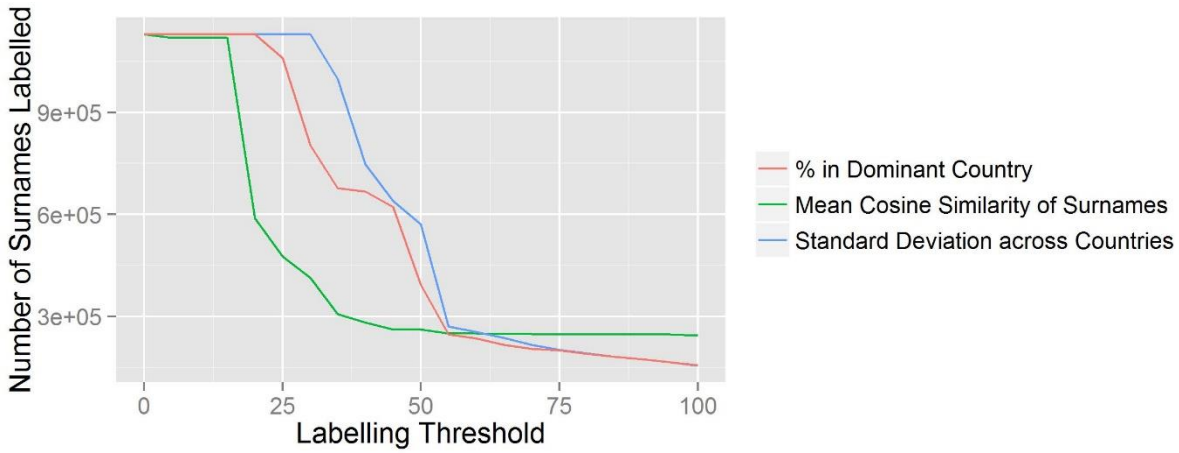


Figure 3. Number of surnames automatically labelled for different % thresholds on the three statistics of ethnic communities.

Properties of individual surnames were investigated by measuring their degree, betweenness and farness. Surnames corresponding to the most extreme values of the three statistics are summarized in Table 1. As discussed in Section 2.3, surnames 'Le', 'Begum' could be labeled as cultural connectors, whereas surnames 'Markovic' are 'Jankovic' represent most culturally or linguistically isolated individuals in the retained data.

Table 1. Surnames with extreme network properties.

Node Property	Highest	Lowest
Degree	Patel, Khan	Rahmani, Minar
Betweenness	Le, Begum	Laib, Bouaka
Farness	Markovic, Jankovic	Patel, Begum

4. Conclusions and Future Work

The paper presents preliminary results of analysing topology of 'naming networks' in order to gain insight into ethnic population structure around the world. The work is still in progress and numerous future directions are possible. In the first instance, validation techniques are needed for the automated labelling presented in this paper.

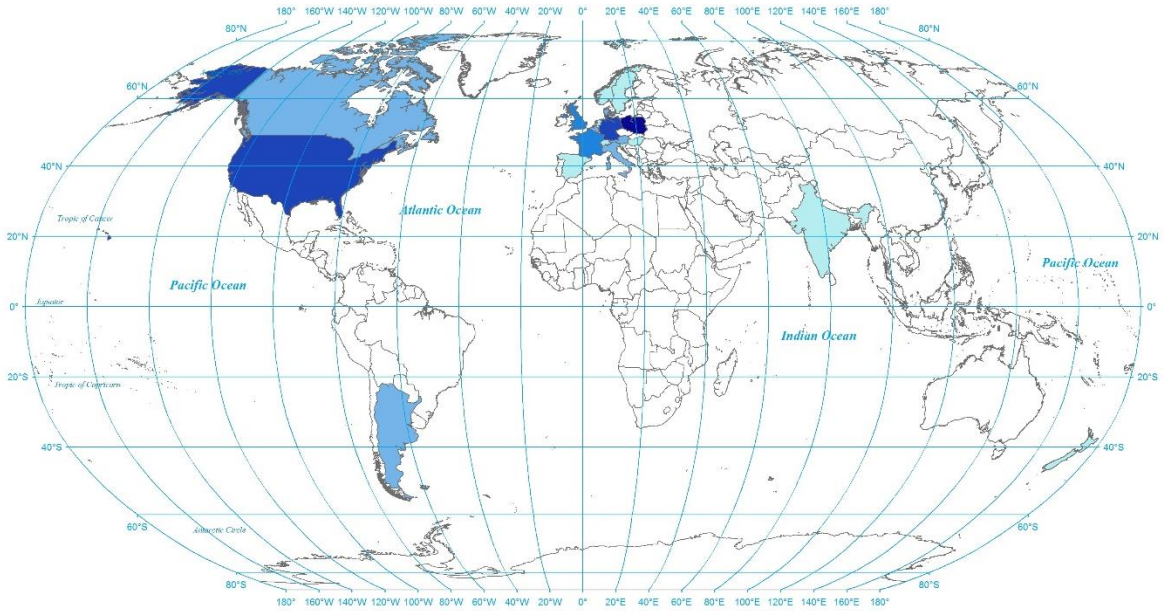


Figure 4. Concentration of surnames classified as Polish when all communities with >40% in their dominant countries are automatically labelled (uncoloured countries are not present in the world names data).

References

Blondel VD, Guillaume J, Lambiotte R and Lefebvre E, 2008, Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Clauset A, Newman MEJ and Moore C, 2004, Finding community structure in very large networks. *Physical Review E70*, 066111.

Mateos P, Longley PA and O'Sullivan D, 2011, Ethnicity and population structure in personal naming networks. *PLoS One*, 6(9): e22943.

Newman MEJ, 2010, *Networks: An Introduction*. Oxford University Press, Oxford, UK.

Newman MEJ and Girvan M, 2004, Finding and evaluating community structure in networks. *Physical Review E69*, 026113.