Personal Name Classification Using Collective Data

Hai H. Nguyen*, Alexandros Alexiou †and Alexander Singleton‡

Department of Geography and Planning, University of Liverpool

January 13, 2017

Summary

This paper describes a semi-supervised approach to personal name classification which includes family names and given names. The analysis uses a plethora of data sources that are integrated into an extensive names database, which is used to construct graphs of the relationships between names among groups of individuals. Centrality measures are calculated for every graph, showing how central a family name is within a specific group. Using outputs from the name ranking technique, a multi-stage methodological approach is developed in order to classify full names into ethnicity and nationality groups.

KEYWORDS: Demographics, Personal Names, Semi-supervised techniques, Graph Theory.

1 Introduction

Naming practices can be used as a mean to reflect the cultural, ethnic and linguistic origins of individuals. Humans intuitively "guess" a person's ethnic/national background by looking at the relationships of a name with their collective geography of names that they have knowledge of. Previous studies showed that family names can be clustered into common origins using their associated personal names. The general methodology of such previous work, e.g., *onomap* (Mateos et al., 2011), is twofold. The first step includes the construction of a large graph structure where nodes represent family names and weighted edges represent shared given names between two family names, where weights are calculated based on the number of occurrences this sharing takes place in the data. An unsupervised network community detection algorithm is then used to derive groups of similar family names. However, this approach has a number of issues. Firstly, it relies heavily on a full-coverage data sources, such as an electoral register or a telephone directory, which is difficult to meet for many countries. Secondly, as a purely unsupervised approach, almost no priori hypothesis can be used during the clustering process. The approach presented here on the other hand is semi-supervised, in that parts of the graphs are assigned a label (group) and utilises much more flexible

*hhn@liverpool.ac.uk

†a.alexiou@liverpool.ac.uk

‡alex.singleton@liverpool.ac.uk

Table 1: Example of data sources

Group	Source URL	Number of records	Source type
Poland	http://www.name-statistics.org/pl/numetara.php?pagina=1	1,681,900	Full name and frequency
Czech Republic	http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx?q=Y2hudW09MQ	279,970	Family names only
Ukraine	https://sites.google.com/site/uaname	100,000	Family names only
India	https://en.wikipedia.org/wiki/Category:Indian_family_names	679	Family names only
Greece	http://www.dimitri.8m.com	377	Family names only

data sources for each potential group of individuals (regions, nationalities, religion, etc.).

2 Data Collection and Integration

Data sources of interests include databases that can be used in the identification of an individuals group, where groups can be either be a country of origin, a geographical region, an ethnic group, or a religion. Candidate family names, given names, and full names are collected from multiple sources, as shown in Table 1.

For some groups it might be only possible to obtain either family names or given names, while others have full names and the associated frequency or count. The numbers of records also range from very limited (e.g. *Greece*) to very large (e.g. *Poland*). In general, raw databases with a large number of records present a lot more noise than smaller ones (for instance, those derived from a personal genealogical collection). Note that at this stage the reliability of web-based data sources is not guaranteed, and hence an extensive cleaning and verifying procedure is needed.

There are two important data types required to build a network of names: a set of full names (for creating the network edges) and the frequency of such pairs (for weighting the edges). However, as seen in Table 1, the majority of data sources contains either family names or given names only, and very often without the frequency. Our solution to this is to use an extract of Facebook users' full names (about 180 millions records) and a few other large database of names (e.g., phone book, electoral roll) to mine the linkages between family names and given names as well as their frequencies. Database cleaning was performed to remove full names pairs from certain population groups, particularly in North America. Some minorities adopt local (English) given names instead of the given names from the country of origin, and would bring noise in the dataset.

3 The Name Network and Name Ranking

Once the list of Given Name-Family Name pairs and their associated frequency were acquired, a family name graph was constructed. Each node of the graph represents a unique family name while a connection (edge) between two nodes represents at least one shared given names between two family names. Similar to the methodology from Mateos et al. (2011), edges are weighted relatively to the frequencies of the number of shared forenames.

Graphs were subsequently cleaned using a network clustering algorithm to detect and remove names that do not belong to the original target group (e.g. ethnic minorities). These algorithms can be very computationally expensive, so a series of different clustering algorithms were tested on sample data. The final method selected was a weighted fast greedy algorithm (Clauset et al., 2004). This process is very important as it helps to expand the original names database when specific ethnic clusters emerge within other groups. For instance, there is an expansive list of native Filipino family names within the Portuguese network.

After the clustering process, final sets of family names associated with population groups were identified by verifying family names within each cluster against name lists and other external sources such as `www.forebears.io`, a large open database of family names. In particular, if a majority (70%) of family names in a cluster is from a specific group, this group is assigned to the cluster. For clusters with mixed names, i.e., clusters with no clear majority, further clustering was applied in order to produce more homogeneous clusters.

The final step of this methodology is the production of a *name ranking* database which is used to approximate the likelihood of each name belonging to a specific group. A weighted Eigenvector centrality is calculated for each family name within the network which assessed the importance of the node within the network (Newman, 2010). Family names with a higher Eigencentrality have more links or are linked with other higher Eigencentrality family names. Compared to other similar centrality measures such as PageRank (Page et al., 1999), Eigencentrality is more suitable for our undirected networks, and also computationally efficient. The output of this process is a list of family names together with their associated non-negative centrality scores for each population group. Similar centrality scores were calculated for given names and two tables are therefore produced, each containing three columns: family name/given name, group, and centrality score.

4 Multi-stage Full Name Classification

Using the database derived from the name ranking process, a pair of given name and family name can be classified into a group. The classification algorithm is multi-stage, as follows:

Stage 1: Consolidated Database Matching Positive matches with given name and family name within groups are used to label the full name. Note that the given name and family name matches do not need to be identical, for example, a Filipino family name match and a Spanish given name match suggests that the full name is a Filipino names, given the history ties between the two groups.

Stage 2: Pattern-based Matching Names without matches in the name ranking tables are then processed with a series of heuristics that apply pattern based matching. For example, names with prefix *VAN DER* will be assigned into *Dutch* group.

Stage 3: Assessing Tokenised Names Family names and given names containing hyphens or spaces are split into separate tokens. Popular tokens such as *DE*, *VAN*, *DAS* are removed, and the remaining are matched against the consolidated database. Tokens that match are assigned to the group, while mixed groups are passed on to the next stage.

Stage 4: Disambiguation of Groups For records with multiple origin matches or made up of multiple tokens of multiple origins, a process of disambiguation is implemented to assign a single origin, otherwise an "Unclassified" tag is returned.

5 Evaluation and Preliminary Results

To evaluate our methodology, we conducted an experiment by classifying UK consumer full names. Since individual-level nationality data are not available, classification results have been compared against the 2011 Census Country of Birth (CoB) dataset, supplied by the Office for National Statistics (ONS). The CoB provides aggregate measures of the amount of people per country of birth at Lower Super Output Area (LSOA) level. Census CoB data are provided based on 15 geographical

Table 2: Comparison between Name Classification and Census CoB

Group	Average Difference	Standard Deviation	Upper Quantile	Lower Quantile
Europe, United Kingdom	-3.84	6.32	-0.90	-6.12
Europe, Ireland	8.38	10.46	12.28	3.17
Europe, EU Countries	1.98	4.60	4.29	0.46
Europe, Rest of	1.25	3.75	2.66	-0.45
Africa, North	-0.23	4.03	0.59	-1.13
Africa, Central and Western	0.44	3.23	0.63	-0.35
Africa, South and Eastern	-0.50	5.26	1.56	-1.57
Asia, Middle East	1.66	7.94	0.92	-0.34
Asia, Central	0.12	3.71	0.00	0.00
Asia, Southern	-0.20	3.62	0.45	-0.63
Asia, South-East	-1.67	3.55	-0.7	-2.52
Asia, Eastern	1.48	3.11	2.48	0.00

groups. However, it is not trivial to evaluate the classification across all 15 groups. For instance, English and North American names are difficult to differentiate; the same holds true for Latin and Spanish names. As such, the geographic areas of "The Americas and the Caribbean" and "Antarctica, Oceania (including Australasia) and other" were excluded from the association. Similarly, the Unclassified cluster was not associated to any specific nationality groups.

For the remainder of the groups, the similarity between the Census and the name classification esti-

mates was evaluated using a consumer dynamics file, supplied by CACI[1]. It includes the names and postcode-level addresses of over 54 million consumers in the UK. Each individual in the dataset was then given a group nationality estimate based on our name classification application. The National Statistics Postcode Lookup, supplied by ONS, was used to geo-code postcodes to a corresponding LSOA.

Using the country grouping methodology by the ONS[2], each of the 81 population groups identified where manually associated into one corresponding Census category, so that each person from the electoral dataset is joined to a Census nationality group based on their full name. Persons were aggregated at the LSOA level and ratios were created for each of the 12 nationality groups. Census CoB data was also appended to the dataset in order to enable comparisons between recorded and estimated area-level aggregates. Since only aggregated data are available, a pairwise differences

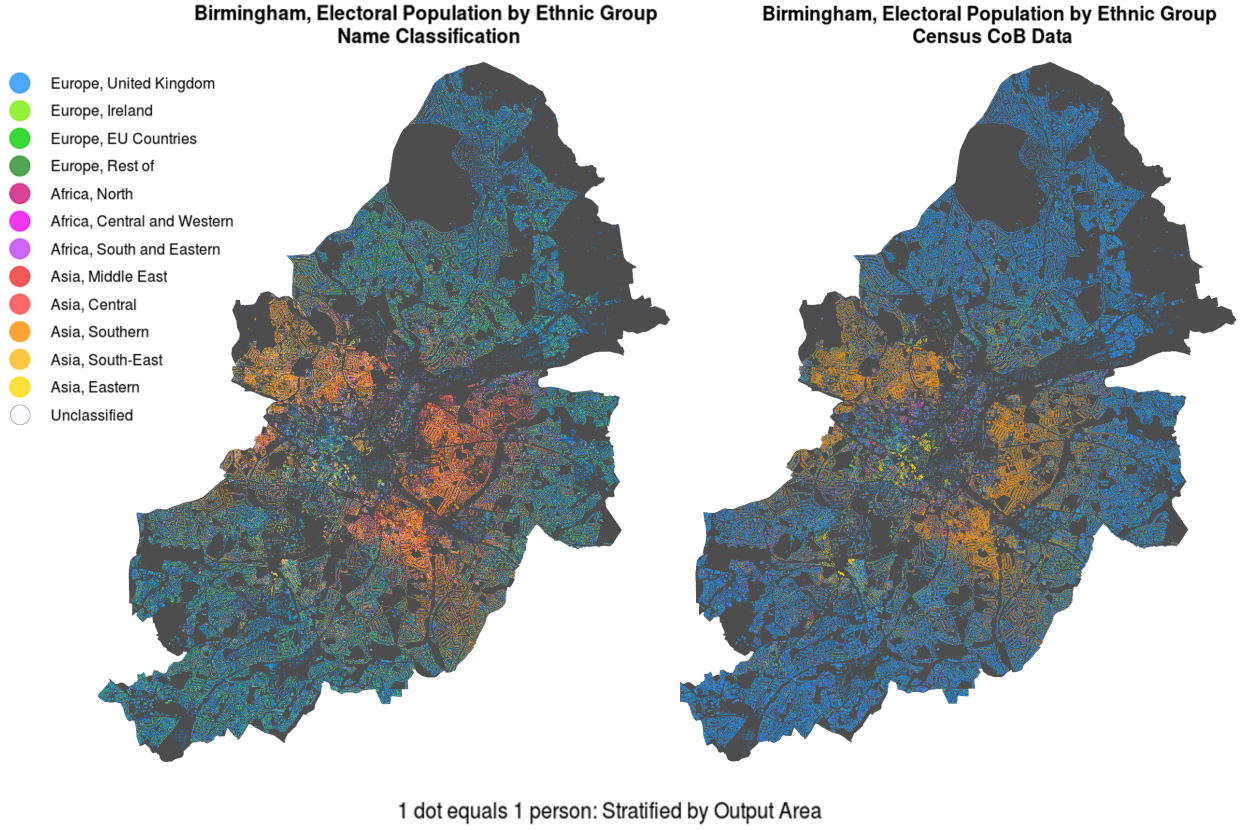

Figure 1: Dot-map presenting results from Name Classification vis-a-vis Census CoB data in Birmingham. Some of the differences may be due to second- or third-generation immigrants that, although have British nationalities, their names indicate otherwise.

[1]https://www.caci.co.uk

[2]Available at `http://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/nationalstatisticscountryclassification`

between ratios of the individual classes at the LSOA level was calculated, presented in Table 2. Positive differences suggest that the name classification overestimates the respective group while negative differences underestimate, with the Irish population being the most overestimated (8.38%). A representation of ethnicity patterns can also be seen using a dot-map presenting Name Classification vis-a-vis Census CoB ratios within the Birmingham local authority (Figure 1).

Although results are preliminary, the names classification seems to respond well to given the complexity of the issue. One of the major issues is the differentiation between groups of populations that have common names but different ethnic backgrounds (e.g. Muslim populations). However, results can be more accurate as more data sources are added in the name ranking database.

6 Acknowledgements

This research is supported by the Economic and Social Research Council (grant number ES/L011840/1).

7 Biography

Hai H. Nguyen is a data scientist at the Consumer Data Research Centre. He holds a PhD in Computer Science from the University of Nottingham and has been working on Linked Data infrastructure and graph-based machine learning.

Alexandros Alexiou is a Research Assistant at the Geographic Data Science Lab and a Data Scientist at the Consumer Data Research Centre. His interests focus on clustering methods, geocomputation and spatial analytics. His work employs programming tools, such as R and Python, open data sources and geographical big data, in order to analyse socio-spatial structure.

Alex Singleton holds a PhD in Geography from University College London (UCL). He previously held research positions at UCL, and is now a Professor in Geographic Information Science at the University of Liverpool. His research interests extend a geographic tradition of area classification and have developed a broad critique of the ways in which geodemographic methods can be refined through modern scientific approaches to data mining, geographic information science and quantitative human geography.

References

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

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

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.