Network analysis on Persons for Official Statistics

Keywords: *Complexity Science, Network analysis, Family networks*

1 Introduction

The census, which is as old as civilization, is the origin of Official Statistics. It measures various sociodemographics properties of the population which were, are and will be of importance for social scientists, historians, policy makers and government. A census is a very valuable source of information but its description of society formed by the connections between people is very limited. It records people living on the same address, but it fails to capture the broader network of relationships: family, friends, neighbors, co-workers and acquaintances. Most demographic statistics describe (aggregates of) properties of inhabitants. If Official Statistics strives to measure society, describing the network of relations between people that form the fabric of society can be a source of interesting demographic statistics. Is the strength of family ties regionally correlated? How diverse are personal networks? Given the current demographic trend in most countries that the average age is increasing, do parents live close to their children or is this distance increasing?
At Statistics Netherlands a research program was formed that tries to use complexity science and network analysis to derive new and additional official statistics. Several projects are started including deriving a enterprise to enterprise network for describing economic networks, but also a project to derive a social network of the Netherlands. This abstract describes the derivation of a directed network with family, dwelling, neighbor, school going children and coworkers relationships for the 17 million inhabitants of the Netherlands and some of its potential use for producing official statistics from it.

2 Methods

A complete and accurate derivation of a network that captures relationships between inhabitants would entail that all people that "know" each other are connected. Although many ingredients for constructing such a network are available, a complete acquaintance graph is in practice not possible. Instead we derive a network of people that are likely to know each other given auxiliary information. The resulting network resembles the real network, having similar network characteristics and statistics. The core of the network is formed by the population register which contains all persons registered in the Netherlands on October 1st 2014. The additional sources defining the edges are:

- The *parent-child register* is used to derive the following family relationships: 'child-parent', 'sibling-sibling', 'grandchild-grandparent', 'nephew/niece-uncle/aunt', 'parent-parent', 'cousin-cousin'.
- The *household register* is used to derive the 'household member' relationship for all person living in the same dwelling.
- The location of each household is used to derive 'neighbor' relationships, for each person living in the ten closest households within 50 meters.

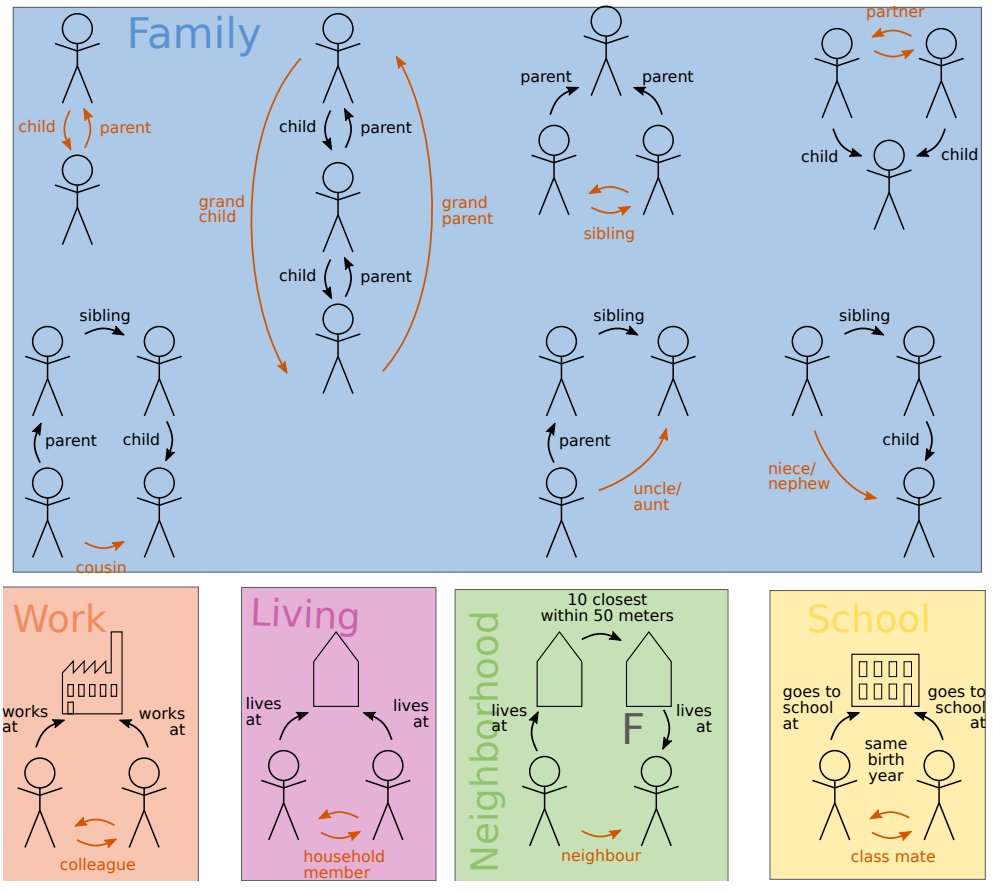

Figure 1: Derivation of person-person network.

- The *employment register* is used to derive 'co-workers'.
- The primary and secondary *education registers* are used to derive 'children going to the same school'. Without more information and to constrain the number of edges only edges between persons of same age (in years) are used.

The resulting network contains 16.9 million vertices and 39.0 billion edges.

3 Results

A direct result of the derived network are family networks. The family network is very likely to be a source for useful official statistics: e.g. geographic distances of family members are of interest to policy makers because older persons are relying more and more on family care.
Current research focuses on measuring 'segregation' in the Netherlands, which is suspected to increase: the working hypothesis is that network communities will be more homogeneous in income, education, ethnicity, neighborhood and schools, indicating segregation. We are planning to apply clustering methods to the network and investigate the dissimilarity between the clusters. Initially, we will be comparing networks of different regions.
Figure 2 shows network communities[1] in the Rotterdam area, in which only inhabitants of Rotterdam and their edges were selected. The Louvain method for community detection [3] using `igraph` R-package [4] was applied to this subgraph with 622 thousand vertices and 171 million edges. The figure displays a scatter plot matrix with ethnic group fractions, in which each dot identifies a community. The colors are determined using k-means clustering [1] on the ethnic fractions. It shows that not all

[1]Communities containing less than 1 thousand persons were removed from the analysis.

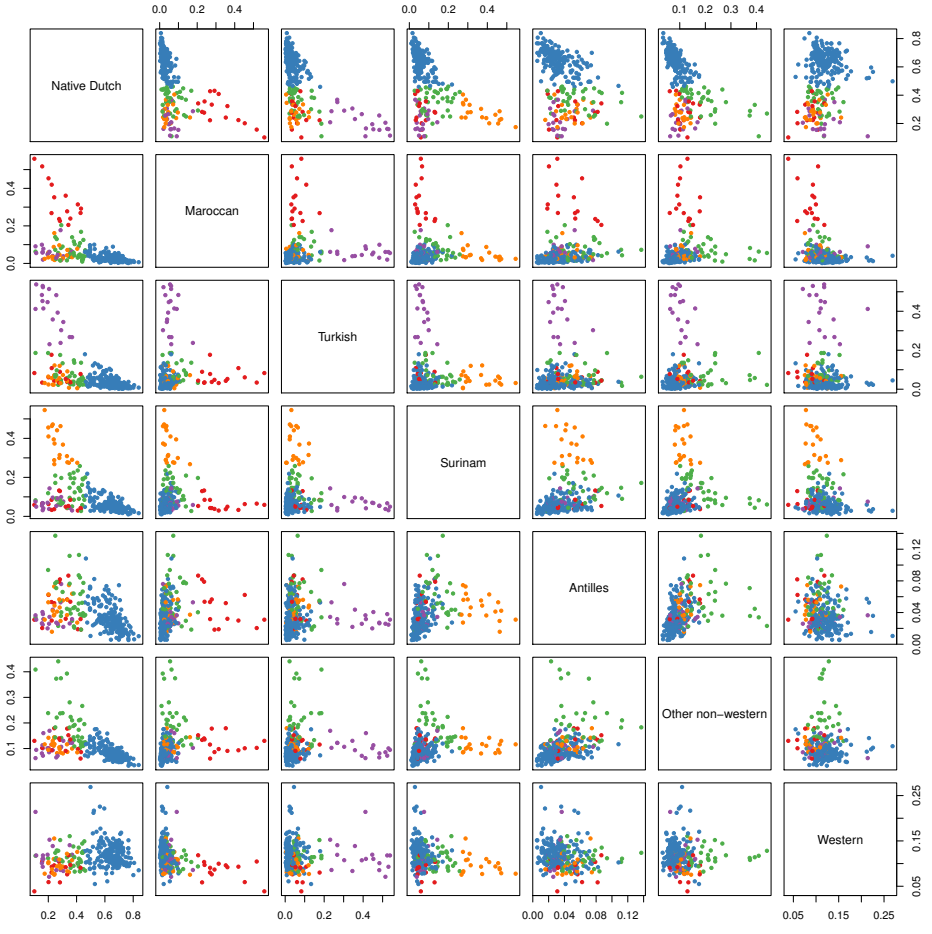

Figure 2: Scatter plot matrix showing the fractions of each of the ethnicity groups in the communities detected in the graph. The colors are derived using k-means clustering.

Table 1: Index of Dissimilarity for each of the ethnic groups.

Ethnicity:	Native Dutch	Moroccan	Turkish	Surinam	Antilles	Other Non-West	Western
Dissimilarity:	0.305	0.439	0.435	0.394	0.219	0.202	0.119

groups are equally 'mixed'. For example, communities with a large number of persons with a Turkish background (purple) have a small fraction of persons with a Moroccan background and vice versa for the clusters with a large number of persons with a Moroccan background (red). The mixing seems to be largest for persons with a Western, Other Non-Western and Antillean background. This is confirmed by the index of dissimilarity[2] calculated for these clusters (see Table 1).

4 Conclusions

Describing societal phenomena using network analysis seems a promising direction for Official Statistics: the essence of many statistics of economy or society is in the interaction of its agents. The results presented are the first steps into investigating the possibilities using (derived) network data for official statistics purposes. The resulting in- and out-degree distributions may be used to model networks of other countries and regions where summary statistics is available, but individual data is scarce. As for using the network to detect communities, further details need to be refined. For example, how does one weight the different types of relationships? How does one

compare community structures of different regions/time periods? Current results do show that using graph analytics methods for analyzing society are promising.

References

[1] Hartigan, J. A. and Wong, M. A.: A K-means clustering algorithm. Applied Statistics 28, 100-108 (1979).

[2] Jahn, Julius, Calvin F. Schmid, and Clarence Schrag: The measurement of ecological segregation. American Sociological Review 12, no. 3 (1947): 293-303.

[3] Blondel V.D., Vincent D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, (Oct 2008)

[4] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org