Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Q1 Inferring cultural regions from correlation networks of given baby names

Q2 Mateusz Pomorski [a], Małgorzata J. Krawczyk [a,*], Krzysztof Kułakowski [a], Jarosław Kwapień [b], Marcel Ausloos [c,d,e]

[a] AGH University of Science and Technology, Faculty of Physics and Applied Computer Science - al. Mickiewicza 30, 30-059 Kraków, Poland

[b] Institute of Nuclear Physics, Polish Academy of Sciences - ul. Radzikowskiego 152, 31-342 Kraków, Poland

[c] GRAPES - rue de la Belle Jardiniere, B-4031 Liege, Federation Wallonie-Bruxelles, Belgium

[d] e-Humanities Group, KNAW - Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands

[e] School of Management, University of Leicester - University Road, Leicester, LE1 7RH, UK

HIGHLIGHTS

- Construction of a network of correlations of frequencies of names in states of USA.
- Communities in this network are stable since 1880 till 1980.
- The structure of these communities matches the administrative regions of USA.
- The Zipf exponents for the names allow to infer about the cultural pressure.
- The interpretation is supported with the related data of Belgium.

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 8 October 2015
Available online xxxx

Keywords:
Networks
Communities
Names

ABSTRACT

We report investigations on the statistical characteristics of the baby names given between 1910 and 2010 in the United States of America. For each year, the 100 most frequent names in the USA are sorted out. For these names, the correlations between the names profiles are calculated for all pairs of states (minus Hawaii and Alaska). The correlations are used to form a weighted network which is found to vary mildly in time. In fact, the structure of communities in the network remains quite stable till about 1980. The goal is that the calculated structure approximately reproduces the usually accepted geopolitical regions: the North East, the South, and the "Midwest + West" as the third one. Furthermore, the dataset reveals that the name distribution satisfies the Zipf law, separately for each state and each year, i.e. the name frequency $f \propto r^{-\alpha}$, where r is the name rank. Between 1920 and 1980, the exponent α is the largest one for the set of states classified as 'the South', but the smallest one for the set of states classified as "Midwest + West". Our interpretation is that the pool of selected names was quite narrow in the Southern states. The data is compared with some related statistics of names in Belgium, a country also with different regions, but having quite a different scale than the USA. There, the Zipf exponent is low for young people and for the Brussels citizens.

[*] Corresponding author.
E-mail address: malgorzata.krawczyk@agh.edu.pl (M.J. Krawczyk).

http://dx.doi.org/10.1016/j.physa.2015.11.003
0378-4371/© 2015 Elsevier B.V. All rights reserved.

Please cite this article in press as: M. Pomorski, et al., Inferring cultural regions from correlation networks of given baby names, Physica A (2015), http://dx.doi.org/10.1016/j.physa.2015.11.003

1. Introduction

Q3

In sociophysics/quantitative sociology, any idea of a research project is inextricably interwoven with an access to related data. The data on babies' names in the United States are available [1] for a long time interval (1880–2011), i.e. the second half of the stretch of history of the USA since the American independence. Thus, they give a unique opportunity to investigate cultural trends in this large country throughout several decades of years. In Ref. [2], some related data has been analyzed from the perspective of the Zipf law, i.e. the name popularity y vs. the name rank r. The Zipf law ($y \propto r^{-\alpha}$) has been found to be valid only in the large r rank range. The time dependence of the parameter α has been found to show a weak and wide maximum in the early 50s, slightly more visible for boys' names. In Ref. [3], the data has been analyzed in terms of the theory of fashion [4]. Two conclusions have been highlighted in Ref. [3]: (i) for many names, the rise of popularity is more abrupt than its fall; (ii) the time interval in which a name is popular is shorter when the set of selected names is richer. The results of the data analysis in Refs. [3,5] indicated that this richness (measured by an index termed 'inequality' in Ref. [5] and 'fragmentation' in Ref. [3]) increases in time; yet, the data are not monotonous again ca. 1950. In Ref. [5], a model has been formulated based on the social impact for a family, resulting from the choice of the name of the newborn baby. It was concluded that this impact decreases in time. In Ref. [6], the same American data has been investigated in different states, in 1910–2012. In particular, the Pearson correlations have been calculated for the names frequencies for all states and for each year. The methods applied in Ref. [6] indicate that the southern and northern states form two uncorrelated clusters, which persist until 1960. Also, the results of a Principal Component Analysis indicated that the difference between the first and second eigenvectors is the largest in 1950; thereafter, the partition is found to be the sharpest. Finally, the time dependence of the popularity, when averaged over the names, confirmed that the popularity decay goes more slowly than its rise (Fig. 6 in Ref. [6]).

In the present report, our aim is to use the results of the correlations of the names popularity between different states, to identify communities in the network of states. These results are combined with the time and state dependent Zipf indices α in the low rank ranges. These tasks are described in the three subsequent sections. Next, we provide some original information on related data on Belgian (BE) names, — Belgium being a country also with different regions, as the USA, but having a quite different smaller population and area size scales than the USA. The BE data, although over a less complete time intervals, allow to validate our conclusions, given in the last section.

2. Correlations between states

For each year in the time interval [1910–2011], the set of most popular $N = 100$ names in 48 USA states (we have no data from Alaska and Hawaii) has been selected [1]. For each of these states and for each of these names, the percentage p of newborn babies with this given name has been derived. Then, the contribution of the a state to the popularity of the name i in the year t is found from,

$$x(i,t,a) = \frac{p(i,t,a)}{\sum\limits_{b=1}^{K} p(i,t,b)}, \tag{1}$$

where $K = 48$ is the number of states. The (name) mean of the variable x as a function of time and state is

$$\langle x(t,a)\rangle = \frac{1}{N}\sum_{i=1}^{N} x(i,t,a). \tag{2}$$

With the deviation from the mean defined as $y(i,t,a) = x(i,t,a) - \langle x(t,a)\rangle$ and its variance $\sigma^2(t,a) = \langle y^2(t,a)\rangle$, the Pearson correlation coefficients ρ of correlations between the states a and b are

$$\rho(t,a,b) = \frac{\sum\limits_{i=1}^{N} y(i,t,a)y(i,t,b)}{N\sqrt{\sigma^2(t,a)\sigma^2(t,b)}}. \tag{3}$$

The variable y has been so defined as to cancel differences in frequencies of the names from the selected set, a fluctuation which can be remarkably large as follows from our perusal of the data.

Recall that our purpose is to emphasize whether there are popular reactions in whatever states on particular name trends. The (a, b) correlation matrix represents a weighted network, where the states are nodes and the correlations give the weights of links. This network is analyzed in the next section.

3. Communities of states

Since the correlation coefficients are in the range $(-1, 1)$, the weighted matrix $w_{ab}(t)$ can be constructed from $w_{ab}(t) = (1 + \rho(t,a,b))/2$ such that $w \in [0, 1]$. The communities of the network of states are identified from the set of differential

Please cite this article in press as: M. Pomorski, et al., Inferring cultural regions from correlation networks of given baby names, Physica A (2015), http://dx.doi.org/10.1016/j.physa.2015.11.003