Evolving Systems manuscript No.
(will be inserted by the editor)

A Genetic Algorithm For Spatiosocial Tensor Clustering

Exploiting TensorFlow Potential

Georgios Drakopoulos · Foteini Stathopoulou · Andreas Kanavos · Michael Paraskevas · Giannis Tzimas · Phivos Mylonas · Lazaros Iliadis

Received: date / Accepted: date

Abstract Tensor clustering is a knowledge management technique which is well known as a major algorithmic and technological driver behind a broad applications spectrum. The latter ranges from multimodal social media analysis and geolocation processing to analytics tailored for large omic data. However, known exact tensor clustering problems when reduced to tensor factorization are provably NP hard. This is attributed in part to the volume of data contained in a tensor, proportional to the product of its dimensions, as well as to the increased interdependency between the tensor entries across its dimensions. One well studied way to circumvent this inherent difficulty is to resort to heuristics. This article presents an enhanced version of a genetic algorithm tailored for community discovery structure in tensors containing spatiosocial data, namely linguistic and geolocation data. The objective function as well as the chromosome fitness functions by design take into account elements of linguistic propagation models. The genetic operators of selection, crossover, and mutation as well as the newly added double mutation operator work directly on the

Georgios Drakopoulos
Cloudminers Inc.

Georgios Drakopoulos and Phivos Mylonas
Department of Informatics, Ionion University
E-mail: {c16drak, fmylonas}@ionio.gr

Foteini Stathopoulou
University of Luxembourg
E-mail: fstathop@uni.lu

Andreas Kanavos
Hellenic Open University
E-mail: kanavos@ceid.upatras.gr

Michael Paraskevas and Giannis Tzimas
Technological and Educational Institution of Western Greece
E-mail: mparask@teiwest.gr, tzimas@teimes.gr

Lazaros Iliadis
Forest Informatics Lab, Democritus University of Thrace
E-mail: iliadis@fmenr.duth.gr

community level. Moreover, various policies for maintaining gene variability across generations are studied in an extensive simulation powered by Google TensorFlow. As with its predecessor, the proposed genetic algorithm has been applied to a dataset consisting of a large number of Tweets and their associated geolocations from the Grand Duchy of Luxembourg, a historically and *de facto* trilingual country. The results are compared with those obtained from the original genetic algorithm and their differences are interpreted.

Keywords Multilingual social networks · Multimodal social networks · Cross cultural communication · Language variation models · Tensor clustering · Google TensorFlow · Genetic algorithms · Gene variability · Geolocation data · Spatiosocial data · Humanistic data · Higher order data

CR Subject Classification H.2.8 · G.2.2 · G.3 · M.1

Mathematics Subject Classification (2000) 05C76 · 05C85 · 05D99 · 62H30 · 91C20 · 91C99

1 Introduction

Two of the most prominent features of current online social media are multimodality and multilinguality. The latter reflects the natural *de jure* or *de facto* state of affairs, itself the result of various social or historical conditions, in a considerable number of countries around the globe including Canada, Switzerland, Cyprus, Belgium, and Luxembourg. In Twitter alone there is no single predominant language, since the volume of tweets in Spanish and in Japanese each equal approximately the number of those in English as shown in Donoso and Sánchez (2017), with Hong et al (2011) maintaining that Indonesian tweets are very close as well. As is the case with any other human activity, language exposure in social media results in almost constant alteration as Croft (2003) claims. This change process according to Eisenstein et al (2014) includes syntax, forms, emoticons, abbreviations, phonetic spellings, and neologisms. Note that different dialects of the same language are treated in this work as related yet distinct languages.

With the advent of multimodal social media the diachronic role of language as the primary human communication vehicle is strongly reinforced with various non-linguistic elements including metatext (i.e. hashtags), memes, short live videos, geolocation, sentiment, and netizen reactions such as the signature Facebook like button. Among the factors influencing the size and shape of online linguistic communities Weinreich et al (1968) singles out location, social status, income, and dialect. Of these factors only the first and the last can be directly quantified, even with a certain degree of uncertainty.

Community structure discovery in multilingual and multimodal social graphs is well known to be among the most challenging tasks due to the high number of semantic, topical, spatial, linguistic, or other type of constraints deriving from the context. Following Drakopoulos et al (2017d) this article augments ordinary Twitter interactions

with linguistic and spatial post similarity metrics. In order to include the additional functionality, it is necessary to represent Twitter netizens and the interaction between them as a multilayer graph, which can be naturally expressed as an adjacency tensor. Given that tensor clustering is an NP hard problem, the possibility of a heuristic solution should be explored, especially for large scale and sparse tensors.

The primary contribution of this article is TENSOR-G2, an enhanced version of the genetic algorithm TENSOR-G proposed in Drakopoulos et al (2017d). The main difference between them is that TENSOR-G2 has a new genetic operator for doubly mutating the same chromosome in a single step as well as a new termination criterion. Moreover, TENSOR-G2 unlike its predecessor is implemented in Google TensorFlow exploiting the inherent parallelism potential of the latter. Two additional notable differences from Drakopoulos et al (2017d) is that Luxembourgish language is now treated separately instead of as a German dialect and that the respective definitions of the two fitness functions have been slightly altered. The original test dataset will serve as a benchmark.

The structure of this work follows. Previous work in the fields of genetic algorithms, tensors, and linguistics is summarized in section 2. The notions underlying the design of both genetic algorithms are explained in section 3, whereas the results obtained from executing them are outlined and interpreted in section 4. Section 5 concludes this article by presenting the main findings as well as potential future research directions. Finally, paper notation is summarized in table 1. Tensors are printed in capital italics and vectors in small boldface.

Table 1 Article notation.

Symbol	Meaning
$\triangleq$	Definition or equality by definition
$\{s_1,\ldots,s_n\}$	Set consisting of elements $s_1,\ldots,s_n$
$\lvert S\rvert$	Set cardinality
$S_1 \setminus S_2$	Asymmetric set difference between sets S_1 and S_2
τ_{S_1,S_2}	Tanimoto similarity coefficient between sets S_1 and S_2
$(s_1,\ldots,s_n)$	Tuple consisting of elements $s_1,\ldots,s_n$
$\lVert\mathscr{T}\rVert_F$	Tensor Frobenius norm
$\circ_n$	Vector outer product along dimension n
$\mathrm{H}(x_1,\ldots,x_n)$	Harmonic mean of $x_1,\ldots,x_n$
$\mathrm{E}[X]$	Mean value of random variable X
$\mathrm{Var}[X]$	Variance of random variable X

2 Previous Work

Current views on the language evolution process as can be found in Matras (2013) where the historical perspective is taken into consideration, in Milroy (1980), and in Kershaw et al (2017) where the broader phenomenon of language is treated. Various language constructs and their change over time are studied in Pakendorf (2014) as well as in Milroy and Milroy (1985), which places an emphasis on diffusion between communities. More studies on various topics of language evolution include Matsumoto (2010) which examines multilingual communities, Dixon (1997) and Androutsopoulos (2011) which claim language evolution is a global social phenomenon, and Labov (2001) and Labov (2007) which explore the relationship between real and online linguistic communities. Quantitative methods for assessing language evolution are developed in Hale (2007), Kershaw et al (2015), and in Michael et al (2014). The link between language change and social change is treated in Trudgill (2011), Nevalainen (2015), Kirk and Mees (2006), and in Djugasvilii (1950). Interaction between multilingual communities and how hybrid expressions and neologisms are created are examined in Eleta and Golbeck (2012) and later in Hale (2014), the latter indicating cross-language awareness through the links between single language blogs. It is of interest that interaction in digital communities tends to be geographically assortative as shown in Backstrom et al (2010) and in Maybaum (2013), a finding which Goel et al (2016) and Kershaw et al (2017) specialize for Twitter.

Tensor analysis, also known as multilinear algebra, as described in numerous papers including Kolda and Bader (2009), Karatzoglou et al (2010), and Dunlavy et al (2011) is the current evolution step of linear algebra in the sense that tensors are multidimensional vectors. Tensors appear naturally in signal processing settings where multiple inputs interact simultaneously with multiple outputs such as MIMO radars explained in Nion and Sidiropoulos (2010), blindly discriminating simultaneous data sources as described in Cardoso (1990), as well as in and biomedical image processing as shown in Westin et al (2002). Social media analysis has also benefited with the introduction of tensors as analytical tools as the latter allow the development of higher order influence analytics as in Drakopoulos et al (2017b), sentiment analysis as in Drakopoulos (2016), or advanced community structure discovery in fuzzy graphs Drakopoulos et al (2017c). In knowledge mining tensors can be used for dimensionality reduction as in De Lathauwer and Vandewalle (2004), in Papalexakis and Doğruöz (2015), and in Shashua and Hazan (2005). Moreover, information retrieval models based on third order tensors have been recently proposed. For instance Drakopoulos and Kanavos (2016) proposes a term-author-document model, while in Drakopoulos et al (2017a) a term-keyword-document is described. A space efficient and persistent scheme for storing compressed tensors as multilayer graphs is presented in Kontopoulos and Drakopoulos (2014).

Genetic algorithms as described in De Jong (1988) and in Holland (1992) constitute a class of numerical optimization algorithms inspired from standard DNA operations and Darwinian evolution as defined in Darwin (1859) and refined in Dawkins (2006), which places heavy emphasis on the concept of fitness. Speficially, these heuristics

consider a large number of candidate solutions named *chromosomes* consisting of a group of *genes*, each coding distinct parts of the solution as shown in Booker et al (1989). The chromosomes are handled at the gene level with operations including *selection*, *crossover*, and *mutation* as explained in Davis (1991). Genetic algorithms have a broad spectrum of applications include disease diagnosis as in Lu et al (2016), sensorineural hearing loss as in Wang et al (2017), electromagnetic field optimization as in Rahmat-Samii and Michielssen (1999), set cover in graphs as in Beasley and Chu (1996), and nonlinear higher order function as in Tanese (1989). Finally, the close connection between genetic algorithms and machine learning are explored in Goldberg and Holland (1988).

3 Implementation

3.1 Tensor Representation

Formally, a tensor is defined as:

Definition 1 (Algebraic tensor definition -from Kolda and Bader (2009)) A p-th tensor $\mathscr{T}$, $p \in \mathbb{Z}^+$, is a linear mapping simultaneously connecting p not necessarily distinct linear spaces $\mathbb{S}_k$, $1 \leq k \leq p$.

The key point in our case is the simultaneous connection of spaces, as TENSOR-G2 operates on third order tensors $\mathscr{T}^{n\times n\times(L_0+2)}$ which represent pairwise elementary spatiosocial interactions between the n netizens. Specifically, the first two dimensions represent the netizen space (twice), while the third dimension denotes the number of ways two netizens can interact. L_0 is the total number of languages taken into consideration, in this case English, German, French, and Luxembourgish. Note that in Drakopoulos et al (2017d) the latter was considered a German dialect but now is treated as a separate language. The two additional interaction ways are the spatial assortativity and the Twitter interaction, expressed through the *follow* relationship.

Each tensor entry signifies whether there is a linguistic, spatial, or online interaction between a given pair pair of netizens.

$$\mathscr{T}[i_1,i_2,i_3] = \begin{cases} 1, & \text{netizens } i_1 \text{ and } i_2 \text{ interact through } i_3 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

One way to partition any p-th order tensor $\mathscr{D} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_p}$ to a sum of r_0 not necessarily overlapping communities with special structure is to use the Kruskal decomposition. The latter states that $\mathscr{D}$ can be rewritten as:

$$\mathscr{D} = \sum_{k=1}^{r_0} \lambda_k \mathscr{D}_k = \sum_{k=1}^{r_0} \lambda_k \mathbf{v}_{k,1} \circ_1 \ldots \circ_{p-1} \mathbf{v}_{k,p}, \qquad \lambda_k > 0, \left\|\mathbf{v}_{k,j}\right\|_2 = 1 \tag{2}$$

where $\circ_k$ denotes the outer tensor product along dimension k. The properties of this product are defined among others in Kolda and Bader (2009). Thus, the original data

tensor $\mathscr{D}$ is rewritten as the sum of r_0 tensors of rank one, which can be thought of as the LSI factors in higher dimensions. However, Kruskal decomposition will not be used as a baseline method as was in Drakopoulos et al (2017d) since r_0 estimation is NP hard, the communities can overlap, and have an excessive structure restriction placed on them. Thus, the performance of TENSOR-G2 will be evaluated using cluster quality metrics. The layered structure of tensor $\mathscr{T}$ is shown in figure 1.

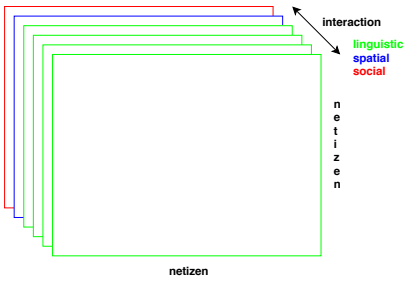

Fig. 1 Tensor layers (linguistic, spatial, and social).

The elements of $\mathscr{T}$ are filled as follows. In the L_0 linguistic layers interaction is taking place when two netizens post a tweet in the same language. In the spatial layer two netizens are close if their corresponding social factor, defined in subsection 3.2, is above a given threshold η_0 (see table 2). Finally, for the last layer of $\mathscr{T}$ any two netizens are considered to interact socially in Twitter if either follows the other.

3.2 Linguistic And Spatial Factors

This subsection introduces the social, linguistic, and spatial factors which assess the degree of various aspects of chromosome compactness and will subsequently serve as building blocks for more complex chromosome fitness functions. Let $L(u)$, $1 \leq u \leq n$, be the language set used by netizen u to interact digitally and also let $\ell_0(u) \in L(u)$ be its predominant language. For any netizen set S let $L(S)$ symbolize the union of all languages used by the netizens beloning to that set

$$L(S) \stackrel{\triangle}{=} \bigcup_{s \in S} L(s) \tag{3}$$

Also let the set of all languages be

$$L_0 \stackrel{\triangle}{=} |L(1) \cup \ldots \cup L(n)| = \left| \bigcup_{k=1}^{n} L(k) \right| \tag{4}$$

Definition 2 (Coherent netizens) Two netizens u and v are coherent if and only if $\ell_0(u) = \ell_0(v)$.

Definition 3 (Coherent and social neighborhoods) Let $\Gamma(u)$ be the set of netizens who interact socially with u and $\Delta(u)$ be the coherent neighbors of u.

The total coherency is a straightforward way to evalaute the average alignment of the social and language neighbourhood of a chromosome. Thus, it is one of the candidate social and linguistic factors of the fitness functions.

Definition 4 (Partial and total coherency) The partial coherency $c(S;L_k)$ of a nonempty set of netizens S with respect to language $L_k \in L(S)$ can be found by averaging the Tanimoto coefficient of $\Gamma(u)$ and $\Delta(u)$. If $L_k \notin L(S)$, then $c(S;L_k)$ equals zero.

$$\begin{aligned} c(S;L_k) &\triangleq \frac{1}{|s \in S \wedge L_k = \ell_0(s)|} \sum_{s \in S \wedge L_k = \ell_0(s)} \tau_{\Gamma(s),\Delta(s)} \\ &= \frac{1}{|s \in S \wedge L_k = \ell_0(s)|} \sum_{s \in S \wedge L_k = \ell_0(s)} \frac{|\Delta(s) \cap \Gamma(s)|}{|\Delta(s) \cup \Gamma(s)|} \\ &= \frac{1}{|s \in S \wedge L_k = \ell_0(s)|} \sum_{s \in S \wedge L_k = \ell_0(s)} \frac{|\Delta(s) \cap \Gamma(s)|}{|\Delta(s)| + |\Gamma(s)| - |\Delta(s) \cap \Gamma(s)|} \end{aligned} \tag{5}$$

The (total) coherency is the maximum coherency over the languages which are used by the netizens comprising S.

$$\varphi(S) \triangleq \max_{L_k} c(S;L_k), \qquad 0 \leq \varphi(S) \leq 1 \tag{6}$$

A purely linguistic factor which reveals linguistic diversity within a chromosome. However, it lacks the social element of coherency.

Definition 5 (Partial and total density) The partial density $d(S)$ of a nonempty set of netizens S with respect to language $L_k \in L(S)$ is defined as the ratio of the number of netizens whose predominant language is L_k to $|S|$. If $L_k \notin L(S)$, then $d(S;L_k) = 0$.

$$d(S;L_k) \triangleq \frac{|s \in S \wedge L_k = \ell_0(s)|}{|S|} \tag{7}$$

The (total) density is the maximum density over the languages which are used by the netizens comprising S.

$$\vartheta(S) \triangleq \max_{L_k} d(S;L_k), \qquad 0 \leq \vartheta(S) \leq 1 \tag{8}$$

The language set of each netizen was determined by observing the frequencies of the languages used. This posed a problem only for very few accounts, most of which were official ones and, hence, trilingual almost by definition. In these isolated cases, ties were broken randomly. Each language was identified through two techniques used in conjunction following the example of Eisenstein (2015). Each post was scanned for words unique to each of the L_0 languages. Additionally, the bigrams and trigrams, namely character sequences, for each tweet were compared to that of established text corpora.

The next factor considers the geolocation aspect of the Twitter data and shows whether a given chromosome is geographically disperse.

Definition 6 (Spatial distance and inverse dispersion) The spatial factor between two netizens u_1 and u_2 is a function of the actual geographic distance $q(u_1,u_2)$ between them as follows:

$$y(u_1,u_2;\eta_0,\delta_0)\stackrel{\triangle}{=}\begin{cases}1, & 0\leq q(u_1,u_2)\leq\delta_0\\ \frac{\delta_0}{q(u_1,u_2)}, & \delta_0<q(u_1,u_2)\leq\eta_0\delta_0\\ 0, & q(u_1,u_2)>\eta_0\delta_0\end{cases} \tag{9}$$

The factor $\psi(S)$, the inverse dispersion, is the ratio of the minimum distance between any netizens divided by the maximum one. In contrast to φ and ϑ, the maximum distance alone is not indicative of the compactness of a chromosome. Moreover, it has the same scale with the other factors. Thus:

$$\psi(S)\stackrel{\triangle}{=}\frac{\min_y y(u_1,u_2)}{\max_y y(u_1,u_2)},\qquad u_1,u_2\in S \tag{10}$$

Each netizen was mapped to the geolocation most frequently associated with his tweets. The latitude and longitude were compared against the national and regional borders of Luxembourg and, if necessary, were clipped to fit frontier areas. The GIS information provided freely by the Global Administrative Areas database (GADM)[1] was highly valuable.

3.3 TensorFlow

TensorFlow is a low level, open source, tensor oriented framework originally developed from the Google Brain team in order to efficiently implement computationally intensive deep learning tasks such as backpropagation and distributed learning such as Adam and AdaGrad in neural networks. It falls under the stateful dataflow graph computational paradigm where the entire computation is represented as a tree with each vertex representing a tensor operation. The latter refers both to a broad spectrum of elementary operations such as Hadamard product and column oriented operations and to advanced operations such as Kruskal decomposition, Tucker factorization, and the computation of tensor eigenvectors. For instance, the Frobenius tensor norm $\|\mathscr{T}\|_F$

$$\|\mathscr{T}\|_F\stackrel{\triangle}{=}\left(\sum_{i_1=1}^{I_1}\cdots\sum_{i_p=1}^{I_p}\mathscr{T}^2[i_1,\ldots,i_p]\right)^{\frac{1}{2}}=\left(\sum_{(i_1,\ldots,i_p)}\mathscr{T}^2[i_1,\ldots,i_p]\right)^{\frac{1}{2}} \tag{11}$$

is efficiently computed in parallel.

[1] www.gadm.org

TensorFlow is designed to exploit the computational potential of multiple CPUs, GPUs, and TPUs. The latter are special hardware processing units whose instruction set comprises of rudimentary tensor operators such as parallel addition, elementwise and Kronecker multiplication, and the computation of unit tensors. When developing the source code, TensorFlow allows the construction of long symbolic expressions through placeholders which are initialized within the context of a session. TensorFlow sessions can be executed serially or in parallel, each with a different context.

3.4 The Algorithm

TENSOR-G2 is outlined in algorithm 1, while its full set of parameters is summarized in table 2. Also, for clarity TENSOR-G2 and its differences from its predecessor are depicted in figure 2.

Algorithm 1 TENSOR-G2 algorithm (double mutations enabled)

Require: Parameter set as in table 2 and termination criterion T
Ensure: Spatiolinguistic structure is heuristically discovered
create chromosome population of size K_0 with J_0 as in (12)
repeat
  evaluate fitness of each chromosome
  retain the $\lceil \alpha_0 K_0 \rceil$ fittest chromosomes
  retain the $\lceil \beta_0 K_0 \rceil$ least fit chromosomes
  crossover the remaining $I_0 = K_0 - \lceil \alpha_0 K_0 \rceil - \lceil \beta_0 K_0 \rceil$ chromosomes
  select the I_0 fittest of the $\Theta\left(I_0^2\right)$ new chromosome pairs
  with probability p_γ **mutate** a chromosome
  with probability p_γ **mutate** another gene of the same chromosome
  if more than L_0 communities exist **then**
    with probability p_ζ:
    for all community pairs in the best fitting chromosome **do**
      if any two communities are spatiolinguistically close **then**
        merge these communities **and update** J_0
      end if
    end for
  end if
until T is **true**
return $\{C_k\}$

The chromosomes C_k have a very simple form, since it suffices that each netizen be assigned to one community. Each such assignment corrsponds to a valid tensor clustering, whose quality of course needs to be checked. Therefore, it suffices to consider as chromosomes vectors of length n, namely the total number of netizens available in the dataset, where each such vector is an integer between 0 and $J_0 - 1$.

This simplifies TENSOR-G2 and at the same time ensures that the genetic operators operate in a sane manner and consistent.

However, as the true number of communities J_0 is unknown, it is estimated based on classical, i.e. non Bayesian, signal estimation techniques and linguistic observations from Kershaw et al (2015) and modeled as a normal random variable where

$$J_0 \sim \mathcal{N}(1+3L_0, L_0) \tag{12}$$

The intuition in favor of the normal distribution is that the latter maximizes the differential entropy in the distribution class having the same variance. Therefore, the normal distribution models the maximum uncertainty for J_0 given all available information about it. Additionally, the properties of the normal distribution limit the range of J_0 to $[1, 1+6L_0]$. However, from a modeling perspective other distributions such as the chi square or the lognormal might also make sense since they both have nonnegative support.

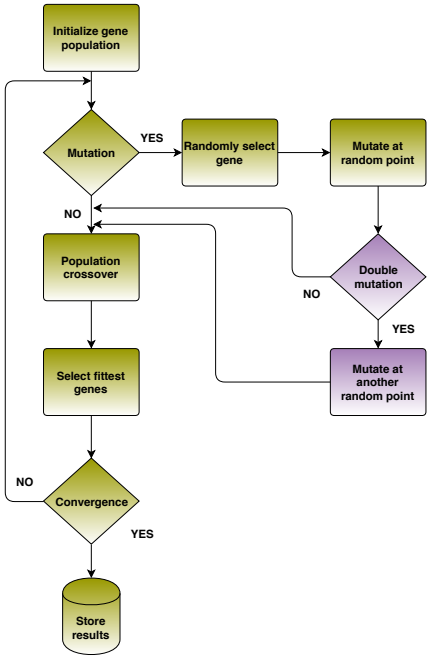

Fig. 2 TENSOR-G flow chart (green) with TENSOR-G2 enhancements (blue).

Based on the building blocks of subsection 3.2, two fitness functions for evaluating the fitness of a chromosome C_k were created. The first is the harmonic mean of coherency and inverse dispersion:

$$g_1(C_k) \triangleq \mathrm{H}(\varphi(C_k), \psi(C_k)) = 2\left(\frac{1}{\varphi(C_k)} + \frac{1}{\psi(C_k)}\right)^{-1} \tag{13}$$

Table 2 Full set of TENSOR-G2 parameters.

Parameter	Meaning	Value
n	Number of tweets	579
α_0	Percentage of best fit clusterings kept in each iteration	0.1
β_0	Percentage of worst fit clusterings kept in each iteration	0.1
γ_0	Threshold of first order difference in T_0	0.05
γ_1	Threshold of second order difference in T_0	0.05
γ_2	Terminating threshold in criterion T_1	0.85
δ_0	Geolocation distance for maximum assortativity	20 Km
η_0	Threshold for geolocation assortativity in terms of δ_0	8
M_0	Minimum number of iterations in criterion T_0	32
M_1	Maximum number of iterations in criterion T_0	1024
N_0	Number of instances of TENSOR-G2 executed	2048
b	Random sample size for merging communities	eq.(17)
L_0	Total number of languages in the tweets	4
J_0	Initial estimation of the true community number	eq.(12)
p_γ	Probability distribution for single and double mutation	Poisson
p_ζ	Probability distribution for agglomeration operation	Poisson

while the second is the harmonic mean of density and inverse dispersion:

$$g_2(C_k) \triangleq \mathrm{H}\left(\vartheta(C_k), \psi(C_k)\right) = 2\left(\frac{1}{\vartheta(C_k)} + \frac{1}{\psi(C_k)}\right)^{-1} \tag{14}$$

The harmomic mean is chosen as the connector of the factors of the fitness functions because of its tendency to be closer, from the above, to the lowest of its arguments, being thus a rather conservative yet relaxed mean. Also, the harmonic mean is known to be robust in outliers.

Having defined the fitness functions, the objective function G of TENSOR-G2 in the j-th iteration is the averaged sum of the chosen fitness function applied to each of the chromosomes:

$$G\left(C^{[j]}\right) = \frac{1}{\left|C_k \in C^{[j]}\right|} \sum_{C_k \in C^{[j]}} g(C_k) \tag{15}$$

where $C^{[j]}$ denotes the set of all chromosomes during the j-th iteration.

Concerning the genetic operations, crossover is easy to implement as it suffices to decide a random place to cut two chromosomes and exchange the corresponding parts so that two new integer vectors of length n are formed. Both the single and the double mutation work equally efficiently. To ensure a certain degree of gene variability, in each iteration the $\lceil \alpha_0 J_0 \rceil$ fittest chromosomes can be chosen to remain intact. Therefore, genes proven to work until that iteration are preserved. However, this policy may result in the entrapment of the genetic algorithm to a local maximum. To avoid this, the $\lceil \beta_0 J_0 \rceil$ least fit chromosomes can be also retained.

This is the intuition behind the last operation of community merge. It comes from clustering theory and from the fact that J_0 is essentially estimated. Thus, its value may vary if the fitness function of the best chromosome points to that direction and as long as there are enough communities to support L_0 languages. With probability p_ζ the best fitting chromosome is decoded and the communities are formed. Then, all possible pairs are inspected and a given pair C_i and C_j is merged if and only if:

$$g(S) \leq \max\left\{g(C_i), g(C_j)\right\} \tag{16}$$

where S is a new community randomly formed from b netizens from each of the C_i and C_j. A successful merge equals a cluster agglomeration. However, this is a potentially expensive step. In our case, b was chosen as:

$$b \triangleq \min\left\{\max\left\{\lceil\log|C_i|\rceil, \lceil\log\left|C_j\right|\rceil\right\}, |C_i|, \left|C_j\right|\right\} \tag{17}$$

With equation (17) two clusters which are comparable in size are logarithmically sampled, wheras a very large cluster is compared to a small one, the latter is taken entirely into consideration.

Two termination conditions were coded into TENSOR-G2, namely T_0 and T_1. The former examines the first and second order absolute discrete differences of the objective function, namely it terminates the execution of the genetic algorithm if the following conditions are simultaneously met:

$$\begin{aligned}\left|G\left(C^{[j]}\right) - G\left(C^{[j-1]}\right)\right| &\leq \gamma_0\\ \frac{1}{2}\left|G\left(C^{[j]}\right) - 2G\left(C^{[j-1]}\right) + G\left(C^{[j-2]}\right)\right| &\leq \gamma_1\end{aligned} \tag{18}$$

The termination criterion T_1 monitors the performance of each individual chromosome. Since each fitness function ranges from 0 to 1 because of the range of the factors ϑ, φ, and ψ and of the properties of the harmonic mean. T_1 terminates when a chromosome C_k achieves:

$$g(C_k) \geq \gamma_2 \tag{19}$$

A secondary failsafe mechanism runs in parallel with the above criteria. Specifically, there is a hardcoded yet parameterisable minimum number of M_0 iterations as well as a maximum of M_1 iterations.

4 Results

4.1 Data Synopsis

The dataset contains information about $n = 579$ Luxemburgian netizens, 217 of whom were identified as predominantly tweeting in English, 163 in French, 152 in German, and 47 in Luxembourgish. With the notable exception of the small Luxembourgish cluster, the remaining sample is quite balanced in terms of language representation. Table 3 contains more information about these netizens.

Table 3 Netizen statistics.

Property	Value
Follows and replies	7571
Spatial connections	1933
min, max, avg degree	1, 31, 17
Monolinguals	29
Bilinguals	196
Trilinguals	354

4.2 Clustering Assessment

Because of the probabilistic selection of J_0, it makes sense to ask whether this leads to compact communities. Figure 3 shows the normalised behavior of two cluster compactness metrics as a function of J_0. As the genetic algorithm is stochastic, each point in this figure represents the mean value of N_0 runs. In order to evaluate the clustering performance of TENSOR-G2, two methodologies were applied. The first, denoted by F_1, is a special case of the *normalised mutual information*. The latter is defined as:

$$F_1 \stackrel{\triangle}{=} \frac{2I_c}{H_c + H_n}, \qquad 0 \leq F_1 \leq 1 \tag{20}$$

The metric F_1 is the ratio of the mutual information I_c between clusters, whereas H_c and H_n are the entropies of the cluster distribution and the netizen distribution respectively. The bigger F_1 is, the better the clustering is in the sense that clusters tend to be mutually exclusive from an information theoretic perspective.

The metric F_2 takes into account the linguistic element and it is a special case of the metric known in the data mining community as the *impurity*. The latter is defined as the mean number of misclassifications, assuming that each data point in the cluster is assigned to the majority class within that cluster. In our case, each netizen is assigned to the majority language of the cluster. Thus, F_2 equals:

$$F_2 \stackrel{\triangle}{=} \frac{1}{\left|c_j \in c^*\right|} \sum_{c_j} (1 - d(c_j)), \qquad 0 \leq F_2 \leq 1 \tag{21}$$

where $c^* = \left\{c_j\right\}$ is the set of tensor clusters and $d(\cdot)$ is the density function of subsection 3.2. A lower value F_2 indicates better clustering in the sense that the clusters are linguistically compact.

Concerning the number of communities from figure 3 follows that the ideal number of clusters which combines a high F_1 value and a low F_2 value lies between 6 and 8, a very narrow range in comparison to the search line of I_0. This relatively low number, which is close to L_0, can be attributed to the dispersion of predominantly French speaking netizens among the more adamant German speakers and the omnipresent English ones. Also, the small Luxembourgish speaking cluster, although very highly concentrated, does not influence the overall clustering result much because of its small size.

From figure 3 also follows that the termination criterion T_2 tends to outperform T_1. One explanation is that T_2 is directly looking for a good clustering expressed as a high scoring chromosome, whereas T_1 is averaging the performance of a large number of chromosomes.

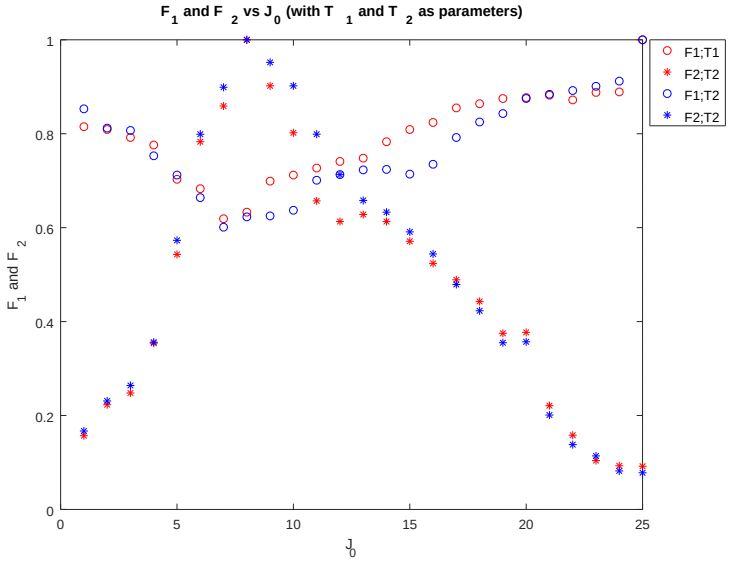

Fig. 3 Compactness metrics F_1 and F_2 vs J_0 (normalised values, T_1 and T_2 are parameters).

4.3 Performance Evaluation

The performance of the proposed genetic algorithm has three major components. The first is the location of the cost components, the second is the analysis of the number of iterations, and the frequency of mutations. Especially the double mutations are of high importance, since they are very rare but also somewhat expensive operation which is also the only way to change the initial cluster estimation I_0.

The main costly operations of TENSOR-G2 are the chromosome population generation, the crossover in each iteration, and the rare double mutation operation, even though it is based on radnom sampling in order to reduce the computational cost. The single mutation operation on the other hand is very efficient operation since it requires only the knowledge of two random numbers. Thus, the total cost of the proposed genetic algorithm can be expressed by the ordered triplet (P_g, P_c, P_d). Knowing these three components the mean value of the total computational cost P, which is a random variable by definition, of TENSOR-G2 is very well approximated as:

$$\mathrm{E}[P] \approx P_g + \mathrm{E}[T] P_c + p_\zeta^2 P_d \tag{22}$$

where $\mathrm{E}[T]$ is the mean number of iterations, which in turn depends on the termination criterion.

Regarding the termination criteria, T_2 seems to systematically lead to a lower number of iterations since:

$$\begin{aligned} \mathrm{E}[T_1] &\approx \frac{M_1}{3} \\ \sqrt{\mathrm{Var}[T_1]} &\approx 24 \end{aligned} \tag{23}$$

whereas:

$$\begin{aligned} \mathrm{E}[T_2] &\approx \frac{M_1}{4} \\ \sqrt{\mathrm{Var}[T_2]} &\approx 18 \end{aligned} \tag{24}$$

Finally, regarding mutations there are two directions of interest. These are the distribution of the total number of single mutations N_γ and double mutations B_γ over the total number of executions of the genetic algorithm as well as the distribution of iterations T_γ to the first mutation from the beginning of the corresponding execution. All distributions are taken conditionally that at least one mutation takes place. In figures 4, 5, and 6 the empirical distributions of N_γ, B_γ, and T_γ are shown respectively.

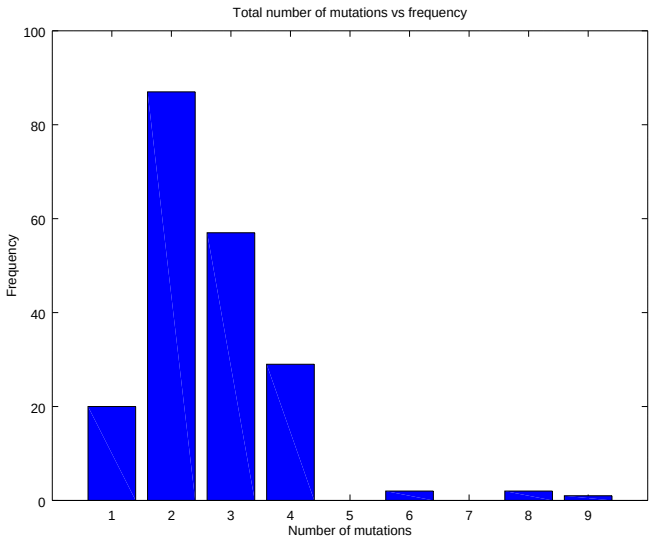

Fig. 4 Empirical conditional distribution of N_γ.

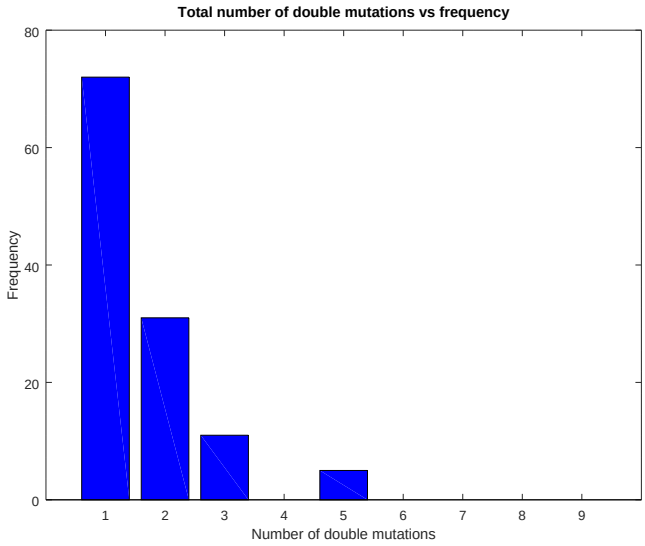

Fig. 5 Empirical conditional distribution of B_γ.

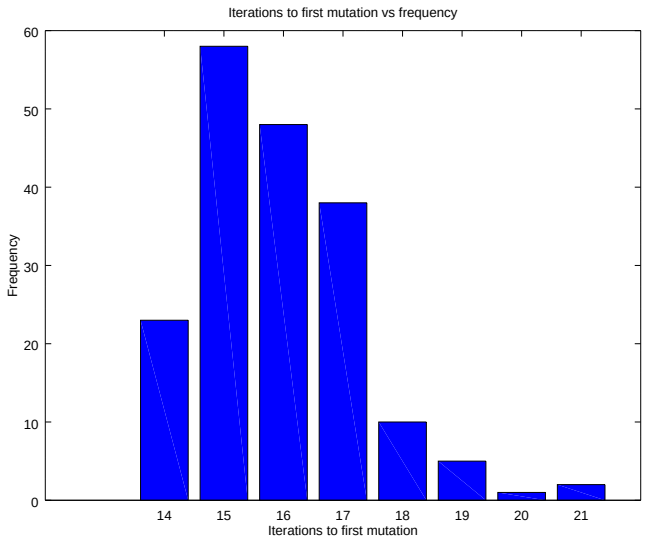

Fig. 6 Empirical conditional distribution of T_γ.

5 Conclusions

This article presents TENSOR-G2, a genetic algorithm for spatiosocial sparse tensor clustering implemented in Google TensorFlow. This type of tensors contain geolocation and linguistic data harvested from tweets from the Grand Duchy of Luxembourg, a country with thriving language communities, even of uneven size, and strong digital presence. The spatiolinguistic communities obtained by TENSOR-G2 using two

different fitness functions were evaluated using two well known metrics from data mining, namely the impurity and the normalised mutual information. Additonally, the mean computational cost can be reduced to three major factors which are probabilistically connected.

This work can be improved in many aspects. Since the total computational cost depend on the termination criterion, an advanced set of parameterless stopping conditions based on competing factors such as the one proposed in Kanavos et al (2017). Additionally, each netizen is assigned to a single spatiosocial cluster, which might not be true, especially for polyglots and polymaths. Thus, a fuzzy clustering scheme might be more desirable in certain scenarios. Moreover, new fitness functions can be develop which not only evaluate the overall chromosome status but they also do it efficiently. Finally, more detailed language change models can be integrated into the fitness function. Finally, since tensors are particularly suited to diffusion phenomena, their application to spatiosocial data in general and to the propagation of language changes should be thoroughly examined.

Acknowledgements This article is part of VALIS.IX, an independent research and innovation project promoting the study of European lingustic diversity and heritage. Moreover, this article is part of project Tensor 451, a long term research initiative whose primary objective is the development of novel, scalable, numerically stable, and interpretable tensor analytics.

References

Androutsopoulos J (2011) Language change and digital media: A review of conceptions and evidence. Standard languages and language standards in a changing Europe

Backstrom L, Sun E, Marlow C (2010) Find me if you can: Improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on World Wide Web, ACM, pp 61–70

Beasley JE, Chu PC (1996) A genetic algorithm for the set covering problem. European Journal of Operational Research 94(2):392–404

Booker LB, Goldberg DE, Holland JH (1989) Classifier systems and genetic algorithms. Artificial intelligence 40(1-3):235–282

Cardoso JF (1990) Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In: ICASSP-90, IEEE, pp 2655–2658

Croft W (2003) Mixed languages and acts of identity: An evolutionary approach. The mixed language debate: Theoretical and empirical advances 145:41

Darwin C (1859) On the origin of species by means of natural selection. John Murray

Davis L (1991) Handbook of genetic algorithms. CUMINCAD

Dawkins R (2006) The selfish gene: Thirtieth anniversary edition. Oxford university press

De Jong K (1988) Learning with genetic algorithms: An overview. Machine learning 3(2):121–138

De Lathauwer L, Vandewalle J (2004) Dimensionality reduction in higher-order signal processing and rank-$(r_1, r_2, \ldots, r_n)$ reduction in multilinear algebra. LAA 391:31–55

Dixon RM (1997) The rise and fall of languages. Cambridge University Press

Djugasvilii JV (1950) Marxism and problems of linguistics. In: Pravda

Donoso G, Sánchez D (2017) Dialectometric analysis of language variation in twitter. arXiv preprint 170206777

Drakopoulos G (2016) Tensor fusion of social structural and functional analytics over Neo4j. In: IISA, IEEE

Drakopoulos G, Kanavos A (2016) Tensor-based document retrieval over Neo4j with an application to PubMed mining. In: IISA, IEEE

Drakopoulos G, Kanavos A, Karydis I, Sioutas S, Vrahatis AG (2017a) Tensor-based semantically-aware topic clustering of biomedical documents. Computation 5(3)

Drakopoulos G, Kanavos A, Mylonas P, Sioutas S (2017b) Defining and evaluating Twitter influence metrics: A higher order approach in Neo4j. SNAM 71(1)

Drakopoulos G, Kanavos A, Tsakalidis K (2017c) Fuzzy random walkers with second order bounds: An asymmetric analysis. Algorithms 10(2)

Drakopoulos G, Stathopoulou F, Tzimas G, Paraskevas M, Mylonas P, Sioutas S (2017d) A genetic algorithm for discovering linguistic communities in spatiosocial tensors with an application to trilingual Luxembourg. In: MHDW

Dunlavy DM, Kolda TG, Acar E (2011) Temporal link prediction using matrix and tensor factorizations. TKDD 5(2):10

Eisenstein J (2015) Sociolinguistic variation in online social media. In: 2015 AAAS Annual Meeting

Eisenstein J, O'Connor B, Smith NA, Xing EP (2014) Diffusion of lexical change in social media. PLoS one 9(11)

Eleta I, Golbeck J (2012) Bridging languages in social networks: How multilingual users of twitter connect language communities? Proceedings of the American Society for Information Science and Technology 49(1):1–4

Goel R, Soni S, Goyal N, Paparrizos J, Wallach H, Diaz F, Eisenstein J (2016) The social dynamics of language change in online networks. In: International Conference on Social Informatics, Springer, pp 41–57

Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning. Machine learning 3(2):95–99

Hale M (2007) Historical linguistics: Theory and method. Wiley-Blackwell

Hale SA (2014) Global connectivity and multilinguals in the Twitter network. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 833–842

Holland JH (1992) Genetic algorithms. Scientific American 267(1):66–73

Hong L, Convertino G, Chi EH (2011) Language matters in Twitter: A large scale study. In: ICWSM

Kanavos A, Drakopoulos G, Tsakalidis A (2017) Graph community discovery algorithms in neo4j with a regularization-based evaluation metric. In: WEBIST

Karatzoglou A, Amatriain X, Baltrunas L, Oliver N (2010) Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In: Proceedings of the fourth ACM conference on Recommender systems, ACM,

pp 79–86
Kershaw D, Rowe M, Stacey P (2015) Language innovation and change in on-line social networks. In: Proceedings of the 26th ACM Conference on Hypertext and Social Media, ACM, pp 311–314
Kershaw D, Rowe M, Noulas A, Stacey P (2017) Birds of a feather talk together: User influence on language adoption. In: Proceedings of the 50th Hawaii International Conference on System Sciences
Kirk NA, Mees B (2006) Stalin, Marr and the struggle for a Soviet linguistics. Verbatim 31(3)
Kolda TG, Bader BW (2009) Tensor decompositions and applications. SIAM Review 51(3):455–500
Kontopoulos S, Drakopoulos G (2014) A space efficient scheme for graph representation. In: ICTAI, IEEE
Labov W (2001) Principles of linguistic change volume 2: Social factors. Language in society 29
Labov W (2007) Transmission and diffusion. Language 83(2):344–387
Lu S, Wang S, Zhang Y (2016) A note on the weight of inverse complexity in improved hybrid genetic algorithm. Journal of medical systems 40(6):1
Matras Y (2013) Languages in contact in a world marked by change and mobility. Revue française de linguistique appliquée 18(2):7–13
Matsumoto K (2010) The role of social networks in the post-colonial multilingual island of Palau: Mechanisms of language maintenance and shift. Multilingua-Journal of Cross-Cultural and Interlanguage Communication 29(2):133–165
Maybaum R (2013) Language change as a social process: Diffusion patterns of lexical innovations in Twitter. In: Annual Meeting of the Berkeley Linguistics Society, pp 152–166
Michael L, Bowern C, Evans B (2014) Social dimensions of language change. In: Bowern C, Evans B (eds) Routledge Handbook of Historical Linguistics, Routledge, pp 484–502
Milroy J, Milroy L (1985) Linguistic change, social network and speaker innovation. Journal of linguistics 21(02):339–384
Milroy L (1980) Language and social networks, 2nd edn. Blackwell Oxford
Nevalainen T (2015) Social networks and language change in Tudor and Stuart London–only connect? English Language and Linguistics 19(2):269–292
Nion D, Sidiropoulos ND (2010) Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar. IEEE Transactions on Signal Processing 58(11):5693–5705
Pakendorf B (2014) Historical linguistics and molecular anthropology. In: Bowern C, Evans B (eds) Routledge Handbook of Historical Linguistics, Routledge
Papalexakis E, Doğruöz AS (2015) Understanding multilingual social networks in online immigrant communities. In: 24th WWW, ACM, pp 865–870
Rahmat-Samii Y, Michielssen E (1999) Electromagnetic optimization by genetic algorithms. Microwave Journal 42(11):232–232
Shashua A, Hazan T (2005) Non-negative tensor factorization with applications to statistics and computer vision. In: ICML, ACM, pp 792–799

Tanese R (1989) Distributed genetic algorithms for function optimization. University of Michigan

Trudgill P (2011) Social structure, language contact and language change. The SAGE Handbook of Sociolinguistics pp 236–249

Wang S, Yang M, Li J, Wu X, Wang H, Liu B, Dong Z, Zhang Y (2017) Texture analysis method based on fractional Fourier entropy and fitness-scaling adaptive genetic algorithm for detecting left-sided and right-sided sensorineural hearing loss. Fundamenta Informaticæ151(1-4):505–521

Weinreich U, Labov W, Herzog MI (1968) Empirical foundations for a theory of language change. University of Texas Press

Westin CF, Maier SE, Mamata H, Nabavi A, Jolesz FA, Kikinis R (2002) Processing and visualization for diffusion tensor MRI. Medical image analysis 6(2):93–108