A Genetic Algorithm For Discovering Linguistic Communities In Spatiosocial Tensors With An Application To Trilingual Luxemburg

Georgios Drakopoulos[4], Fotini Stathopoulou[1], Giannis Tzimas[2], Michail Paraskevas[3], Phivos Mylonas[4], and Spyros Sioutas[4]

[1] University of Athens, Faculty of English Language and Literature
Zografou Campus, Athens 15784, Hellas
fstathop@uoa.gr
[2] Technological and Educational Institution of Western Greece
Mesolonghi Campus, Hellas
tzimas@teimes.gr
[3] Technological and Educational Institution of Western Greece
Patras Campus, Hellas
mparask@teiwest.gr
[4] Ionion University, Department of Informatics
Tsirigoti Square 7, Kerkyra 49100, Hellas
{drakop, fmylonas, sioutas}@ionio.gr

Abstract. Multimodal social networks are omnipresent in Web 2.0 with virtually every human communication action taking place there. Nonetheless, language remains by far the main premise such communicative acts unfold upon. Thus, it is statutory to discover language communities especially in social data stemming from historically multilingual countries such as Luxemburg. An adjacency tensor is especially suitable for representing such spatiosocial data. However, because of its potentially large size, heuristics should be developed for locating community structure efficiently. Linguistic structure discovery has a plethora of applications including digital marketing and online political campaigns, especially in case of prolonged and intense cross-linguistic contact. This conference paper presents TENSOR-G, a flexible genetic algorithm for approximate tensor clustering along with two alternative fitness functions derived from language variation or diffusion properties. The Kruskal tensor decomposition serves as a benchmark and the results obtained from a set of trilingual Luxemburgian tweets are analyzed with linguistic criteria.

Keywords: language variation, multilingual social networks, cross cultural communication, geolocation edges, tensor algebra, multilayer graphs, functional analytics, genetic algorithms, heuristics, spatiosocial data

1 Introduction

Diachronically language remains the primary communication vehicle. Thus, not only is by definition a complex social phenomenon, but also a major generator

of massive and structured, or often semistructured, humanistic data. The latter is evident in the case of multimodal social media like Twitter, Facebook, and LinkedIn where language tends to be short and informal with non-linguistic elements such as memes and hashtags replacing a sentence or a part thereof. Although text is largely displaced by images in Instagram or by video in Snapchat, aided by the deployment of 4G mobile networks [22], language is far from expunged from the digital sphere.

Multilingualism is present for various social or historical reasons in various countries such as Canada, Switzerland, Belgium, and Luxemburg to name a few. This is strongly reflected in Twitter netizen interaction, as there is no single *lingua franca* in terms of data volume. Instead, in this sense Japanese and Spanish are roughly equivalent with English and Indonesian follows closely [13][27]. This operational frame leads to two major effects. First, language undergoes a ceaseless alteration driven by external factors [7]. This flux of linguistic changes includes syntax, forms, emoticons, abbreviations, phonetic spellings, and neologisms [20]. Second, digital linguistic communities are formed whose compactness depends heavily on social, spatial, and linguistic factors such as overall status, region, and dialect respectively [48]. This form of online activity which consists of social, namely linguistic, and spatial components is termed *spatiosocial*.

Traditionally, Twitter interaction is represented as a *follow* or a *reply* graph or a combination thereof. In the latter case edge weight or valence is determined by the intensity of these two network functions. However, for complex network functionality, such as that between multilingual and geographically dispersed netizens, a sophisticated representation is required. One solution is multilayer graphs, namely graphs whose edges have one and only one label and each vertex pair can be connected with more than one edges as long as these edges have pairwise distinct labels. The algebraic counterpart of a multilayer graph, which is of combinatorial nature, is an *adjacency tensor*, which is the analogous of adjacency matrices for ordinary social graphs.

The primary contribution of this conference paper is TENSOR-G, a genetic algorithm tailored for locating linguistic communities in multilayer graphs containing spatiosocial data and represented as compressed adjacency tensors. Two alternative fitness functions based on sociolinguistic notions, specifically of how resistant to change a language is with social networks acting as an explanatory framework, are outlined as part of TENSOR-G, though other appropriate ones can be selected. The proposed genetic algorithm has been applied to Twitter data from Luxemburg, a trilingual country with rich online activity.

The structure of this conference paper follows. The scientific literature is reviewed in section 2. Fundamental sociolinguistic concepts necessary to develop and evalute the performance of TENSOR-G are introduced in section 3, whereas the heuristic algorithm itself is outlined in section 4. The conference paper concludes with section 6 where the groundwork for future work is laid. Finally,

paper notation is summarized in table 1. Tensors are printed in capital italics and vectors in small boldface.

Table 1. Paper notation.

Symbol	Meaning
$\triangleq$	Definition or equality by definition
$\{s_1, \ldots, s_n\}$	Set consisting of elements $s_1, \ldots, s_n$
$\lvert S \rvert$ or $\lvert \{s_1, \ldots, s_n\} \rvert$	Set cardinality
$S_1 \setminus S_2$	Asymmetric set difference S_1 minus S_2
τ_{S_1,S_2}	Tanimoto similarity coefficient between sets S_1 and S_2
$\langle x_k \rangle$	Sequence of elements x_k
$(s_1, \ldots, s_n)$	Tuple of elements $s_1, \ldots, s_n$
$\lVert \mathcal{T} \rVert_F$	Tensor Frobenius norm
$\circ_n$	Vector outer product along dimension n
$\mathcal{H}(x_1, \ldots, x_n)$	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

A mainstay of sociolinguistics is the language evolution process [35][40]. The latter is treated as crucial to understanding language itself [28]. As [43] states some linguistic patterns may only make sense with knowledge from outside the discipline. Change diffusion among communities by correlating linguistic variation with social factors is examined in [39]. The mechanisms of language maintenance and change in the multilingual community of Palau are studied in [36]. The universality of language change as a social phenomenon is treated in [12][1]. Since speech communities and their digital reflections differ widely, linguistic change is expected to be universally modeled [33][34]. Moreover, the latter can be cast in quantitative terms [25]. The propagation and diffusion of this change through a speech community is influenced by the structural and social properties of that community [47]. In other words, the processes of change might be the same, but the social conditions may be different enough to render variation-change-diffusion models non-transferable across languages [38]. Finally, in certain historical cases lingustics are the focus of sociopolitical dispute as suggested for instance in [30] which attributes to [45] considerable changes in the scientific administration of the former USSR.

Multilingual digital interaction is a related yet distinct line of research [25]. The ways netizens arbitrate among language groups in their social networks, focusing on social network properties, language choice, and information diffusion are the

focus of [21]. The study [26] analyzed topic-based cross-language linkings among blogs and concluded that designing for cross-cultural awareness has an impact on the underlying network structure. The connections in online social media tend to be geographically assortative [2][37]. A probabilistic characterization of macro-scale linguistic connections with respect to demographic and geographic predictors is given in [29]. The valence of Twitter connections in conjunction with linguistic changes is the focus of [23]. The diffusion of linguistic changes and their social correlations comparatively in Tudor and Stuart London is explored in [41].

Multilinear algebra or tensor analysis is the current evolution step of linear algebra [31][18]. Signal processing applications of tensor algebra include MIMO radars [42], blind source separation [6], and biomedical image analysis [49][46]. In data mining tensors have been applied to dimensionality reduction [11] to [44]. Within the context of social network analysis, multilayer graphs were the tool for sentiment analysis in Twitter [14]. In information retrieval third order tensors extended the term-document linear algebraic model to term-author-document [15] and to term-keyword-document [16] models. Finally, higher order statistics, which are closely associated with tensors, have been used to assess the performance of operating system level process scheduling policies [17].

Genetic algorithms are an offshoot of numerical optimization and machine learning [10]. This class of heuristics imitates Darwinian evolution [9][8]. Operations include *selection*, *crossover*, and *mutation* and are applied on candidate problem solutions termed *genes* [5]. Finally, the close ties between genetic algorithms and machine learning are overviewed in [24].

3 Linguistic Notions and Spatiosocial Data

In order to facilitate further discussion as well as the analysis of TENSOR-G, some preliminary notions should be outlined.

Definition 1. *Spatiosicial data have spatial and social components at minimum.*

Definition 2. *A multilayer graph G is the ordered quintuple*

$$G \triangleq (V, E, Q, \Sigma, f) \tag{1}$$

where V is the vertex set, $E \subseteq V \times V \times Q$ the edge set, Q the label set, and Σ the edge value set. The function $f : E \to \Sigma$ assigns to edges a value.

Thus, (u_1, u_2, q) denotes that u_1 and u_2 are connected by an edge whose label is q, whereas $(u_1, u_2, *)$ means that there exists at least one edge connecting u_1 and u_2 regardless of its label.

Let $L(u)$ denote the language set of vertex u, where $|L(u)| \geq 1$. Also, let $\ell(v) \in L(v)$ be the *predominant language*, namely the language more often used in online communication. Finally, if there are n vertices in total, then the total number of languages L_0 is

$$L_0 \triangleq |L(1) \cup \ldots \cup L(n)| \; = \; \left| \bigcup_{k=1}^{n} L(k) \right| \tag{2}$$

Assumption 1 *Each vertex u has a single predominant language.*

This does not imply that a netizen is obliged to post only in one language, but indicates instead which language is the most frequent.

Definition 3. *Two vertices u_1 and u_2 are coherent if and only if $\ell(u_1) = \ell(u_2)$.*

Definition 4. *The set of neighbors of u is denoted as $\Gamma(u)$, while that of coherent neighbors as $\tilde{\Gamma}(u)$. Then $0 \leq \left|\tilde{\Gamma}(u)\right| \leq |\Gamma(u)|$*

Definition 5. *The coherency $c_\ell(S)$ of a set of vertices $S \subseteq V$, $S \neq \emptyset$ for a language $\ell \in \bigcup_{s \in S} L(s)$ is the average ratio of the coherent neighbors to the total number of neighbors.*

$$c_\ell(S) \triangleq \frac{1}{|s \in S \; \wedge \; \ell \; = \; \ell_0(s)|} \sum_{s \in S \wedge \ell = \ell_0(s)} \frac{\left|\tilde{\Gamma}(s)\right|}{|\Gamma(s)|}, \qquad 0 \; \leq \; c_\ell(S) \; \leq \; 1 \tag{3}$$

Definition 6. *The density $d_\ell(S)$ of a set of vertices $S \subseteq V$, $S \neq \emptyset$ for a languge $\ell \in \bigcup_{s \in S} L(s)$ is ratio of the vertices whose predominant language is ℓ to $|S|$.*

$$d_\ell(S) \triangleq \frac{|s \in S \; \wedge \; \ell \; = \; \ell_0(s)|}{|S|}, \qquad 0 \; \leq \; d_\ell(S) \; \leq \; 1 \tag{4}$$

Next a social, in particular linguistic, and a spatial factor are introduced, upon which the fitness function of TENSOR-G will be built.

Definition 7. *Factor $\phi_\ell(u)$ expresses how easy is for a linguistic change to spread from u for language ℓ. The contagion depends on the ratio of the number of coherent neighbors to that of noncoherent ones. Thus*

$$\phi_\ell(u) \triangleq \begin{cases} \frac{1}{2} + \frac{1}{2\pi} \arctan\left(\frac{|\tilde{\Gamma}(u)|}{|\Gamma(u) \setminus \tilde{\Gamma}(u)|} \right), & \tilde{\Gamma}(u) \; \subset \; \Gamma(u) \\ 1, & \Gamma(u) \; = \; \tilde{\Gamma}(u) \end{cases} \tag{5}$$

The continuous function of $\arctan(\cdot)$ has been selected since it monotonously maps $\mathbb{R}$ to $\left[-\frac{\pi}{2}\,,\,\frac{\pi}{2}\right]$. Alternatively, a related metric $\phi'_\ell(u)$ is defined as

$$\phi'_\ell(u) \triangleq \tau_{\Gamma(u),\tilde{\Gamma}(u)} \triangleq \frac{\left|\Gamma(u) \cap \tilde{\Gamma}(u)\right|}{\left|\Gamma(u) \cup \tilde{\Gamma}(u)\right|} \; = \; \frac{\left|\Gamma(u) \cap \tilde{\Gamma}(u)\right|}{|\Gamma(u)| + \left|\tilde{\Gamma}(u)\right| - \left|\Gamma(u) \cap \tilde{\Gamma}(u)\right|} \tag{6}$$

Although both $\phi_\ell(u)$ and $\phi'_\ell(u)$ return a value in $[0\,,\,1]$, the latter is preferrable due to its superior numerical stability, especially when $\tilde{\Gamma}(u) \ll \Gamma(u)$.

Definition 8. *Factor $\psi(u_1, u_2)$ is a function of the geographic distance $d(u_1, u_2)$ between two netizens u_1 and u_2 as follows*

$$\psi(u_1, u_2) \triangleq \begin{cases} 1, & (u_1, u_2, *) \in E \wedge 0 \leq d(u_1, u_2) \leq \delta_0 \\ \frac{\delta_0}{d(u_1, u_2)}, & (u_1, u_2, *) \in E \wedge d(u_1, u_2) > \delta_0 \\ 0, & (u_1, u_2, *) \notin E \end{cases} \tag{7}$$

Finally, the following definition will be valuable in assessing the performance of the proposed genetic algorithm.

Definition 9. *When $\phi_\ell(u)$ or $\phi'_\ell(u)$ exceed a threshold η_0, then u is uncontested.*

4 Genetic Algorithm Tensor Clustering

4.1 Tensor and Multilayer Graph Representations

The spatiosocial multilayer graph was constructed as follows from Twitter data by a language sampling approach as described in the next section. First, the label set was decided to contain five elements, and thus $|Q| = 5$, namely

$$Q \triangleq \{\text{:location}, \text{:german}, \text{:french}, \text{:english}, \text{:social}\} \tag{8}$$

Notice that the spatial component is expressed by the geolocation distance, whereas there are four spatial components, namely the three languages most commonly spoken in Luxemburg, and one dimension for online interaction as indicated by the *follow* and *reply* functions. The edge labels follow the Neo4j notation [32]. The following two criteria determined whether there is digital interaction between any two netizens:

- If netizen u follows v or vice versa, then interaction is considered to exist.
- Alternatively, if either of netizens u and v mention the other, then interaction is also considered to take place.

Then, each netizen was mapped to one vertex of the graph so $|V| = n$. The connectivity conditions for any netizen pair u and v were the following:

- If u and v are interacting, then edge $(u, v, \text{:social})$ is added.
- If $d(u, v) < 8\delta_0$, then edge $(u, v, \text{:location})$ is added.
- If $\ell_0(u) = \ell_0(v)$, then the appropriately labeled edge is added.

The geolocation of each vertex was determined by the location Twitter metadata field and it was compared to the bounds of a rectangle circumscribing the borders of Luxembourg. Furthermore, the latitude and longitude pair of each tweet was checked against both the national and regional borders of Luxembourg as encoded by GIS files publicly available through the Global Administrative Areas database (GADM)[5].

[5] www.gadm.org

Concerning the language set of each vertex, two methods were used. First, the corresponding tweets were analyzed as in [19]. The starting points were words unique to a specific language. Then, by dividing the findings into linguistic classes such as phonetic spellings and lexical words a profile for each candidate language was built and it was compared to that of the three languages under investigation. Second, the frequencies of digrams and trigrams, namely byte sequences encoding Unicode characters, were compared to these from standard corpora.

Finally, $\Sigma \triangleq \{0, 1\}$, as TENSOR-G focuses on edge labels, and Σ contains token values indicating whether an edge exists or not.

As is the case with ordinary social graphs, multilayer graphs can be also represented algebraically through adjacency tensors. A *tensor* is a multidimensional vector meaning that each entry is indexed by a tuple of p non permutable integers $(i_1 \ldots i_p)$ where $1 \leq i_k \leq I_k$. Formally

Definition 10. *A p-th tensor $\mathcal{T}$, $p \in \mathbb{Z}^+$, is a linear mapping simultaneously connecting p not necessarily distinct linear spaces $\mathbb{S}_k$, $1 \leq k \leq p$.*

In this specific case the linear spaces combined to create the adjacency tensor are the netizen space, in fact twice, and the label space. Thus, the corresponding adjacency tensor $\mathcal{T}$ is of third order, specifically

$$\mathcal{T} \in \Sigma^{|V| \times |V| \times |Q|} = \{0, 1\}^{n \times n \times (L_0 + 2)} \tag{9}$$

4.2 Algorithmic Aspects

The proposed genetic algorithm TENSOR-G is outlined in algorithm 1. As with any such scheme, its development is more an art then science. For instance, being a heuristic, TENSOR-G relies heavily on random numbers which largely decide the outcome of selection and crossover operators. Finally, observe that algorithm 1 has a low conditional Kolmogorov complexity, since once the data, namely the tensor $\mathcal{T}$ and the random number sequence, are factored out, then the algorithm proper is of constant size and can be thus efficiently coded in the universal Turing machine. The use of $\langle \rho_k \rangle$ is implied throughout the algorithm.

Since the initial number of communities J_0 is unknown, it is selected semirandomly based on knowledge from linguistics and the Gaussian distribution

$$f_{J_0}(j_0) = \frac{1}{4L_0\sqrt{2\pi}} \exp\left(-\left(\frac{j_0 - 3L_0}{4L_0}\right)^2\right) \tag{10}$$

Namely, $\mathrm{E}[J_0] = 3L_0$ and $\mathrm{Var}[J_0] = 16L_0^2$. These parameters were selected based on statistical observations from [29], while the Gaussian distribution itself was chosen because it has the maximum differential entropy among all distributions with the same finite variance giving, thus, an upper limit to the variation of J_0.

Algorithm 1 Proposed Genetic Algorithm (TENSOR-G)

Require: Termination criterion τ_0, random sequence $\langle \rho_k \rangle$, tensor $\mathcal{T}$
Ensure: Linguistic communities are approximately discovered
pick number of communities J_0 semirandomly
partition $\mathcal{T}$ by assigning vertices to each community C_k, $1 \leq k \leq J_0$
repeat
  evaluate fitness of each community C_k
  retain the $\lceil \alpha_0 J_0 \rceil$ fittest communities with probability p_α
  retain the $\lceil \beta_0 J_0 \rceil$ least fit communities with probability p_β
  crossover the remaining m communities to create each possible pair
  select the m fittest of the $\Theta\left(m^2\right)$ new pairs
  choose a community pair with probability p_γ **and** mutate the pair
  with probability p_ζ:
  for all community pairs **do**
    if any two communities are spatiosocially close **then**
      merge these communities **and** update J_0
    end if
  end for
until τ_0 is **true**
return $\{C_k\}$

Each of the J_0 communities can have an arbitrary number of vertices as long as it is not empty. Note that the n netizens can be distributed to J_0 with a very large number of ways, specifically

$$\binom{n}{L_0} \;=\; \frac{\partial^{L_0}}{\partial x^{L_0}}(1+x)^n\bigg|_{x=0} \;\approx\; n^{L_0}, \qquad L_0 \;\leq\; n \tag{11}$$

To avoid this, the vertices are randomly assigned to communities. Although this may lead to less than satisfactory fitness, it is computationally efficient and its effects are eventually nullified over some iterations.

There are two fitness functions for evaluating spatiosocial communities. The first is the harmonic mean of coherency and $\bar{\psi}$, the mean value of factor ψ

$$g_1(C_k) \triangleq \mathcal{H}\left(c_\ell(C_k), \bar{\psi}\right) \;=\; 2\left(\frac{1}{c_\ell(C_k)} + \frac{|C_k|}{\sum_{u_1,u_2 \in C_k} \psi(u_1,u_2)}\right)^{-1} \tag{12}$$

while the second is the harmonic mean of density and $\bar{\psi}$

$$g_2(C_k) \triangleq \mathcal{H}\left(d_\ell(C_k), \bar{\psi}\right) \;=\; 2\left(\frac{1}{d_\ell(C_k)} + \frac{|C_k|}{\sum_{u_1,u_2 \in C_k} \psi(u_1,u_2)}\right)^{-1} \tag{13}$$

In each iteration a portion $\lceil \alpha_0 J_0 \rceil$ of the fittest communities is kept unchanged with probability p_α. This is done in order to preserve a possibly very good solution. On the other hand, it entails the potential danger that TENSOR-G is

trapped to a local maximum. For this reason, with probability p_β a segment of the $\lceil \beta_0 J_0 \rceil$ least fit communities is also retained in order to provide a (counter-intuitive) escape from such a trap.

Since TENSOR-G is designed for tensor clustering, it is imperative that each netizem is assigned to one and only one language community and that no communities become void. To this end, the crossover, selection, and mutation operations were designed to uphold these constraints in spite of their inherent randomness. Specifically, the crossover operation creates selects each possible pair of the m communities. Inside each of the $\Theta\left(m^2\right)$ pairs a number of netizens, which may be random or determinstic, is selected to be swapped. Note that m is a random variable whose values and their associated probabilities are

$$m = \begin{cases} J_0 - \lceil \alpha_0 J_0 \rceil, & p_\alpha \\ J_0 - \lceil \beta_0 J_0 \rceil, & p_\beta \\ J_0 - \lceil \alpha_0 J_0 \rceil - \lceil \beta_0 J_0 \rceil, & p_\alpha p_\beta \\ J_0, & 1 - p_\alpha - p_\beta - p_\alpha p_\beta \end{cases} \tag{14}$$

This potentially large number of community pairs is then evaluated by either g_1 or g_2 and the m fittest are selected. Finally, with probability p_γ a random pair of the m new ones is selected and only one vertex is swapped between them.

The last and optional operation of community merge may come as a surprise to the reader familiar with genetic algorithms, the reason being that merge is a typical clustering operation and not part of the standard functions a genetic algorithn performs. However, since TENSOR-G is essentially a clustering algorithm, it is worth incorporating a clustering element with probability p_ζ. Thus, TENSOR-G can partially work as an agglomerative algorithm. The condition which determines the spatiosocial proximity of two communities is that they have the same predominant language and also that a random sample of b netizen pairs has low overall value for the ψ factor. In this case

$$b \triangleq \max\left\{\log\left|C_i\right|, \log\left|C_j\right|\right\} \tag{15}$$

The primary termination criterion τ_0 was a combination of a hardcoded maximum number of M_0 iterations, with a minimum of μ_0 iterations, and of a condition that the average total fitness sum should exceed a threshold γ_0 during the past five iterations. Also, there was a secondary termination criterion τ_1 that stopped TENSOR-G when a partition achieved a fitness of γ_1.

4.3 Kruskal Decomposition

At this point the reference method, Kruskal tensor decomposition, is introduced. Given a p-th order tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_p}$ and an integer $r_0 \leq p$, the Kruskal

numer of the tensor rank, r_0 rank one tensors $\mathcal{G}_k$ and positive normalization scalars λ_k are computed such that

$$\mathcal{T} \;=\; \sum_{k=1}^{r_0} \lambda_k \, \mathcal{G}_k, \qquad \lambda_k > 0 \tag{16}$$

A rank one tensor $\mathcal{G}_p \in \mathbb{R}^{I_1 \times \ldots \times I_p}$ is one that can be written as a series of $p-1$ vector outer products [31], namely

$$\mathcal{G}_p \triangleq \mathbf{v}_1 \circ_1 \mathbf{v}_2 \circ_2 \mathbf{v}_3 \ldots \mathbf{v}_{p-1} \circ_{p-1} \mathbf{v}_p, \qquad \mathbf{v}_k \in \mathbb{R}^{I_k}, \|\mathbf{v}_k\|_2 \;=\; 1 \tag{17}$$

Note that this is a direct generalization of a rank one matrix, essentially a second order tensor $\mathcal{G}_2$, which is written as

$$\mathcal{G}_2 \triangleq \mathbf{v}_1 \circ_1 \mathbf{v}_2 \;=\; \begin{bmatrix} \mathbf{v}_1[1]\mathbf{v}_2[1] & \ldots & \mathbf{v}_1[1]\mathbf{v}_2[I_2] \\ \vdots & \ddots & \vdots \\ \mathbf{v}_1[I_1]\mathbf{v}_2[1] & \ldots & \mathbf{v}_1[I_1]\mathbf{v}_2[I_2] \end{bmatrix}, \qquad \|\mathbf{v}_1\|_2 \;=\; \|\mathbf{v}_2\|_2 \;=\; 1 \tag{18}$$

Along similar lines, a third order tensor $\mathcal{G}_3$ is defined as

$$\mathcal{G}_3 \triangleq \mathbf{v}_1 \circ_1 \mathbf{v}_2 \circ_2 \mathbf{v}_3, \mathcal{G}_3[i_1, i_2, i_3] = \mathbf{v}_1[i_1]\mathbf{v}_2[i_2]\mathbf{v}_3[i_3], \qquad \mathbf{v}_k \in \mathbb{R}^{I_k}, \|\mathbf{v}_k\|_2 = 1 \tag{19}$$

However, both computing r_0 and the exact Kruskal decomposition are NP-hard problems. Therefore, for various estimates $\hat{r}_0$ the approximate decomposition is computed instead

$$\min_{\hat{r}_0, \lambda_k, \mathcal{G}_k} \left\| \mathcal{T} - \sum_{k=1}^{\hat{r}_0} \lambda_k \mathcal{G}_k \right\|_F \;=\; \min_{\hat{r}_0, \lambda_k, \mathbf{v}_{k,j}} \left\| \mathcal{T} - \sum_{k=1}^{\hat{r}_0} \lambda_k \mathbf{v}_{k,1} \circ_1 \mathbf{v}_{k,2} \ldots \mathbf{v}_{k,p} \right\|_F \tag{20}$$

where the Frobenius norm $\|\mathcal{T}\|_F$ of a real valued tensor $\mathcal{T}$ is defined as

$$\|\mathcal{T}\|_F \triangleq \left(\sum_{i_1=1}^{I_1} \ldots \sum_{i_p=1}^{I_p} \mathcal{T}^2[i_1, \ldots, i_p] \right)^{\frac{1}{2}} \;=\; \left(\sum_{(i_1, \ldots, i_p)} \mathcal{T}^2[i_1, \ldots, i_p] \right)^{\frac{1}{2}} \tag{21}$$

4.4 Implementation Aspects

Regarding implementation, TENSOR-G was implemented in MATLAB. To conserve memory, the tensor was stored in quadruples of the form (i_1, i_2, i_3, ℓ) and the netizen properties such as their geolocation and predominant language were separately stored. In other words, the tensor was compressed with the coordinate scheme which requires four integers for every non-zero entry. Although more efficient tensor compression schemes exist [3], the coordinate method is balanced between memory conservation and simplicity. Each gene of TENSOR-G was encoded as a list quadruples for efficient manipulation. The Kruskal decomposition was already implemented in the MATLAB tensor toolbox [4].

Finally, the data was obtained by a Twitter crawler implemented in Python using the *tweepy*[6] library. The latter uses the OAuth authentication protocol in conjunction with the quadruple of Twitter generated tokens. Also, it is subject to the constraints placed by Twitter for batch data harvest.

5 Results

5.1 Data Synopsis

The tensor contains information about $n = 579$ Luxemburgian netizens, 217 of whom were identified as predominantly tweeting in English, 199 in German, and 163 in French. In overall this is a fairly balanced sample in terms of language representation. Table 2 contains more information about these netizens.

Table 2. Netizen statistics.

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

As it can be deduced, the above network is sparse since the average degree is 17. That justifies the coordinate compression scheme for the spatiosocial tensor.

5.2 Performance

The values and the characteristics of the thresholds and parameters used in TENSOR-G are summarized in table 3.

The number of communities which achieved the better overall fitness in TENSOR-G for both fitness functions was 5. Using a range of values of ± 3 around this number as $\hat{r}_0$, the lowest Frobenius difference norm was achieved for 7 communities. These number of uncontested vertices was computed using ϕ'_ℓ. For the English, German, and French respectively the clustering obtained by TENSOR-G returned 201, 170, and 126 , whereas the Kruskal decomposition yielded 192, 163, and 109. This can be attributed to the dispersion of predominantly French speaking netizens among the more adamant German speakers and the omnipresent

[6] www.tweepy.org

Table 3. TENSOR-G parameters.

Parameter	Meaning	Value
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 that must be exceeded in τ_0 to continue	0.15
γ_1	Terminating threshold in criterion τ_1	0.85
δ_0	Geolocation distance for maximum assortativity	25 Km
η_0	Threshold for declaring a vertex uncontested	0.65
M_0	Maximum number of iterations in criterion τ_0	1024
μ_0	Minimum number of iterations in criterion τ_0	32
N_0	Number of instances of TENSOR-G executed	2048
b	Random sample size for merging communities C_i and C_j	eq.(15)
L_0	Total number of languages in the tweets	3
p_α	Probability distribution for retaining best clusterings	Binomial
p_β	Probability distribution for retaining worst clusterings	Binomial
p_γ	Probability distribution for mutation	Poisson
p_ζ	Probability distribution for agglomeration check	Poisson

English ones. Also, notice that Kruskal decomposition was desgined with another minimization property in mind. Specifically, the constraint for rank 1 tensors led to more communities which are more compact but left many vertices in a contested state.

6 Conclusions

This conference paper presents TENSOR-G, a genetic algorithm for spatiosocial sparse tensor clustering. The latter contains trilingual Twitter data in English, French, and German from Luxemburg, a country with thriving language communities and strong digital presence. The communities obtained by TENSOR-G using two different fitness functions based on languistic criteria were compared to those obtained by Kruskal tensor decomposition. Although the proposed methodology is slower and more memory intensive than the benchmark, the communities of TENSOR-G were more compact from a linguistic viewpoint and also make more sense in gelocation terms.

This work can be improved in many aspects. A more detailed description of digital interaction would include separate labels for *follow* and *mention* options and possibly additional layers for other Twitter functions. Also, bidirectional connections between netizens would reveal more communication patterns, for instance how differs the communication between netizens and between a netizen and an institution or a company and whether Dunbar's number is a loose bound or not in the digital sphere. Moreover, better fitness functions can be designed

utilizing observable and measurable languge variations such as change in individual words and their spelling compared to template corpora. However, it is not necessarily true that language change proceeds horizontally in the different domains. Thus, any research of language change needs to incorporate both the similarities and the differences in mechanisms across different domains [23].

Regarding future research directions, language change results from the differential propagation of linguistic variants distributed among the linguistic repertoires of communicatively interacting netizens. From this it follows that language change is socially-mediated in two important ways. First, language is a social epidemiological process that takes place by propagating some aspect of communicative practice across a network and the organization of the social group in question can affect how a variant propagates. Second, sociocultural factors such as language ideologies, can encourage the propagation of particular variants at the expense of others in particular context. Variant selection leads to language change when it forms part of larger scale processes of differential variant propagation within the speech community. 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 conference paper has been developed within the framework of the project "Strengthening the Research Activities of the Directorate of the Greek School Network and Network Technologies", financed by the own resources of the Computer Technology Institute and Press "Diophantos" (project code 0822/001).

References

1. Androutsopoulos, J.: Language change and digital media: A review of conceptions and evidence. Standard languages and language standards in a changing Europe (2011)
2. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: Improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on World Wide Web. pp. 61–70. ACM (2010)
3. Bader, B.W., Kolda, T.G.: Efficient MATLAB computations with sparse and factored tensors. SIAM Journal on Scientific Computing 30(1), 205–231 (2007)
4. Bader, B.W., Kolda, T.G., et al.: MATLAB tensor toolbox version 2.5 (2012)
5. Booker, L.B., Goldberg, D.E., Holland, J.H.: Classifier systems and genetic algorithms. Artificial intelligence 40(1-3), 235–282 (1989)
6. Cardoso, J.F.: Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In: ICASSP-90. pp. 2655–2658. IEEE (1990)
7. Croft, W.: Mixed languages and acts of identity: An evolutionary approach. The mixed language debate: Theoretical and empirical advances 145, 41 (2003)

8. Darwin, C.: On the origin of species by means of natural selection. John Murray (November 1859)
9. Dawkins, R.: The selfish gene: Thirtieth anniversary edition. Oxford university press (2006)
10. De Jong, K.: Learning with genetic algorithms: An overview. Machine learning 3(2), 121–138 (1988)
11. De Lathauwer, L., Vandewalle, J.: Dimensionality reduction in higher-order signal processing and rank-$(r_1, r_2, \ldots, r_n)$ reduction in multilinear algebra. LAA 391, 31–55 (2004)
12. Dixon, R.M.: The rise and fall of languages. Cambridge University Press (1997)
13. Donoso, G., Sánchez, D.: Dialectometric analysis of language variation in twitter. arXiv preprint 1702.06777 (2017)
14. Drakopoulos, G.: Tensor fusion of social structural and functional analytics over Neo4j. In: Proceedings of the 6th International Conference of Information, Intelligence, Systems, and Applications. IISA 2016, IEEE (July 2016)
15. Drakopoulos, G., Kanavos, A.: Tensor-based document retrieval over Neo4j with an application to PubMed mining. In: Proceedings of the 6th International Conference of Information, Intelligence, Systems, and Applications. IISA 2016, IEEE (July 2016)
16. Drakopoulos, G., Kanavos, A., Karydis, I., Sioutas, S., Vrahatis, A.G.: Tensor-based semantically-enhanced PubMed retrieval. Computation (May 2017), accepted
17. Drakopoulos, G., Megalooikonomou, V.: An adaptive higher order scheduling policy with an application to biosignal processing. In: SSCI 2016. IEEE (December 2016)
18. Dunlavy, D.M., Kolda, T.G., Acar, E.: Temporal link prediction using matrix and tensor factorizations. TKDD 5(2), 10 (2011)
19. Eisenstein, J.: Sociolinguistic variation in online social media. In: 2015 AAAS Annual Meeting (2015)
20. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: Diffusion of lexical change in social media. PLoS one 9(11) (2014)
21. Eleta, I., Golbeck, J.: 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 (2012)
22. Ge, X., Cheng, H., Guizani, M., Han, T.: 5G wireless backhaul networks: challenges and research advances. IEEE Network 28(6), 6–11 (2014)
23. Goel, R., Soni, S., Goyal, N., Paparrizos, J., Wallach, H., Diaz, F., Eisenstein, J.: The social dynamics of language change in online networks. In: International Conference on Social Informatics. pp. 41–57. Springer (2016)
24. Goldberg, D.E., Holland, J.H.: Genetic algorithms and machine learning. Machine learning 3(2), 95–99 (1988)
25. Hale, M.: Historical linguistics: Theory and method. Wiley-Blackwell (2007)
26. Hale, S.A.: Global connectivity and multilinguals in the Twitter network. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 833–842. ACM (2014)
27. Hong, L., Convertino, G., Chi, E.H.: Language matters in Twitter: A large scale study. In: ICWSM (2011)
28. Kershaw, D., Rowe, M., Noulas, A., Stacey, P.: Birds of a feather talk together: User influence on language adoption. In: Proceedings of the 50th Hawaii International Conference on System Sciences (2017)

29. Kershaw, D., Rowe, M., Stacey, P.: Language innovation and change in on-line social networks. In: Proceedings of the 26th ACM Conference on Hypertext and Social Media. pp. 311–314. ACM (2015)
30. Kirk, N.A., Mees, B.: Stalin, Marr and the struggle for a Soviet linguistics. Verbatim 31(3) (2006)
31. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51(3), 455–500 (2009)
32. Kontopoulos, S., Drakopoulos, G.: A space efficient scheme for graph representation. In: Proceedings of the 26th International Conference on Tools with Artificial Intelligence. pp. 299–303. ICTAI 2014, IEEE (November 2014)
33. Labov, W.: Principles of linguistic change volume 2: Social factors. Language in society 29 (2001)
34. Labov, W.: Transmission and diffusion. Language 83(2), 344–387 (2007)
35. Matras, Y.: Languages in contact in a world marked by change and mobility. Revue française de linguistique appliquée 18(2), 7–13 (2013)
36. Matsumoto, K.: 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 (2010)
37. Maybaum, R.: Language change as a social process: Diffusion patterns of lexical innovations in Twitter. In: Annual Meeting of the Berkeley Linguistics Society. pp. 152–166 (2013)
38. Michael, L., Bowern, C., Evans, B.: Social dimensions of language change. In: Bowern, C., Evans, B. (eds.) Routledge Handbook of Historical Linguistics, pp. 484–502. Routledge (2014)
39. Milroy, J., Milroy, L.: Linguistic change, social network and speaker innovation. Journal of linguistics 21(02), 339–384 (1985)
40. Milroy, L.: Language and social networks. Blackwell Oxford, 2nd edn. (1980)
41. Nevalainen, T.: Social networks and language change in Tudor and Stuart London–only connect? English Language and Linguistics 19(2), 269–292 (2015)
42. Nion, D., Sidiropoulos, N.D.: Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar. IEEE Transactions on Signal Processing 58(11), 5693–5705 (2010)
43. Pakendorf, B.: Historical linguistics and molecular anthropology. In: Bowern, C., Evans, B. (eds.) Routledge Handbook of Historical Linguistics. Routledge (2014)
44. Papalexakis, E., Doğruöz, A.S.: Understanding multilingual social networks in online immigrant communities. In: 24th WWW. pp. 865–870. ACM (2015)
45. Stalin, J.V.: Marxism and problems of linguistics. In: Pravda (May 1950)
46. Tagkalakis, F., Papagiannaki, A., Drakopoulos, G., Megalooikonomou, V.: Augmenting fMRI-generated brain connectivity with temporal information. In: Proceedings of the 6th International Conference of Information, Intelligence, Systems, and Applications. IISA 2016, IEEE (July 2016)
47. Trudgill, P.: Social structure, language contact and language change. The SAGE Handbook of Sociolinguistics pp. 236–249 (2011)
48. Weinreich, U., Labov, W., Herzog, M.I.: Empirical foundations for a theory of language change. University of Texas Press (1968)
49. Westin, C.F., Maier, S.E., Mamata, H., Nabavi, A., Jolesz, F.A., Kikinis, R.: Processing and visualization for diffusion tensor MRI. Medical image analysis 6(2), 93–108 (2002)