Economica

*Economica* (2011) **78**, 723–750
doi:10.1111/j.1468-0335.2010.00844.x

# A Generalized Index of Fractionalization

By Walter Bossert†, Conchita D'Ambrosio‡ and Eliana La Ferrara††

†*University of Montréal and CIREQ* ‡*Università di Milano-Bicocca, DIW Berlin and Università Bocconi* ††*Università Bocconi and IGIER*

Final version received 13 July 2009.

This paper characterizes an index that is informationally richer than the commonly used ethno-linguistic fractionalization (ELF) index. Our measure of fractionalization takes as a primitive the individuals, as opposed to ethnic groups, and uses information on the similarities among them. Compared to existing indices, our measure does not require that individuals are pre-assigned to exogenously determined categories or groups. We provide an empirical illustration of how our index can be operationalized and what difference it makes as compared to the standard ELF index. This application pertains to the pattern of fractionalization in the USA.

## Introduction

The role of ethnic and cultural diversity has received increasing attention by economists in recent years. Numerous contributions have analysed the relationship between ethnic heterogeneity and socioeconomic outcomes, including public good provision, growth, corruption and social capital. The transmission of cultural traits and the formation of heterogeneity have also been studied theoretically and empirically. Among the first group of studies, ethnic diversity has been shown to be associated with lower growth rates (Easterly and Levine 1997), more corruption (Mauro 1995), lower contributions to local public goods (Alesina *et al.* 1999), lower participation in groups and associations (Alesina and La Ferrara 2000) and a higher propensity to form jurisdictions to sort into homogeneous groups (Alesina *et al.* 2004). For a review of contributions on the relationship between ethnic diversity and economic performance, see Alesina and La Ferrara (2005). For the formation and transmission of cultural traits see, among others, Bisin and Verdier (2000), Fernandez *et al.* (2004) and Giuliano (2007). The growing interest in these topics is likely attributable to the upward trend in migration flows and the fact that many societies are becoming increasingly heterogeneous from a cultural point of view.

Yet the economics literature does not seem to have advanced very far in the measurement of ethnic and cultural diversity. This contrasts with the breadth of the literature on the measurement of income inequality, the traditional notion of heterogeneity employed by economists. While we can rely on a variety of indices of economic inequality, and these indices have been axiomatically characterized from a theoretical point of view, the economic literature on the measurement of categorical heterogeneity is much less developed. Virtually every empirical contribution on the topic uses the so-called index of ethno-linguistic fractionalization (ELF), which is a decreasing transformation of the Herfindahl concentration index applied to population shares. The ELF index measures the probability that two randomly drawn individuals from the overall population belong to different (predefined) ethnic groups. While ELF has the advantage of being simple to compute and easy to interpret, its economic underpinnings seem inadequate. To our knowledge, the only paper that attempts to provide a theoretical background for the use of ELF is the one by Vigdor (2002). He proposes a behavioural interpretation of ELF in a model where individuals display differential altruism. He assumes that an individual's willingness to spend on local public goods depends partly on the benefits that other members of the community derive from the good, and that the weights of this altruistic component vary depending on how many members of the community share the same ethnicity of that individual. Notice that our goal here is to provide a characterization, rather than a behavioural interpretation, of a new index of fractionalization.

The implicit contention in economic models is often that different ethnic groups may have different preferences and this would generate conflicts of interest in economic decisions. It is hard to believe that population shares would be enough to capture the extent of divergence in preferences among society's members. Presumably, people of different culture or ethnicity will feel differently about each other depending on how similar they are in other dimensions. A second channel through which ethnic or cultural diversity may affect economic performance is the existence of possible skill complementarities among different types. But again, it is unlikely that simple population shares will capture the nature and extent of skill complementarities among groups.

If the rationale for including ethnic diversity effects in economic models lies in preferences or technological features, then measuring fractionalization purely as a function of population shares seems a severe limitation. Similarity between individuals should play a role. This similarity could depend, for example, on language spoken, age, educational background or employment status, to mention just a few attributes. If preferences might be induced by these other characteristics, then considering similarities between individuals will give a better understanding of the potential conflict in economic decisions. Providing a measure of fractionalization that accounts for the degree of similarity among agents therefore seems an important task.

The goal of this paper is to characterize a *generalized fractionalization index* (GELF) that takes as primitive the individuals and uses information on their similarities to measure fractionalization. We propose to use as a building block a *similarity matrix* containing pairwise similarity values $\{s_{ij}\}$ among any two individuals $i$ and $j$ in society. An entry equal to 1 in the matrix represents perfect similarity among individuals, and an entry equal to 0 represents complete dissimilarity. We then rely on four axioms to characterize GELF. The first axiom is normalization, requiring that in a society with maximal similarity our diversity index takes value 0, and in a society with maximal dissimilarity it takes a positive value. The second axiom, anonymity, requires that individuals are treated impartially; that is, our diversity measure is invariant with respect to permutations. The third axiom, additivity, imposes a separability property on our index. The fourth and last axiom, replication invariance, requires the index to be invariant with respect to replications of the population. We prove that a diversity measure satisfies these four axioms if and only if it is a positive multiple of the difference between 1 and the sum of similarity values in the matrix divided by the square of the population size. We call this generalized index GELF and show that it is a natural extension of ELF. More generally, depending on the metric used to measure similarity among individuals and on the level of aggregation of the information (that is, similarity among individuals or among groups), our index nests a number of indices used in the literature. In the limit case where the information is purely categorical (that is, similarity is either 0 or 1), our measure reduces to ELF. In richer information settings where measuring the distance among individuals is feasible and meaningful, our index conveys a broader measure of diversity. The flexibility of our formulation and its suitability to being applied in very different informational environments are advantages of the measure that we propose. Another advantage is that our index does not require that individuals are pre-assigned to exogenously determined categories or groups. Our theoretical framework can actually be used to determine an endogenous partition of society into groups. Relevant groups may be constituted by clusters of individuals who have perfect (or very high) similarity among themselves, and share the same (or very close) similarity values *vis-à-vis* the rest of society.

We also provide an empirical illustration of how GELF can be operationalized and what difference it makes as compared to the standard ELF index. This application pertains to the pattern of fractionalization in the USA. Using individual level data from the 1990 Census, we compute the two indices for all states. We find that the ranking of several states changes significantly when we use GELF rather than ELF. For example, in 1990 Hawaii was the most diverse state in terms of ethnic diversity (ELF) and California was the fifth. When we compute GELF, embedding information on similarity in income, education and employment, as well as ethnicity, Hawaii moves to 42nd place and California to 30th. This is because economic opportunities in these states are relatively more equal across races than they are in other states. The District of Columbia, on the other hand, is the second most fractionalized on the basis of ELF and becomes the most fractionalized—by a wide margin—when we use GELF. Finally, we compute *grouped* versions of the GELF index and show how each variable contributes to the pattern of similarity among races.

Our paper is related to several strands of the literature. First, it naturally relates to the economics literature on ethnic diversity and its economic effects. While the bulk of this literature does not focus on the specific issue of measurement, a few contributions do. As the majority of applications have used language as a proxy for ethnicity, some authors have criticized the use of ELF on the grounds that linguistic diversity may not correspond to ethnic diversity. Among these, Alesina *et al.* (2003) have proposed a classification into groups that combines information on language with information on skin colour. These authors propose three measures of fractionalization: one purely linguistic, one related to religion, and one that broadly defines ethnicity by combining language and skin colour. Note that this approach differs from ours because it defines ethnic (or linguistic, or religious) categories on the basis of certain criteria and then applies the ELF formula to the resulting number of groups.

Other authors, in particular Fearon (2003), have criticized standard applications of ELF on the grounds that they would fail to account for the salience of ethnic distinctions in different contexts. For example, the same two ethnic groups may be allies in one country and opponents in another, and using simply their shares in the population would fail to capture this. We share Fearon's concerns on this point, and indeed we hope that our index can be a first step towards incorporating issues of salience in the measurement of fractionalization, albeit in a simplistic way. In particular, if one thinks that differences in income, or education, or any other measurable characteristic, may be the reason why ethnicity matters only in certain contexts, our GELF index already weighs ethnic categories by their salience.

Turning to the notion of distance among ethnic groups, relatively little has been done. Using a heuristic approach, Laitin (2000) and Fearon (2003) rely on measures of distance between languages to assess how different linguistic groups are across countries. In particular, in his 2003 contribution Fearon proposes a measure of 'cultural fractionalization' that adapts Greenberg's (1956) formula by weighing population shares with a resemblance factor that depends on the number of shared classifications between any two languages. This measure intuitively captures the expected cultural distance between two people drawn at random from the population. As we show below, this measure can be derived as a special case of our GELF index. Caselli and Coleman (2006) stress the importance of ethnic distance in a theoretical model and propose to measure it using surveys of anthropologists. Finally, a few recent contributions underline the correlation between genetic distance and pairwise income differences, trust and trade flows (Giuliano *et al.* 2006; Spolaore and Wacziarg 2009; Guiso *et al.* 2009).

Second, the paper relates to the literature on ethnic polarization. In her original contribution, Reynal-Querol (2002) adapts the measure of polarization developed by Esteban and Ray (1994) to the case of categorical variables, such as ethnicity or religion, and proposes an index of ethnic polarization, RQ, that captures how far the distribution of ethnic groups is from the bipolar case. Montalvo and Reynal-Querol (2005) show that the RQ index is a more powerful predictor of the incidence of civil wars than ELF. The authors also show that RQ is highly correlated with ELF at low levels, uncorrelated at intermediate levels, and negatively correlated at high levels. In a recent contribution (Montalvo and Reynal-Querol 2008), the same authors analyse the theoretical properties of RQ and show that the explanatory power of RQ for the incidence of wars is greater the higher the intensity of the conflict. Desmet *et al.* (2005) focus on ethno-linguistic conflict that arises between a dominant central group and peripheral minority groups. To this end, the authors propose an index of peripheral ethno-linguistic diversity, PD, that can capture the notion of both diversity and polarization. The relationship between RQ, PD and GELF is discussed in Section III.

Third, the paper is related to the theoretical economics literature on the measurement of diversity. For example, Weitzman (1992) suggests an index that is primarily intended to measure biodiversity. Moreover, the measurement of diversity has become an increasingly important issue in the recent literature on the ranking of opportunity sets in terms of freedom of choice, where opportunity sets are interpreted as sets of options available to a decision-maker. Examples of such studies include Weitzman (1998), Pattanaik and Xu (2000), Nehring and Puppe (2002) and Bossert *et al.* (2003). A fundamental difference between the above-mentioned contributions and the approach followed in this paper is the informational basis employed, which results in a very different set of axioms that are suitable for a measure of diversity. Both Weitzman's (1992) seminal paper and the literature on incorporating notions of diversity in the context of measuring freedom of choice proceed by constructing a ranking of *sets* of objects, interpreted as sets of species in the case of biodiversity and as sets of available options in the context of freedom of choice. On the other hand, we operate in an informationally richer environment: the measure of fractionalization may be influenced not only by whether a group is present, but also by the relative population shares of these groups along with the pairwise similarities among them. We are interested in capturing a different aspect of diversity than Nehring and Puppe (2002), namely the instrumental —as opposed to intrinsic—value of diversity, where the number of individuals plays a key role.

Finally, ELF is also used in the literature on network formation as a measure of heterogeneity in the underlying population, where distances in characteristics translate into distances in connections in the network (see, for example, Moody 2001).

The remainder of the paper is organized as follows. In Section I we introduce the notion of a similarity matrix, present the formula of our diversity index, and provide some examples to show how it compares with the ELF index and how our framework can be used to derive an endogenous partition of society into groups. Section II contains our main theoretical result, namely, the axiomatic characterization of GELF. The relationship between GELF and alternative measures that appear in the literature is discussed in Section III. Section IV provides an empirical illustration, and Section V concludes with a summary of our results and possible extensions.

## I. Similarity and Fractionalization: Notation and Examples

In this section we introduce the notion of a similarity matrix, which is the building block of our index. We then present our proposed diversity measure, GELF, and show that the commonly employed ELF is a special case of our index. Finally, we briefly illustrate how our framework can be used to partition the population into groups.

### *Similarity*

While the existing literature on the measurement of fractionalization relies on exogenous partitions of the population into groups, our starting point is a society composed of individuals. We believe that a measure of fractionalization of a society should take as primitive the individual and consider attributes such as ethnicity like any other personal characteristic in determining the similarity among individuals. In our informal discussion, we shall occasionally refer to ethnic groups in order to be in line with the literature to which we aim to contribute. Similarly, the empirical application will also make use of ethnic categories for comparison purposes with standard indices. However, the characterization result that we provide in this paper is very general and we do not need to impose any predefined partition of the population into groups.

Our reasoning proceeds as follows. Imagine a society composed of individuals with personal characteristics, whatever they might be. Any two individuals may be perfectly identical according to the characteristics under consideration, completely dissimilar, or similar to different degrees. For simplicity, we normalize the similarity values to be in the interval $[0, 1]$, assign the value 1 to perfect similarity and a value of 0 to maximum dissimilarity. If the society is composed of $n$ individuals, the comparison process will generate $n^2$ similarity values. These values are collected in a matrix that we call the *similarity matrix*. Each row $i$ of this matrix contains the similarity values of individual $i$ with respect to all members of society. Naturally, all entries on the main diagonal of such a matrix—the entries representing the similarity of each individual to itself—are equal to 1. Furthermore, a similarity matrix is symmetric: the similarity between individuals $i$ and $j$ is equal to that between $j$ and $i$. The possibility of including non-symmetric similarity matrices is discussed later.

Let $\mathbb{N}$ denote the set of positive integers, and let $\mathbb{R}$ denote the set of all real numbers. The set of all non-negative real numbers is $\mathbb{R}_+$, and the set of positive real numbers is $\mathbb{R}_{++}$. For $n\in\mathbb{N}$, $\mathbb{R}^n$ is Euclidean $n$-space and $\Delta^n$ is the $n$-dimensional unit simplex. Furthermore, $\mathbf{0}^n$ is the vector consisting of $n$ zeros. A *similarity matrix of dimension* $n\in\mathbb{N}\setminus\{1\}$ is an $n \times n$ matrix $S = (s_{ij})_{i,j\in\{1,\ldots,n\}}$ such that:

(a) $s_{ij}\in[0, 1]$ for all $i,j\in\{1, \ldots, n\}$;
(b) $s_{ii} = 1$ for all $i\in\{1, \ldots, n\}$;
(c) $[s_{ij} = 1 \Rightarrow s_{ik} = s_{kj}]$ for all $i,j,k\in\{1, \ldots, n\}$.

The three restrictions on the elements of a similarity matrix have very intuitive interpretations. (a) is consistent with a normalization requiring that complete dissimilarity is assigned a value of 0 and full similarity is represented by 1. Clearly, this requires that each individual has a similarity value of 1 when assessing the similarity to itself, as stipulated in (b). Condition (c) requires that if two individuals are fully similar, it is not possible to distinguish between them as far as their similarity to others is concerned. Because $i = j$ is possible in (c), the conjunction of (b) and (c) implies that a similarity matrix is symmetric. Finally, (c) implies that full similarity is transitive, in the sense that if $s_{ij} = s_{ji} = s_{jk} = s_{kj} = 1$, then $s_{ik} = s_{ki} = 1$ for all $i, j, k \in \{1, \ldots, n\}$. Our characterization result remains valid if restriction (c) is dropped—that is, our index can be characterized on a larger domain where the notion of similarity is not necessarily symmetric, as may be the case if the similarity values are obtained from people's subjective views on the degree to which they differ from others. We state our main result with restriction (c) to emphasize that we do not need non-symmetric similarity matrices and, thus, our characterization is not dependent on an artificially large domain. See the Appendix for details.

### *Measuring diversity: GELF and ELF*

Let $\mathcal{S}^n$ be the set of all $n$-dimensional similarity matrices, where $n \in \mathbb{N} \setminus \{1\}$ and $\mathcal{S} = \bigcup_{n \in \mathbb{N} \setminus \{1\}} \mathcal{S}^n$. A *diversity measure* is a function $D : \mathcal{S} \to \mathbb{R}_+$. The specific measure that we propose in this paper is what we call the *generalized fractionalization* (GELF) index $G$. It is defined as

$$(1) \qquad G(S) = 1 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}$$

for all $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}^n$ (or any positive multiple—clearly, multiplying the index value by $\gamma \in \mathbb{R}_{++}$ leaves all diversity comparisons unchanged). GELF is the expected dissimilarity between two individuals drawn at random. If the dissimilarity between two individuals $i$ and $j$ is given by the absolute difference between their incomes, $G$ may be interpreted as a variant of the well-known Gini coefficient adapted to our environment. There are axiomatizations of the Gini index and some of its generalizations in the context of income inequality measurement (see, for instance, Donaldson and Weymark 1980; Weymark 1981; Bossert 1990). However, these characterizations are fundamentally different from ours because we operate on a different domain—a domain of similarity matrices rather than a domain of income vectors. In accordance with this difference, we formulate our axioms in terms of the primitives of the framework—the similarity matrices—rather than in terms of income vectors. This difference distinguishes our axiomatic approach from those that appear in income inequality measurement. Furthermore, because we are interested in measuring fractionalization based on similarities rather than in measuring the inequality of income distributions, our set of axioms does not include a 'transfer' property.

As an example, suppose that a three-dimensional similarity matrix is given by

$$S = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & 0 & 1 \end{pmatrix}.$$

The corresponding value of $G$ is given by

$$G(S) = 1 - \frac{1}{9}\left[1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{2} + 1 + 0 + \frac{1}{4} + 0 + 1\right] = \frac{1}{2}.$$

It is easy to show that $G(S)$ is indeed a generalization of the commonly employed ethno-linguistic fractionalization (ELF) index. The application of ELF is restricted to an environment where the only information available is the vector $p = (p_1, \ldots, p_K) \in \Delta^K$ of population shares for $K \in \mathbb{N}$ predefined groups. No partial similarity values are taken into consideration—individuals are either fully similar or completely dissimilar, that is, $s_{ij}$ can assume the values 1 and 0 only. Letting $\Delta = \bigcup_{K \in \mathbb{N}} \Delta^K$, the ELF index $E : \Delta \to \mathbb{R}_+$ is defined by letting

$$E(p) = 1 - \sum_{k=1}^{K} p_k^2$$

for all $K \in \mathbb{N}$ and $p \in \Delta^K$. Thus ELF is 1 minus the well-known Herfindahl index of concentration.

In our setting, the ELF environment can be described by a subset $\mathcal{S}_{01} = \bigcup_{n \in \mathbb{N} \setminus \{1\}} \mathcal{S}_{01}^n$ of our class of similarity matrices, where $s_{ij} \in \{0, 1\}$ for all $n \in \mathbb{N} \setminus \{1\}$, $S \in \mathcal{S}_{01}^n$ and $i$, $j \in \{1, \ldots, n\}$. By properties (b) and (c), it follows that within this subclass of matrices, the population $\{1, \ldots, n\}$ can be partitioned into $K \in \mathbb{N}$ non-empty and disjoint subgroups $N_1, \ldots, N_K$ with the property that, for all $i, j \in \{1, \ldots, n\}$,

$$s_{ij} = \begin{cases} 1 & \text{if there exists } k \in \{1, \ldots, K\} \text{ such that } i, j \in N_k, \\ 0 & \text{otherwise.} \end{cases}$$

Letting $n_k \in \mathbb{N}$ denote the cardinality of $N_k$ for all $k \in \{1, \ldots, K\}$, it follows that $\sum_{k=1}^{K} n_k = n$ and $p_k = n_k / n$ for all $k \in \{1, \ldots, K\}$. For $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}_{01}^n$, we obtain

$$G(S) = 1 - \frac{1}{n^2} \sum_{k=1}^{K} n_k^2 = 1 - \sum_{k=1}^{K} p_k^2 = E(p).$$

For example, suppose that

$$S = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

that is, we are analysing a society composed of three individuals. Two of them (individuals 1 and 2) are fully similar: the similarity values $s_{12}$ and $s_{21}$ are equal to 1 and, furthermore, they have the same degree of similarity—zero—with respect to the remaining member of society (individual 3). Because individual 3 is not completely similar to anyone else, it forms a group on its own. The corresponding value of $G$ is given by

$$G(S) = 1 - \frac{1}{9}[1 + 1 + 0 + 1 + 1 + 0 + 0 + 0 + 1] = \frac{4}{9}.$$

Because $S \in \mathcal{S}_{01}^3$, we can alternatively calculate this diversity value using ELF. We have $K = 2$, $N_1 = \{1, 2\}$, $N_2 = \{3\}$, $p_1 = \frac{2}{3}$ and $p_2 = \frac{1}{3}$. Thus

$$E(p) = 1 - \left[ \left( \frac{2}{3} \right)^2 + \left( \frac{1}{3} \right)^2 \right] = \frac{4}{9} = G(S).$$

*Partitioning society into groups*

Our framework allows us to obtain population subgroups endogenously from similarity matrices even if similarity values can assume values other than 0 and 1. A plausible method of doing so is the following. Any two individuals $i$ and $j$ belong to the same group if the similarity between $i$ and $j$ is equal to 1 and, moreover, the similarities of $i$ with respect to all other individuals $k$ are the same as those of $j$. Using this process, a group partition emerges naturally from the similarity matrix without having to impose it in advance. This method has several advantages: (i) it releases the researcher from the choice of the one characteristic that determines fractionalization in the society of interest; (ii) it makes it possible to consider simultaneously multiple characteristics; (iii) it allows group formation across characteristics; (iv) it considers the intensity of similarities between groups.

Formally, we define a partition of $\{1,\ldots,n\}$ into $K\in\mathbb{N}$ non-empty and disjoint subgroups $N_1,\ldots,N_K$. By properties (b) and (c), these subgroups are such that for all $k\in\{1,\ldots,K\}$, $i,j\in N_k$ and $h\in\{1,\ldots,n\}$, we have $s_{ij}=s_{ji}=1$ and $s_{ih}=s_{hi}=s_{hj}=s_{jh}$. Thus for all $k,\ell\in\{1,\ldots,K\}$, we can unambiguously define $\bar{s}_{k\ell}=s_{ij}$ for some $i\in N_k$ and some $j\in N_\ell$. Again using $n_k\in\mathbb{N}$ to denote the cardinality of $N_k$ for all $k\in\{1,\ldots,K\}$, it follows that $\sum_{k=1}^{K} n_k=n$ and $p_k=n_k/n$ for all $k\in\{1,\ldots,K\}$. For $n\in\mathbb{N}\setminus\{1\}$ and $S\in\mathcal{S}^n$, we obtain

$$(2)\qquad G(S)=1-\frac{1}{n^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K} n_k n_\ell \bar{s}_{k\ell}=1-\sum_{k=1}^{K}\sum_{\ell=1}^{K} p_k p_\ell \bar{s}_{k\ell}.$$

Clearly, the ELF index $E$ is obtained for the case where all off-diagonal entries of $S$ are equal to 0.

To provide a numerical illustration of this case, let

$$S=\begin{pmatrix}1 & 1 & \frac{1}{2}\\ 1 & 1 & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2} & 1\end{pmatrix},$$

that is, we consider another society of three individuals. Again, two of them (individuals 1 and 2) are fully similar: the similarity values $s_{12}$ and $s_{21}$ are equal to 1 and, furthermore, they have the same degree of similarity with respect to the remaining member of society (individual 3). This time, however, the similarity between the members of the first group and the remaining individual is equal to $\frac{1}{2}$ rather than 0. Individual 3 is not completely similar to anyone, thus is in a group by itself. The corresponding index value is

$$G(S)=1-\frac{1}{9}\left[1+1+\frac{1}{2}+1+1+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+1\right]=\frac{2}{9}.$$

According to the method outlined above, we can alternatively partition the population $\{1,2,3\}$ into two groups $N_1=\{1,2\}$ and $N_2=\{3\}$. The population shares of these groups are $p_1=\frac{2}{3}$ and $p_2=\frac{1}{3}$. We obtain the intergroup similarity values $\bar{s}_{11}=\bar{s}_{22}=s_{11}=s_{22}=s_{12}=s_{21}=1$ and $\bar{s}_{12}=\bar{s}_{21}=s_{i3}=s_{3i}=\frac{1}{2}$ for $i\in\{1,2\}$, which, using (2), leads to the index value

$$G(S)=1-\left[\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2+\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}+\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}\right]=\frac{2}{9}.$$

## II. A Characterization of GELF

We now turn to a characterization of GELF. Our characterization relies on four axioms, which we proceed to illustrate in order. We then state and prove the main theorem containing the formula of our diversity index.

*Axiom 1: Normalization*

Let $I^n$ denote the $n \times n$ identity matrix, and let $\mathbf{1}^n$ denote the $n \times n$ matrix all of whose entries are equal to 1. Clearly, both of these matrices are in $\mathcal{S}^n$ and they represent extreme cases within this class. $I^n$ can be thought of as having maximal diversity: any two distinct individuals are completely dissimilar and, therefore, each individual is in a group by itself. $\mathbf{1}^n$, on the other hand, represents maximal concentration (and thus minimal diversity) because there is but a single group in the population all members of which are fully similar. Our first axiom is a straightforward normalization property. It requires that the value of $D$ at $\mathbf{1}^n$ is equal to 0, and the value of $D$ at $I^n$ is positive for all $n \in \mathbb{N} \setminus \{1\}$.

Given that the matrix $\mathbf{1}^n$ is associated with minimal diversity, it is a very plausible restriction to require that $D$ assumes its minimal value for these matrices. Note that this minimal value is the same across population sizes. This is plausible because, no matter what the population size $n$ might be, there is but a single group of perfectly similar individuals, and thus there is no diversity at all.

In contrast, it would be much less natural to require that the value of $D$ at $I^n$ be identical for all population sizes $n$. It is quite plausible to argue that having more distinct groups each of which consists of a single individual leads to more diversity than a situation where there are fewer groups containing one individual each. Our first axiom can thus be formalized as follows.

*Normalization*. For all $n \in \mathbb{N} \setminus \{1\}$,

$$D(\mathbf{1}^n) = 0 \text{ and } D(I^n) > 0.$$

*Axiom 2: Anonymity*

Our second axiom is very uncontroversial as well. It requires that individuals are treated impartially, paying no attention to their identities. For $n \in \mathbb{N} \setminus \{1\}$, let $\Pi^n$ be the set of permutations of $\{1, \ldots, n\}$, that is, the set of bijections $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$. For $n \in \mathbb{N} \setminus \{1\}$, $S \in \mathcal{S}^n$ and $\pi \in \Pi^n$, $S_\pi$ is obtained from $S$ by permuting the rows and columns of $S$ according to $\pi$. Anonymity requires that $D$ be invariant with respect to permutations.

*Anonymity*. For all $n \in \mathbb{N} \setminus \{1\}$, $S \in \mathcal{S}^n$ and $\pi \in \Pi^n$,

$$D(S_\pi) = D(S).$$

*Axiom 3: Additivity*

Our third axiom is additivity. As the role of this axiom may be less obvious compared to the two previous ones, it is worth providing a brief motivation for its use. Notions of additivity are among the most commonly used properties employed in axiomatic approaches, not only in the context of social index numbers but also in numerous other areas within economics. For instance, additivity plays a major role in fields such as game theory (Shapley 1953), the analysis and design of cost sharing methods (Moulin 1995, 2002; Moulin and Vohra 2003; Moulin and Sprumont 2005) and bargaining theory (Peters 1992), to name but a few. While the exact formulation of additivity depends, of course, on the objects to be studied (here, measures of fractionalization), the intuition behind the principle itself is shared by the variations that can be found in the literature. Roughly speaking (we will be more precise below), additivity means that the value of a function (here, the value of a fractionalization index) calculated at the sum of two values of its variables (in our case, the sum of two similarity matrices) can be obtained on the basis of the sum of the values of the function at the two values of its variables.

A crucial feature of additivity is that it entails a *separability* property: the contribution of any variable to the overall index value can be assessed in isolation, without having to know the values of the other variables. Thus additivity properties are often linked to *independence* conditions of various forms. To take a simple example, consider a function $f$ of three real-valued variables. $f$ can be interpreted as an index number assigning welfare or inequality values to income distributions involving three individuals, for instance. Suppose that we want to compare the value of $f$ at $(x, y, z)$ to the value of $f$ at $(x', y', z)$; that is, we want to examine which of the values $f(x, y, z)$ and $f(x', y', z)$ exceeds the other or if the two values are the same. Note that the value of the third variable is the same in the two vectors to be compared—namely, $z$. Independence demands that the comparison of the values of $f$ at $(x, y, z)$ and at $(x', y', z)$ be independent of the value of $z$—that is, if $z$ is replaced with $z'$ *in both vectors to be compared*, the relative ranking of the vectors according to $f$ should not change. In consequence, the influence of the first two variables on $f$ can be assessed without reference to the value of the third variable. For instance, independence requires that $f(2, 3, 1)$ is greater than or equal to $f(6, 1, 1)$ if and only if $f(2, 3, 13)$ is greater than or equal to $f(6, 1, 13)$: as long as the third variable has the same value in the vectors to be compared (in our example, 1 or 13, but it could be any value whatsoever), it does not matter what this value is. This type of independence property is a consequence of additivity, suitably formulated for the environment under consideration. See, for instance, Blackorby *et al.* (1978) for a detailed discussion of separability properties and their link to notions of additivity.

With a minor qualification to be discussed shortly, additivity means in our context that fractionalization is additive in similarity matrices in the sense that the fractionalization of the normalized sum of two matrices is the sum of the fractionalization values for the original matrices, corrected to account for the normalization in the definition of the sum. Note that an additive structure is occasionally imposed without formulating it explicitly as an axiom on the primitives of the problem. See, for instance, Esteban and Ray (1994), who consider polarization measures that can be expressed as additive functions of entities that do not represent the primitives of the setup, and then impose further axioms on these derived entities rather than the primitives. In contrast, our formulation imposes all axioms on the primitives of the problem. Our additivity property is standard, except that we have to respect the restrictions imposed by the definition of $\mathscr{S}^{\mathrm{n}}$. In particular, we cannot simply add two similarity matrices $S$ and $T$ of dimension $n$ because, according to ordinary matrix addition, all entries on the diagonal of the sum $S + T$ will be equal to 2 rather than 1, and therefore $S + T$ is not an element of $\mathscr{S}^{\mathrm{n}}$. For that reason, we define the operation $\oplus$ on the sets $\mathscr{S}^{\mathrm{n}}$ by letting, for all $n \in \mathbb{N} \setminus \{1\}$ and $S, T \in S^n$, $S \oplus T = (s_{ij} \oplus t_{ij})_{i,j \in \{1,\ldots,n\}}$ with

$$s_{ij} \oplus t_{ij} = \begin{cases} 1 & \text{if } i = j, \\ s_{ij} + t_{ij} & \text{if } i \neq j. \end{cases}$$

The standard additivity axiom has to be modified in another respect in our setting. Because the diagonal is unchanged when moving from $S$ and $T$ to $S \oplus T$, it would be questionable to require the value of $D$ at $S \oplus T$ to be given by the sum of $D(S)$ and $D(T)$ because, in doing so, we would double count the diagonal elements in $S$ and in $T$. Therefore this sum has to be corrected by the value of $D$ at $I^n$, and we obtain the following axiom.

*Additivity*. For all $n \in \mathbb{N} \setminus \{1\}$ and $S, T \in \mathcal{S}^n$ such that $(S \oplus T) \in \mathcal{S}^n$,

$$D(S \oplus T) = D(S) + D(T) - D(I^n).$$

*Axiom 4: Replication invariance*

With the partial exception of the normalization condition (which implies that our diversity measure assumes the same value for the matrix $\mathbf{1}^n$ for all population sizes $n$), the first three axioms apply to diversity comparisons involving fixed population sizes only. Our last axiom imposes restrictions on comparisons across population sizes. We consider specific replications and require the index to be invariant with respect to these replications. The scope of the axiom is limited to what we consider clear-cut cases and, therefore, represents a rather mild variable population requirement. In particular, consider the $n$-dimensional identity matrix $I^n$. As argued before, this matrix represents an extreme degree of diversity: each individual is in a group by itself and shares no similarities with anyone else. Now consider a population of size $nm$ where there are $m$ copies of each individual $i \in \{1, \ldots, n\}$ such that within any group of $m$ copies, all similarity values are equal to 1 and all other similarity values are equal to 0. Thus this particular replication has the effect that instead of $n$ groups of size 1 that do not have any similarity to other groups, now we have $n$ groups each of which consists of $m$ identical individuals and, again, all other similarity values are equal to 0. As before, the population is divided into $n$ homogeneous groups of equal size. Adopting a relative notion of diversity, it would seem natural to require that diversity has not changed as a consequence of this replication. To provide a precise formulation of the resulting axiom, we use the following notation. For $n, m \in \mathbb{N} \setminus \{1\}$, we define the matrix $R^n_m = (r_{ij})_{i,j \in \{1,\ldots,nm\}} \in \mathcal{S}^{nm}$ by

$$r_{ij} = \begin{cases} 1 & \text{if } \exists h \in \{1, \ldots, n\} \text{ such that } i, j \in \{(h-1)m + 1, \ldots, hm\}, \\ 0 & \text{otherwise.} \end{cases}$$

Now we can define our replication invariance axiom.

*Replication invariance*. For all $n, m \in N \setminus \{1\}$,

$$D(R^n_m) = D(I^n).$$

*Characterization*

The four axioms introduced earlier in this section characterize GELF, as we state in the following theorem.

*Theorem*. A diversity measure $D : \mathscr{S} \to \mathbb{R}_+$ satisfies normalization, anonymity, additivity and replication invariance if and only if $D$ is a positive multiple of

$$G(S) = 1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}$$

for all $n \in \mathbb{N}\setminus\{1\}$ and $S \in \mathscr{S}^n$.

*Proof*. That any positive multiple of $G$ satisfies the axioms is straightforward to verify. Conversely, suppose that $D$ is a diversity measure satisfying normalization, anonymity, additivity and replication invariance. Let $n \in \mathbb{N}\setminus\{1\}$, and define the set $\mathcal{X}^n \subseteq \mathbb{R}^{n(n-1)/2}$ by

$$\mathcal{X}^n = \{x = (x_{ij})_{\substack{i \in \{1,\ldots,n-1\} \\ j \in \{i+1,\ldots,n\}}} \mid \exists S \in S^n \text{ such that}$$
$$s_{ij} = x_{ij} \text{ for all } i \in \{1,\ldots,n-1\} \text{ and } j \in \{i+1,\ldots,n\}\}.$$

Define the function $F^n : \mathcal{X}^n \to \mathbb{R}$ by letting, for all $x \in \mathcal{X}^n$,

(3) $$F^n(x) = D(S) - D(I^n),$$

where $S \in \mathscr{S}^n$ is such that $s_{ij} = x_{ij}$ for all $i \in \{1,\ldots,n-1\}$ and $j \in \{i+1,\ldots,n\}$. This function is well defined because $\mathscr{S}^{\text{n}}$ contains symmetric matrices with ones on the main diagonal only. Because $D$ is bounded below by 0, it follows that $F^n$ is bounded below by $-D(I^n)$. Furthermore, the additivity of $D$ implies that $F^n$ satisfies Cauchy's basic functional equation

(4) $$F^n(x+y) = F^n(x) + F^n(y)$$

for all $x, y \in \mathcal{X}^n$ such that $(x+y) \in \mathcal{X}^n$; see Aczél (1966, Section 2.1). We have to address a slight complexity in solving this equation because the domain $\mathcal{X}^n$ of $F^n$ is not a Cartesian product, which is why we provide a few further details rather than invoking the corresponding standard result immediately.

Fix $i \in \{1,\ldots,n-1\}$ and $j \in \{i+1,\ldots,n\}$, and define the function $f^n_{ij} : [0,1] \to \mathbb{R}$ by

$$f^n_{ij}(x_{ij}) = F^n(x_{ij}; \mathbf{0}^{n(n-1)/2-1})$$

for all $x_{ij} \in [0,1]$, where the vector $(x_{ij}; \mathbf{0}^{n(n-1)/2-1})$ is such that the component corresponding to $ij$ is given by $x_{ij}$ and all other entries (if any) are equal to 0. Note that this vector is indeed an element of $\mathcal{X}^n$, and therefore $f^n_{ij}$ is well defined. The function $f^n_{ij}$ is bounded below because $F^n$ is and, as an immediate consequence of (4), it satisfies the Cauchy equation

(5) $$f^n_{ij}(x_{ij} + y_{ij}) = f^n_{ij}(x_{ij}) + f^n_{ij}(y_{ij})$$

for all $x_{ij}, y_{ij} \in [0,1]$ such that $(x_{ij} + y_{ij}) \in [0,1]$. Because the domain of $f^n_{ij}$ is an interval containing the origin and $f^n_{ij}$ is bounded below, the only solutions of (5) are linear functions; see Aczél (1966, Section 2.1). Thus there exists $c^n_{ij} \in \mathbb{R}$ such that

(6) $$F^n(x_{ij}; \mathbf{0}^{n(n-1)/2-1}) = f^n_{ij}(x_{ij}) = c^n_{ij} x_{ij}$$

for all $x_{ij} \in [0,1]$.

Let $S \in \mathscr{S}^n$. By additivity, the definition of $F^n$ and (6),

$$F^n\left((s_{ij})_{\substack{i\in\{1,\ldots,n-1\}\\ j\in\{i+1,\ldots,n\}}}\right) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} F^n(s_{ij};\mathbf{0}^{n(n-1)/2-1})$$

$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f_{ij}^n(s_{ij}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^n s_{ij},$$

and, defining $d^n = D(I^n)$ and substituting into (3), we obtain

$$D(S) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}^n s_{ij} + d^n. \tag{7}$$

Now fix $i,k\in\{1,\ldots,n-1\}$, $j\in\{i+1,\ldots,n\}$ and $\ell\in\{k+1,\ldots,n\}$, and let $S\in\mathcal{S}^n$ be such that $s_{ij}=s_{ji}=1$ and all other off-diagonal entries of $S$ are equal to 0. Let the bijection $\pi\in\Pi^n$ be such that $\pi(i)=k$, $\pi(j)=\ell$, $\pi(k)=i$, $\pi(\ell)=j$ and $\pi(h)=h$ for all $h\in\{1,\ldots,n\}\backslash\{i,j,k,\ell\}$. By (7), we obtain

$$D(S) = c_{ij}^n + d^n \text{ and } D(S_\pi) = c_{k\ell}^n + d^n,$$

and anonymity implies $c_{ij}^n = c_{k\ell}^n$. Therefore there exists $c^n\in\mathbb{R}$ such that $c_{ij}^n = c^n$ for all $i\in\{1,\ldots,n-1\}$ and $j\in\{i+1,\ldots,n\}$, and substituting into (7) yields

$$D(S) = c^n \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij} + d^n$$

for all $n\in\mathbb{N}\backslash\{1\}$ and $S\in\mathcal{S}^n$.

Normalization requires

$$D(\mathbf{1}^n) = c^n\frac{n(n-1)}{2} + d^n = 0,$$

and therefore $d^n = -c^n n(n-1)/2$ for all $n\in\mathbb{N}\backslash\{1\}$. Using normalization again, we obtain

$$D(I^n) = -c^n\frac{n(n-1)}{2} > 0,$$

which implies $c^n < 0$ for all $n\in\mathbb{N}\backslash\{1\}$. Thus

$$D(S) = c^n\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij} - c^n\frac{n(n-1)}{2} \tag{8}$$

for all $n\in\mathbb{N}\backslash\{1\}$ and $S\in\mathcal{S}^n$.

Let $n$ be an even integer greater than or equal to 4. By replication invariance and (8),

$$D(R_{n/2}^2) = c^n\frac{n}{2}\left(\frac{n}{2}-1\right) - c^n\frac{n(n-1)}{2} = -c^2 = D(I^2).$$

Solving, we obtain

$$c^n = 4\frac{c^2}{n^2}. \tag{9}$$

Now let $n$ be an odd integer greater than or equal to 3. Thus $q = 2n$ is even, and the above argument implies

$$(10) \qquad c^q = 4\frac{c^2}{q^2} = \frac{c^2}{n^2}.$$

Furthermore, replication invariance requires

$$D(R_2^n) = D(R_2^{q/2}) = c^q\frac{q}{2} - c^q\frac{q(q-1)}{2} = -c^n\frac{n(n-1)}{2} = D(I^n).$$

Solving for $c^n$ and using the equality $q = 2n$, it follows that $c^n = 4c^q$ and, combined with (10), we obtain (9) for all odd $n \in \mathbb{N}\backslash\{1\}$ as well.

Substituting into (8), simplifying and defining $\gamma = -2c^2 > 0$, it follows that for all $n \in \mathbb{N}\backslash\{1\}$ and $S \in \mathscr{S}^n$,

$$\begin{aligned}
D(S) &= 4\frac{c^2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij} - 2\frac{c^2}{n^2}n(n-1) \\
&= 2\frac{c^2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} s_{ij} - 2c^2 + 2\frac{c^2}{n} \\
&= -2c^2\left[1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} s_{ij} - \frac{1}{n}\right] \\
&= -2c^2\left[1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\right] \\
&= \gamma G(S). \qquad \square
\end{aligned}$$

## III. Alternative and Related Approaches

In this section we discuss the differences between GELF and related indices proposed in various literatures. We start briefly with the linguistics and statistical literature and compare GELF with Greenberg's (1956) index and with the quadratic entropy index (QE). We then proceed with the economics literature, focusing on the indices of ethnic polarization (RQ) and peripheral diversity (PD).

What is known in the economics literature as ELF is in the statistical literature known as the Gini–Simpson index, introduced first by Gini (1912) and then by Simpson (1949) as a measure of diversity of the multinomial distribution. The same index has been proposed by the linguist Greenberg (1956) who termed it the '$A$ index'. In his 1956 article, Greenberg suggested a way to measure the degree of resemblance among $K$ languages. Indicating by $r_{k\ell} \geq 0$ the resemblance between languages $k$ and $\ell$, the proposed $B$ index is

$$B = 1 - \sum_{k=1}^{K}\sum_{\ell=1}^{K} p_k p_\ell r_{k\ell}.$$

This is the index used by Fearon (2003) in his empirical contribution on cultural fractionalization.

In an independent contribution, Rao (1982) suggested exactly the same generalization of ELF, the *quadratic entropy* index (QE), in order to take into account different distance values $d_{k\ell} \geq 0$ of different pairs of categories $k, \ell$. As opposed to Greenberg (1956), Rao (1984) and Rao and Nayak (1985) provide various axiomatizations of the measure. QE is an index that, rewritten in the settings of our paper, considers distances other than 0 and 1 between individuals belonging to different groups; that is,

$$\mathrm{QE} = \sum_{k=1}^{K} \sum_{\ell=1}^{K} p_k p_\ell d_{k\ell}.$$

Recall the definition of $\bar{s}_{k\ell}$ in Section I and the formula for GELF (2). Letting $d_{k\ell} = 1 - \bar{s}_{k\ell}$, we immediately see that GELF is QE, and hence $B$, when the population is partitioned *ex ante* into groups on the basis of a characteristic.

The inspection of the indices $B$ and QE gives further insights into the relationship between GELF and ELF. As we said above, GELF is the expected dissimilarity between two individuals drawn at random from the population. ELF is the likelihood that two randomly drawn individuals belong to different (exogenous) categories. Rewriting ELF as $1 - \sum_{k=1}^{K} p_k p_k$, we see that ELF can be interpreted as 1 minus a weighted sum of population shares $p_k$, where the weights are these shares themselves. GELF, on the other hand, is its natural generalization: it can be written as 1 minus a weighted sum of the population shares. However, the weight assigned to $p_k$ is now not merely $p_k$ itself but a considerably more refined expression that takes account of the similarities of the group members to the individuals in other groups. In calculating GELF, each individual counts in two capacities. Through its membership in its *own* group, an individual contributes to the population share of the group. In addition, there is a secondary contribution via the similarities to individuals of *other* groups.

It should be noted that when the distance values are differences in income, QE is twice the well-known absolute Gini coefficient. The latter, when normalized by mean income, is among the most popular indices of income inequality.

In economics, the index of ethnic polarization RQ (see Reynal-Querol 2002; Montalvo and Reynal-Querol 2005) shares a structure similar to that of ELF and of GELF. It is defined by

$$\mathrm{RQ} = 1 - \sum_{k=1}^{K} \left( \frac{\frac{1}{2} - p_k}{\frac{1}{2}} \right)^2 p_k.$$

As is the case for ELF, RQ employs a weighted sum of population shares. The weights employed in RQ capture the deviation of each group from the maximum polarization share $\frac{1}{2}$ as a proportion of $\frac{1}{2}$. Analogously to ELF, underlying the formula for RQ is the implicit assumption that any two groups are either completely similar or completely dissimilar, and thus the weights depend on population shares only.

The index of peripheral diversity PD (see Desmet *et al.* 2005) is a specification of the original Esteban and Ray (1994) polarization index. It is derived from the alienation–identification framework proposed by Esteban and Ray (1994), applied to distances between languages spoken rather than to income distances as in Esteban and Ray (1994). Desmet *et al.* (2005) distinguish between the effective alienation felt *vs* the dominant group by the minorities. In particular, expressed in the setting of our paper, the index is defined by

$$\mathrm{PD} = \sum_{k=1}^{K} \left[ p_k^{1+\alpha}(1 - \bar{s}_{0k}) + p_k p_0^{1+\alpha}(1 - \bar{s}_{0k}) \right],$$

where $\alpha \in \mathbb{R}$ is a parameter indicating the importance given to the identification component, 0 is the dominant group, and the other $K$ are minority groups. When $\alpha < 0$, PD is an index of peripheral diversity; when $\alpha > 0$, PD is an index of peripheral polarization. The structure of this index is different from that of those previously discussed. As is the case for GELF, it does incorporate a notion of dissimilarity between groups, given by the complement to one of the similarity value. On the other hand, as opposed to the previous indices, the identification component plays a crucial role, enhancing (when $\alpha > 0$) or diminishing (when $\alpha < 0$) the alienation produced by distances between groups. An additional difference to the other indices discussed in this section is the distinction between the dominant groups and the minorities.

## IV. An Empirical Illustration

In this section we provide an application of GELF to the pattern of diversity in the USA. Our goal is to compare the extent of diversity across states, taking into account different dimensions of similarity among individuals; in particular, we consider racial identity, household income, education and employment status.

### *Method and results*

The dataset used is the 5% IPUMS from the 1990 Census. We use individual-level information on all household heads in the sample and record the following characteristics:

(a) **Race.** Each individual is attributed to one of five racial groups: (i) White, (ii) Black, (iii) American Indian, Eskimo or Aleutian, (iv) Asian or Pacific Islander, (v) Other. The last category includes any race other than the four mentioned. The 1990 Census does not identify Hispanic as a separate racial category. However, Alesina *et al.* (1999), who construct ELF from the same five categories, report that the category Hispanic (obtained from a different source) has a correlation of more than 0.9 with the category Other in the Census data.
(b) **Income.** Total household income.
(c) **Education.** The years of education of the individual.
(d) **Employment.** Each individual is attributed to one of four categories: (i) civilian employed or armed forces, at work, (ii) civilian employed or armed forces, with a job but not at work, (iii) unemployed, (iv) not in labour force.

Drawing on the above information, we construct GELF in several ways. The first, and most general, is an implementation of formula (1) that takes into account all four dimensions at the same time without imposing an exogenous partition into groups. The second and third approaches rely on an *ex ante* partition of the population and implement the grouped version of GELF, expression (2). Specifically, the five racial groups described under (a) are used in our illustration to generate the partition. We then measure the similarity among these groups along the remaining dimensions. The choice of race as the exogenously given category is purely instrumental to comparing our results to the widely used ELF index that relies exclusively on racial shares. Obviously, depending on the specific application, the grouping could be done on the cleavage that is most relevant for the phenomenon under study. The idea underlying this second set of results is to propose a way to compute GELF that is less data-intensive and to see whether the qualitative pattern of results differs from that obtained using the full similarity matrix. This second set of results, in turn, is obtained under two alternative methods. The first requires the availability of the entire distribution of individual characteristics, and can be used when individual survey data are available. The second relies only on aggregate data on *mean* characteristics by group.

### *Similarity of individuals*

To implement our index (1), we start from the variables (a) to (d) and apply principal component analysis. To extract principal components, we employ a polychoric correlation matrix to take into account the fact that some of our variables are categorical (see for example, Kolenikov and Angeles 2009). Specifically, we extract for each individual $i$ a synthetic measure $x_i$, the first principal component, that we employ to compute pairwise distances among all individuals living in the same state, that is, $|x_i - x_j|$. To generate similarity values $s_{ij}$ that are bounded between 0 and 1, we normalize this distance by the difference between the maximum and the minimum value of the $x_i$ in the entire sample, and we subtract the resulting value from 1. Once we have the full matrix of similarity values $(s_{ij})_{i,j \in \{1,\ldots,n\}}$, computation of (1) is straightforward. We refer to this index as GELF with no further specifications.

The main result of our empirical analysis is summarized in Figure 1. On the horizontal axis we plot values of ethno-linguistic fractionalization (ELF) for all states in 1990. The vertical axis reports the corresponding value of GELF. While the two are positively correlated, their relationship is far from linear: the correlation coefficient is only 0.59. In particular, states like Hawaii, California and Nevada are much more heterogeneous if one only looks at racial shares than if all dimensions are considered jointly. This is because in these states the distribution of income, education and employment is relatively more similar among races than in other states. At the opposite end we have states like Alaska, Kentucky, Rhode Island, Massachusetts and in general New England, where diversity measured in terms of racial shares is relatively low, but

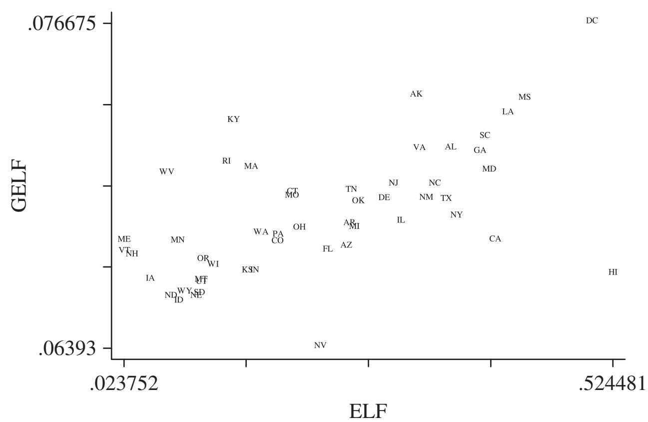


FIGURE 1. GELF and ELF in the USA.

different races differ in the distribution of the remaining characteristics to such an extent that they are actually more diverse when the full similarity GELF is employed.

Table 1 provides the counterpart to the graphical analysis, as it reports the full set of states listed in decreasing order of ethno-linguistic fractionalization, the corresponding values of ELF and GELF, and the difference in ranks between ELF and GELF for each state. We prefer to rely on a comparison of ranks because the absolute values of the two indices are not comparable. In particular, in the last column of Table 1 we report the difference ELF rank – GELF rank, so that negative values indicate that a given state is less fractionalized according to GELF than according to ELF, while positive values indicate the opposite. The magnitude of the difference gives a rough approximation of how big a difference it makes for a particular state to use one index over the other, in terms of relative rankings.

*Similarity of distributions*

Once the population is exogenously partitioned into racial groups, we can assess the distance among these groups by comparing the distributions of individual characteristics such as income, education and employment. We refer to the indices as Grouped $\text{GELF}_{\text{income}}$ etc. Consider income, for example. We first estimate non-parametrically the distributions of household income by race of the head of the household, $\widehat{f^k}(y)$. The estimation method applied in the paper is the *adaptive* or *variable* kernel modified to take into account sample weights attached to each observation in each group. After estimating the densities of household income by race, we measure the overlap among them, implying that two racial groups whose income distributions perfectly overlap are considered perfectly similar. The measure of overlap of distributions applied is the Kolmogorov measure of variation distance:

$$\text{Kov}_{k\ell} = \frac{1}{2}\int \left|\widehat{f^k}(y) - \widehat{f^\ell}(y)\right| \mathrm{d}y.$$

$\text{Kov}_{k\ell}$ is a measure of the lack of overlap between groups $k$ and $\ell$. It ranges between 0 and 1, taking value 0 if $\widehat{f^k}(y) = \widehat{f^\ell}(y)$ for all $y \in \mathbb{R}$, and 1 if $\widehat{f^k}(y)$ and $\widehat{f^\ell}(y)$ do not overlap at all. The distance is sensitive to changes in the distributions only when both take positive values, being insensitive to changes whenever one of them is 0. It will not change if the distributions move apart, provided that there is no overlap between them or that the overlapping part remains unchanged. The resulting measure of similarity between any two groups $k$ and $\ell$, that we employ to implement formula (2) for grouped GELF, is

$$\bar{s}_{k\ell} = 1 - \text{Kov}_{k\ell}.$$

This method is also applied on the distribution of the synthetic measure $x_i$ obtained for each individual in each group by principal component analysis. In this case we estimate $\widehat{f^k}(x)$, the distribution of the synthetic measure by race, and compute the Kolmogorov measure of variation distance and the measure of similarity as described above. The results are displayed in Table 2.

States in Table 2 are listed in decreasing order of GELF, and two additional indices (with the corresponding ranks) are reported. The first index, which we denote as Grouped $\text{GELF}_d$, employs the Kolmogorov distance among distributions of the synthetic index to compute similarity values that are then used in formula (2). The second index, denoted as Grouped GELF, is simpler in that only the *average* value of the synthetic index for each racial group is used when computing distances (differences).

TABLE 1
GELF AND ELF IN THE USA

| State name | State abb. | ELF | ELF rank | GELF | GELF rank | Difference (ELF rank – GELF rank) |
|---|---|---|---|---|---|---|
| Hawaii | HI | 0.5245 | 1 | 0.0668 | 42 | – 41 |
| District of Columbia | DC | 0.5032 | 2 | 0.0767 | 1 | 1 |
| Mississipi | MS | 0.4344 | 3 | 0.0737 | 3 | 0 |
| Louisiana | LA | 0.4165 | 4 | 0.0731 | 4 | 0 |
| California | CA | 0.4042 | 5 | 0.0681 | 30 | – 25 |
| Maryland | MD | 0.3975 | 6 | 0.0709 | 12 | – 6 |
| South Carolina | SC | 0.3940 | 7 | 0.0722 | 6 | 1 |
| Georgia | GA | 0.3885 | 8 | 0.0716 | 9 | – 1 |
| New York | NY | 0.3644 | 9 | 0.0690 | 23 | – 14 |
| Alabama | AL | 0.3577 | 10 | 0.0717 | 7 | 3 |
| Texas | TX | 0.3534 | 11 | 0.0697 | 21 | – 10 |
| North Carolina | NC | 0.3425 | 12 | 0.0703 | 15 | – 3 |
| New Mexico | NM | 0.3332 | 13 | 0.0698 | 19 | – 6 |
| Virginia | VA | 0.3259 | 14 | 0.0717 | 8 | 6 |
| Alaska | AK | 0.3225 | 15 | 0.0738 | 2 | 13 |
| Illinois | IL | 0.3069 | 16 | 0.0688 | 24 | – 8 |
| New Jersey | NJ | 0.3005 | 17 | 0.0703 | 14 | 3 |
| Delaware | DE | 0.2904 | 18 | 0.0697 | 20 | – 2 |
| Oklahoma | OK | 0.2640 | 19 | 0.0696 | 22 | – 3 |
| Michigan | MI | 0.2591 | 20 | 0.0686 | 26 | – 6 |
| Tennessee | TN | 0.2566 | 21 | 0.0701 | 16 | 5 |
| Arkansas | AR | 0.2546 | 22 | 0.0688 | 25 | – 3 |
| Arizona | AZ | 0.2509 | 23 | 0.0679 | 34 | – 11 |
| Florida | FL | 0.2324 | 24 | 0.0677 | 35 | – 11 |
| Nevada | NV | 0.2248 | 25 | 0.0639 | 51 | – 26 |
| Ohio | OH | 0.2037 | 26 | 0.0686 | 27 | – 1 |
| Connecticut | CT | 0.1967 | 27 | 0.0700 | 17 | 10 |
| Missouri | MO | 0.1958 | 28 | 0.0698 | 18 | 10 |
| Pennsylvania | PA | 0.1821 | 29 | 0.0683 | 29 | 0 |
| Colorado | CO | 0.1815 | 30 | 0.0680 | 33 | – 3 |
| Washington | WA | 0.1637 | 31 | 0.0684 | 28 | 3 |
| Indiana | IN | 0.1574 | 32 | 0.0669 | 41 | – 9 |
| Massachusetts | MA | 0.1535 | 33 | 0.0710 | 11 | 22 |
| Kansas | KS | 0.1501 | 34 | 0.0669 | 40 | – 6 |
| Kentucky | KY | 0.1354 | 35 | 0.0728 | 5 | 30 |
| Rhode Island | RI | 0.1290 | 36 | 0.0712 | 10 | 26 |
| Wisconsin | WI | 0.1145 | 37 | 0.0671 | 39 | – 2 |
| Oregon | OR | 0.1054 | 38 | 0.0673 | 38 | 0 |
| Utah | UT | 0.1033 | 39 | 0.0664 | 45 | – 6 |
| Montana | MT | 0.1027 | 40 | 0.0665 | 44 | – 4 |
| South Dakota | SD | 0.1015 | 41 | 0.0660 | 47 | – 6 |
| Nebraska | NE | 0.0980 | 42 | 0.0659 | 49 | – 7 |
| Wyoming | WY | 0.0856 | 43 | 0.0661 | 46 | – 3 |
| Idaho | ID | 0.0797 | 44 | 0.0657 | 50 | – 6 |
| Minnesota | MN | 0.0788 | 45 | 0.0681 | 32 | 13 |
| North Dakota | ND | 0.0718 | 46 | 0.0659 | 48 | – 2 |
| West Virginia | WV | 0.0674 | 47 | 0.0708 | 13 | 34 |

TABLE 1
CONTINUED

| State name | State abb. | ELF | ELF rank | GELF | GELF rank | Difference (ELF rank – GELF rank) |
|---|---|---|---|---|---|---|
| Iowa | IA | 0.0503 | 48 | 0.0666 | 43 | 5 |
| New Hampshire | NH | 0.0321 | 49 | 0.0675 | 37 | 12 |
| Vermont | VT | 0.0240 | 50 | 0.0677 | 36 | 14 |
| Maine | ME | 0.0238 | 51 | 0.0681 | 31 | 20 |

While the use of means or of the entire distribution yields very similar results, the comparison with GELF suggests that for some states the exogenous definition of racial categories does make a difference: these are the same states for which the difference between ELF and GELF in Figure 1 was more pronounced. In this sense, and not surprisingly, the Grouped GELF index calculated according to (2) is more similar to ELF than the GELF index (1) calculated on the full similarity matrix.

### *Similarity of means*

As an alternative to the distance among distributions, we compute a crude measure of similarity based on the *expected value* of the distribution of the characteristic analysed. This is to illustrate the performance of GELF in case of grouped data or poor availability of information in the dataset.

We can measure similarity with respect to continuous or to categorical variables. For continuous variables, such as household income or education, we indicate by $\lambda^k$ the sample mean of the distribution for group $k$, by $\lambda_{\mathrm{Max}}$ the maximum mean value among all groups in all states, and by $\lambda_{\mathrm{Min}}$ the minimum. Then we can compute $\bar{s}_{k\ell}$ for each state as

$$(11) \qquad \bar{s}_{k\ell} = 1 - \left| \frac{\lambda^k - \lambda^\ell}{\lambda_{\mathrm{Max}} - \lambda_{\mathrm{Min}}} \right|.$$

Note that expression (11) is bounded between 0 and 1 by construction.

For categorical variables like employment, we create a dummy variable that assumes the value 1 if the household head is employed, and 0 if unemployed or not in the labour force. We have also experimented with a different definition where 1 corresponds to households whose head is employed or not in the labour force, and 0 where the head is unemployed. The results were not significantly affected and are available from the authors. Indicating by $\delta^k$ the sample means of this variable for group $k$ (that is, the share of the population assuming value 1), similarity between any two groups $k$ and $\ell$ is

$$\bar{s}_{k\ell} = 1 - \left|\delta^k - \delta^\ell\right|.$$

Again, sample weights are used in the computations for these variables.

Results are contained in Table 3. We denote by Grouped $\mathrm{GELF}_{\mathrm{income}}$ the index obtained when distance among racial groups is measured solely in terms of differences in average income, by Grouped $\mathrm{GELF}_{\mathrm{edu}}$ that when differences are in average years of education, and by Grouped $\mathrm{GELF}_{\mathrm{empl}}$ that when the difference is in the share of people employed. For each index, we report the value and the rank, and states are still listed in decreasing order of the full similarity GELF. The results are quite informative and are more easily visualized through Figure 2. Panel A of the figure plots the original values of

TABLE 2
GELF AND GROUPED GELF (KOLMOGOROV AND AVERAGE) IN THE USA

| State | GELF | GELF rank | Grouped $GELF_d$ | Grouped $GELF_d$ rank | Grouped GELF | Grouped GELF rank |
|---|---|---|---|---|---|---|
| HI | 0.0668 | 42 | 0.0917 | 6 | 0.0588 | 13 |
| DC | 0.0767 | 1 | 0.2306 | 1 | 0.1864 | 1 |
| MS | 0.0737 | 3 | 0.1161 | 2 | 0.1061 | 2 |
| LA | 0.0731 | 4 | 0.1070 | 3 | 0.0951 | 3 |
| CA | 0.0681 | 30 | 0.0793 | 9 | 0.0586 | 14 |
| MD | 0.0709 | 12 | 0.0675 | 14 | 0.0564 | 16 |
| SC | 0.0722 | 6 | 0.0974 | 4 | 0.0879 | 4 |
| GA | 0.0716 | 9 | 0.0850 | 7 | 0.0758 | 6 |
| NY | 0.0690 | 23 | 0.0701 | 12 | 0.0617 | 10 |
| AL | 0.0717 | 7 | 0.0810 | 8 | 0.0727 | 7 |
| TX | 0.0697 | 21 | 0.0783 | 10 | 0.0646 | 8 |
| NC | 0.0703 | 15 | 0.0701 | 13 | 0.0613 | 12 |
| NM | 0.0698 | 19 | 0.0656 | 17 | 0.0614 | 11 |
| VA | 0.0717 | 8 | 0.0752 | 11 | 0.0637 | 9 |
| AK | 0.0738 | 2 | 0.0966 | 5 | 0.0875 | 5 |
| IL | 0.0688 | 24 | 0.0664 | 15 | 0.0576 | 15 |
| NJ | 0.0703 | 14 | 0.0661 | 16 | 0.0541 | 18 |
| DE | 0.0697 | 20 | 0.0545 | 19 | 0.0510 | 20 |
| OK | 0.0696 | 22 | 0.0365 | 26 | 0.0303 | 26 |
| MI | 0.0686 | 26 | 0.0558 | 18 | 0.0516 | 19 |
| TN | 0.0701 | 16 | 0.0432 | 24 | 0.0356 | 25 |
| AR | 0.0688 | 25 | 0.0543 | 20 | 0.0475 | 21 |
| AZ | 0.0679 | 34 | 0.0440 | 23 | 0.0558 | 17 |
| FL | 0.0677 | 35 | 0.0448 | 22 | 0.0388 | 23 |
| NV | 0.0639 | 51 | 0.0363 | 27 | 0.0285 | 29 |
| OH | 0.0686 | 27 | 0.0392 | 25 | 0.0364 | 24 |
| CT | 0.0700 | 17 | 0.0481 | 21 | 0.0407 | 22 |
| MO | 0.0698 | 18 | 0.0291 | 31 | 0.0252 | 31 |
| PA | 0.0683 | 29 | 0.0324 | 29 | 0.0294 | 27 |
| CO | 0.0680 | 33 | 0.0348 | 28 | 0.0293 | 28 |
| WA | 0.0684 | 28 | 0.0257 | 35 | 0.0184 | 37 |
| IN | 0.0669 | 41 | 0.0284 | 32 | 0.0241 | 32 |
| MA | 0.0710 | 11 | 0.0310 | 30 | 0.0256 | 30 |
| KS | 0.0669 | 40 | 0.0261 | 34 | 0.0216 | 33 |
| KY | 0.0728 | 5 | 0.0202 | 41 | 0.0146 | 42 |
| RI | 0.0712 | 10 | 0.0247 | 36 | 0.0197 | 36 |
| WI | 0.0671 | 39 | 0.0272 | 33 | 0.0213 | 34 |
| OR | 0.0673 | 38 | 0.0168 | 44 | 0.0125 | 45 |
| UT | 0.0664 | 45 | 0.0222 | 39 | 0.0180 | 39 |
| MT | 0.0665 | 44 | 0.0222 | 38 | 0.0184 | 38 |
| SD | 0.0660 | 47 | 0.0243 | 37 | 0.0201 | 35 |
| NE | 0.0659 | 49 | 0.0182 | 43 | 0.0148 | 41 |
| WY | 0.0661 | 46 | 0.0189 | 42 | 0.0157 | 40 |
| ID | 0.0657 | 50 | 0.0214 | 40 | 0.0145 | 43 |
| MN | 0.0681 | 32 | 0.0157 | 46 | 0.0114 | 46 |
| ND | 0.0659 | 48 | 0.0164 | 45 | 0.0128 | 44 |
| WV | 0.0708 | 13 | 0.0123 | 47 | 0.0097 | 47 |

TABLE 2
CONTINUED

| State | GELF | GELF rank | Grouped $GELF_d$ | Grouped $GELF_d$ rank | Grouped GELF | Grouped GELF rank |
|---|---|---|---|---|---|---|
| IA | 0.0666 | 43 | 0.0085 | 48 | 0.0058 | 48 |
| NH | 0.0675 | 37 | 0.0052 | 49 | 0.0040 | 49 |
| VT | 0.0677 | 36 | 0.0044 | 50 | 0.0020 | 51 |
| ME | 0.0681 | 31 | 0.0043 | 51 | 0.0030 | 50 |

ELF on the horizontal axis against Grouped $GELF_{income}$ on the vertical. The two measures are closely correlated with two extreme outliers: Hawaii is much less fractionalized when we use Grouped $GELF_{income}$ than when we use ELF, while the opposite occurs for the District of Columbia. The intuition is similar to that provided when commenting on Figure 1: in states like Hawaii or California, average income levels are relatively more similar among races than they are in DC or in Connecticut, for example. A similar picture is offered in panel B with respect to years of education. Interestingly, however, when we look at employment levels (panel C) the relationship between the two indices becomes hump-shaped. The maximum value of diversity according to Grouped $GELF_{empl}$ corresponds to *intermediate* levels of ethnic fractionalization; on the other hand, very low or very high levels of ELF translate into middle range values of diversity when both race and similarity in employment status are taken into account. A possible interpretation of this result is that sizeable differences in employment status (such as high unemployment levels for minorities) may be politically difficult to sustain in states where a relatively high fraction of the population is non-white. On the other hand, the same does not hold for income, as if income differences were more easily acceptable compared to the universal right of access to employment.

While only suggestive and illustrative, the above analysis highlights some of the potential benefits that may derive from the use of fractionalization indices that do not rely simply on population shares, but also try to incorporate information on other dimensions along which individuals may differ.

## V. CONCLUDING REMARKS

The main purpose of this paper is to provide a theoretical foundation and an empirical illustration of a new measure of ethnic diversity. Unlike the most commonly used ELF index, our generalized version GELF makes use of a broader informational base. Instead of limiting the relevant variables to the population shares of predefined groups, we start out with a notion of similarity among individuals and calculate our index value accordingly. It is possible to derive a partition into groups endogenously, and the standard ELF index emerges as a special case when no partial similarity is allowed. The results of our empirical application suggest that accounting for the extent of similarity among individuals in observable dimensions other than race may indeed alter the picture of ethnic diversity in the USA. In places like New England or Washington DC, racial fractionalization is magnified when similarity in income, education or employment is taken into account; in places like California, the opposite occurs.

Before concluding, we would like to stress an important methodological point. While in this paper we characterize GELF on the basis of similarities among individuals, our approach is silent on how these similarities should be defined. In particular, our approach is fully

TABLE 3
GELF AND GROUPED GELF (INCOME, EDUCATION, EMPLOYMENT) IN THE USA

| State | GELF rank | Grouped $GELF_{income}$ | Rank | Grouped $GELF_{edu}$ | Rank | Grouped $GELF_{empl}$ | Rank |
|---|---|---|---|---|---|---|---|
| DC | 1 | 0.2880 | 1 | 0.2436 | 1 | 0.1426 | 29 |
| AK | 2 | 0.0936 | 10 | 0.0825 | 8 | 0.2261 | 5 |
| MS | 3 | 0.1181 | 2 | 0.1048 | 2 | 0.1748 | 23 |
| LA | 4 | 0.1181 | 3 | 0.0836 | 6 | 0.1960 | 16 |
| KY | 5 | 0.0249 | 36 | 0.0063 | 46 | 0.1169 | 34 |
| SC | 6 | 0.1054 | 5 | 0.0905 | 4 | 0.1943 | 17 |
| AL | 7 | 0.0937 | 9 | 0.0618 | 14 | 0.2023 | 11 |
| VA | 8 | 0.0938 | 8 | 0.0675 | 12 | 0.2100 | 9 |
| GA | 9 | 0.1161 | 4 | 0.0690 | 11 | 0.2021 | 12 |
| RI | 10 | 0.0285 | 33 | 0.0190 | 35 | 0.1151 | 36 |
| MA | 11 | 0.0401 | 25 | 0.0259 | 26 | 0.1332 | 32 |
| MD | 12 | 0.1028 | 6 | 0.0581 | 16 | 0.2117 | 8 |
| WV | 13 | 0.0135 | 44 | 0.0039 | 49 | 0.0631 | 47 |
| NJ | 14 | 0.0936 | 11 | 0.0596 | 15 | 0.2140 | 7 |
| NC | 15 | 0.0827 | 13 | 0.0568 | 18 | 0.2081 | 10 |
| TN | 16 | 0.0555 | 22 | 0.0236 | 29 | 0.1845 | 20 |
| CT | 17 | 0.0709 | 17 | 0.0435 | 21 | 0.1615 | 25 |
| MO | 18 | 0.0353 | 30 | 0.0174 | 37 | 0.1575 | 26 |
| NM | 19 | 0.0625 | 19 | 0.0751 | 9 | 0.2300 | 4 |
| DE | 20 | 0.0735 | 16 | 0.0533 | 19 | 0.2004 | 14 |
| TX | 21 | 0.0840 | 12 | 0.0830 | 7 | 0.2346 | 3 |
| OK | 22 | 0.0396 | 26 | 0.0254 | 28 | 0.2013 | 13 |
| NY | 23 | 0.0967 | 7 | 0.0674 | 13 | 0.2349 | 2 |
| IL | 24 | 0.0778 | 15 | 0.0576 | 17 | 0.2156 | 6 |
| AR | 25 | 0.0564 | 20 | 0.0389 | 22 | 0.1849 | 19 |
| MI | 26 | 0.0629 | 18 | 0.0356 | 23 | 0.1917 | 18 |
| OH | 27 | 0.0475 | 24 | 0.0257 | 27 | 0.1617 | 24 |
| WA | 28 | 0.0230 | 37 | 0.0209 | 34 | 0.1411 | 30 |
| PA | 29 | 0.0396 | 27 | 0.0229 | 30 | 0.1499 | 28 |
| CA | 30 | 0.0779 | 14 | 0.0851 | 5 | 0.2522 | 1 |
| ME | 31 | 0.0028 | 51 | 0.0025 | 51 | 0.0233 | 51 |
| MN | 32 | 0.0169 | 40 | 0.0067 | 45 | 0.0737 | 45 |
| CO | 33 | 0.0361 | 29 | 0.0330 | 25 | 0.1531 | 27 |
| AZ | 34 | 0.0556 | 21 | 0.0733 | 10 | 0.1965 | 15 |
| FL | 35 | 0.0552 | 23 | 0.0480 | 20 | 0.1790 | 22 |
| VT | 36 | 0.0031 | 50 | 0.0034 | 50 | 0.0235 | 50 |
| NH | 37 | 0.0047 | 49 | 0.0051 | 48 | 0.0313 | 49 |
| OR | 38 | 0.0129 | 46 | 0.0156 | 39 | 0.0963 | 38 |
| WI | 39 | 0.0264 | 35 | 0.0153 | 40 | 0.1033 | 37 |
| KS | 40 | 0.0265 | 34 | 0.0209 | 33 | 0.1304 | 33 |
| IN | 41 | 0.0304 | 32 | 0.0178 | 36 | 0.1337 | 31 |
| HI | 42 | 0.0325 | 31 | 0.0940 | 3 | 0.1166 | 35 |
| IA | 43 | 0.0054 | 48 | 0.0056 | 47 | 0.0482 | 48 |
| MT | 44 | 0.0175 | 39 | 0.0131 | 41 | 0.0935 | 40 |
| UT | 45 | 0.0168 | 41 | 0.0211 | 32 | 0.0945 | 39 |
| WY | 46 | 0.0150 | 43 | 0.0171 | 38 | 0.0796 | 43 |
| SD | 47 | 0.0188 | 38 | 0.0084 | 43 | 0.0922 | 41 |

TABLE 3
CONTINUED

| State | GELF rank | Grouped $GELF_{income}$ | Rank | Grouped $GELF_{edu}$ | Rank | Grouped $GELF_{empl}$ | Rank |
|---|---|---|---|---|---|---|---|
| ND | 48 | 0.0131 | 45 | 0.0073 | 44 | 0.0674 | 46 |
| NE | 49 | 0.0162 | 42 | 0.0119 | 42 | 0.0899 | 42 |
| ID | 50 | 0.0117 | 47 | 0.0224 | 31 | 0.0745 | 44 |
| NV | 51 | 0.0394 | 28 | 0.0330 | 24 | 0.1811 | 21 |

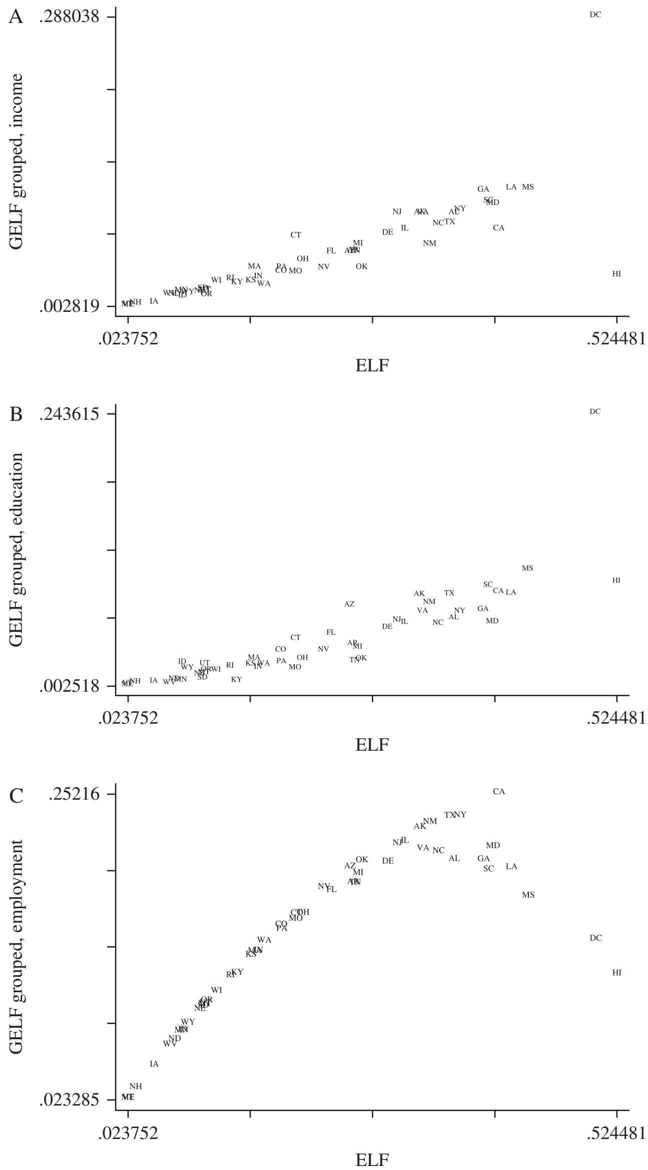


FIGURE 2. Grouped GELF and ELF in the USA.

compatible with a setting in which the notion of continuous distance does not apply (that is, individuals are either fully similar or fully dissimilar, in which case our primitives $s_{ij}$ take on the values 0 or 1 only), as well as with a setting in which it is meaningful to think of similarity among individuals in a continuous way. In addition, our index allows us to incorporate a multidimensional concept of similarity, as opposed to a single dimension. We view this flexibility as an advantage of our approach, and one that makes our index applicable in many different settings. Our choice in the empirical illustration was guided by the attempt to compare our results with well-known patterns in the economics literature on ethnic fractionalization in the USA. We chose as dimensions of similarity ethnicity, household income, education and employment status since we believe that these are important aspects of the US economy that could influence the behaviour of individuals. However, the choice of variables to be employed in the measurement of similarity could include very different aspects and should be guided by the specific application that one has in mind.

Finally, the application of our index is not limited to studies involving ethnolinguistic fractionalization. The generalized index that we propose may be applied to various areas in economics, including, for example, industrial organization. GELF is an index of diversity, and the difference between 1 and the index value can be interpreted as an index of concentration. Embedding information on similarity among firms in a concentration index may yield results different from the traditional Herfindahl index, which is purely based on market shares.

## APPENDIX

Here we illustrate that our characterization result is unchanged if the set of similarity matrices $\mathcal{S}^n$ consists of all $n \times n$ matrices $S$ satisfying conditions (a) and (b) of Section I, but not necessarily (c). This is achieved by some straightforward modifications of the definitions used in the proof of the Theorem.

That any positive multiple of $G$ satisfies the axioms on the larger domain as well is, again, straightforward to verify. Conversely, suppose that $D$ is a diversity measure defined on the larger domain satisfying normalization, anonymity, additivity and replication invariance. Let $n \in \mathbb{N} \setminus \{1\}$, and define the set $\mathcal{X}^n \subseteq \mathbb{R}^{n(n-1)/2}$ by

$$\mathcal{X}^n = \{x = (x_{ij})_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,n\}\setminus\{i\}}} \mid \exists S \in \mathcal{S}^n \text{ such that}$$
$$s_{ij} = x_{ij} \text{ for all } i \in \{1,\ldots,n\} \text{ and } j \in \{1,\ldots,n\}\setminus\{i\}\}.$$

Define the function $F^n : \mathcal{X}^n \to \mathbb{R}$ by letting, for all $x \in \mathcal{X}^n$,

$$F^n(x) = D(S) - D(I^n), \tag{A1}$$

where $S \in \mathcal{S}^n$ is such that $s_{ij} = x_{ij}$ for all $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,n\}\setminus\{i\}$. Because $D$ is bounded below by 0, it follows that $F^n$ is bounded below by $-D(I^n)$. Furthermore, the additivity of $D$ implies that $F^n$ satisfies Cauchy's basic functional equation

$$F^n(x+y) = F^n(x) + F^n(y) \tag{A2}$$

for all $x, y \in \mathcal{X}^n$ such that $(x+y) \in \mathcal{X}^n$; see Aczél (1966, Section 2.1).

Fix $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,n\}\setminus\{i\}$, and define the function $f^n_{ij} : [0,1] \to \mathbb{R}$ by

$$f^n_{ij}(x_{ij}) = F^n(x_{ij}; 0^{n(n-1)-1})$$

for all $x_{ij} \in [0,1]$, where the vector $(x_{ij}; 0^{n(n-1)-1})$ is such that the component corresponding to $ij$ is given by $x_{ij}$, and all other entries (if any) are equal to zero. The function $f^n_{ij}$ is bounded below because $F^n$ is and, as an immediate consequence of (A2), it satisfies the Cauchy equation

$$f^n_{ij}(x_{ij} + y_{ij}) = f^n_{ij}(x_{ij}) + f^n_{ij}(y_{ij}) \tag{A3}$$

for all $x_{ij}, y_{ij} \in [0, 1]$ such that $(x_{ij} + y_{ij}) \in [0, 1]$. Because the domain of $f_{ij}^n$ is an interval containing the origin, and $f_{ij}^n$ is bounded below, the only solutions to (A3) are linear functions; see Aczél (1966, Section 2.1). Thus there exists $c_{ij}^n \in \mathbb{R}$ such that

$$\text{(A4)} \qquad F^n(x_{ij}; 0^{n(n-1)-1}) = f_{ij}^n(x_{ij}) = c_{ij}^n x_{ij}$$

for all $x_{ij} \in [0, 1]$.

Let $S \in S^n$. By additivity, the definition of $F^n$ and (A4),

$$F^n\left( (s_{ij})_{\substack{i\in\{1,\ldots,n\} \\ j\in\{1,\ldots,n\}\setminus\{i\}}} \right) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} F^n(s_{ij}; \mathbf{0}^{n(n-1)-1})$$

$$= \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} f_{ij}^n(s_{ij}) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} c_{ij}^n s_{ij}$$

and, defining $d^n = D(I^n)$ and substituting into (A1), we obtain

$$\text{(A5)} \qquad D(S) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} c_{ij}^n s_{ij} + d^n.$$

Now fix $i, k \in \{1, \ldots, n\}$, $j \in \{1, \ldots, n\} \setminus \{i\}$ and $\ell \in \{1, \ldots, n\} \setminus \{k\}$, and let $S \in \mathcal{S}^n$ be such that $s_{ij} = 1$ and all other off-diagonal entries of $S$ are equal to 0. Let the bijection $\pi \in \Pi^n$ be such that $\pi(i) = k$, $\pi(j) = \ell$, $\pi(k) = i$, $\pi(\ell) = j$ and $\pi(h) = h$ for all $h \in \{1, \ldots, n\} \setminus \{i, j, k, \ell\}$. By (A5), we obtain

$$D(S) = c_{ij}^n + d^n \text{ and } D(S_\pi) = c_{k\ell}^n + d^n,$$

and anonymity implies $c_{ij}^n = c_{k\ell}^n$. Therefore there exists $c^n \in \mathbb{R}$ such that $c_{ij}^n = c^n$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n\} \setminus \{i\}$, and substituting into (A5) yields

$$D(S) = c^n \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} + d^n$$

for all $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}^n$.

Normalization requires

$$D(\mathbf{1}^n) = c^n n(n-1) + d^n = 0,$$

and therefore $d^n = -c^n n(n-1)$ for all $n \in \mathbb{N} \setminus \{1\}$. Using normalization again, we obtain

$$D(I^n) = -c^n n(n-1) > 0,$$

which implies $c^n < 0$ for all $n \in \mathbb{N} \setminus \{1\}$. Thus

$$\text{(A6)} \qquad D(S) = c^n \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} s_{ij} - c^n n(n-1)$$

for all $n \in \mathbb{N} \setminus \{1\}$ and $S \in \mathcal{S}^n$.

Let $n$ be an even integer greater than or equal to 4. By replication invariance and (A6),

$$D(R_{n/2}^2) = c^n n\left(\frac{n}{2} - 1\right) - c^n n(n-1) = -c^2 = D(I^2).$$

Solving, we obtain

$$\text{(A7)} \qquad c^n = 2\frac{c^2}{n^2}.$$

Now let $n$ be an odd integer greater than or equal to 3. Thus $q = 2n$ is even, and the above argument implies

$$c^q = 2\frac{c^2}{q^2} = \frac{c^2}{2n^2}. \tag{A8}$$

Furthermore, replication invariance requires

$$D(R_2^n) = D(R_2^{q/2}) = c^q q - c^q q(q-1) = -c^n n(n-1) = D(I^n).$$

Solving for $c^n$ and using the equality $q = 2n$, it follows that $c^n = 4c^q$ and, combined with (A8), we obtain (A7) for all odd $n \in \mathbb{N}\setminus\{1\}$ as well.

Substituting into (A6), simplifying and defining $\gamma = -2c^2 > 0$, it follows that for all $n \in \mathbb{N}\setminus\{1\}$ and $S \in \mathcal{S}^n$,

$$\begin{aligned} D(S) &= 2\frac{c^2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} s_{ij} - c^n n(n-1) \\ &= 2\frac{c^2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} - 2\frac{c^2}{n^2}n - 2\frac{c^2}{n^2}n(n-1) \\ &= -2c^2\left[1 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\right] \\ &= \gamma G(S). \end{aligned}$$

## ACKNOWLEDGMENTS

We are grateful to Itzhak Gilboa for extremely useful suggestions. We thank the referees, Paola Ballon, Vincent Buskens, Joan Esteban, Michele Pellizzari, Debraj Ray and seminar participants at CORE, Dalhousie University, Università di Milano, Università di Pavia, the 2005 Polarization and Conflict Workshop, the 2006 EURODIV, SCW and IARIW Conferences for helpful comments. Silvia Redaelli provided outstanding research assistance. Financial support from the Polarization and Conflict Project CIT-2-CT-2004-506084 funded by the European Commission-DG Research Sixth Framework Programme and the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

## REFERENCES

ACZÉL, J. (1966). *Lectures on Functional Equations and Their Applications*. New York: Academic Press.

ALESINA, A. F. and LA FERRARA, E. (2000). Participation in heterogeneous communities. *Quarterly Journal of Economics*, **115**, 847–904.

——— and ——— (2005). Ethnic diversity and economic performance. *Journal of Economic Literature*, **43**, 762–800.

———, BAQIR, R. and EASTERLY, W. (1999). Public goods and ethnic divisions. *Quarterly Journal of Economics*, **114**, 1243–84.

———, ——— and HOXBY, C. (2004). Political jurisdictions in heterogeneous communities. *Journal of Political Economy*, **112**, 348–96.

———, DEVLEESCHAUWER, A., EASTERLY, W., KURLAT, S. and WACZIARG, R. (2003). Fractionalization. *Journal of Economic Growth*, **8**, 155–94.

BISIN, A. and VERDIER, T. (2000). Beyond the melting pot: cultural transmission, marriage, and the evolution of ethnic and religious traits. *Quarterly Journal of Economics*, **115**, 955–88.

BLACKORBY, C., PRIMONT, D. and RUSSELL, R. R. (1978). *Duality, Separability, and Functional Structure: Theory and Economic Applications*. Amsterdam: North-Holland.

BOSSERT, W. (1990). An axiomatization of the single-series Ginis. *Journal of Economic Theory*, **50**, 82–92.

———, PATTANAIK, P. K. and XU, Y. (2003). Similarity of options and the measurement of diversity. *Journal of Theoretical Politics*, **15**, 405–21.

CASELLI, F. and COLEMAN, W. J. (2006). On the theory of ethnic conflict. NBER Working Paper no. 12125.
DESMET, K., ORTUÑO-ORTÍN, I. and WEBER, S. (2005). Peripheral diversity and redistribution. CEPR Discussion Paper no. 5112.
DONALDSON, D. and WEYMARK, J. A. (1980). A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory*, **22**, 67–86.
EASTERLY, W. and LEVINE, R. (1997). Africa's growth tragedy: policies and ethnic divisions. *Quarterly Journal of Economics*, **111**, 1203–50.
ESTEBAN, J.-M. and RAY, D. (1994). On the measurement of polarization. *Econometrica*, **62**, 819–51.
FEARON, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, **8**, 195–222.
FERNANDEZ, R., FOGLI, A. and OLIVETTI, C. (2004). Mothers and sons: preference transmission and female labor force dynamics? *Quarterly Journal of Economics*, **119**, 1249–99.
GINI, C. (1912). Variabilità e mutabilità. *Studi Economico-Giuridici della Facoltà di Giurisprudenza dell'Università di Cagliari*, **a.III**, parte II.
GIULIANO, P. (2007). Living arrangements in western Europe: does cultural origin matter? *Journal of the European Economic Association*, **5**, 927–52.
———, SPILIMBERGO, A. and TONON, G. (2006). Genetic, cultural and geographical distances. IZA Discussion Paper no. 2229.
GREENBERG, J. H. (1956). The measurement of linguistic diversity. *Language*, **32**, 109–15.
GUISO, L., SAPIENZA, P. and ZINGALES, L. (2009). Cultural biases in economic exchange? *Quarterly Journal of Economics*, **124**, 1095–131.
KOLENIKOV, S. and ANGELES, G. (2009). Socioeconomic status measurement with discrete proxy variables: is principal component analysis a reliable answer? *Review of Income and Wealth*, **55**, 128–65.
LAITIN, D. (2000). What is a language community? *American Journal of Political Science*, **44**, 142–54.
MAURO, P. (1995). Corruption and growth. *Quarterly Journal of Economics*, **110**, 681–712.
MONTALVO, J. G. and REYNAL-QUEROL, M. (2005). Ethnic polarization, potential conflict, and civil wars. *American Economic Review*, **95**, 796–816.
——— and ——— (2008). Discrete polarization with an application to the determinants of genocides. *Economic Journal*, **118**, 1835–65.
MOODY, J. (2001). Race, school integration, and friendship segregation in America. *American Journal of Sociology*, **107**, 679–716.
MOULIN, H. (1995). On additive methods to share joint costs. *Japanese Economic Review*, **46**, 303–32.
——— (2002). Axiomatic cost and surplus sharing. In K. J. Arrow, A. K. Sen and K. Suzumura (eds), *Handbook of Social Choice and Welfare*, Vol. 1, Amsterdam: Elsevier, pp. 543–96.
——— and SPRUMONT, Y. (2005). On demand responsiveness in additive cost sharing. *Journal of Economic Theory*, **125**, 1–35.
——— and VOHRA, R. (2003). A representation of additive cost sharing methods. *Economics Letters*, **80**, 399–407.
NEHRING, K. and PUPPE, C. (2002). A theory of diversity. *Econometrica*, **70**, 1155–98.
PATTANAIK, P. K. and XU, Y. (2000). On diversity and freedom of choice. *Mathematical Social Sciences*, **40**, 123–30.
PETERS, H. (1992). *Axiomatic Bargaining Game Theory*. Dordrecht: Kluwer.
RAO, R. C. (1982). Diversity: its measurement, decomposition, apportionment and analysis. *Sankhyā*, **44A**, 1–22.
——— (1984). Convexity properties of entropy functions and analysis of diversity. In Y. L. Tong (ed.), *Inequalities in Statistics and Probability*, IMS Lecture Notes 5. Hayward, CA: Institute of Mathematical Statistics, pp. 68–77.
——— and NAYAK, T. K. (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, **IT-31**, 589–93.
REYNAL-QUEROL, M. (2002). Ethnicity, political systems and civil war. *Journal of Conflict Resolution*, **46**, 29–54.
SHAPLEY, L. S. (1953). A value for *n*-person games. In H. W. Kuhn and A. W. Tucker (eds), *Contributions to the Theory of Games*. Princeton, NJ: Princeton University Press, pp. 307–17.
SIMPSON, E. H. (1949). Measurement of diversity. *Nature*, **163**, 688.
SPOLAORE, E. and WACZIARG, R. (2009). The diffusion of development. *Quarterly Journal of Economics*, **124**, 469–529.
VIGDOR, J. L. (2002). Interpreting ethnic fragmentation effects. *Economics Letters*, **75**, 271–6.
WEITZMAN, M. (1992). On diversity. *Quarterly Journal of Economics*, **107**, 363–405.
——— (1998). The Noah's ark problem. *Econometrica*, **66**, 1279–98.
WEYMARK, J. A. (1981). Generalized Gini inequality indices. *Mathematical Social Sciences*, **1**, 409–30.