1. Introduction

The measurement of polarization has received increasing attention in recent years (see, amongst others, Foster and Wolfson 1992, Esteban and Ray 1994 (henceforth ER), Wolfson 1994, Wang and Tsui 2000, Chakravarty and Majumder 2001, Rodriguez and Salas 2003, Duclos, Esteban and Ray 2004 (henceforth DER) and Bossert and Schworm 2008). One of the principal reasons for this interest is the effect that polarization has on a number of social, economic and political phenomena, and in particular those related to social tensions and conflict. However, while most researchers have focused their attention on the measurement of 'income polarization' alone, that is on clustering around local means of the income distribution, only relatively few have attempted to analyze what might be broadly referred to as 'social polarization' (see, for example, D'Ambrosio 2001, Zhang and Kanbur 2001, DER, and Montalvo and Reynal-Querol 2005). Social polarization refers to the situation where the factors determining individuals' identities, and therefore social groups, are culturally, ideologically, biologically or socially driven and do not depend solely on income (classic examples are ethnic, racial, religious and political polarization). The measurement of social polarization is clearly relevant, as in many circumstances the distribution of income is not the only pertinent cause of social conflict (see Easterly and Levine 1997, Esteban and Ray 1999, Montalvo and Reynal-Querol 2005, Collier and Hoeffler 2004 for empirical and theoretical contributions on the existing links between polarization and conflict and other related issues).

Traditional income-polarization measures are implicitly or explicitly based on the assumption that the individuals who are clustered around certain income levels form a cohesive *group* that might potentially express its unrest via social action or revolt. Following ER, individuals are assumed to feel 1) *identified* with other individuals possessing the same income level as them, and 2) *alienated* from individuals with different incomes. Within the bipolarization framework, measures are also implicitly constructed under the assumption that the problems of a society with a declining middle class derive from the presence of large and cohesive 'poor' and 'rich' classes. However, there are obviously a number of other salient characteristics (such as race, ethnicity and gender) that exert considerable influence in the definition of individuals' sense of identity. As argued by Dasgupta and Kanbur (2007, p.1816), "*[…] the nominal distribution of income could give a misleading picture of tensions in society, both within and across communities. Ideologies of community solidarity may well trump those of class solidarity because of the implicit sharing of community resources brought about by community-specific public goods*".

The only polarization measure that, to the best of our knowledge, explicitly accounts for the distribution of groups along ethnic or religious lines is the Reynal-Querol index (henceforth *RQ*). This index is only defined on the basis of the population-weights that these groups represent as it is an indicator of ethnic diversity in the population. It disregards differences in their economic status which, in our opinion, do very often play a role in the polarization process experienced by societies.

One of the aims of the present contribution is to define a social-polarization index which combines the intuition of both of the approaches described above: on the one hand, the partition of the society into groups is performed on the basis of salient social characteristics (race and ethnicity, to mention just two); and, on the other hand, we take into account the extent to which these groups are clustered in certain regions of an attribute's distribution. Alternative approaches with similar aims are described in D'Ambrosio (2001) and Zhang and Kanbur (2001). Our contribution differs regarding the type of attribute variables taken into consideration, which are here qualitative in nature (see below).

Many of the variables that seem relevant for the computation of polarization in existing datasets are categorical or ordinal in nature. In this case, the polarization indices proposed for quantitative variables, such as income, have to be modified. A recent contribution by Apouey (2007) proposes an index for ordinal data, measuring polarization in the distribution of self-assessed health status (SAH). Apouey's polarization measure is an extension of traditional income bipolarization measures, and, as such, does not include information on any salient social characteristics in the analysis. The implicit assumption in Apouey is that the individuals in the same area of the (health) distribution form a cohesive group and no other characteristics matter.

We believe that there are a number of limitations to the approaches in *RQ* and Apouey, which can be illustrated by the following example. Let us assume that the population is divided up into two racial groups (for simplicity, Blacks and Whites) and that there are five self-reported health statuses: Very Poor (VP), Poor (P), Fair (F), Good (G) and Very Good (VG). Consider the following pair of self-assessed health distributions.

[[[Figure 1 around here]]]
[[[Figure 2 around here]]]

Both the *RQ* index and Apouey's measure suggest that polarization is the same in Figures 1 and 2. In the case of the *RQ* index, this is because the calculation only takes into account the population proportion of Blacks and Whites, and disregards the distribution of health, while Apouey's measure takes into account the distribution of health but not the existence of identity groups (Blacks and Whites). However, it seems intuitively clear that the scenario in Figure 1 (where all Blacks are underprivileged and all Whites are privileged) is more polarized than that in Figure 2 (where neither Blacks nor Whites are privileged relative to each other) *if* it is the case that Race is salient in defining individuals' identity.

In some cases we might also be interested in defining the notion of polarization in the context of categorical/nominal data. To the best of our knowledge, these kinds of measures do not yet exist in the literature, even though it is not difficult to imagine circumstances in which the particular distribution of social groups across the different categories of a nominal variable would affect the calculation of social polarization. Consider, as above, a population split in two racial groups, Blacks and Whites, with 50:50 population shares, and a categorical variable, such as place of residence or employment category. The *RQ* index will produce the highest possible level of polarization (equal to 1) irrespective of the distribution of the two groups across the different categories. However, we contend that polarization should be sensitive to the fact that in some cases both racial groups are equally represented across the different categories, whereas in other cases the groups are totally segregated. Intuitively we can argue that the level of social polarization is higher in the latter than in the former case.

This paper thus proposes polarization measures which combine the classic income and social-polarization approaches by exploring the extent to which different social groups, defined on the basis of salient characteristics, are clustered in certain 'privileged or underprivileged regions' of an attribute's distribution measured on a categorical or ordinal scale. For that purpose, we will make use of the Identification-Alienation approach (henceforth IA), which postulates that polarization is proportional to the sum of effective antagonisms existing between individuals. Even if this approach is not based on the primitives of the problem, it has been used in different well-known studies among which we highlight those of ER and DER. The measures are characterized axiomatically in order to provide a normative basis for the appropriateness of their use.

2. The measures

We assume that the population under analysis is composed of N individuals and is partitioned into k exogenously-given groups $G^{(k)}:=\{G_1,\ldots,G_k\}$. We require that this partition accurately reflect individuals' feelings of identity: i.e., we assume that the members of any group feel identified with their peers within their group but alienated with respect to others. The lines along which such groups are defined are typically ethnic or religious, but many other partitions are possible depending on the society under consideration. In the examples above the population was divided up into two racial groups: Blacks and Whites. The population shares of these groups are denoted by $\pi_1,\ldots,\pi_k$ respectively and their absolute size by $N_1,\ldots,N_k$.

In the contexts of both categorical and ordinal data, we assume that individuals belong to C different categories (with ordinal data, these categories are ordered according to a certain criterion). In the examples above there were five self-reported health statuses: Very Poor (VP), Poor (P), Fair (F), Good (G) and Very Good (VG). To describe the distribution of the k groups across the C categories, we define $p_{G_i,c}$ as the share of group G_i in category c. By definition,

$$\sum_{c=1}^{C} p_{G_i,c} = 1 .$$

There are M_c individuals in each category c, so that

$$M_c = \sum_{i=1}^{k} N_i p_{G_i,c} \quad \text{and} \quad \sum_{c=1}^{C} M_c = N .$$

According to the IA approach, individuals are assumed to identify with members of their own group and but feel alienated towards members of the other groups. Our underlying assumption is that each group constitutes a homogeneous body whose members cannot be distinguished from each other when measuring social tension.[1] On the one hand – following ER and DER – the identification component for each member of the group depends on the size of the group to which she belongs (N_i). On the other hand, and given that the members of any group are indistinguishable from each other, the alienation component for every individual in a particular group is assumed to be the same. For this reason, and to keep the exposition simple, we refer to alienation between groups rather than alienation between individuals. Alienation between groups is therefore measured by a function

$$d : \bigcup_{k=2}^{\infty} \left(G^{(k)} \times G^{(k)} \right) \to \Re_+$$

where $\Re_+$ is the set of non-negative real numbers. The polarization literature has proposed a varied number of candidates for $d(G_i,G_j)$. In the context of social polarization, Montalvo and Reynal-Querol (2005) simply assume that $d(G_i,G_j)=1$ for all $G_i, G_j \in G^{(k)}$ with $i \neq j$. In the case of income polarization, ER propose the function $|x-y|$, where x, y are the (log of)

[1] It is of course possible to introduce more sophisticated hypotheses concerning the specification of individuals' feelings of identification and alienation. For instance, one might want to consider alienation between members of the same group G_i but belonging to different categories. Notwithstanding, the main contribution of this paper is not on the modelling of such behavioral-related traits but rather in defining polarization measures on the basis of ordinal and categorical data. Alternative specifications of the identification-alienation hypotheses could be easily incorporated in the framework presented here, an issue that might be attempted in future research.

income levels of different individuals. When defining alienation functions, other authors (see, for instance, D'Ambrosio 2001, and Anderson *et al*. 2010) introduce the overlap coefficient between two income distributions $f_i(x)$, $f_j(x)$, which is defined as

$$\theta_{ij} = \int_{-\infty}^{\infty} \min\{f_i(x), f_j(x)\}dx.$$

In the context of categorical and ordinal data, the overlap coefficient between groups G_i,G_j can be rewritten as

$$\theta_{ij} = \sum_{c=1}^{C} \min\{p_{G_i,c}, p_{G_j,c}\}.$$

This coefficient lies between 0 (disjoint groups) and 1 (perfectly-overlapping groups). Alienation is then defined as $1-\theta_{ij}$, taking the value 0 when the groups overlap completely and 1 when the groups are completely disjoint.[2] Alienation therefore depends on the extent to which the respective groups' representations in the different categories overlap, but *not* on the size of the corresponding groups. The coefficients θ_{ij} only measure the extent to which the distributions of the groups among the different categories are different or not. The greater the degree of overlap, the more similar are the groups, and hence the less the degree of alienation between them. At the other extreme, the lack of overlap between two groups reflects a greater difference and alienation between them. The same reasoning lies behind segregation indices, which typically compare distributions between women and men, or between Blacks and Whites (see, for instance, Charles and Grusky 1995). When one group is concentrated in certain categories where the other is absent, and vice versa, the degree of animosity/alienation between them is greater.

The overlap coefficient implies that the alienation felt between groups is symmetric ($\theta_{ij} = \theta_{ji}$). For ordinal variables, an additional feature can be included in the alienation measure. It has been argued that feelings of alienation between groups should not necessarily be reciprocal. Consider, say, a comparison between a poor and a rich individual: while the poor person has good reason to feel animosity towards the rich person, the opposite might not hold. In this context, alienation between groups '*i*' and '*j*' can be defined as a function of:

$$A_{ij} = \frac{\sum_{s=1}^{N_i}\sum_{t=1}^{N_j}\delta_{st}}{N_i N_j}$$

where δ_{st} equals 1 if individual '*i*' from group *i* is ranked below individual '*t*' from group *j* and 0 otherwise. This procedure yields an asymmetric function ($A_{ij}\neq A_{ji}$), consistent with the alienation felt from underprivileged towards more privileged groups not necessarily being reciprocated[3] (this contrasts with traditional income-polarization measures, where alienation is always symmetric). The value of A_{ij} measures the extent to which group G_i is

[2] It might also be possible to introduce other conceptually-related measures, like the Kolmogorov measure of variation distance (which was used among others by Bossert et al. 2011) or other overlap measures (e.g.: as in Anderson et al. 2010). The choice of such alternative measures, however, would not alter the substantive contributions of this paper.

[3] There are many alternative ways in which an asymmetric function measuring alienation can be defined. However, we have preferred to work with a simple function that faithfully reflects the intuitions on asymmetric alienation put forward in ER and DER – adapted to the ordinal context. Alternative asymmetric functions would not alter substantially the results presented in this paper.

underprivileged with respect to group G_j. When A_{ij}=1, all of the members of group G_i are ranked below any member of group G_j with respect to the ordinal attribute we take into consideration: this is the case of maximal alienation. At the other extreme, A_{ij}=0 when no member of group G_i is ranked below any member of group G_j, which refers to minimal alienation. Recall that, by construction, $A_{ij}+A_{ji}\leq 1$. When there is absolutely no overlap between groups G_i,G_j, then $A_{ij}+A_{ji}=1$. Alternatively, when some members of G_i and G_j belong to the same ordinal category, $A_{ij}+A_{ji}<1$.

According to IA, the effective antagonism felt between two individuals is basically the same as the feeling of alienation, but the individual's feeling of identification influences the effective voicing of their alienation. Antagonism is assumed to be measurable with a function $-T(i,a)-$ that depends on identification and alienation. T is continuous, increasing in its second argument and $T(i,0)=T(0,a)=0$. Finally, according to IA, total polarization is postulated to be proportional to the sum of all effective antagonisms, that is:

$$P(G^{(k)}) \equiv \sum_{s=1}^{k}\sum_{t=1}^{k} N_s N_t T(i(G_s), d(G_s, G_t)). \qquad (1)$$

While this expression is a bit of a black box it serves as the starting point in ER and DER, among others. Equation (1) is a very general expression which is easily adaptable to different contexts. For categorical or ordinal data with symmetric alienation and under the assumptions above, (1) can be rewritten as

$$P_S(G^{(k)}) \equiv \sum_{s=1}^{k}\sum_{t=1}^{k} N_s N_t T(N_s, 1-\theta_{st}). \qquad (2)$$

In the context of ordinal data with asymmetric alienation, (1) becomes

$$P_A(G^{(k)}) \equiv \sum_{s=1}^{k}\sum_{t=1}^{k} N_s N_t T(N_s, A_{st}). \qquad (3)$$

2.1. Axioms and characterization results.

The following properties are used to give a specific functional form to the polarization measure. These axioms share some similarity with those proposed by ER where the variable of interest is income.

Axiom 1. *Consider a two-group society at time t_0 with π_1 being greater than π_2, two categories c=1,2 (in the ordinal case we assume that the second category represents a higher achievement level than the first) and $p_{G_1,1}=1, p_{G_1,2}=0, p_{G_2,1}=d, p_{G_2,2}=1-d$, for some $0<d<1$. Assume now that after some time t_1 the second group splits into two equally-sized groups $\tilde{G}_2,\tilde{G}_3$, with $p_{\tilde{G}_2,1}=d-\varepsilon, p_{\tilde{G}_2,2}=1-d+\varepsilon, p_{\tilde{G}_3,1}=d+\varepsilon, p_{\tilde{G}_3,2}=1-d-\varepsilon$, for some arbitrarily small $\varepsilon<d$. Polarization should not increase following this split.*

The intuition behind this axiom is the following. Before the distributional change, there is a large group in only one category (the 'poor category' in the ordinal case) and a smaller group that is distributed between the first category (with a small presence there) and the second category (the 'rich category' in the ordinal case). After some time, the small group G_2 breaks down in two equally-sized groups $\tilde{G}_2,\tilde{G}_3$ in a way such that the average animosity from the large group G_1 towards the new subgroups is the same as the original animosity with respect to G_2. Given that average animosity remains the same and that the opposition that G_2 might have created against the larger group G_1 has been diluted by its division in two smaller subgroups, we might expect polarization to decrease. Figure 3 illustrates this axiom.

[[[Figure 3 around here]]]

Axiom 2. *Consider a three-group society at time t_0 with π_1 being greater than π_2 and π_3, two categories c=1,2 and* $p_{G_1,1}=1, p_{G_1,2}=0, p_{G_2,1}=d, p_{G_2,2}=1-d, p_{G_3,1}=0, p_{G_3,2}=1$, *with d<0.5. Assume now that, after some time t_1, the distribution of groups in the two categories is* $p_{G_1,1}=1, p_{G_1,2}=0, p_{G_2,1}=d-\varepsilon, p_{G_2,2}=1-d+\varepsilon, p_{G_3,1}=0, p_{G_3,2}=1$, *for some arbitrarily small* $\varepsilon < d$ *. Polarization should not fall following this split.*

Figure 4 illustrates this axiom. There is initially a large group G_1 and two smaller groups, G_2 , G_3, so that the animosity of G_1 towards G_2 is greater than the animosity of G_2 towards G_3. After some time, the animosity of G_1 towards G_2 becomes even larger and that of G_2 towards G_3 even smaller. Consequent to this change we expect polarization to increase, as the smaller groups G_2, G_3 can be seen as a more cohesive opposition to the larger group G_1.

[[[Figure 4 around here]]]

Axiom 3. *If $P(G_1^{(k)}) \geq P(G_2^{(k)})$ and q>0, then $P(qG_1^{(k)}) \geq P(qG_1^{(k)})$, where $qG_1^{(k)}$, $qG_2^{(k)}$ represent population scalings of $G_1^{(k)}$, $G_2^{(k)}$ respectively.*

This is a common invariance axiom in the poverty, inequality and polarization measurement literatures. It states that if populations are scaled up or down (for instance: if all groups are doubled in size but their relative share in the different categories remains unaltered), the comparisons between societies should remain the same. In particular, this allows us to carry out meaningful comparisons between societies of different absolute sizes.

The following set of axioms have to be stated in both symmetric (S) and asymmetric (A) alienation versions (*i.e.*, when alienation is measured by $1-\theta_{ij}$ and A_{ij} respectively). Their underlying meaning is identical.

Axiom 4S. *Assume symmetric alienation. For any population of fixed size N and an arbitrarily large number of categories c=1,2,… consider a distribution where $p_{G_i,i}=1$ for all $G_i \in \{G_1,\ldots,G_k\}$ (so that $p_{G_i,j}=0$ for all j≠i). Then, an increase in the values of k will not increase polarization.*

Axiom 4A. *Assume asymmetric alienation. For any population of fixed size N and any distribution where $p_{G_i,c}=p_{G_j,c}$ for any c, all $G_i,G_j \in \{G_1,\ldots,G_k\}$ and $p_{G_i,c}>0$ for at least two different categories, an increase in the value of k will not increase polarization.*

These axioms capture the broad idea that, other things being equal, the greater the number of groups, the lower is the corresponding polarization. The intuition behind these axioms is as follows: as the different groups become smaller (since the size of the population is fixed and the number of groups has risen) and the alienation between these groups remains constant, their members have less power to effectively voice their unrest, thus reducing the level of social tension. Some authors have used this idea, or very similar ones, in the analysis of conflict and polarization (see, for example, Esteban and Ray 1994, 1999, or Montalvo and Reynal-Querol 2005, who trace this idea from the seminal works of Horowitz 1985). It is important to recall that this axiom would be inappropriate if the purpose were to measure bipolarization, as is the case, for example, of the Wolfson index (see Wolfson 1994).

The following axiom has again to be stated in symmetric and asymmetric alienation versions.

Axiom 5S. *Assume symmetric alienation. Consider a three-group distribution $\{G_1,G_2,G_3\}$ with respective sizes $N_1>N_2=N_3>0$ and $p_{G_i,i}=1$ for $G_i \in \{G_1,G_2,G_3\}$. Then a population mass*

transfer from G_1 to G_2 and G_3 of the same amount without altering the rank of the sizes of the groups will not reduce polarization.

Axiom 5A. *Assume asymmetric alienation. Consider a three-group distribution $\{G_1,G_2,G_3\}$ with respective sizes $N_1>N_2=N_3>0$ and $p_{G_i,c} = p_{G_j,c}$ for any c, all $G_i,G_j \in \{G_1,G_2,G_3\}$ and $p_{G_i,c} > 0$ for at least two different categories. Then a population mass transfer from G_1 to G_2 and G_3 of the same amount without altering the rank of the sizes of the groups will not reduce polarization.*

The underlying intuition behind these axioms is the same. As population mass is transferred from the larger to the smaller groups, groups gradually become more similar, thus equalizing their relative forces and increasing the tension between them. It seems reasonable to say that, other things being equal, a distribution with three equally-populated and equidistant groups exhibits greater levels of social tension than one in which one of the groups is much more populated than the others.

With these axioms we can now present our characterization results.

Theorem 1. *A social-polarization measure as defined in equation (2) satisfies axioms 1, 2, 3, 4S, and 5S if and only if it is proportional to*

$$P_S(G^{(k)}) \equiv \sum_{s=1}^{k}\sum_{t=1}^{k} \pi_s^{1+\alpha}\pi_t(1-\theta_{st}) \qquad (4)$$

where $\alpha \in [\alpha^*,1]$, *with* $\alpha^* = \dfrac{2-\log_2 3}{\log_2 3-1} \approx 0.71.$

Theorem 2. *A social-polarization measure as defined in equation (3) satisfies axioms 1, 2, 3, 4A and 5A if and only if it is proportional to*

$$P_A(G^{(k)}) \equiv \sum_{s=1}^{k}\sum_{t=1}^{k} \pi_s^{1+\alpha}\pi_t A_{st} \qquad (5)$$

where $\alpha \in [\alpha^*,1]$, *with* $\alpha^* = \dfrac{2-\log_2 3}{\log_2 3-1} \approx 0.71.$

Proofs of Theorems 1 and 2: See the Appendix.

These theorems axiomatically characterize our new polarization measures. The indices can be normalized to take values between [0,1] by multiplying them by an appropriate constant.[4] Parameter alpha reflects the degree of polarization sensitivity: if alpha were allowed to take the value of zero (which is *not* the case for the range of admissible values obtained in Theorems 1 and 2), the indices would be equivalent to a fractionalization index (see Bossert et al. 2011). Note that when $\alpha = 1$, and when there is no overlap between the different groups (so that $\theta_{ij} = 0$), $P_S(G^{(k)})$ reduces to the well-known *RQ* index.

Having defined the new polarization measures $P_S(G^{(k)})$ and $P_A(G^{(k)})$ it is of interest to compare their values with those of other indices (such as the *RQ* index or Apouey's index) for the examples used for motivation in the Introduction. Recall the radically-different scenarios depicted in Figures 1 and 2: in the former all Whites (Blacks) declare very good (very poor) health, whereas in the latter, half of Blacks and Whites declare the worst health status and the other half very good health. It is straightforward to check that both Apouey's index and the

[4] In the case of $P_S(G^{(k)})$, the maximum is achieved when there are only two groups of the same size with no overlapping. In that case, $P_S(G^{(k)}) = (1/2)^{1+\alpha}$, so $P_S(G^{(k)})$ can be normalized by multiplying by $2^{1+\alpha}$. In the case of $P_A(G^{(k)})$, the maximum is achieved when there are only two groups G_1,G_2 with $\pi_1=2/3$, $\pi_2=1/3$ and where all members of G_1 are ranked below all members of G_2. In that case, $P_A(G^{(k)}) = ((2/3)^{1+\alpha})/3$, so $P_A(G^{(k)})$ can be normalized by multiplying by $3(3/2)^{1+\alpha}$. The proof of these statements is omitted but is available upon request.

RQ index yield the same level of polarization in both cases. However, both $P_S(G^{(k)})$ and $P_A(G^{(k)})$ reduce their values when moving from the first to the second scenario, a direction that is in line with our intuition on how a polarization index should behave under such circumstances.

3. Concluding remarks

In this paper we have introduced new polarization indices which are appropriate in the context of categorical or ordinal data. This is particularly useful as cardinal information on the phenomena of interest is very often not available. These new measures have been axiomatically characterized, thus providing a normative justification that can be used to gauge their appropriateness vis-à-vis other polarization measures which are currently available in the literature, such as the Montalvo and Reynal-Querol index (2005), the income-polarization indices proposed by ER and DER, and Apouey's (2007) index. There is still much room for further research on polarization on the basis of ordinal and cardinal data. Our treatment of the question in this paper is just a beginning that can be expanded in many directions.

Appendix

Proof of Theorem 1

The proof of Theorem 1 is lengthy and technically involved. It is very similar in structure to the proof of Theorem 2 below, and so will not be presented here to avoid over-lengthening the appendix. It is available to any interested reader upon request.

Proof of Theorem 2

We start with the necessity part of the theorem, that is: if $P_A(G^{(k)})$ is of the form expressed in equation (3) and if axioms *1*, *2*, *3*, *4A* and *5A* hold, then $P_A(G^{(k)})$ must be proportional to $\sum_{s=1}^{k}\sum_{t=1}^{k}\pi_s^{1+\alpha}\pi_t A_{st}$ with $\alpha \in [\alpha^*,1]$, where $\alpha^* = \frac{2-\log_2 3}{\log_2 3 - 1} \approx 0.71$.

Necessity: From equation (3), polarization is proportional to $\sum_{s=1}^{k}\sum_{t=1}^{k} N_s N_t T(N_s, A_{st})$. We start by showing that axioms 1 and 2 imply that T must be linear in alienation. Consider the distribution described in axiom 1. In the initial distribution, the groups are assumed to have population masses N_1, N_2, with $N_1 > N_2$. Total polarization can be written as

$$P_1 \equiv N_1 N_2 [T(N_1, 1-d)].$$

When the second group is split into two equally-sized groups, polarization can be written as

$$P_2 \equiv ((N_1N_2)/2)[T(N_1,\tilde{A}_{12}) + T(N_1,\tilde{A}_{13}) + T(N_2,\tilde{A}_{21}) + T(N_2,\tilde{A}_{31})] + (N_2^2/4)[T(N_2,\tilde{A}_{23}) + T(N_2,\tilde{A}_{32})],$$

where, by construction, $\tilde{A}_{21}=\tilde{A}_{31}=0$, $\tilde{A}_{12}=1-d-\varepsilon$, $\tilde{A}_{13}=1-d+\varepsilon$, $\tilde{A}_{23}=(d+\varepsilon)(1-d+\varepsilon)$, $\tilde{A}_{32}=(d-\varepsilon)(1-d-\varepsilon)$.

According to axiom 1, we should have $P_1 \geq P_2$. This means that

$$N_1N_2[T(N_1,1-d)] \geq$$
$$((N_1N_2)/2)[T(N_1,1-d-\varepsilon) + T(N_1,1-d+\varepsilon)] + (N_2^2/4)[T(N_2,(d+\varepsilon)(1-d+\varepsilon)) + T(N_2,(d-\varepsilon)(1-d-\varepsilon))].$$

In the last term, N_2 can be made arbitrarily small. In the limit, as $N_2 \to 0$, we have

$$T(N_1,1-d) \geq [T(N_1,1-d-\varepsilon) + T(N_1,1-d+\varepsilon)]/2$$

as $T(0,x)=0$ and T is continuous. From the previous expression we can conclude that $T(m,.)$ is

concave in alienation for each $m>0$.
Consider now the distribution presented in axiom 2. The population masses of groups 1, 2 and 3 will be written as m,n,k, where $m \geq n,k$. We write total polarization in terms of ε:

$$P(\varepsilon) \equiv mk[T(m, A_{13})+T(k, A_{31})]+mn[T(m, 1\text{-d}+\varepsilon)+T(n, A_{21})]+ nk[T(n, \text{d}-\varepsilon)+T(k, A_{32})]$$

Given the fact that $A_{13}=1$ and $A_{31}= A_{21}= A_{32}=0$, the last term can be rewritten as

$$P(\varepsilon) \equiv mk[T(m, 1)]+mn[T(m, 1\text{-d}+\varepsilon)]+ nk[T(n, \text{d}-\varepsilon)].$$

According to axiom 2, $P(\varepsilon) \geq P(0)$, so we must have that

$$mn[T(m, 1\text{-d}+\varepsilon)]+ nk[T(n, \text{d}-\varepsilon)] \geq mn[T(m, 1\text{-d})]+ nk[T(n, \text{d})].$$

Letting $n,k \rightarrow m$, the last expression can be rewritten as

$$[T(m, 1-d+\varepsilon)]- [T(m, 1-d)] \geq [T(m,d)]- [T(m, d-\varepsilon)].$$

Since $1-d>0.5>d$, the final expression shows that $T(m,.)$ is convex in alienation for each $m>0$. Hence, $T(m,d)$ must be linear in d for all $m>0$, so $T(m,d)$ can be written as $\varphi(m)d$.
We next show that $\varphi(m)=km^{\alpha}$ for some $k,\alpha>0$. Consider the following two-group distribution: $\{G_1,G_2\}$ with $A_{12}=d_1$, $A_{21}=d_2$, and population masses m,n respectively. Total polarization is proportional to $mn(\varphi(m)d_1+ \varphi(n)d_2)$. Now, for each m' it is possible to find n' such that $mn(\varphi(m)d_1+ \varphi(n)d_2)= m'n'(\varphi(m')d_1+ \varphi(n')d_2)$. By axiom 3 it follows that for any $\lambda>0$,

$$\lambda^2 mn(\varphi(\lambda m)d_1+ \varphi(\lambda n)d_2)= \lambda^2 m'n'(\varphi(\lambda m')d_1+ \varphi(\lambda n')d_2).$$

Combining the last two expressions we have that

$$\frac{\varphi(m)d_1 + \varphi(n)d_2}{\varphi(\lambda m)d_1 + \varphi(\lambda n)d_2} = \frac{\varphi(m')d_1 + \varphi(n')d_2}{\varphi(\lambda m')d_1 + \varphi(\lambda n')d_2}$$

Taking the limits $n,n' \rightarrow 0$, we have that $\varphi(n), \varphi(n') \rightarrow 0$, so the final expression becomes

$$\frac{\varphi(m)}{\varphi(\lambda m)} = \frac{\varphi(m')}{\varphi(\lambda m')}$$

Setting $\lambda=1/m$ and $l=m'/m$ the final expression is equivalent to the fundamental Cauchy equation $\varphi(m)\varphi(l)=\varphi(ml)\varphi(1)$, where φ is a continuous function. According to Aczél (1963, p. 41, Theorem 3), the solutions of this equation are always of the form $\varphi(m)=km^{\alpha}$ for some k,α. Depending on the sign of k and α, $\varphi(m)$ can be an increasing or decreasing function. We will prove that it must here be an increasing function. We again take the distribution presented in axiom 2 expressed in terms of ε:

$$P(\varepsilon) \equiv mk\varphi(m)+mn[\varphi(m)(1\text{-d}+\varepsilon)]+ nk[\varphi(n)(\text{d}-\varepsilon)].$$

Since $P(\varepsilon) \geq P(0)$ we must have that

$$mn\varphi(m)(1\text{-d}+\varepsilon)+ nk\varphi(n)(\text{d}-\varepsilon) \geq mn\varphi(m)(1\text{-d})+ nk\varphi(n)\text{d}.$$

Therefore

$$m\varphi(m) \geq k\varphi(n).$$

Taking $k \rightarrow m$, we obtain $\varphi(m) \geq \varphi(n)$, where $m \geq n$. Therefore φ is increasing, so we must have that $k,\alpha>0$.

Hence, $P_A(G^{(k)})$ is proportional to $\sum_{s=1}^{k}\sum_{t=1}^{k} N_s^{1+\alpha} N_t A_{st}$ for some $\alpha>0$. We will now show that axioms *4A* and *5A* force α to belong to the interval $[\alpha^*,1]$.
Consider the situation described in axiom 4A. The total population mass (N) is divided into k equally-populated subgroups. Polarization can be written in terms of k as

$$P(k) \equiv \sum_i \sum_{j \neq i} \left(\frac{N}{k}\right)^{2+\alpha} \delta = \delta \left(\frac{N}{k}\right)^{2+\alpha} k(k-1)$$

where δ is the distance between any two groups (by construction, all of the distances between groups are the same). Axiom 4A says that $P(k)$ should be a non-increasing function of k. If we impose the restriction that $P(k) \geq P(k+1)$, we have that $k^{2+\alpha} \leq (k-1)(k+1)^{1+\alpha}$ must hold true

for any $k \geq 2$. Taking the corresponding logarithms and after some basic algebraic manipulation, the previous restriction can be rewritten as:

$$\alpha \geq \frac{\log_k(k+1)[1+\log_{k+1}(k-1)]-2}{1-\log_k(k+1)} =: f(k)$$

Consider now the function $F(x)$ defined as the continuous extension of $f(k)$ to the set of real numbers greater than 1. It is now straightforward to check that $(\partial F/\partial x)(x) \leq 0$ for all $x \geq 2$. This implies that $f(k)$ is a non-increasing function. Hence, a necessary and sufficient test case to find the set of admissible values for α is to impose that k takes the smallest possible value in the previous expression (which is $k=2$). We therefore obtain the lower bound for α as

$$\alpha \geq \frac{\log_2(3)-2}{1-\log_2(3)} = f(2) = \alpha^*$$

.

We now impose the condition stated in axiom 5A to find the upper bound for the admissible values for α. Without loss of generality we will assume that the whole population mass (N) is normalized to 1. The population mass distribution for the three groups is then simply $1-2x$, x, x, with $x \leq 1-2x$ (that is, $x \leq 1/3$). Since we do not allow the transfer process to reverse the size rank of the groups, x cannot be greater than 1/3. Polarization can be written in terms of x as:

$$P(x) \equiv 2\delta[(1-2x)^{1+\alpha}x + x^{1+\alpha}(1-2x) + x^{2+\alpha}]$$

where $\delta > 0$ is the alienation between any two groups. According to axiom 5A, $P(x)$ must be a non-decreasing function in x for all $x \in [0,1/3]$. We now compute the first derivative:

$$(\partial P/\partial x) \equiv (1-2x)^{1+\alpha} + \alpha x^{1+\alpha} + (1+\alpha)[x^{\alpha}(1-2x) - 2x(1-2x)^{\alpha}].$$

We wish to identify the values of α for which $(\partial P/\partial x)(x)$ is non-negative for all $x \in [0,1/3]$. We now define the following function:

$$F(x) := \frac{\frac{\partial P}{\partial x}(x)}{(1-2x)^{1+\alpha}} = 1 + \alpha\left(\frac{x}{1-2x}\right)^{1+\alpha} + (1+\alpha)\left[\left(\frac{x}{1-2x}\right)^{\alpha} - 2\left(\frac{x}{1-2x}\right)\right].$$

Recall that if $x \in [0,1/3], (1-2x)^{1+\alpha} > 0$ the sign of $(\partial P/\partial x)(x)$ is the same as that of $F(x)$. Renaming variables, we define $z := x/(1-2x)$, so that $F(x)$ can be rewritten in terms of z as

$$G(z) = 1 + \alpha z^{1+\alpha} + (1+\alpha)[z^{\alpha} - 2z].$$

It is straightforward to check that for $x \in [0,1/3]$, the function $x/(1-2x)$ increases monotonically and is bounded between 0 and 1. We then wish to identify the values of α for which $G(z)$ is non-negative for all $z \in [0,1]$.

Note that $G(z)$ is a continuous function with $G(0)=1$ and $G(1)=0$. Moreover, we have that $(\partial G/\partial z) = \alpha(\alpha+1)z^{\alpha} + (\alpha+1)[\alpha z^{\alpha-1} - 2]$. In particular, $(\partial G/\partial z)(1) = 2(\alpha^2 - 1)$. This means that for $\alpha > 1$, $(\partial G/\partial z)(1) > 0$. By the continuity of G, for some arbitrarily small $\varepsilon > 0$, we must have that $G(1-\varepsilon) < 0$. Hence, axiom 5A does not hold whenever α is strictly larger than 1.

It is straightforward – but tedious – to show that when $\alpha \in (0,1], G(z) \geq 0$ (the proof is available upon request), so $(\partial P/\partial x)(x)$ is non-negative for all $x \in [0,1/3]$. This is the range of values of α for which axiom 5A holds. Combining the restrictions contained in axioms 4A and 5A, we conclude that α must lie in the interval $[(\log_2(3)-2)/(1-\log_2(3)), 1]$.

This proves the necessity part of the Theorem.

Sufficiency: Consider the polarization index $P_A(G^{(k)}) \equiv \sum_{s=1}^{k}\sum_{t=1}^{k} \pi_s^{1+\alpha} \pi_t A_{st}$. It is straightforward to check that $P_A(G^{(k)})$ satisfies axioms *1, 2, 3, 4A* and *5A* (the proof is available upon request).

This completes the proof of the theorem.

Q.E.D.

References

Alesina, A. and E. Spolaore (1997) "On the number and size of nations" *Quarterly Journal of Economics* **113**, 1027-1056.

Anderson, G., G. Ying and L. Wah (2010) "Distributional overlap: simple, multivariate, parametric and non-parametric tests for alienation, convergence and general distributional difference issues" *Econometric Reviews* **29**, 247-275.

Apouey, B. (2007) "Measuring health polarization with self-assessed health data" *Health Economics* **16**, 875-894.

Bossert, W. and W. Schworm (2008) "A Class of Two-Group Polarization Measures" *Journal of Public Economic Theory* **10**, 1169-1187.

Bossert, W., C. D'ambrosio and E. la Ferrara (2011) "A Generalized Index of Fractionalization" *Economica* **78**, 723-750.

Chakravarty, S. and A. Majumder (2001) "Inequality, polarization and welfare: theory and applications" *Australian Economic Papers* **40**, 1-13.

Charles, M. and D. Grusky (1995) "Models for Describing the Underlying Structure of Sex Segregation" *American Journal of Sociology* **100**, 931-71.

Collier, P. and A. Hoeffler (2004) "Greed and grievance in civil war" *Oxford Economic Papers* **56**, 563-595.

D'Ambrosio, C. (2001) "Household characteristics and the distribution of income in Italy: an application of social distance measures" *Review of Income and Wealth* **47**, 43-64.

Dasgupta, I. and R. Kanbur (2007) "Community and class antagonism" *Journal of Public Economics* **91**, 1816-1842.

Duclos, J-Y., J. Esteban and D. Ray (2004) "Polarization: Concepts, Measurement, Estimation" *Econometrica* **72**, 1737-72.

Easterly, W. and R. Levine (1997) "Africa's growth tragedy: policies and ethnic divisions" *Quarterly Journal of Economics* **112**, 1203-50.

Esteban J. and D. Ray (1994) "On the Measurement of Polarization" *Econometrica* **62**, 819-852.

Esteban, J. and D. Ray (1999) "Conflict and distribution" *Journal of Economic Theory* **87**, 379-415.

Foster, J. and M. Wolfson (1992) "Polarization and the decline of the middle class: Canada and the US", Mimeo, Vanderbilt University (now published in the *Journal of Economic Inequality* (2010) **8**, 247-273).

Montalvo, J. and M. Reynal-Querol (2005) "Ethnic polarization, potential conflict and civil wars" *American Economic Review* **95**, 796-815.

Rodriguez, J. and R. Salas (2003) "Extended bi-polarization and inequality measures" *Research on Economic Inequality* **9**, 69–83.

Wang, Y. and K. Tsui (2000) "Polarization orderings and new classes of polarization indices" *Journal of Public Economic Theory* **2**, 349-363.

Wolfson, M. (1994) "When inequalities diverge" *American Economic Review, Papers and Proceedings* **84**, 353-358.

Zhang, X. and R. Kanbur (2001) "What difference do polarization measures make? An application to China" *Journal of Development Studies* **37**, 85-98.