Supporting Information

Chodrow 10.1073/pnas.1708201114

1. Visualization of Demographic Distributions
The text makes heavy use of a map visualization, which color-codes tracts according to the group that is most overrepresented there relative to the region-wide distribution. Formally, for each spatial location \( x \), the hue is the argument \( y \) such that

\[
y^* = \arg \max_y p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{p_Y(y)}. \tag{S1}
\]

and the saturation corresponds to the value of the maximization.

2. Generality of Bregman Information

2.1. Smoothing-Based Measures Through Bregman Divergences. The purpose of this section is to show that many state of the art measures of segregation according to their information-theoretic structure.

The following functions \( f : \mathcal{P} \rightarrow \mathbb{R} \) are strictly convex on their domain and continuously differentiable on \( \text{int} \mathcal{P} \).

- Euclidean norm: \( f_1(p) = \|p\|^2 \).
- Negative entropy: \( f_2(p) = \sum_i p_i \log p_i \).
- Cumulative entropy: \( f_3(p) = (g_3 \circ \sigma)(p) \), where \( g_3(c) = \sum_i c_i \log c_i \).
- Cumulative variance: \( f_4(p) = (g_4 \circ \sigma)(p) \), where \( g_4(c) = -4 \sum_j c_j (1 - c_j) \).
- Cumulative root variance: \( f_5(p) = (g_5 \circ \sigma)(p) \), where \( g_5(c) = -2 \sum_j \sqrt{c_j (1 - c_j)} \).
- Square mean: Assume additionally that the alphabet \( \mathcal{Y} \) is an alphabet of integers, and let \( f_6(p) = \left( \sum_i p_i y_i \right)^2 \).

The function \( \sigma : \mathcal{P} \rightarrow \mathbb{R}_+^n \) is the cumulative summation map given by \( \sigma(p) = \sum_{i=1}^n b_i \).

- Cumulative entropy: The map \( \sigma \) is linear, and \( f_1 \) is, therefore, strictly convex if \( g_1 \) is strictly convex on \( \mathbb{R}_+^n \). We note that

\[
\frac{\partial^2 g_1(c)}{\partial c_i \partial c_j} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}.
\]

We, therefore, have that \( H g_1(c) \) is diagonal and consists of positive entries everywhere in \( \mathbb{R}_+^n \) and that it is, therefore, positive definite as required.

- Cumulative variance: Convexity is unaffected by linear terms. Since \( g_3(c) - 4 e^c = 4 \|c\|^2 \) is strictly convex on \( \mathbb{R}_+^n \), so is \( g_3 \).

As with \( g_3(c) \), \( H g_3 \) is, therefore, diagonal and consists of positive entries, and it is, therefore, positive definite.

- Square mean: The map \( y \mapsto \sum_i p_i y_i \) is linear, and the squaring operation is strictly convex.

Corollary 1. The spatial exposure index between group \( m \) and group \( n \) is defined by ref. 6 as

\[
P_{mn} = \int p(m|x)p(n|x)dP_X(x). \tag{3}
\]

The spatial exposure index satisfies

\[
\sum_{m \neq n} P_{mn} = 1 + H(Y|X),
\]

where the divergence generating function is the Euclidean norm \( f_1(p) = \|p\|^2 \).

Corollary 2. The Divergence Index \( D \) of ref. 26 satisfies

\[
D = I(X, Y),
\]

where the divergence generating function is the entropy \( f_2(p) = \sum_i p_i \log p_i \).

Proof: Equation 7 of ref. 6, for example, defines \( \tilde{H} \) as

\[
\tilde{H} = \frac{H(Y) - H(Y|X)}{H(Y)},
\]

from which the result follows by the standard identity \( I(X, Y) = H(Y) - H(Y|X) \).

Corollary 3. The Information Theory Index \( \tilde{H} \) of refs. 6, 17, and 28 satisfies

\[
\tilde{H} = \frac{I(X, Y)}{H(Y)},
\]

where the divergence generating function is the entropy \( f_2(p) = \sum_i p_i \log p_i \).

Proof: Equation 7 of ref. 6, for example, defines \( \tilde{H} \) as

\[
\tilde{H} = \frac{H(Y) - H(Y|X)}{H(Y)},
\]

from which the result follows by the standard identity \( I(X, Y) = H(Y) - H(Y|X) \).

Corollary 4. The Ordinal Information Theory Index \( H^O \), the Ordinal Variation Ratio \( RO \), and the Ordinal Square Root Index \( S^O \) of ref. 28 are all of the form

\[
I(X, Y) \quad \tilde{I}(N, X) \quad \hat{I}(N, X) \quad \bar{I}(N, X) \quad \hat{I}(N, X) \quad \bar{I}(N, X)
\]

where the divergence-generating functions are \( f_6, f_4, f_5 \), and \( f_3 \) respectively.

Corollary 5. Let \( N \) be a random variable with \( H(N|X) = 0 \); that is, \( N \) is completely determined by \( X \). The Generalized Neighborhood Sorting Index \( G \) of ref. 27 satisfies

\[
G = \sqrt{\frac{\tilde{I}(N, Y)}{\tilde{I}(X, Y)}},
\]

where the divergence generating function is \( f_6 \) and where \( \hat{I}(N, X) \) is computed with respect to the spatially smoothed distribution \( \Phi(p) \), in which \( \Phi \) is the uniform ego-network smoother of order \( n \).

Proof: Since we have already proved the convexity and continuous differentiability of \( f_6 \), the only task required is to cast equation 4 of ref. 27 in the claimed form. The numerator of this equation may be written

\[
\int d\tilde{p}(\tilde{p}(-|n), \tilde{p}) d\mathcal{N}(n),
\]
and the denominator may be written
\[ \int d_j (\tilde{p}(x), \tilde{p}) dP_X(x), \]
as required.

3. Algorithms

Let \( c : \mathcal{X} \to \{1, \ldots, k\} \) be a function that assigns to each location \( x \) a cluster label \( c(x) \). We regard \( C = c(\mathcal{X}) \) as a random variable and aim to choose \( c \), such that the aggregation that it induces captured segregation at large spatial scales. The Chain Rule of Bregman information (29) offers a decomposition of the form
\[ I_f(X, Y) = I_f(C, Y) + I_f(X, Y|C). \]  

The term \( I_f(C, Y) \) gives the segregation captured at the aggregate spatial scale, and \( I_f(X, Y|C) \) is the residual segregation at lower scales. A good labeling function will tend to make the first term large. This motivates the following problem:
\[ \begin{align*}
\argmax_{\mathcal{E}} &\ I_f(c(X), Y) \\
\text{subject to} &\ \text{spatial constraints.}
\end{align*} \]  

We consider two approaches to this problem. Agglomerative hierarchical clustering directly addresses this problem using a stage-wise greedy approach, while spectral partitioning solves a related, first-order approximation to this problem in an approximate but global fashion.

3.1. Agglomerative Partitioning. Given a Bregman divergence, the Jensen–Bregman divergence (39) between locations is defined by the formula
\[ d_{\beta}(x_1, x_2) = p_X(x_1)d_{\beta}(p|X=x_1, p|X=x_2) \\
+ p_X(x_2)d_{\beta}(p|X=x_2, p|X=x_1). \]

The Jensen–Bregman divergence \( d_{\beta}(x_1, x_2) \) measures the Bregman information loss associated with merging spatial units \( x_1 \) and \( x_2 \) into a single unit and combining their demographics. When the divergence is small, \( x_1 \) and \( x_2 \) have similar demographic characterstics, and we may cease to distinguish them with only small information loss. Greedy regionalization proceeds stage-wise. At each stage, we identify two adjacent spatial units that can be merged with minimal information loss; formally, we solve
\[ x_1^*, x_2^* = \argmin_{x_1, x_2 \in \mathcal{E}} d_{\beta}(x_1, x_2). \]  

We then merge \( x_1^* \) and \( x_2^* \), repeating until only the desired number of spatial units remains. This procedure is formalized in Algorithm 1.

**Algorithm 1:** Agglomerative partitioning

1: function AGGLOMPARTITION(\( R \), \( k \))
2: while |\( R \)| > \( k \) do
3: \( x_1^*, x_2^* = \argmin_{x_1, x_2 \in \mathcal{E}} d_{\beta}(x_1, x_2). \)
4: aggregate(\( x_1^*, x_2^* \))
5: end while
6: return \( R \)
7: end function

3.2. Spectral Partitioning. Spectral graph partitioning according to ref. 40 aims to approximately solve the normalized cut problem. Given an undirected graph \( G \) with weights \( w_{ij} \) between nodes \( i \) and \( j \), spectral graph partitioning seeks cuts in \( G \), such that the subsets defined by the cuts have strong connections within themselves but only weak ones between them. Formally, it aims to approximately solve the normalized cut problem:
\[ \begin{align*}
\argmin_{c : \mathcal{G} \to \{ \ldots, k \}} &\ \sum_{\ell} \frac{\text{cut}(c, \ell)}{\text{vol}(c, \ell)},
\end{align*} \]  

where \( \text{cut}(c, \ell) = \sum_{c(i) = \ell, c(j) \neq \ell} w_{ij} \) and \( \text{vol}(c, \ell) = \sum_{c(i) = \ell, c(j) \neq \ell} w_{ij} \). To use spectral partitioning, we construct the entries of the weight matrix \( W \) in terms of the Jensen–Bregman divergence:
\[ w_{ij} = \begin{cases} 
-\frac{d_{\beta}(x_i, x_j)}{\sigma} & \quad (i, j) \in \mathcal{E} \\
0 & \quad \text{otherwise.}
\end{cases} \]  

To approximately solve Eq. S5, we form the “random walk” normalized Laplacian (41) given by
\[ L = D^{-1}(D - W), \]
where \( D = \text{diag}(W) \) is the diagonal matrix of row sums of \( W \). The eigenvectors of \( L \) then form a feature space, in which each additional dimension contains increasingly detailed information about the cut structure of \( G \). Performing \( k \)-means clustering on the \( k \) eigenvectors corresponding to the \( k \) smallest eigenvalues then yields the partition. This procedure is formalized in Algorithm 2.

When \( d_{\beta} \) is the Kullback–Leibler divergence, \( I_f(X, Y) \) is the Shannon information between \( X \) and \( Y \). In this case, the weight matrix \( W \), the hyperparameter \( \sigma \), and normalized cut problem [S5] all have attractive probabilistic interpretations. In this case, the Chernoff bound implies that \( w_{ij} \) is asymptotic equal to the probability of randomly selecting \( x_i \) and \( x_j \)’s residents from area unit \( x_1 \) with demographic distribution equal to \( p_{\mathcal{X}}(x_1) \) and then similarly selecting \( x_i \) and \( x_j \)’s residents from area unit \( x_2 \) with empirical distribution equal to \( p_{\mathcal{X}}(x_1) \). This probability is small when \( x_1 \) and \( x_2 \) have very different demographic distributions. The parameter \( \sigma \) may be viewed as a regularizer; when it is large, the “effective populations” of each spatial unit are small, and the weights \( w_{ij} \) therefore, tend to be more uniform. The expressions \( \text{cut}(c, \ell) \) and \( \text{vol}(c, \ell) \) are then interpretable as first-order approximations to the probability of confusing two units within region \( \ell \) and confusing two units along the boundary of region \( \ell \), respectively.

**Algorithm 2:** Spectral partitioning

1: function SPECTRALPARTITION(\( R \), \( \sigma \), \( k \))
2: for \( i, x_i \in \mathcal{E} \) do
3: \( A_{ij} \leftarrow \exp[-d_{\beta}(x_i, x_j)/2\sigma] \)
4: end for
5: \( D \leftarrow \text{diag}(A) \)
6: \( L \leftarrow D^{-1}(D - A) \)
7: \( Y \leftarrow [x_1, \ldots, x_k] \) w.r.t. the \( \ell \)th eigenvector of \( L \)
8: return \( k \mbox{means}(Y, k) \)
9: end function

4. Sample Eigenvectors for Detroit

Spectral partitioning inspects the eigenvectors of the normalized Laplacian \( L \) to find regional structure. We show in Fig. S1 a sampling of eigenvectors in Detroit to illustrate the information that they contain. The \( k \)-means algorithm in the eigenspace is used to optimally aggregate this information into a final regionalization (42).

5. Spectral Partitioning Without Hierarchical Postprocessing

In the text, we show the results of spectral partitioning with hierarchical postprocessing to generate final regionalizations for Detroit, Chicago, and Philadelphia. For illustrative purposes, we show in Fig. 2 the intermediate stage of spectral partitioning using \( \sigma = 30 \). The values of \( k \) may be chosen by inspecting the spectrum of the normalized Laplacian; additional details are in ref. 41. Fig. S24 may be usefully contrasted with Fig. 1.

6. Manifold View of Local Scale

6.1. The Information Manifold and the Metric Tensor. A geometric view of the metric tensor \( g \) may be obtained from a manifold view of spatial compositional data analysis. Fig. S3
illustrates this framework in a selected subregion of Detroit using just three ethnoracial categories for visualization purposes. Fig. S3A shows the spatial units as provided by the US Census dataset. In Fig. S3B, we construct the adjacency network of spatial units and explicitly show the demographic composition of each one. By mapping each spatial unit to its demographic distribution, we obtain a set of points on the probability simplex $P$, with edges between geographically adjacent units (Fig. S3C). We view this construction as approximating a smooth submanifold $M$ of $P$. Distances between points in this space are measured according to the metric tensor $g$, which is fully determined by the Bregman divergence $f$ as discussed in the text. The role of the metric tensor is to provide a method of translating between geographic distances and demographic distances. This underlies its role as a local scale measure: when one can travel in geographic space without demographic change, this implies that spatial variation, if present, must be locally on larger scales. In particular, the geodesic distance between points on $M$ is defined as the shortest path along $M$ between those points under $g$: \[
abla(x_1, x_2) = \arg\min_{\gamma \in C_M(x_1, x_2)} \int_0^1 \sqrt{g(\gamma(t), \gamma(t))} dt. \] [S8]

$C_M(x_1, x_2)$ is the set of unit-speed curves along $M$, such that, for any $\gamma \in C$, $\gamma(0) = x_1$ and $\gamma(1) = x_2$. We may interpret $\delta(p_1, p_2)$ as the minimal amount of demographic change one would undergo in traveling from location $x_1$ to location $x_2$.

6.2. Proof of Theorem 1. We now present a proof of the fully general statement of Theorem 1. Define the local information $j(x_0)$ as

\[
j(x_0) \triangleq \lim_{r \to 0} \frac{1}{r^2} \int_{B_r(x_0)} I(X, Y | X \in B_r(x_0)).
\] [S9]

where $B_r(x_0) = \{x \in R^2 | \|x - x_0\|^2 \leq r^2\}$. For simplicity, fix $x_0$, and let $B_r = B_r(x_0)$. In this section, we will prove the following theorem.

Theorem. Let $R \subset \mathbb{R}^n$ be compact and of nonzero measure under a finite measure $\mu$ that is absolutely continuous with respect to the Lebesgue measure $\lambda$. Assume that the map $\alpha : x \mapsto p_Y | X = x$ is smooth. Then, the local information $j(x)$ exists at all $x \in \text{int}(R)$ and satisfies

\[
j(x_0) = \frac{1}{2n + 2} \text{tr} \; g_x.
\] [S10]

The proof proceeds essentially by direct calculation; to structure the calculation in a coherent way, we divide it into a series of lemmas.

Recall that, by hypothesis, $P_X$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^n$ and therefore, has a Radon–Nikodym derivative (probability density function) $p_X$, so that

\[
\int f \, dP_X = \int f p_X \, d\lambda
\]

for any $f$. We also assumed that $p_X$ is smooth and will proceed to take derivatives and Taylor expansions accordingly.

Let $V_n(r)$ be the volume of the $n$ ball of radius $r$ and $S_{n-1}(r)$ be the volume of the $n - 1$ sphere of radius $r$. We will abbreviate $V_n = V_n(1)$ and $S_n = S_n(1)$. We have $V_n(r) = r^n V_n$ and $S_{n-1}(r) = r^{n-1} S_n$.

We use the standard symbol $o(f(r))$ to denote terms satisfying

\[\lim_{r \to 0} \frac{o(f(r))}{f(r)} = 0.\]

Lemma 1. We have

\[
\mathbb{P}(X \in B_r) = \mathbb{P}(X = x_0) V_n(r) + o(r^{n+1}).
\]

Proof: We compute

\[
\mathbb{P}(x \in B_r) = \int_{B_r} dP_X
\]

\[
= \int_{B_r} p_X(x) \, d\lambda(x)
\]

\[
= \int_{B_r} [p_X(x_0) + Dp_X(x_0)(x - x_0)
\]

\[
+ o(\|x - x_0\|) \, d\lambda(x).
\]

The middle term in the integral vanishes via spherical symmetry, and we obtain

\[
\mathbb{P}(x \in B_r) = \int_{B_r} [p_X(x_0) + o(\|x - x_0\|)] \, d\lambda(x)
\]

\[
= p_X(x_0) V_n(r) + \int_{B_r} o(r) \, d\lambda(x)
\]

\[
= p_X(x_0) V_n(r) + o(r^{n+1})
\]

as was to be shown.

For notational convenience, let $\alpha(x) = p(\cdot | x)$ and $a = \mathbb{P}(X \in B_r)$.

Lemma 2. We have

\[a = \alpha(x_0) + o(r).\]

Proof: We compute

\[a = \mathbb{E}[\alpha(X) | X \in B_r]
\]

\[= \int_{B_r} \alpha(x) \, dP_X | X \in B_r(x)
\]

\[= \int_{B_r} \alpha(x) p(x | X \in B_r) \, d\lambda(x)
\]

\[= \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} \alpha(x) p(x) \, d\lambda(x).
\]

Since $\alpha(x)$ and $p(x)$ are both smooth, their product is as well, and we may Taylor expand about $x = x_0$ to obtain

\[a = \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} [\alpha(x_0) p(x_0) + T(x - x_0)
\]

\[+ o(\|x - x_0\|)] \, d\lambda(x),
\]

where $T$ stands for a linear map that we need not calculate, since this term vanishes in the integral through spherical symmetry. Thus,
\[ a = \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} \left[ \alpha(x_0) p(x_0) + o(\|x - x_0\|) \right] d\lambda(x) \]

\[ = \frac{\alpha(x_0) p(x_0) V_n(r) + o(r^{n+1})}{\mathbb{P}(X \in B_r)} \]

\[ = \frac{\alpha(x_0) p(x_0) V_n(r) + o(r^{n+1})}{p(x_0) V_n(r) + o(r^{n+1})} \quad \text{(Lemma 1)} \]

\[ = \alpha(x_0) + o(r), \] as was to be shown.

**Lemma 3.** For any \( x \in B_r, \)

\[ d_f(p(\cdot|x), p(\cdot|X \in B_r)) = d_f(p(\cdot|x), p(\cdot|x_0)) + o(r^2). \]  

**Proof:** The proof proceeds by exploiting the local quadratic structure of the Bregman divergence \( d_f: \)

\[ d_f(q + \delta, q) = \frac{1}{2} \mathcal{H}_{f_a}(\delta, \delta) + o(\|\delta\|^2). \]  

Using the same notation \( \alpha(x) = p(\cdot|x) \) and \( a = p(\cdot|X \in B_r) \) from Eq. S12, we obtain

\[ d_f(p(\cdot|x), p(\cdot|X \in B_r)) = d_f(\alpha(x), a) \]

\[ = \mathcal{H}_{f_a}(\alpha(x) - a, \alpha(x) - a) + o(\|a(x) - a\|^2). \]

Since \( p(y|x) \) is smooth as a function of \( x, \) so is \( \alpha(x), \) and the final term is, therefore, \( o(\|a\|^2). \) Rearranging terms, we can write

\[ d_f(p(\cdot|x), p(\cdot|X \in B_r)) = \mathcal{H}_{f_a}(\alpha(x) - a, \alpha(x) - a) + (\mathcal{H}_{f_a} - \mathcal{H}_{f_a}(\alpha)) (\alpha(x) - a, \alpha(x) - a) + o(\|a(x) - a\|^2). \]

Since \( f \) is smooth, the components of the tensor \( \mathcal{H}_{f_a} - \mathcal{H}_{f_a}(\alpha) \) are \( o(r). \) Furthermore, by Lemma 2, \( \alpha(x) - a = \alpha(x) - \alpha(x_0) + o(r) = O(r) + o(r). \) The entire second term is, therefore, \( o(r^2). \)

Turning to the first term, we have

\[ \mathcal{H}_{f_a}(\alpha(x) - a, \alpha(x) - a) \]

\[ = \mathcal{H}_{f_a}(\alpha(x) - \alpha(x_0) + o(r), \alpha(x) - \alpha(x_0) + o(r)) \]

\[ = \mathcal{H}_{f_a}(\alpha(x) - \alpha(x_0), \alpha(x) - \alpha(x_0) + o(r^2)) \]

\[ = d_f(\alpha(x), \alpha(x_0)) + o(r^2), \]

where in the second line, we have used the fact that \( \alpha(x) - \alpha(x_0) = O(r). \) This completes the proof.

**Lemma 4.** Let \( T \) be a real, symmetric bilinear form and \( \Delta_2 \) be the diagonal operator. Then,

\[ \int_{B_r(0)} T \circ \Delta_2 d\lambda = \frac{\rho^{n+2} S_{n-1}}{n(n+2)} \text{tr} \left( T \right). \]

**Proof:** By the spectral theorem, there exists an orthonormal basis, in which the matrix \( A \) of \( T \) is diagonal, and its entries are the eigenvalues \( \{ \lambda_i \} \) of \( T. \) Since \( B_r(0) \) is radially symmetric about the origin, we may integrate in this basis instead, obtaining

\[ \int_{B_r(0)} T \circ \Delta_2 d\lambda = \int_{B_r(0)} v^T A v d\lambda(v). \]

Since \( A \) is diagonal,

\[ \int_{B_r(0)} v^T A v d\lambda(v) = \int_{B_r(0)} \sum_i \lambda_i v_i^2 d\lambda(v) = \sum_i \lambda_i \int_{B_r} v_i^2 d\lambda. \]

By spherical symmetry, the integrals inside the sum are all equal to \( 1/n \int_{B_r(0)} \|v\|^2 d\lambda(v), \) and we obtain

\[ \sum_i \lambda_i \int_{B_r} v_i^2 d\lambda = \left( \sum_i \frac{\lambda_i}{n} \right) \int_{B_r(0)} \|v\|^2 d\lambda(v) \]

\[ = \frac{\text{tr} (T)}{n} \int_{B_r(0)} \|v\|^2 d\lambda(v) \]

\[ = \frac{\text{tr} (T)}{n} \int_{\rho \in [0,r]} \rho^2 S_{n-1}(\rho) d\rho \]

\[ = \frac{(n+1)}{n} \text{tr} (T), \]

(\text{polar coordinate transform})

\[ = \frac{\rho^{n+2}}{n(n+2)} \text{tr} (T), \]

as was to be shown.

We are now prepared to prove the theorem. We have

\[ I(X, Y|X \in B_r) = \int_{B_r} d_f(p(\cdot|x), p(\cdot|X \in B_r)) d\mathbb{P}_X \cdot d\mathbb{P}_Y \in B_r, \]

\[ = \int_{B_r} d_f(\alpha(x), a) d\mathbb{P}_X \cdot d\mathbb{P}_Y \in B_r, \]

\[ = \int_{B_r} d_f(\alpha(x), a) p(x|X \in B_r) d\lambda(x) \]

\[ = \frac{1}{\mathbb{P}(X \in B_r)} \int_{B_r} d_f(\alpha(x), a) p(x) d\lambda(x) \]

\[ = \frac{1}{p(x_0) V_n(r + o(r^{n+1}))} \int_{B_r} d_f(\alpha(x), a) p(x) d\lambda(x). \]  

(\text{Lemma 1})

Focusing now on the integral, we have

\[ \int_{B_r} d_f(\alpha(x), a) p(x) d\lambda(x) \]

\[ = \int_{B_r} [d_f(\alpha(x), \alpha(x_0)) + o(r^2)] p(x) d\lambda(x) \]

\[ = \frac{1}{2} \int_{B_r} [\mathcal{H}_{f_a}(\alpha(x) - \alpha(x_0), \alpha(x) - \alpha(x_0) + o(r^2)] \]

\[ \times p(x) d\lambda(x). \]

Since \( \alpha(x) - \alpha(x_0) = D_{\alpha(x_0)}(x - x_0) + o(r), \) we can rewrite the integrand, obtaining

\[ \int_{B_r} [\mathcal{H}_{f_a}(\alpha(x) - \alpha(x_0), \alpha(x) - \alpha(x_0) + o(r^2)] \]

\[ \times p(x) d\lambda(x). \]

\[ \int_{B_r} [\mathcal{H}_{f_a}(\Delta_2 \circ D_{\alpha(x_0)})(x - x_0) + o(r^2)] p(x) d\lambda(x). \]
Since $p(x)$ is smooth, $p(x) - p(x_0) = o(r)$, and we can further rewrite the integral as

$$
I(x, y|X \in B_r) = \frac{1}{p(x_0)V_n(r) + o(r^{n+1})} \times \left[ p(x_0) \frac{S_{n-1}r^{n+2}}{2n(n+2)} \text{tr}(g_0) + o(r^{n+2}) \right].
$$

We can cancel terms using the fact that $S_{n-1}/V_n = n$, and we obtain

$$
I(x, y|X \in B_r) = \frac{1}{2n+2} \text{tr}(g_0) + o(r^2).
$$

Dividing through by $r^2$ and taking the limit as $r \to 0$ proves the theorem.

### 6.3. Numerical Computation of Spatial Derivatives

Computation of the metric tensor $g_0$, requires the estimation of the derivative $D\alpha$ of the attribute map at $x$ as well as the computation the Hessian tensor $Hf$ at $x$. For most common choices of $f$, the Hessian may be computed analytically, and therefore, we focus on the computation of the derivative $D\alpha$. Since the direct computation of different quotients typically leads to numerical instability, we instead use a more robust method based on weighted linear regression. The fundamental idea is to regress the attribute differences $\alpha(x_i) - \alpha(x)$ on the geographic displacements $x_i - x$; the regression coefficients will then approximate the components of the derivative $D\alpha$. Let $E(x)$ denote the ego network of node $x$ in the geographic network of Fig. S2B. Our approximation formula is

$$
D\alpha_x \approx (X^T W x)^{-1} X^T W y_x,
$$

and $X, W$, and $Y_x$ are defined below.

$X$ is the matrix with the $i$th row that is the difference $x_i - x$ for each $x_i \in E(x)$.

$W$ is a diagonal weighting matrix that prioritizes tracts closer to the origin $x$. We used a Gaussian radial basis weight, yielding

$$
W_{ij} = \begin{cases} 
\exp \left[ -\frac{\|x_i - x\|^2}{2\sigma} \right] & i = j \\
0 & \text{otherwise,}
\end{cases}
$$

where $\sigma$ is a tunable characteristic length scale set to 10 km in our computations, corresponding to very weak weighting.

$Y_x$ is the matrix with the $i$th row that is the vector $p_{Y|x=x_i} - p_{Y|x=x}$. 

www.pnas.org/cgi/content/short/1708201114
Fig. S1. Illustrative eigenvectors in Detroit. Each eigenvector contains information about spatial demographic boundaries, with divisions deemed more important by the algorithm highlighted by eigenvectors corresponding to smaller $k$. In these visualizations, $\sigma = 30$.

Fig. S2. Spectral partitions of Detroit (A), Philadelphia (B), and Chicago (C) before hierarchical postprocessing.

Fig. S3. Construction of the approximate information manifold $M$. Starting with spatial data (A), we construct an adjacency network $G$ (B) which approximates the topology of a compact submanifold in $\mathbb{R}^2$. The map $\alpha$ embeds $G$ in $P(C)$. 
<table>
<thead>
<tr>
<th>Measure</th>
<th>Refs.</th>
<th>Functional form</th>
<th>Divergence generator $f(p)$</th>
<th>Smoothing kernel $\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Exposure Index</td>
<td>6</td>
<td>$1 + H(Y</td>
<td>X)$</td>
<td>$|p|^2$</td>
</tr>
<tr>
<td>Divergence Index</td>
<td>26</td>
<td>$I(X, Y)$</td>
<td>$\sum p_i \log p_i$</td>
<td>Quartic</td>
</tr>
<tr>
<td>Information Theory Index</td>
<td>6, 17</td>
<td>$I(X, Y)/H(Y)$</td>
<td>$\sum p_i \log p_i$</td>
<td>General</td>
</tr>
<tr>
<td>Ordinal Information Theory Index</td>
<td>28</td>
<td>$I(X, Y)/H(Y)$</td>
<td>$\sum c_i \log c_i$</td>
<td>General</td>
</tr>
<tr>
<td>Ordinal Variation Ratio</td>
<td>28</td>
<td>$I(X, Y)/H(Y)$</td>
<td>$-4 \sum c_i(1 - c_i)$</td>
<td>General</td>
</tr>
<tr>
<td>Ordinal Square Root Index</td>
<td>28</td>
<td>$I(X, Y)/H(Y)$</td>
<td>$-2 \sum \sqrt{c_i(1 - c_i)}$</td>
<td>General</td>
</tr>
<tr>
<td>Neighborhood Sorting Index</td>
<td>27</td>
<td>$\sqrt{I(N, Y)/I(X, Y)}$</td>
<td>$(\sum p_i y_i)^2$</td>
<td>Ego network</td>
</tr>
</tbody>
</table>

In the three ordinal measures of ref. 28, $c_i = \sum_{i=1}^{k} p_i$. The Neighborhood Sorting Index is designed for ordinal variables, like income, in which $y_i$ is an income level. In this measure, $I(N, Y)$ is the mutual information between smoothed neighborhoods and demographics.