Supplementary Information for
A social perspective on perceived distances reveals deep community structure

Kenneth S. Berenhaut, Katherine E. Moore, Ryan L. Melvin

Kenneth S. Berenhaut
Email: berenhks@wfu.edu

This PDF file includes:
Supplementary text
Figures S1 to S14
Tables S1 to S8
SI References
Table of Contents

1. Properties of cohesion
2. Instructive example
   a. **Fig S1**: Illustration of conflict foci
   b. **Table S1**: Matrix of cohesion values for data in Fig 2
3. **Fig S2**: Local depth for cholera fatalities
4. Two-dimensional Euclidean data
   a. **Fig. S3**: Further 2-dimensional Euclidean data sets
   b. **Fig. S4**: Results for clustering approaches for data in Fig 4F
   c. **Fig. S5**: Results for clustering approaches for data in Fig 4
5. Cognate language data set
   a. **Fig. S6**: Embedding and dendrogram comparisons for cognate language data
   b. **Table S2**: Cluster assignment comparisons for cognate language data
   c. **Fig. S7**: Local depth and cohesion values for cognate language data
   d. **Fig. S8**: Community structure for cognate data restricted to three large language groups
   e. **Table S3**: Normalized mutual information values comparing clustering results
   f. **Fig. S9**: Community structure using cosine similarity
   g. **Fig. S10**: Mutual $k$-nearest neighbors graphs for the language cognate data
6. Gene expression data set
   a. **Fig. S10**: Mutual $k$-nearest neighbors graphs for the gene expression data
   b. **Fig. S11**: Within-group distributions of distance and cohesion for the gene expression data
   c. **Table S4**: Results for clustering approaches for the gene expression data
   d. **Fig. S12**: Classification via cohesion for the gene expression data
7. Cultural distance data set
   a. **Table S5**: Results for clustering approaches for the cultural distance data
   b. **Table S6**: Distances associated with strong ties in India and weak ties in US
   c. **Table S7**: Normalized mutual information values comparing clustering results
8. Further applications
   a. **Fig. S13**: Local and global time series smoothing results for PaLD and LOWESS
   b. **Table S8**: Average number of strongly cohesive relationships
9. **Fig. S14**: Pseudo-code for the algorithmic implementation of PaLD
Properties of Cohesion

Throughout this section, the underlying set $S$ is partitioned as $S = A \cup B$ and $Y$ and $Z$ are selected as in the definition of cohesion in Eq. 3.

We show that cohesion satisfies three properties reasonable for approaches which convey strength of community connections. The properties are stated in a limiting sense; in the case of PaLD, a finite condition which guarantees each property is provided explicitly.

For the first two properties, the limiting nature considered is one in which $S = A \cup B$ and the sets $A$ and $B$ are moving away from one another in the sense that $d(a, b) \to \infty$ (resp. $d(b, a) \to \infty$) for all $a \in A$ and $b \in B$, while pairwise distances within $A$ (resp. $B$) are maintained.

(a) **Separation under increasing distance.**

Suppose $S = A \cup B$ and $A$ and $B$ are mutually separated in the sense that

$$\max\{d(a, a')| a, a' \in A\} < \min\{d(a, b), a \in A \text{ and } b \in B\},$$

and

$$\max\{d(b, b')| b, b' \in A\} < \min\{d(b, a), a \in A \text{ and } b \in B\}.$$  

Then the between-set cohesion values are zero, i.e.,

$$C_{a,b} = C_{b,a} = 0$$ for any $a \in A$ and $b \in B$.

**Proof.** Suppose that $a \in A$ and $b \in B$.  Partitioning according to the location of $Y$, and employing the definition in Eq. 3,

$$C_{a,b} = P(Z = b, d(Z, a) < d(Z, Y), Y \in A)$$

$$+P(Z = b, d(Z, a) < d(Z, Y), Y \in B).

(S1)

In the case that $Y \in A$, since $A$ is separated from $B$, $U_{a,Y} \subseteq A$, hence $Z \in A$ (whereas $b \in B$).  Therefore, the first term on the righthand side of Eq. S1 is equal to zero.  If $Y \in B$, since $B$ is separated from $A$, we have

$$d(b, Y) < d(b, a).$$

Therefore, the second term is also equal to zero.  We conclude from Eq. S1 that $C_{a,b} = 0$.  Similarly,

$$C_{b,a} = 0.$$  □
(b) Limiting irrelevance of density.
Suppose that \( A = \{a_1, a_2, \ldots, a_{n_1}\} \) and \( A' = \{a_1', a_2', \ldots, a_{n_1}'\} \) have the same ordinal structure in the sense that for any \( 1 \leq i, j, k \leq n_1, \)
\[
d(a_k, a_i) < d(a_k, a_j) \text{ if and only if } d(a_k', a_i') < d(a_k', a_j').
\]
Suppose additionally that \( S = A \cup B \) (resp. \( S' = A' \cup B \)) for some set \( B = \{b_1, b_2, \ldots, b_{n_2}\} \) with the property that \( A \) and \( B \) (resp. \( A' \) and \( B \)) are mutually separated as in the statement of (a) above. Then for any \( 1 \leq i, j \leq n_1 \), then \( c_{a_ia_j} = c_{a_i'a_j'} \), i.e., the corresponding (within-set) pair-wise cohesion values are equal.

\[
\begin{align*}
\text{Proof.} & \text{ Suppose that } 1 \leq i, j \leq n_1 \text{ are fixed and set } x = a_i \text{ and } w = a_j \text{ (resp. } x' = a_i' \text{ and } w' = a_j'). \\
& \text{As in Eq. S1,}
\end{align*}
\]
\[
c_{x,w} = P(Z = w, d(Z, x) < d(Z, Y), Y \in A) + P(Z = w, d(Z, x) < d(Z, Y), Y \in B). \tag{S2}
\]

In the case \( Y \in A \), by the separation of \( A \) (resp. \( A' \)), we have \( U_{x,Y} \subseteq A \) (resp. \( U_{x',Y} \subseteq A' \)), and hence
\[
P(Z = w, d(Z, x) < d(Z, Y) \mid Y \in A) = P(Z = w', d(Z, x') < d(Z, Y) \mid Y \in A'), \tag{S3}
\]
since both terms only depend on the common ordinal structure of \( A \) and \( A' \), respectively.

On the other hand, if \( Y \in B \), since \( A \) and \( B \) are mutually separated, \( U_{x,Y} = S \) and hence \( d(w, x) < d(w, Y) \). Therefore,
\[
P(Z = w, d(Z, x) < d(Z, Y) \mid Y \in B) = P(Z = w', d(Z, x') < d(Z, Y), Y \in B) = \frac{1}{n} \tag{S4}
\]
Since \( P(Y \in A) = P(Y \in A') \) the result now follows from Eqs. S2, S3 and S4. \( \square \)
For the third property, the limiting nature is one in which the set \( S = A \cup B \) becomes concentrated within the set \( B \), in the sense that \( |A|/|B| \) tends to zero.

(c) **Separation under increasing concentration.**
Suppose that \( S = A \cup B \) (so that \( A = B^c \)) and that \( B \) is concentrated with respect to \( A \) in the sense that
\[
\max\{d(b, b') \mid b, b' \in B\} < \min\{d(b, a) \mid a \in A, b \in B\},
\]
and for any \( a, a' \in A \), either (i) \( d(b, a) < d(a', a) \) for all \( b \in B \), or (ii) \( d(b, a) > d(a', a) \) for all \( b \in B \).

Then for any \( a \in A \) and \( b \in B \), the cohesion
\[
C_{a,b} < \frac{1}{n} |B^c|,
\]
and hence, if \( |B| \) is sufficiently large relative to \( |B^c| \), the relationship between \( a \) and \( b \) is not particularly strong (see Eq. 4).

**Proof.** Suppose that \( a \in A \) and \( b \in B \). In the case that \( Y \in B \), since \( B \) is concentrated with respect to \( A \), we have \( d(z, b) < d(z, a) \) for all \( z \in B \). In particular,
\[
P(Z = b, d(Z, a) < d(Z, Y), Y \in B) = 0. \tag{S5}
\]

In the case that \( Y \in A \), we observe that if \( b \in U_{a,Y} \), since \( B \) is concentrated with respect to \( A \), \( B \subseteq U_{a,Y} \) and thus \( |U_{a,Y}| \geq |B \cup \{a\}| = |B| + 1 \). It follows that
\[
P(Z = b, d(Z, a) < d(Z, Y) \mid Y \in A) \leq 1/(|B| + 1). \tag{S6}
\]

Since \( P(Y \in A) = (|A| - 1)/(n - 1) \), by Eq. S5 and S6,
\[
C_{a,b} = P(Z = b, d(Z, a) < d(Z, Y), Y \in B) + P(Z = b, d(Z, a) < d(Z, Y), Y \in A) \leq \frac{1}{|B| + 1} \left( \frac{|A| - 1}{n - 1} \right) < \frac{1}{n} \left( \frac{|A|}{|B|} \right).
\]

Since \( C_{x,x} > C_{x,w} \) for all \( w \neq x \) and the sum over all pairwise cohesions is \( n/2 \) we have
\[
\frac{n}{2} = \sum_{x,w} C_{x,w} \leq n \sum_{x} C_{x,x}.
\]

Hence, half of the average of the diagonal entries in the cohesion matrix (and thus the threshold for distinguishing particularly strong relationships) is at least \( 1/(4n) \) and the result follows. \( \square \)
As mentioned in the text, we also have the following properties which follow directly from the definitions in Eq. 2 and 3:

(d) The sum of the cohesion to a given $x$ is equal to the local depth of $x$, $\ell(x)$, i.e.,

$$\sum_w C_{x,w} = \ell(x).$$

(e) The cohesion to a given $x$ is maximized by $x$ itself, i.e., $C_{x,x} > C_{x,w}$ for all $w \neq x$.

(f) The total cohesion among all points is equal to $n/2$, i.e,

$$\sum_x \sum_w C_{x,w} = n/2.$$

(g) The threshold in Eq. 4 can be computed as half the average of the diagonal of the matrix, $[C_{x,w}]$, of cohesion values (assuming $d(x, x) < d(x, y)$ for all $y \neq x$).

(h) The average local depth is equal to $1/2$.

(i) Cohesion and local depths are invariant under monotone transformations of dissimilarities.
Fig. S1. For the 16-point Euclidean data presented in Fig. 1, we display several conflict foci, $U_{x,Y}$ for a particular point $x$. 
|    | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ |
|----|------|------|------|------|------|------|------|------|
| $s_1$ | 0.166 | 0.120 | 0.120 | 0.018 | 0.000 | 0.000 | 0.000 | 0.018 |
| $s_2$ | 0.131 | 0.177 | 0.018 | 0.120 | 0.077 | 0.020 | 0.000 | 0.000 |
| $s_3$ | 0.119 | 0.020 | 0.165 | 0.105 | 0.000 | 0.000 | 0.018 | 0.038 |
| $s_4$ | 0.018 | 0.104 | 0.104 | 0.179 | 0.000 | 0.062 | 0.095 | 0.062 |
| $s_5$ | 0.000 | 0.080 | 0.000 | 0.000 | 0.151 | 0.036 | 0.018 | 0.000 |
| $s_6$ | 0.000 | 0.018 | 0.000 | 0.060 | 0.036 | 0.187 | 0.122 | 0.080 |
| $s_7$ | 0.000 | 0.000 | 0.020 | 0.077 | 0.018 | 0.160 | 0.207 | 0.160 |
| $s_8$ | 0.018 | 0.000 | 0.036 | 0.056 | 0.000 | 0.080 | 0.122 | 0.187 |

**Table S1.** The matrix of cohesion values for the dataset depicted in Fig. 1 (rounded to three decimal places). The row sums yield the associated local depth values which are: 0.44, 0.54, 0.47, 0.62, 0.28, 0.50, 0.64, and 0.50. Values greater than the threshold (0.089) describing particularly strong relationships are indicated in bold.

Note that the local depth for the point $s_8$ is given by

$$\ell_{s_8}(s_7) = P(d(Z, s_7) < d(Z, Y)) = \frac{1}{7} \left( \frac{5}{8} + \frac{4}{8} + \frac{3}{4} + \frac{5}{7} + \frac{2}{3} + \frac{2}{3} \right) = 0.642.$$

Similarly, the contribution of $s_6$ to the local depth of $s_7$ is

$$C_{s_7, s_6} = \frac{1}{7} \left( \frac{1}{8} + \frac{1}{7} + \frac{1}{4} + \frac{1}{7} + 0 + \frac{1}{3} \right) = 0.160.$$
**Fig S2.** We consider the locations of cholera fatalities in London in August of 1854 (1, 2). The ability to use walking distance allows incorporation of the informative underlying geography in detection of hotspots (3) via local depth. Note that the largest depth values (in red) occur near the contaminated pump (indicated in black).

**Fig. S3.** We consider the cluster network, $G^*$, for three further two-dimensional Euclidean data sets. In A, we see that the presence of individual points between groups can, at times, serve as bridges between closely-knit groups of points (see Discussion and Conclusion for further consideration of noise in data). In B, we see that, in contrast to some common partitioning-based approaches to clustering, the cluster network may naturally consist of a single component. In C, sub-cluster structure can be seen within connected components of the (community) cluster network.
Fig. S4. For the data in Fig 4F, we display the results of a few other common clustering approaches. In A is the cluster assignment given by PaLD and HDBSCAN (for minPts = 3, 4, 5, 6) and for PAM (k = 8). In B and C, the results for $k$-means are displayed for parameters $k = 6$ and $k = 8$. The average silhouette widths for the partitions in A, B and C are 0.765, 0.759; and 0.67, respectively. Next, we consider the results for several hierarchical methods. The dendrogram for the Ward method is shown in D, the average silhouette width is maximized when $k = 8$, giving the result in A. In E-J, we provide the dendrograms and clusters (for $k = 8$) given by hierarchical clustering for several linkage methods.
Fig. S5. For comparison purposes, we display the resulting clusterings for various methods. Parameter values are included. In A and B, we consider 6-clusterings using hierarchical clustering and average and complete linkage (see (4, 5)); compare with Fig. 4. In C, D, E and G, we consider hierarchical clustering for further data sets employed in Fig. 4. In F, we display the results of HDBSCAN for data selected uniformly at random from the unit square with minPts = 5; note that points classified as noise are indicated with a “+”. Finally, in H and I, we employ DBSCAN for two values of epsilon.
Fig. S6. For the cognate language data set, the results of principle component analysis and non-metric multidimensional scaling (7) are given in A and B, respectively. For purposes of illustration and comparison, vertices are colored according to the clusters identified by PaLD. In C, the result of hierarchical clustering (using complete linkage) is displayed. Vertical lines indicate the cut-points which optimize the Calinski-Harabasz index (8) \((k = 8)\) and average silhouette width (9) \((k = 14)\), respectively. Again, for the purpose of comparison, names are colored according to the clusters identified by PaLD. See Table S3 for numerical comparisons via normalized mutual information.
| PaLD | HDBSCAN  | HDBSCAN  | Mutual k-NN | Mutual k-NN |
|------|----------|----------|-------------|-------------|
|      | minPts = 4 | minPts = 5 | k = 5       | k = 6       |
| Noise: | Noise: | Noise: | Noise: | Noise: |
| No languages labeled as noise. | Irish_A | Irish_A | Irish_A | Irish_A |
| | Irish_B | Irish_B | Irish_B | Irish_B |
| | English_ST | English_ST | Lithuanian_O | Lithuanian_O |
| | Takitaki | Lithuanian_ST | Latvian | Khalka |
| | Lithuanian_ST | Armenian_Mod | Armenian_Mod | Armenian_Mod |
| | Armenian_List | Afghan | Afghan | Welsh |
| | Afghan | Waziri | Waziri | Tadziki |
| | Waziri | TOCHARIAN_A | TOCHARIAN_A | TOCHARIAN_B |
| | TOCHARIAN_B | Romanian_List | Romanian_List | Romanian_List |
| | Romanian_List | Vlach | Vlach | Vlach |
| | Vlach | Italian | Italian | Italian |
| | Italian | Ladin | Ladin | Ladin |
| | Ladin | Provencal | Provencal | Provencal |
| | Provencal | French | French | French |
| | French | Walloon | Walloon | Walloon |
| | Walloon | French_Creole_C | French_Creole_C | French_Creole_C |
| | French_Creole_C | French_Creole_D | French_Creole_D | French_Creole_D |
| | French_Creole_D | Sardinian_N | Sardinian_N | Sardinian_N |
| | Sardinian_N | Sardinian_L | Sardinian_L | Sardinian_L |
| | Sardinian_L | Sardinian_C | Sardinian_C | Sardinian_C |
| | Sardinian_C | Spanish | Spanish | Spanish |
| | Spanish | Portuguese_ST | Portuguese_ST | Portuguese_ST |
| | Portuguese_ST | Brazilian | Brazilian | Brazilian |
| | Brazilian | Catalan | Catalan | Catalan |
| | Catalan | Romanian_List | Romanian_List | Romanian_List |
| | Romanian_List | Vlach | Vlach | Vlach |
| | Vlach | Italian | Italian | Italian |
| | Italian | Ladin | Ladin | Ladin |
| | Ladin | Provencal | Provencal | Provencal |
| | Provencal | French | French | French |
| | French | Walloon | Walloon | Walloon |
| | Walloon | French_Creole_C | French_Creole_C | French_Creole_C |
| | French_Creole_C | French_Creole_D | French_Creole_D | French_Creole_D |
| | French_Creole_D | Sardinian_N | Sardinian_N | Sardinian_N |
| | Sardinian_N | Sardinian_L | Sardinian_L | Sardinian_L |
| | Sardinian_L | Sardinian_C | Sardinian_C | Sardinian_C |
| | Sardinian_C | Spanish | Spanish | Spanish |
| | Spanish | Portuguese_ST | Portuguese_ST | Portuguese_ST |
| | Portuguese_ST | Brazilian | Brazilian | Brazilian |
| | Brazilian | Catalan | Catalan | Catalan |
| | Catalan | Swedish_Up | Swedish_Up | Swedish_Up |
| | Swedish_Up | Swedish_VL | Swedish_VL | Swedish_VL |
| | Swedish_VL | Swedish_List | Swedish_List | Swedish_List |
| | Swedish_List | Danish | Danish | Danish |
| | Danish | Riksmal | Riksmal | Riksmal |
| | Riksmal | Icelandic_ST | Icelandic_ST | Icelandic_ST |
| | Icelandic_ST | Faroese | Faroese | Faroese |
| | Faroese | German_ST | German_ST | German_ST |
| | German_ST | Penn_Dutch | Penn_Dutch | Penn_Dutch |
| | Penn_Dutch | Dutch_List | Dutch_List | Dutch_List |
| | Dutch_List | Afrikaans | Afrikaans | Afrikaans |
| | Afrikaans | Flemish | Flemish | Flemish |
| | Flemish | Frisian | Frisian | Frisian |
| | Frisian | Swedish_Up | Swedish_Up | Swedish_Up |
| | Swedish_Up | Swedish_VL | Swedish_VL | Swedish_VL |
| | Swedish_VL | Swedish_List | Swedish_List | Swedish_List |
| | Swedish_List | Danish | Danish | Danish |
| | Danish | Riksmal | Riksmal | Riksmal |
| | Riksmal | Icelandic_ST | Icelandic_ST | Icelandic_ST |
| | Icelandic_ST | Faroese | Faroese | Faroese |
| | Faroese | German_ST | German_ST | German_ST |
| | German_ST | Penn_Dutch | Penn_Dutch | Penn_Dutch |
| | Penn_Dutch | Dutch_List | Dutch_List | Dutch_List |
| | Dutch_List | Afrikaans | Afrikaans | Afrikaans |
| | Afrikaans | Flemish | Flemish | Flemish |
| | Flemish | Frisian | Frisian | Frisian |
| | Frisian | English_ST | English_ST | English_ST |
| | English_ST | Takitaki | Takitaki | Takitaki |
| English_ST | Greek_ML | Greek_MD | Greek_Mod | Greek_D | Greek_K |
|-----------|---------|---------|----------|--------|--------|
| English_ST | Greek_ML | Greek_MD | Greek_Mod | Greek_D | Greek_K |
| Greek_K     | Slovenian_L | Lusatian_U | Czech | Slovak | Czech_E |
|             | Ukrainian | Byelorussian | Polish | Russian | Macedonian |
|             | Bulgarian | Serbocroatian |   |        |        |
| Lithuanian_O | Lithuanian_ST | Lithuanian_ST | Latvian |        |        |
| Latvian      | Gypsy_Gk | Singhalese | Kashmir | Marathi | Gujarati |
|              | Panjabi_ST | Lahnda | Hindi | Bengali | Nepali_List |
|              | Khaskura | Ossetic | Persian_List | Baluchi | Wakhi |
|              | HITTITE | TOCHARIAN_A | TOCHARIAN_B |        |        |
| Armenian_Mod | Armenian_List | Nepali_List | Khaskura |        |        |
| Waziri       | Afghan |        |        |        |        |
|     | Irish_A | Irish_B | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST |
|-----|---------|---------|---------|---------|-------------|-----------|-----------|---------|---------|-------------|-----------|-----------|---------|---------|-------------|-----------|-----------|---------|---------|-------------|-----------|-----------|
|     | Irish_A | Irish_B | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST | Welsh_N | Welsh_C | Breton_List | Breton_SE | Breton_ST |
|     |         |         |         |         |             |           |           |         |         |             |           |           |         |         |             |           |           |         |         |             |           |           |

Table S2. For the purposes of comparison, we provide cluster assignments for the 2665-dimensional cognate language data set. We display those obtained from PaLD together with results from HDBSCAN (for minPts = 4, 5) and the connected components of the mutual k-nearest neighbor graph (k = 5, 6). Note the variations on classical language families. Several languages are classified as “noise” by HDBSCAN, and Scandinavian and Germanic language families have been merged when minPts=5. PaLD provides potentially valuable inter- and intra-cluster structure, see Fig 4. See Table S3 for numerical comparisons via normalized mutual information.
Fig. S7. In A, we display a histogram of local depths (along with some specific values) for the Indo-European languages depicted in Fig. 5. Note that several ancient languages have large depth values. In B, a histogram of cohesion values for the Indo-European cognate data is displayed along with several values of $C_{x,w}$ (the contribution of $w$ to the local depth of $x$). The vertical dashed line indicates the threshold value of 0.0164 (see Eq. 4).
Fig. S8. We further consider the Indo-European cognate data. Here we restrict (individually) to three of the larger language cluster groups, revealing further internal cluster structure.
Table S3. We provide normalized mutual information values (10) comparing partitions provided by a variety of clustering methods for the cognate language data set. Note that, when necessary (for methods other than PaLD), to obtain concrete partitions we select parameters according to the optimization of Average Silhouette Width (9) and Calinski-Harabasz score (8). HDBSCAN and DBSCAN may identify some languages as noise (the number of such points is indicated in parentheses; note that a default HDBSCAN run identified 13 out of 87 languages as noise) (4). In the case of PaLD, we obtained two clusters of size one (Albanian_G) and the centrally positioned (Greek_K), and 5 clusters of size two. For discussion of noise, see Discussion and Conclusions.
Fig. S9. We illustrate resulting community structure in the cognate language data set employing (negative) cosine similarity in place of Euclidean distance.
For comparison, we present the mutual $k$-nearest neighbors graphs (6) for the language cognate and gene expression data considered in Fig. 5-6. For purposes of illustration, vertices in A and B are colored as derived from PaLD, and in C and D according to tissue type; compare with Fig. 5-6. In A and B, we set $k = 18$ and $k = 9$, respectively. In C and D, we set $k = 30$ and $k = 11$, respectively.
**Fig. S11.** Histograms of within-group distances (resp. cohesion) for tissue data in Fig. 6. Note the similar distributions for cohesion values which are greater than the threshold (i.e., particularly strong ties). Weak ties (for cohesion) are indicated in gray.
### K-Means ($k = 7$)

| Tissue       | 0 | 24 | 2 | 0 | 0 | 0 | 0 | 0 |
|--------------|---|----|---|---|---|---|---|---|
| Liver        | 0 | 24 | 2 | 0 | 0 | 0 | 0 | 0 |
| Kidney       | 0 | 0  | 2 | 37| 0 | 0 | 0 | 0 |
| Endometrium  | 0 | 0  | 0 | 15| 0 | 0 | 0 | 0 |
| Colon        | 0 | 0  | 0 | 0 | 34| 0 | 0 | 0 |
| Placenta     | 6 | 0  | 0 | 0 | 0 | 0 | 0 | 0 |
| Cerebellum   | 0 | 0  | 2 | 0 | 0 | 0 | 5 | 31|
| Hippocampus  | 0 | 0  | 0 | 0 | 0 | 0 | 31| 0 |

### HDBSCAN ($minpts = 5$)

| Tissue       | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 |
|--------------|---|----|---|---|---|---|---|---|
| Liver        | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 |
| Kidney       | 0 | 0  | 0 | 0 | 0 | 0 | 15| 6 |
| Endometrium  | 0 | 0  | 0 | 0 | 0 | 15| 0 | 0 |
| Colon        | 0 | 0  | 0 | 0 | 34| 0 | 0 | 0 |
| Placenta     | 6 | 0  | 0 | 0 | 0 | 0 | 0 | 0 |
| Cerebellum   | 0 | 0  | 5 | 0 | 31| 0 | 0 | 0 |
| Hippocampus  | 0 | 0  | 31| 0 | 0 | 0 | 0 | 0 |

### HCLUST ($k = 7$)

| Tissue       | 0 | 0  | 0 | 0 | 0 | 24| 2 | 0 |
|--------------|---|----|---|---|---|----|---|---|
| Liver        | 0 | 0  | 0 | 0 | 0 | 24| 2 | 0 |
| Kidney       | 37| 0  | 0 | 0 | 0 | 0 | 2 | 0 |
| Endometrium  | 15| 0  | 0 | 0 | 0 | 0 | 0 | 0 |
| Colon        | 0 | 0  | 0 | 34| 0 | 0 | 0 | 0 |
| Placenta     | 0 | 0  | 0 | 0 | 0 | 0 | 0 | 6 |
| Cerebellum   | 0 | 0  | 0 | 36| 0 | 0 | 2 | 0 |
| Hippocampus  | 0 | 12 | 19| 0 | 0 | 0 | 0 | 0 |

### HCLUST ($h = 120$)

| Tissue       | 0 | 0  | 0 | 0 | 0 | 24| 0 | 2 |
|--------------|---|----|---|---|---|----|---|---|
| Liver        | 0 | 0  | 0 | 0 | 0 | 24| 2 | 0 |
| Kidney       | 9 | 18 | 0 | 0 | 10| 0 | 2 | 0 |
| Endometrium  | 0 | 0  | 0 | 0 | 0 | 0 | 0 | 15|
| Colon        | 0 | 0  | 0 | 34| 0 | 0 | 0 | 0 |
| Placenta     | 0 | 0  | 0 | 0 | 0 | 0 | 0 | 2 |
| Cerebellum   | 0 | 0  | 0 | 31| 0 | 0 | 2 | 0 |
| Hippocampus  | 0 | 12 | 19| 0 | 0 | 0 | 0 | 0 |

### PaLD

| Tissue       | 0 | 0  | 0 | 0 | 2 | 17| 0 | 0 |
|--------------|---|----|---|---|---|----|---|---|
| Liver        | 0 | 0  | 0 | 0 | 2 | 17| 0 | 7 |
| Kidney       | 36| 0  | 0 | 0 | 0 | 0 | 0 | 0 |
| Endometrium  | 15| 0  | 0 | 0 | 0 | 0 | 0 | 0 |
| Colon        | 0 | 0  | 0 | 0 | 33| 1 | 0 | 0 |
| Placenta     | 0 | 2  | 0 | 0 | 0 | 0 | 0 | 0 |
| Cerebellum   | 0 | 2  | 0 | 0 | 1 | 4 | 2 | 1 |
| Hippocampus  | 0 | 0  | 31| 0 | 0 | 0 | 0 | 0 |
Table S4. In the above tables, we give the cluster assignments (in columns) against the type (in rows) for PaLD together with the results of k-Means, HDBSCAN and hierarchical clustering for a few choices of parameters. It is important to keep in mind that, in the case of PaLD, the set of partition labels is a small part of the information provided through the cohesion matrix and, as always, the method does not involve a search over an underlying parameter space.

| PaLD         | HDBSCAN  | Mutual k-NN | DBSCAN  | k-Mediods (PAM) |
|--------------|----------|-------------|---------|----------------|
|              | minPts = 5 | k = 6       | eps = .05 | k = 4          |
|              |          |             |         |                |
| Bulgaria     |          |             |         |                |
| Hungary      |          |             |         |                |

| Bulgaria     | Sweden   | Cyprus      | Finland  | Italy         |
| Sweden       | Cyprus   | Finland     |          |              |
|              | Italy    | Poland      | Slovenia |              |
|              | Poland   | Slovenia    | Spain    |              |
|              | Sweden   | Cyprus      | Finland  |              |
|              | France   | Germany     | Netherlands |            |
| Great Britain| France   | Spain       |          |              |
|              |         | US: West North Central |         |
|              |         | US: East South Central |         |
|              |         | US: South Atlantic |         |
|              |         | US: Middle Atlantic States |         |
|              |         | US: East North Central |         |
|              |         | US: New England |         |
|              |         | US: West South Central |         |
|              |         | US: Rocky Mountain state |         |
|              |         | US: California |         |
|              |         | Hungary      |         |
|              |         | Estonia      |         |
|              |         | US: West North Central |         |
|              |         | US: East South Central |         |
|              |         | US: South Atlantic |         |

| US: West North Central | US: East South Central | US: South Atlantic | Finland Great Britain |
|------------------------|------------------------|--------------------|-----------------------|
| US: West North Central | US: East South Central | US: South Atlantic | US: West North Central |
| US: East South Central | US: South Atlantic | Finland Great Britain | US: West South Central |
| US: South Atlantic | US: West North Central | US: East South Central | US: Rocky Mountain state |
| US: Middle Atlantic States | US: East North Central | US: New England | US: California |
| US: East North Central | US: West South Central | US: Rocky Mountain state | US: Pacific |
| US: New England | US: West South Central | US: California | US: West North Central |
| US: West South Central | US: Rocky Mountain state | US: Pacific | US: East South Central |
| US: Rocky Mountain state | US: California | US: West North Central | US: East South Central |
| US: California | US: Pacific | US: West North Central | US: East South Central |
| US: Middle Atlantic States | US: East North Central | US: New England | US: West South Central | US: Rocky Mountain state | US: California | US: Pacific |
|----------------------------|-----------------------|----------------|-----------------------|-------------------------|---------------|------------|
| US: East North Central     | US: New England       | US: West South Central | US: Rocky Mountain state | US: California | US: Pacific |
| US: California             |                       |                |                       |                         |               |            |
| CN: Liaoning Province      | CN: Heilongjiang Province | CN: Shanxi Province | CN: Shaannxi Province | CN: Hebei Province    | CN: Beijing   |
| CN: Beijing                | CN: Jiangsu Province  | CN: Shandong Province | CN: Zhejiang Province | CN: Guangdong Province | CN: Hubei Province |
| CN: Henan Province         | CN: Shanghai          | CN: Fujian Province | CN: Shandong Province | CN: Guangxi Province  | CN: Hubei Province |
| CN: Hunan Province         | CN: Guangdong Province| CN: Hunan Province | CN: Hubei Province    | CN: Guangxi Province  | CN: Hubei Province |
| CN: Henan Province         | CN: Anhui Province    | CN: Jiangxi Province | CN: Guizhou Province  | CN: Guangxi Province  | CN: Gansu     |
| CN: Anhui Province         | CN: Guizhou Province  | CN: Guangxi Province | CN: Gansu             | CN: Hebei Province    | CN: Hunan Province |
| CN: Guangxi Province       |                       |                |                       |                         |               |            |
| CN: Gansu                  | CN: Hebei Province    |                |                       |                         |               |            |
| CN: Hebei Province         | CN: Hunan Province    |                |                       |                         |               |            |
| CN: Hunan Province         | CN: Shanghai          |                |                       |                         |               |            |
| CN: Guangdong Province     | CN: Guizhou Province  |                |                       |                         |               |            |
| CN: Guangxi Province       |                       |                |                       |                         |               |            |
| CN: Gansu                  | CN: Hebei Province    |                |                       |                         |               |            |
| CN: Gansu                  | CN: Hunan Province    |                |                       |                         |               |            |
| CN: Shanghai               | CN: Guangdong Province|                |                       |                         |               |            |
| CN: Guangxi Province       | CN: Guizhou Province  |                |                       |                         |               |            |
| CN: Guangxi Province       | CN: Gansu             |                |                       |                         |               |            |
| IN: Kerala                 |                       |                |                       |                         |               |            |
| CN: Gansu                  |                       |                |                       |                         |               |            |
Table S5. For the purposes of comparison, we provide cluster assignments for the cultural distance data obtained from the World Values Survey. We display cluster assignments obtained from PaLD together with results from HDBSCAN (for MinPts = 5), the connected components of the mutual k-nearest neighbor graph (k = 4), DBSCAN (for eps = 0.05), and PAM (for k = 4). Note that several regions in India are classified as noise by HDBSCAN. Lastly, PAM places Finland and Great Britain in the U.S. cluster, and IN: Kerala in the identified China cluster. PaLD also provides potentially valuable inter- and intra-cluster structure, see Fig 6. See Table S6 for numerical comparisons via normalized mutual information.
To illustrate the accounting for varying density, we provide a comparison of the strong ties (i.e., those above the threshold of 0.0217) within India and the weak ties within the United States:

| Strong Ties within India          | Distance | Weak Ties within United States       | Distance |
|----------------------------------|----------|-------------------------------------|----------|
| Uttar Pradesh                    | Bihar    | 0.0426                              | West South Central | Mid Atlantic | 0.0129 |
| Maharashtra                      | Andhra Pradesh | 0.0596                          | New England | East North Central | 0.0143 |
| Tamil Nadu                       | Karnataka | 0.0639                              | California | Rocky Mountain State | 0.0160 |
| Maharashtra                      | Bihar    | 0.0650                              | East South Central | Mid Atlantic | 0.0163 |
| Uttar Pradesh                    | Madhya Pradesh | 0.0653                          | California | West North Central | 0.0190 |
| West Bengal                      | Uttar Pradesh | 0.0699                          | California | South Atlantic | 0.0192 |
| Tamil Nadu                       | Maharashtra | 0.0704                          | West South Central | New England | 0.0210 |
| West Bengal                      | Bihar    | 0.0720                              | California | East North Central | 0.0216 |
| Maharashtra                      | Karnataka | 0.0736                              | East South Central | Pacific | 0.0238 |
| Uttar Pradesh                    | Andhra Pradesh | 0.0740                          | New England | East South Central | 0.0252 |
| Bihar                            | Andhra Pradesh | 0.0769                          | California | West South Central | 0.0258 |
| Uttar Pradesh                    | Tamil Nadu | 0.0782                              | California | East South Central | 0.0273 |
| Tamil Nadu                       | Karnataka | 0.0793                              |             |                |        |
| West Bengal                      | Bihar    | 0.0809                              |             |                |        |
| Andhra Pradesh                   | Karnataka | 0.0845                              |             |                |        |
| Bihar                            | Karnataka | 0.0855                              |             |                |        |
| West Bengal                      | Maharashtra | 0.0858                          |             |                |        |
| Tamil Nadu                       | Andhra Pradesh | 0.0904                          |             |                |        |
| Rajasthan                        | Maharashtra | 0.0934                          |             |                |        |
| Andhra Pradesh                   | Orrisa    | 0.0939                              |             |                |        |
| Bihar                            | Punjab    | 0.1002                              |             |                |        |
| West Bengal                      | Orrisa    | 0.1009                              |             |                |        |
| Madhya Pradesh                   | Bihar    | 0.1019                              |             |                |        |
| Orrisa                           | Kerala    | 0.1038                              |             |                |        |
| Uttar Pradesh                    | Punjab    | 0.1042                              |             |                |        |
| Andhra Pradesh                   | Assam     | 0.1043                              |             |                |        |
| Kerala                           | Assam     | 0.1045                              |             |                |        |
| West Bengal                      | Kerala    | 0.1057                              |             |                |        |
| Kerala                           | Punjab    | 0.1081                              |             |                |        |
| Karnataka                        | Assam     | 0.1096                              |             |                |        |
| Orrisa                           | Punjab    | 0.1103                              |             |                |        |
| Rajasthan                        | Madhya Pradesh | 0.1133                          |             |                |        |

*Table S6.* To illustrate the accounting for varying density, we provide a comparison of the strong ties (i.e., those above the threshold of 0.0217) within India and the weak ties within the United States.
States. Notice that even the most culturally distant regions (California and East South Central) within the United States are more similar than all strongly cohesive pairs in India (with minimum distance 0.0426).

| Parameter | Population Labels | PaLD | HDBSCAN | HDBSCAN | k-NN | k-NN | PAM | DBSCAN | DBSCAN | Hierarch. Complete | Hierarch. Single | Hierarch. Single |
|-----------|-------------------|------|---------|---------|------|------|-----|--------|--------|-------------------|----------------|-----------------|
|           | N/A               | minPts = 4 | minPts = 5 | k = 5 | k = 4 | eps = 0.68 | eps = 0.70 | k = 4 | k = 4 | k = 4 | k = 7 |
| Population Labels | 1.00 | 0.93 | 0.79 | 0.79 | 0.81 | 0.84 | 0.88 | 0.8 | 0.71 | 1.00 | 0.47 | 0.80 |
| PaLD      | 0.93 | 1.00 | 0.75 | 0.75 | 0.67 | 0.91 | 0.82 | 0.76 | 0.68 | 0.93 | 0.48 | 0.79 |
| HDBSCAN (4) | 0.79 | 0.75 | 1.00 | 0.66 | 0.68 | 0.76 | 0.67 | 0.74 | 0.79 | 0.54 | 0.88 |
| HDBSCAN (5) | 0.79 | 0.75 | 0.96 | 1.00 | 0.65 | 0.67 | 0.75 | 0.69 | 0.76 | 0.79 | 0.52 | 0.86 |
| k-NN (5) | 0.81 | 0.87 | 0.66 | 0.65 | 1.00 | 0.96 | 0.71 | 0.71 | 0.64 | 0.81 | 0.40 | 0.72 |
| k-NN (6) | 0.84 | 0.91 | 0.68 | 0.67 | 0.96 | 1.00 | 0.74 | 0.71 | 0.62 | 0.84 | 0.42 | 0.72 |
| PAM (4) | 0.88 | 0.82 | 0.76 | 0.75 | 0.71 | 0.74 | 1.00 | 0.71 | 0.67 | 0.88 | 0.53 | 0.80 |
| DBSCAN (0.35) | 0.80 | 0.76 | 0.67 | 0.69 | 0.71 | 0.71 | 0.71 | 1.00 | 0.84 | 0.8 | 0.33 | 0.66 |
| DBSCAN (0.40) | 0.71 | 0.68 | 0.74 | 0.76 | 0.64 | 0.62 | 0.67 | 0.84 | 1.00 | 0.71 | 0.39 | 0.73 |
| Hier. Complete (4) | 1.00 | 0.93 | 0.79 | 0.79 | 0.81 | 0.84 | 0.88 | 0.80 | 0.71 | 1.00 | 0.47 | 0.80 |
| Hier. Single (4) | 0.47 | 0.48 | 0.54 | 0.52 | 0.4 | 0.42 | 0.53 | 0.33 | 0.39 | 0.47 | 1.00 | 0.66 |
| Hier. Single (7) | 0.80 | 0.79 | 0.88 | 0.86 | 0.72 | 0.72 | 0.80 | 0.66 | 0.73 | 0.80 | 0.66 | 1.00 |

Table S7. We provide normalized mutual information values (10) comparing partitions provided by a variety of clustering methods for the cultural distance data.
Fig. S12. For the tissue data set included in Fig. 6, we classify using strong contributions and compare results to that obtained by $k$-NN \((11, 12)\). We consider mean classification errors over randomly selected labeled subsets of varying sizes (50 trials). To have comparable results, in the case that $x$ has no strongly cohesive points, we assign the class label of the most cohesive point. Note that in comparison with $k$-NN, PaLD is competitive without need for selection of the parameter $k$. 
Fig. S13. Local and global time series smoothing results for PaLD and LOWESS are given for a series of Southern Oscillation Index values (see (13)). In this case, the curves given by PaLD and LOWESS (14) are similar, without the necessity to select a bandwidth (e.g., 5 and 66 percent), degree of local polynomials (e.g., linear), nor weight function (e.g., tri-cube).

|     | n   | number of strong ties | average degree |
|-----|-----|-----------------------|----------------|
| **Fig. 4A** | 150 | 591                   | 3.9            |
| **Fig. 4B** | 788 | 8189                  | 10.4           |
| **Fig. 4C** | 500 | 4122                  | 8.2            |
| **Fig. 4D** | 49  | 126                   | 2.6            |
| **Fig. 4E** | 230 | 4122                  | 17.9           |
| **Fig. 4F** | 240 | 1392                  | 5.8            |
| **Fig. 5**  | 87  | 219                   | 2.5            |
| **Fig. 6**  | 189 | 1064                  | 5.6            |
| **Fig. 7**  | 59  | 154                   | 2.6            |

Table S8. We provide number of points (n) for each data set considered in the Applications section together with the total number of particularly strong relationships (i.e., the number of edges in $G^*_s$) and average degree of $G^*_s$. 
Algorithm 1 Computing the matrix of partitioned local depths given a matrix $D = \{d(i, j)\}_{i,j=1}^n$ of dissimilarities

1: function PALD($D$)
2:     $A \leftarrow \{0\}_{i,j=1}^n$
3:     for $i = 1$ to $n$ do
4:         for $j = 1$ to $n$ satisfying $j \neq i$ do
5:             $U_{i,j} \leftarrow \{k \mid d(k, i) \leq d(j, i) \text{ or } d(k, j) \leq d(i, j)\}$
6:                 for $l$ in $U_{i,j}$ do
7:                     if $d(l, i) < d(l, j)$ then
8:                         $A_{i,l} \leftarrow A_{i,l} + 1/\text{size}(U_{i,j})$
9:                     if $d(l, i) = d(l, j)$ then
10:                        $A_{i,l} \leftarrow A_{i,l} + (1/2)(1/\text{size}(U_{i,j}))$
11:                $C \leftarrow A/(n-1)$
12:         return $C$

Fig. S14. Pseudo-code for the algorithmic implementation of partitioned local depths (PaLD). The output is the matrix of cohesion values describing pair-wise relationship cohesion. Local depths can be obtained from the row sums of the output matrix, $C$. 
SI References

1. J. Snow, *On the Mode of Communication of Cholera* (John Churchill, 1855).
2. P. Li, Cholera, Version 0.7.0 (R Package, 2019); https://CRAN.R-project.org/package=cholera
3. A. B. Lawson, Hotspot detection and clustering: ways and means. *Environ. Ecol. Stat.*** 17, 231-245 (2010).
4. R. J. G. B. Campello, P. Kröger, J. Sander, A. Zimek, Density-based clustering, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*** 10, 1343 (2020).
5. B. S. Everitt, S. Landau, M. Leese, D. Stahl, *Cluster analysis*, Fifth ed., (John Wiley & Sons, Ltd, 2011).
6. M. R. Brito, E. L. Chavez, A. J. Quiroz, J. E. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detections. *Stat. Probab. Lett.*** 35, 33-42 (1997).
7. I. Borg, P. J. F. Groenen, *Modern Multidimensional Scaling*, (Springer, 2005).
8. T. Caliński, J. Harabasz. A dendrite method for cluster analysis. *Commun. Stat. Theory Methods*** 3, 1-27 (1974).
9. P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*** 20, 53-65 (1987).
10. A. L. N. Fred, A. K. Jain. Robust data clustering. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003 Proceedings*** 2, 128—133 (2003).
11. K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is “nearest neighbor” meaningful? *Int. Conf. Database Theory* 217-235 (1999).
12. T. Cover, P. Hart, Nearest neighbor pattern classification. *IEEE T Inform Theory*** 13, 21-27 (1967).
13. D. Stoffer, Applied Statistical Time Series Analysis, Version 1.9 (R Package 2019); https://CRAN.R-project.org/package=astsa
14. W. S. Cleveland, Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*** 74, 829-836 (1979).