A social perspective on perceived distances reveals deep community structure

Kenneth S. Berenhaut[a,1], Katherine E. Moore[a,2], and Ryan L. Melvin[a,b,3]

[a]Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, NC 27109; and [b]Department of Physics, Wake Forest University, Winston-Salem, NC 27109

Edited by David Donoho, Department of Statistics, Stanford University, Stanford, CA; received May 21, 2020; accepted November 30, 2021

Community structure, including relationships between and within groups, is foundational to our understanding of the world around us. For dissimilarity-based data, leveraging social concepts of conflict and alignment, we provide an approach for capturing meaningful structural information resulting from induced local comparisons. In particular, a measure of local (community) depth is introduced that leads directly to a probabilistic partitioning conveying locally interpreted closeness (or *cohesion*). A universal choice of threshold for distinguishing strongly and weakly cohesive pairs permits consideration of both local and global structure. Cases in which one might benefit from use of the approach include data with varying density such as that arising as snapshots of complex processes in which differing mechanisms drive evolution locally. The inherent recalibrating in response to density allows one to sidestep the need for localizing parameters, common to many existing methods. Mathematical results together with applications in linguistics, cultural psychology, and genetics, as well as to benchmark clustering data have been included. Together, these demonstrate how meaningful community structure can be identified without additional inputs (e.g., number of clusters or neighborhood size), optimization criteria, iterative procedures, or distributional assumptions.

community structure | networks | social perspective | local depth | cohesion

APPLIED MATHEMATICS

Deriving structure from data is fundamental to our understanding of the world around us. Although clustering into groups (with implied relationships between and within groups) is a crucial component of human perception, there are relatively few methods of proximity-based data analysis that harness the richness of a social perspective. Here, a nature-inspired concept of conflict provides a foundation from which to consider mathematical construction of *local* (*conflict*) *foci* and resulting measures of local depth, cohesion, and (particularly) strong ties. As in other socially and biologically inspired work (1–3), insights are derived through a representation of the data within a latent space with rich internal structure. Formalized social concepts anchor the approach and facilitate the interpretation and communication of results.

The method, referred to here as partitioned local depth (PaLD), provides a framework for a holistic consideration of the structure of data. In particular, we begin by introducing local (community) depth, a concept which builds on existing approaches to (global) depth in which geometric constructions allow one to express features of centrality as interpretable probabilities which are free of parameters and robust to outliers (see, for instance, refs. 4 and 5). Here, a direct approach to localization which arises from a concept of opposition (i.e., "conflict") provides for local interpretation of depth.

Partitioning probabilities defining local depth, we obtain a quantity referred to as *cohesion*, which can be understood as a measure of locally perceived closeness that captures features of relative positioning and, as a result, accounts for density variation. The motivating social framework gives rise to a natural threshold for distinguishing strongly and weakly cohesive pairs and provides an alternative perspective for the concept of near neighbors. Topological features of the data can be considered via the resulting networks; we refer to the connected components of the graph of strong ties as community clusters. Mathematical results together with applications to real-world and benchmark clustering data have been included. Together, these demonstrate how meaningful community structure can be identified without additional inputs (e.g., number of clusters or neighborhood size), optimization criteria, iterative procedures, or distributional assumptions, in the presence of varying density.

It is crucial to note the importance of accounting for varying local density, particularly in applications involving complex evolutionary processes (see, for instance, refs. 6–9). In this context, here, relative positioning is observed entirely through distance comparisons within triples of points, and thus, the methods introduced may also be valuable in nonmetric and high-dimensional settings. The inherent recalibrating in response to density allows one to sidestep the need for localizing parameters (such as neighborhood size) common to many existing methods. The overarching perspective provides a framework for a holistic consideration of the structure of data, which integrates concepts of local depth, cohesion, strong ties, and network structure.

Significance

Community structure arising through relationships and interactions is essential to our understanding of the world around us. Leveraging social concepts of conflict and support, we introduce a method to transform input dissimilarity comparisons into output pairwise relationship strengths (or *cohesion*) and resulting weighted networks. The introduced perspective may be particularly valuable for data with varying local density such as that arising from complex evolutionary processes. Mathematical results, together with applications in linguistics, genetics, and cultural psychology as well as to benchmark data, have been included. Together, these demonstrate how meaningful community structure can be identified without additional inputs (e.g., number of clusters or neighborhood size), optimization criteria, iterative procedures, or distributional assumptions.

Author contributions: K.S.B. and R.L.M. collaborated on the early development of local depth; K.S.B. and K.E.M. developed PaLD; and K.S.B. and K.E.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: berenhks@wfu.edu.

[2]Present address: Department of Mathematics and Statistics, Amherst College, Amherst, MA 01002.

[3]Present address: Department of Anesthesiology and Perioperative Medicine, University of Alabama at Birmingham, Birmingham, AL 35233.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2003634119/-/DCSupplemental.

Published January 21, 2022.

The paper proceeds as follows. We first discuss the social framework from which the perspective of PaLD arises and then move on to the mathematical constructions of (community) local depth and the resulting concept of cohesion (partitioned depth). Next, we consider a theoretical threshold for distinguishing strong and weak relationships and the resulting networks. Some theoretical results regarding cohesion are also provided. We then turn toward applications to benchmark, linguistic, genetic, and cultural values data. We conclude with a consideration of cohesion in high-dimensional settings and a discussion of performance, benefits, and implementation.

Social Framework

In this section, we provide a brief discussion of the important underlying social framework and the associated social latent space. While this provides important context for the concepts that follow, the reader may choose to move directly to the mathematical development in *Local Community Depth*, *Cohesion*, and *Particularly Strong Cohesion*.

The PaLD perspective arises from social concepts of opposition, support, and strength of alignment. In particular, representing data points within a latent social space and considering the relative positioning of individuals, we can obtain a measure of local prominence (or power) of an individual via the support received when in direct opposition with others. Specifically, suppose (social) individuals x and y are in conflict. The set of individuals, z, with particular impetus for involvement in the dispute could, in some sense, be viewed to be those with more knowledge of either x or y than x and y have of one another. In this social context, we refer to such sets as *(local) conflict foci*, expressed formally in Eq. **1** and denoted by $U_{x,y}$.

The local prominence of an individual may be reflected by considering the level of "support" that the subject has in the foci they induce. Specifically, for a fixed x, the local depth can be formulated as the proportion of members in randomly selected (induced) foci which are closer to x than to the respective opposing individual. This is expressed formally in Eq. **2**.

As an individual becomes locally prominent due to the support from those around them, the contribution to local depth from others can be seen as expressing strength of alignment, here referred to as cohesion. Note that the chance of selection of a particular individual in a conflict focus is inversely related to the size of that focus, which aligns with the concept that individuals have more significant social presence within smaller groups (10). Cohesion is defined formally in Eq. **3**.

The concepts of strong and weak ties have been studied extensively in the sociological literature (11–13). We consider the cohesion between two points to be "particularly strong" when mutual cohesion is greater than that expected of a random point involved in conflict (Eq. **4**). The network structure provided by the distinction between (particularly) strong and weak ties can provide insight into varying roles of relationships in overall structural organization.

Here, we take the perspective that strong ties are the foundations of communities, and thus, a (community) cluster is a connected group with no strong ties to individuals outside the group. The network of (community) clusters does not sever strong ties and can provide both evidence for spectrums and contiguity of a population in addition to clear evidence for the existence of distinct groups. Discussion of group formation in the sociological literature can be found in refs. 14–16. For discussion of how the study of clustering algorithms can provide insights into the agglomeration and division of social groups, see refs. 17 and 18.

The presence of varying local density is common in social settings (e.g., urban versus rural), influences the (physical) distances at which interactions between individuals occur, and plays an important role in group structure and formation. Accounting for density variation is an important feature in what follows.

We now turn to a mathematical treatment of local depth, cohesion, and related concepts.

Local Community Depth

We begin by introducing the concept of local (community) depth, which results from an approach to localization that incorporates interactions occurring at varying scales.

Suppose $S = \{s_1, s_2, \ldots, s_n\}$ is a finite set of data points, and write $x \in S$ to indicate membership in S. For any points $x, y \in S$, let $d(x, y)$ denote the dissimilarity of y from x. For the general theoretical framework, the dissimilarity does not need to satisfy (metric) properties such as symmetry or the triangle inequality (see *Properties of Cohesion*).

For any pair of points (x, y), define the local focus induced by x and y, $U_{x,y}$, to be the set of points which are as close to either x or y as x and y are to one another. That is,

$$U_{x,y} = \{\, z \in S \mid d(z, x) \leq d(y, x) \text{ or } d(z, y) \leq d(x, y) \,\}. \qquad \textbf{[1]}$$

The local (community) depth of a point x is then the probability that, when an opposing point, y, is selected (uniformly) at random, a second randomly selected point within the resulting induced focus is closer to x than to y (with ties broken by a coin flip). Formally, for a fixed $x \in S$, select Y uniformly at random from $S\backslash\{x\}$ (i.e., satisfying $Y \neq x$) and then select $Z = Z_{x,Y}$ uniformly at random from the focus $U_{x,Y}$. The local depth of the point x, denoted $\ell_S(x)$, is then defined via

$$\ell_S(x) = P(d(Z, x) < d(Z, Y)) + \frac{1}{2}P(d(Z, x) = d(Z, Y)). \qquad \textbf{[2]}$$

For clarity of exposition, in what remains, the term resolving ties will be suppressed. In all cases, the average local depth over the set S is equal to $1/2$.

The resulting measure of relative support is local (community) depth; the term "depth" is used here since, as with common employment of the term, we seek to capture geometric features of centrality. Several existing measures of depth are given as interpretable probabilities which are based on geometric constructions including half-spaces, simplices, and lenses (see, for instance, refs. 4 and 5). Two recent approaches to depth in which varying degrees of localization may be considered are introduced in refs. 19 and 20. Note that in the definition of local depth, relative positioning is observed entirely from the perspective of distance comparisons (within triples of points); this provides a desired sense of robustness and permits consideration of depth in high-dimensional settings (see *Applications* and *Performance Considerations*). In *Theorem 2: Limiting Irrelevance of Density*, we show that local depth accounts for varying density in the sense that, provided subsets are sufficiently separated, the local depth of a point is maintained as within-subset distances are contracted or dilated (see *Properties of Cohesion*).

For instructive purposes, in Fig. 1, we consider a small two-dimensional Euclidean dataset. In Fig. 1*A*, a single local focus is indicated, and the computation of local depth for a selected point is displayed. In Fig. 1*B*, as a prelude to what follows, we give the community structure revealed by partitioning local depth, with corresponding depth values indicated adjacent to the respective nodes. Edges colored are those for which the contribution to local depth (i.e., mutual cohesion) is greater than the interpretable threshold given in Eq. **4** (see *Cohesion* and *SI Appendix*, Table S1).

When the input distances correspond to between-node separation in a given network, local depth can be viewed as a new measure of network centrality; for a discussion of extant measures of network centrality, see, for instance, refs. 21 and 22 and the references therein. To provide a simple yet instructive example of an

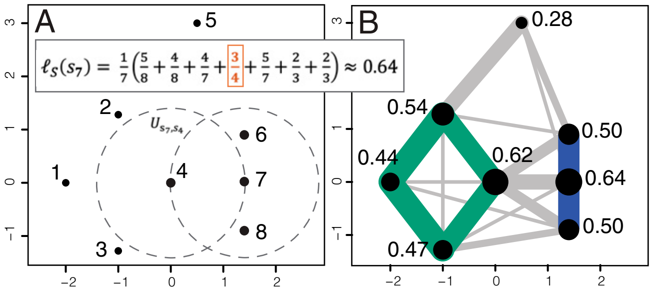

Fig. 1. We consider a small two-dimensional Euclidean data set. In *A*, we depict a conflict focus, U_{s_7,s_4}, and the calculation of local depth for a selected point, s_7 (here labeled 7); see also *SI Appendix*, Table S1. In *B*, local depth values are provided and illustrated by vertex size. Contribution to local depth (see *Cohesion*) from the surrounding points are depicted via edges in the overlayed graph; mutual contributions greater than the threshold are colored.

application of local depth for non-Euclidean dissimilarities, in *SI Appendix*, Fig. S2, we provide local depths associated with cholera fatalities in London in 1854 (23) using walking distance. Throughout, since distances are only considered indirectly, the concept of local (community) depth may be valuable for detecting hotspots (24) in the presence of varying density. To provide insight into varying degrees of localization, one could choose to alter the manner in which Y is selected (via some kernel), say.

In the next section, we introduce the concept of cohesion, which results from consideration of contributions to local depth.

Cohesion

A point becomes locally deep because of those around it, and thus, a measure of pairwise cohesion can be given in terms of the contribution of a point to the local depth of another. In particular, the concept of cohesion considered here is partitioned local depth. For a discussion of social motivation, see *Social Framework*.

Formally, for two points x and w in S, select $Y \in S\backslash\{x\}$ uniformly at random and $Z = Z_{x,Y}$ uniformly at random from $U_{x,Y}$. The cohesion (or contribution to local depth) of w to x, denoted $C_{x,w}$, is then defined via

$$C_{x,w} \;=\; P(\, Z = w \text{ and } d(Z,x) < d(Z,Y)\,). \qquad \textbf{[3]}$$

Since, from Eq. **3**, cohesion is simply partitioned local depth, the sum of the cohesion of all points to x is equal to the local depth of x. Cohesion is not a uniform rescaling (or transformation) of distance but rather a transformation that captures features of relative positioning via the geometric construction of the induced local foci $\{U_{x,y}\}$. Note again that, in a particular focus, the influence of a given point is inversely related to the cardinality of the respective focus (see *Social Framework*). The manner in which cohesion responds to varying local density is considered further in *Properties of Cohesion* as well as *Performance Considerations*.

Particularly Strong Cohesion

We now present a theoretical criterion for a pairwise relationship to be considered strong; this will be integral to the determination of community structure. The criterion is a global threshold, provided as an interpretable probability, that incorporates information regarding relationship strength across the space. The resulting definition of particularly strong ties can provide an alternative perspective to that of near neighbors, which reflects aspects of relative position.

Specifically, for fixed x and w, viewing the associated cohesion value as the probability of support for x from w, we consider $C_{x,w}$ to be "particularly strong" if the impact of w for x is greater than that expected of a random *focus point* W for a random point X. More formally, select X, Y in S (with $Y \neq X$) uniformly at random and $Z = Z_{X,Y}$ as before from $U_{X,Y}$. Finally, select $W = W_{X,Y}$ uniformly from $U_{X,Y}$ to represent a typical focus point. Then, the relationship between x and w is said to be "particularly strong" whenever

$$\min\{\, C_{x,w},\ C_{w,x} \,\} \;\geq\; P(\, Z = W \text{ and } d(Z,X) < d(Z,Y)\,); \qquad \textbf{[4]}$$

compare Eq. **4** to Eq. **3**. The concept of a relationship implicitly suggests mutual cohesion, which is the essential motivation for the use of the minimum function. In every example in this paper, the threshold distinguishing "particularly strong" relationships is precisely that given in Eq. **4** (see Figs. 1, 2, and 4–8).

We remark that with the reasonable added assumption that x is closer to itself than it is to any other point (i.e., $d(x,x) < d(x,y)$ for all $y \neq x$), we have that

$$P(Z = W) = P(Z = X) = \frac{1}{n}\sum_x P(Z = x)$$
$$= \frac{1}{n}\sum_x P(\, Z = x, d(Z,x) < d(Z,Y)\,). \qquad \textbf{[5]}$$

Leveraging symmetry in the selection of X and Y (and referring to the definition of $C_{x,x}$ via Eq. **3**), the threshold in Eq. **4**, which distinguishes particularly strong relationships, can be computed simply as half the average of the diagonal of the matrix of cohesion values. Evidence in support of the value of the given threshold can also be found in applications beyond community structure (see *Local and Global Considerations*).

For a simple example, we consider the two-dimensional data set in Fig. 2*A*. Here, we have a tightly knit group of five points at the lower left and a collection of seven points at much larger distances at the upper right. Particularly strong relationships are colored according to connected component of the implied graph of strong ties (see *Community Structure*). Fig. 2*B* provides the corresponding plot of cohesion against distance; for further discussion of the information conveyed by consideration of distance–cohesion pairs, see *Performance Considerations*. Identified groups may be held together by ties at distances that are many times larger than those separating them. Intracluster distances are as small as 87.8, while the distance from the point at the top right to its closest fellow cluster member is 238.5. The two weakly connected points on the far right of Fig. 2*A* exhibit the asymmetry of pairwise cohesion (0.061 and 0.065, at distance ∼250; only one is greater than the threshold). Note in Fig. 2*B* that the two main communities (at the lower left and upper right) with differing densities have comparable cohesion values. This will be a common feature exhibited in *Applications* and discussed further in *Properties of Cohesion* and *Performance Considerations*.

Extensive applications of cohesion are considered in *Applications* and *Performance Considerations*. We now provide some theoretical properties of cohesion and a discussion of the resulting networks.

Properties of Cohesion

In this section, we provide some fundamental properties of cohesion which are reasonable for approaches which convey the strength of community connections. These properties highlight the value of considering distance comparisons (rather than absolute distance values) and of incorporating interactions occurring across a variety of scales.

As we have seen throughout, cohesion and local depth values only depend on within-triplet dissimilarity comparisons. This can be particularly valuable when one wishes, for instance, to incorporate dissimilarity information provided by humans (e.g.,

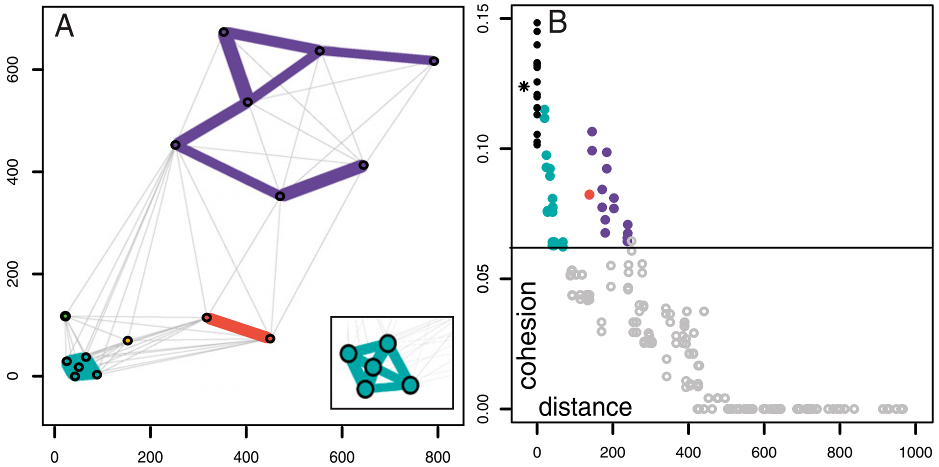

Fig. 2. We consider a small two-dimensional Euclidean dataset with varying local density. In *A*, edges between pairs whose mutual cohesion is greater than the threshold distinguishing strong relationships are bolded and indicated by colors. The perspective reveals three distinct (community) clusters and two isolated points. (*Inset*) The highly connected nature of the cluster in *Bottom Left* is displayed. In *B*, note that cohesion is more than a simple direct transformation of distance (the threshold of 0.062 is indicated by a horizontal line).

via crowdsourcing). Studies have suggested that one can often more reliably provide distance comparisons than exact numerical evaluations (5, 25). One example of a query regarding within-triplet distances could be of the form "Is z more similar to x than it is to y?" (Fig. 3*A*); this highlights the potential value in leveraging information of this form when there is uncertainty in a measure of dissimilarity. Fig. 3*A* provides an illustration of a triple for which an exact numeric evaluation of distance may be challenging.

In the following theorems, we provide formalizations of key properties of local depth and cohesion (see the illustrations in Fig. 3 *B–D*). Provided that sets are sufficiently separated, cohesion values are invariant under contraction and dilation of within-set distances (see *Theorem 2: Limiting Irrelevance of Density*). We also show how between-group ties weaken when the number of points in a concentrated region increases (see *Theorem 3: Separation under Increasing Concentration*). All properties are phrased asymptotically; the explicit distances at which these results hold is provided in each statement. The effect of such properties can be seen in the data considered throughout the paper and in *Applications* and *Performance Considerations*.

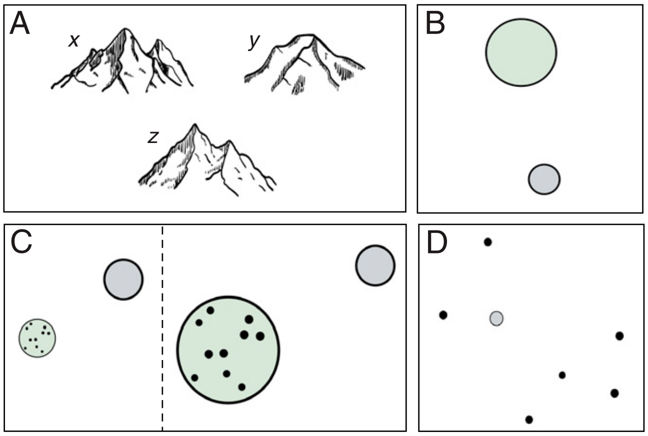

Fig. 3. An illustration for the setting of distance comparisons as well as three properties reasonable for approaches which convey the strength of community connections. In *A*, within-triple distance comparisons are determined by asking for instance: "Is *z* more similar to *x* or to *y*?" In *B*, *separation under increasing distance;* in *C*, *limiting irrelevance of density*; and in *D*, *separation under increasing concentration*.

Theorem 1: Separation under Increasing Distance. Suppose $S = A \cup B$ and A and B are mutually separated in the sense that $\max\{d(a, a')|a, a' \in A\} < \min\{d(a, b), a \in A, B \in B\}$ (respectively for B). Then, the between-set cohesion values are zero (i.e., $C_{a,b} = C_{b,a} = 0$ for any $a \in A$ and $b \in B$).

For proof, see *SI Appendix*.

Theorem 2: Limiting Irrelevance of Density. Suppose that A and A' have the same ordinal structure in the sense that, for a_i, a_j, $a_k \in A$ and a'_i, a'_j, $a'_k \in A'$, $d(a_k, a_i) < d(a_k, a_j)$ if and only if $d(a'_k, a'_i) < d(a'_k, a'_j)$. Suppose additionally that $S = A \cup B$ (respectively $S' = A' \cup B$) for some set B with the property that A and B (resp. A' and B) are mutually separated. Then, $\ell_S(a_i) = \ell_{S'}(a'_i)$ and $C_{a_i, a_j} = C_{a'_i, a'_j}$ for any i, j.

For proof, see *SI Appendix*.

Theorem 3: Separation under Increasing Concentration. Suppose that $S = A \cup B$, and B is concentrated with respect to A in the sense that $\max\{d(b, b')|b, b' \in B\} < \min\{d(b, a)|\, a \in A, b \in B\}$, and for any a, $a' \in A$, either 1) $d(b, a) < d(a', a)$ for all $b \in B$, or 2) $d(b, a) > d(a', a)$ for all $b \in B$. If $|B|$ is sufficiently large relative to $|S|$, then for any $a \in A$ and $b \in B$, the relationship between a and b is not particularly strong.

For proof, see *SI Appendix*.

Note: One may observe that the probabilities defining cohesion (see Eq. **3**) can be computed directly according to the finite sum

$$C_{x,w} = \frac{1}{n-1} \sum_{\substack{y \in S \\ y \neq x}} \frac{\mathbf{1}(d(w, x) < d(w, y),\ w \in U_{x,y})}{|U_{x,y}|}.$$

Furthermore, since cohesion is partitioned local depth, local depth can be expressed as $\ell(x) = \sum_{w \in S} C_{x,w}$, (see Eq. **2**). A summary of properties and pseudocode for computing cohesion are included in *SI Appendix*.

Graphical Community Structure

Valuable information about the structure of data can be communicated via the weighted network whose edge weights express mutual cohesion. The subgraph consisting of particularly strong ties may reveal distinct features of group separation and can provide a perspective on the inherent shapes of communities, which may be valuable in applications.

Specifically, the community structure of S can be analyzed via the associated (undirected) *cohesion network* G_S, whose vertex set is S and whose edges (x, w) are weighted by $\min\{C_{x,w}, C_{w,x}\}$ (Figs. 1*B* and 2*A*). As in the discussion following Eq. **4**, the minimum function is employed to express mutual cohesiveness [see also mutual k-nearest neighbor graphs (26)]. The (community) cluster network, denoted G_S^*, is a subnetwork of G_S consisting of the edges (x, w) for which the relationship between x and w is particularly strong. We refer to the connected components of G_S^*, including inherent internal network structure, as the (community) clusters in this setting (see Figs. 1, 2, and 4–8).

The weak ties in G_S can provide information about the relative positioning of individuals within communities (see *Applications*, and, in particular, see Figs. 5–7). Specifically, positions of points in low-dimensional embeddings of the cohesion network, G_S, can give insight into contextual similarity, complementing the perspective of other embedding techniques (see, for instance, refs. 27 to 29 and the references therein). The full cohesion matrix and the associated directed network may convey additional structural information (including potential relational asymmetries).

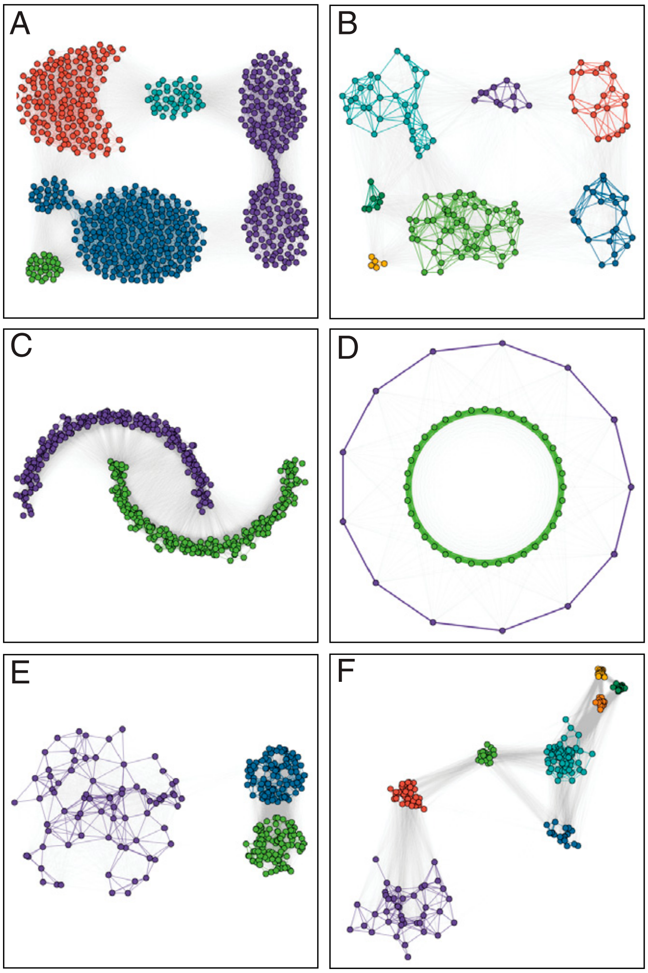

Fig. 4. Examples of two-dimensional benchmark Euclidean datasets with overlayed cluster networks are provided in *A*–*F*. In *D*, the two closest points to any on the outer ring are located within the inner ring. In *E* and *F*, note the detection of clusters in the absence of constant within-cluster density. Recall that no parametric assumptions or optimization criteria are employed. Points are plotted at proportionally accurate distances.

The (community) cluster network maintains all particularly strong relationships. Nevertheless, as seen in *Applications*, the connected components of the cluster network do align with commonly accepted partitions of benchmark clustering data (Fig. 4). Cohesion and the resulting networks give an approach for analyzing structure, which can provide evidence for both spectrums and the existence of distinct groups. Since PaLD is defined without the use of optimization criteria or extraneous inputs, the results obtained from this perspective can complement that provided by other methods.

PaLD and Existing Methods

We now discuss PaLD and its place alongside other methods and approaches, in particular, the potential value provided by recalibrating in response to density variation (see *Theorem 2: Limiting Irrelevance of Density* and *Performance Considerations*), which removes the need for localizing parameters.

Particularly strong ties can be viewed as providing a complementary perspective on the concept of near neighbors that captures features of relative positioning beyond those provided by the distance to or ranking of one's neighbors. The number of neighbors varies across the graph, as points with larger local depth will typically have more strong ties in G_S^* than those with smaller local depth (see Eqs. **3** and **4**).

Related work in topological data analysis seeks to address aspects of shape, connectivity, and structure in data via constructed graphs (see, for instance, ref. 30 and the references therein). Though here we will only focus on connected components, methods such as (persistent) homology may be valuable in a further analysis of structural information provided by community networks. Here, the community and cluster networks G_S and G_S^* capture aspects of relative positioning and do not require additional parameters. As such, a structural analysis of community graphs (which reflect cohesion rather than distance) can complement that obtained from graphs created using neighborhood radii and k-nearest neighbors.

Density-based algorithms, including density-based spatial clustering of applications with noise (DBSCAN) (and related techniques, including hierarchical density-based spatial clustering of applications with noise [HDBSCAN]; see, for instance, refs. 7 and 8), can give insight into varying density throughout the underlying space (via tuning parameters), often with the intention to remove "noise" or identify high-density regions separated by lower-density regions. The manner in which PaLD accounts for density variation (see *Properties of Cohesion*) could potentially be valuable in initial computations associated with current algorithms.

PaLD may reveal individuals between groups, and their role as bridges between select individuals in distinct communities. When there is a strict desire to partition into groups, one may wish to use another method in conjunction with PaLD to carefully remove points which are considered to potentially be noise. Since community clusters require a complete absence of strong relationships with points outside, the existence of disjoint groups is a strong signal for separation (see also *Applications* and *SI Appendix*, Fig. S3). Although not pursued here, when additional partitioning of the community graph is desired, community detection methods for networks, such as spectral clustering or the Louvain algorithm (31–33), could be applied directly to G_S^* (or G_S).

Applications

In this section, we consider applications of the method to uncover community structure in benchmark examples as well as high-dimensional data arising in the study of linguistics, genetics, and cultural psychology. In each example in what follows, cohesion values are computed to determine edge weights, and the resulting weighted network is provided (edges which correspond to particularly strong ties are colored in the

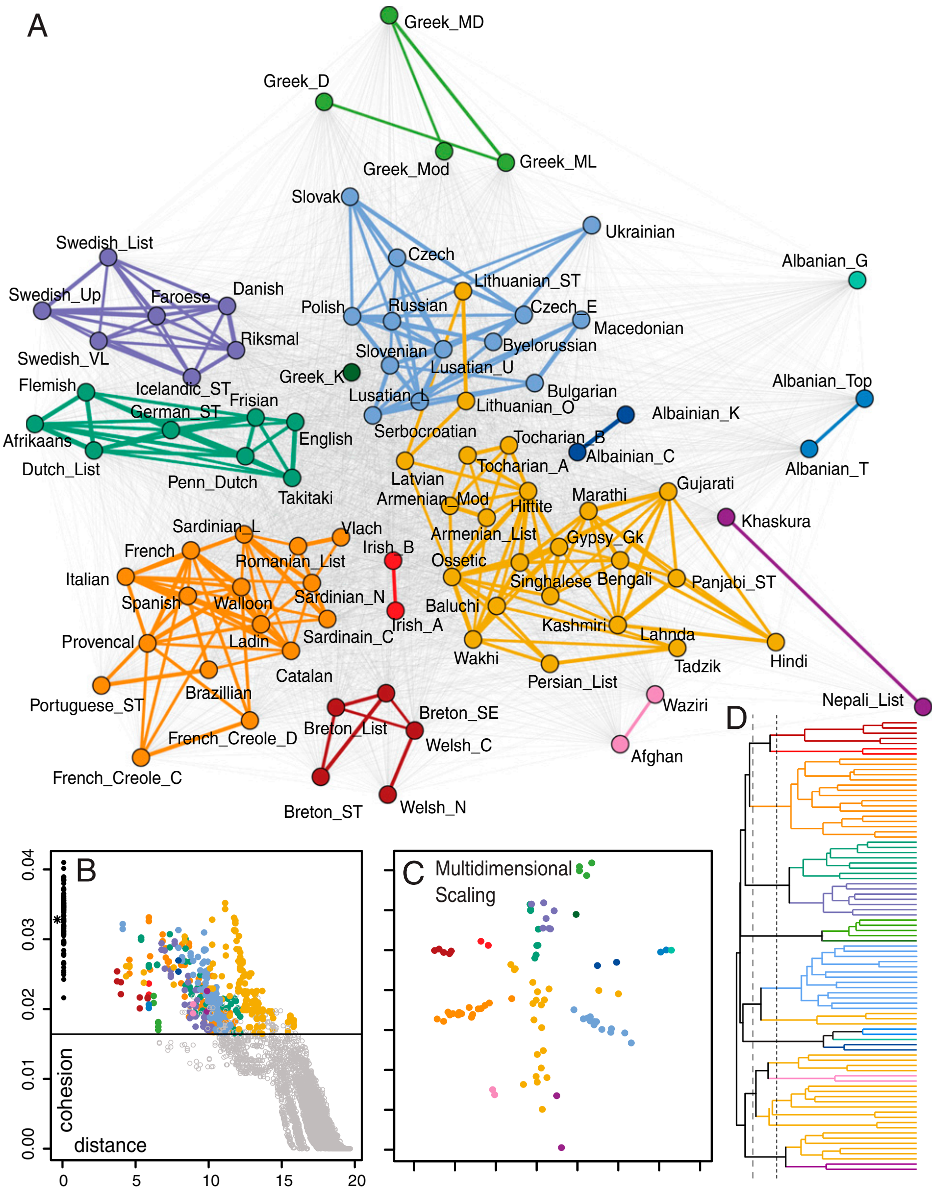

Fig. 5. We consider community structure for 87 Indo-European languages employing cognate information coded via 2,665-dimensional binary vectors. In *A*, note commonly identifiable language clusters and corresponding inter- and intracluster structure. Several ancient languages are centrally located. In *B*, we display cohesion against distance; the threshold delineating particularly strong relationships is indicated by a horizontal line. In *C*, a two-dimensional embedding provided by multidimensional scaling is given (35). In *D*, hierarchical clustering of the distance data is displayed, with dashed vertical lines indicating cuts given by maximizing the Calinski–Harabasz index (53) (k = 8) and average silhouette width (54) (k = 14), respectively. In *B*–*D*, the coloring is according to that given by PaLD.

accompanying plot). Throughout, PaLD does not require user inputs beyond a collection of pairwise distances, is not iterative, and is defined independently of any external optimization criteria or loss function specifying a priori necessitated properties of clusters. The results obtained in the presence of the simplicity and transparency of the approach suggest that cohesion is potentially valuable in overall considerations of data structure.

To highlight the relationship between community structure and existing methods, we include results from clustering methods (such as hierarchical, k-means, DBSCAN, and HDBSCAN) and low-dimensional embeddings [such as principal component analysis and multidimensional scaling (34, 35)]. For further comparisons, see *SI Appendix*, Figs. S3–S6 and S10–S12 and Tables S2–S6.

We now consider PaLD applied to some benchmark data as well as high-dimensional examples in linguistics, genetics, and cultural psychology.

Benchmark Data. In Fig. 4, we display the community cluster networks for six structured two-dimensional Euclidean datasets along with overlayed community graphs; for the generation of the data employed in Fig. 4 *A*–*C*, see refs. 36 and 37. In Fig. 4 *A*–*C*, note that the connected components of the (thresholded) cluster graphs coincide with commonly accepted clustering of such sets without leveraging external cost functions or requiring further inputs (such as the number of clusters or neighborhood size). In Fig. 4*D*, the nested ring structure of the set is identified despite the fact that the two closest points to any point within the outer ring are located in the inner ring. In Fig. 4 *E* and *F*, within-cluster density varies; in both examples (as in Fig. 2), there are within-cluster ties which span larger distances than the minimum distance between clusters. In each example, the graphical community structure conveys valuable additional information that is not obtained from methods whose focus is on the assignment of class labels. In all examples in this paper, the threshold employed is that provided in Eq. **4**.

We now consider applications to high-dimensional data in linguistics, genetics, and cultural psychology.

Languages. In Fig. 5, we display the community cluster network for 87 Indo-European languages arising from cognate information, coded using 2,665-dimensional binary vectors (38, 39). In Fig. 5*A* and in the remainder of the paper, we use the Fruchterman–Reingold (FR) force-directed graph drawing algorithm (40) to display the weighted network G_S, and indicate the edges of G_S^* using colors which highlight the inherent community structure. The only input required to obtain the corresponding matrix of cohesion values was simply a collection of distances; Euclidean distance was employed here.

One may note the commonly identifiable language clusters and that, under a slight rotation, some of the underlying geography is mirrored in the plot. Fig. 5*B* provides a plot of cohesion versus distance with the threshold from Eq. **4** indicated by a horizontal line. A discussion of the information available through consideration of distance–cohesion pairs can be found in *Performance Considerations*. Observe that large values of cohesion can appear over a wide range of distances. For instance, the community of ancient and central languages (including Tocharian, Hittite, and Armenian) exhibits some of the largest pairwise cohesion values despite being at relatively high pairwise distances (above half the maximum distance); see *SI Appendix*, Fig. S7 for a histogram of cohesion values. Subcommunities can be revealed via internal structure in G_S^* (see, for instance, the cluster of Romance languages in the lower left and Slavic languages toward the upper right in Fig. 5*A*) as in the case of hierarchical approaches. Note that the community graph provides rich structural information beyond that given by low-dimensional embeddings or conveyed by trees and cluster labels (see Fig. 5*A* in contrast to Fig. 5 *C* and *D*). Additional results from PaLD and a variety of complementary perspectives, including further low-dimensional visualizations and near neighbor networks, are included in *SI Appendix* (*SI Appendix*, Figs. S6–S10 and Tables S2–S6).

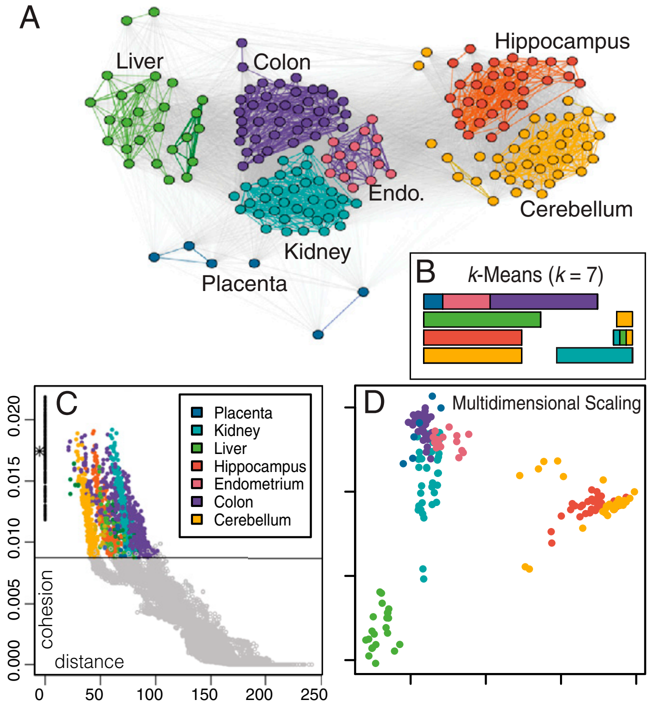

Fig. 6. We consider a 189 *x* 22, 215 array of gene expression measurements from various tissues; vertices are colored according to tissue type. In *A*, colored edges indicate components in the network of "particularly strong" relationships. Again, the FR force-directed graph-drawing algorithm was employed for network visualization. In *B*, the partition obtained from *k*-means (with $k = 7$) is provided. In *C* is a display of cohesion against distance; note the vertical swaths of color. In *D*, a two-dimensional embedding of the distance data using multidimensional scaling is provided.

Gene Expression Data. In Fig. 6, we display the community cluster network for a collection of 22,215-dimensional gene expression data from 189 tissue samples obtained from a variety of individuals (41). In Fig. 6*A*, the analysis reveals community structure among tissue types. Coloring in the plot is according to tissue type for vertices and connected components for edges. As in Fig. 5, we use Euclidean distance here.

Note in particular the graphical positioning of brain and lower abdominal tissues. The vertical bands of color in Fig. 6*C* illustrate how varying densities among groups are brought to comparable levels of cohesion (see also *Performance Considerations*). The community graph provides rich local and global information beyond that given by class labels and low-dimensional embeddings (see Fig. 6*A* in contrast to Fig. 6 *B* and *D*). For further results using standard partitioning methods, see *SI Appendix*, Table S4.

Cultural Psychology. In Fig. 7, we consider cultural distance information obtained in ref. 42 from two recent waves of the World Values Survey (2005 to 2009 and 2010 to 2014) (43). Distances are computed using the cultural fixation index (CF_{ST}), which is a measure built on the framework of fixation indices from population biology (44, 45). Recall that the foundation of PaLD in within-triplet comparisons allows for the employment of application-dependent and non-Euclidean measures of dissimilarity.

In Fig. 7*A*, we display community structure for regions within the United States, China, India, and the European Union. As highlighted in ref. 42, the United States is relatively homogeneous compared to Europe and India. Again, community clusters of varying density are brought to comparable levels of cohesion (Fig. 7 *B* and *C*). Specifically, note that the regions at the largest cultural distance in the United States (East South Central and California, at a distance of 0.027, with no particularly strong connection) is less than that between all strongly connected pairs within India (having minimum distance 0.043) (*SI Appendix*, Table S7). This mirrors common local cultural perspectives. In Fig. 7*D*, we provide a two-dimensional nonmetric multidimensional scaling plot (35) based on the pairwise cultural fixation index values. Results of k-medoids ($k = 4$) (46) and HDBSCAN applied to the distance data are provided in Fig. 7*E* and *SI Appendix*, Tables S5 and S6. Numerical comparisons of partitions obtained from clustering methods (via normalized mutual information) are provided in *SI Appendix*, Tables S3 and S6.

The complexity of the datasets in this section can be seen through the consideration of results obtained via existing clustering methods (*SI Appendix*, Figs. S4–S6 and S10 and Tables S2–S6). It may be noted that methods such as k-nearest neighbor and hierarchical approaches can provide challenges in parameter, cutoff, and optimization criteria selection (*SI Appendix*, Figs. S4–S6 and S10 and Tables S2–S6). Even quite complex and novel density-based methods such as HDBSCAN may identify locally central points as noise in applications. In particular, languages, including variants of the ancient Tocharian and Armenian languages (as well as English for minPts = 4), are classified by HDBSCAN as noise, despite standing as central (Fig. 5; the associated local depth values are 0.65, 0.68, and 0.66, respectively). In addition, for the cultural data in Fig. 7, the widely spread Indian regions are broken apart by HDBSCAN with six classified as noise, and the European and

https://doi.org/10.1073/pnas.2003634119

US regions are agglomerated into one single cluster (aside from Romania [noise], *SI Appendix*, Table S5).

Performance Considerations

In this section, we discuss the performance of partitioned local depth in accounting for groups with varying local density. In particular, we consider increasing dimensions as a setting conducive to complex structure wherein the within-group variance of distances is diminishing, whereas the variance of cohesion is maintained. We will discuss how, in line with the theoretical results, the performance of partitioned local depth can be observed via consideration of cohesion in relation to distance (Figs. 2*B*, 5*B*, 6*B*, 7*C*, and 8).

In Fig. 8, we consider randomly generated data with inherent group structure and varying density over increasing dimensions. In particular, 100 points were selected uniformly at random from each of three balls of radii 0.5, 1, and 2.5; the corresponding distribution centroids are at constant distances 3, 4, and 5 (Fig. 8*A*). In Fig. 8*B*, we present the associated plots of cohesion versus distance together with small representations of the associated cluster graphs (employing the FR graph-drawing algorithm). Observe that as dimensionality increases, the variability of within-cluster distances decreases dramatically. On the other hand, the variability of cohesion values is maintained, with comparable ranges across clusters regardless of distance and density. In the 10,000-dimensional example in Fig. 8 *B* and *C*, the distances between points in Groups 1 and 2 (∼3.2) is less than the within-cluster distances in Group 3 (of ∼3.5). For a general discussion of dimensionality in data, see, for instance, refs. 47 and 48. In the histograms in Fig. 8*C*, it can be seen quite dramatically that within-group cohesion distributions (particularly for strong ties) equate well across groups despite vast differences in within-group distance distributions. In fact, the average total variation distance (TVD; see, for instance, ref. 49) between the (binned) distributions of within-group distances is 1, whereas that of within-group cohesions is 0.27. In the case of the cultural distance data in Fig. 7*B*, the corresponding averages for TVD are 0.78 (for distance) and 0.42 (for cohesion). A similar analysis for tissue types is provided in *SI Appendix*, Fig. S11. Recall that, as discussed in *Applications*, the recalibrating of distance via cohesion provides adaptation to local perspectives. Note that similar behavior can be observed in the distance–cohesion pairs displayed in Figs. 2*B*, 5*B*, 6*B*, and 7*C*.

As with all methods, care must be taken that input distances reflect the sense of inherent proximity for a given application; the lack of requisite (metric) properties of symmetry or the triangle inequality does provide greater flexibility/freedom in this selection. Additionally, the foundation in distance comparisons means that any monotone transformation of the dissimilarity function (e.g., log) will provide the same results. This feature may be particularly valuable for applications involving non-Euclidean dissimilarities (see, for instance, the applications to cultural distance in *Applications* and walking distance in *SI Appendix*, Fig. S2).

Local and Global Considerations

Particularly cohesive relationships give meaning to the concept of "local" within a given domain, an idea which underlies locality-based approaches to data analysis. Cohesiveness can be employed in place of variants of local kernel approaches (see, for instance, chapter 6 in ref. 50) in classification, imputation, anomaly detection, regression, smoothing, and elsewhere. Specifically, we can obtain meaningful global (or local) weightings by normalizing across all contributions (or only those which are particularly strong). Note that the kernels that result from PaLD vary according to relative positioning and density across the underlying space. Examples of applications to the smoothing of time series and classification are given in *SI Appendix*, Figs. S12 and S13. Additionally, the threshold in Eq. **4**, when

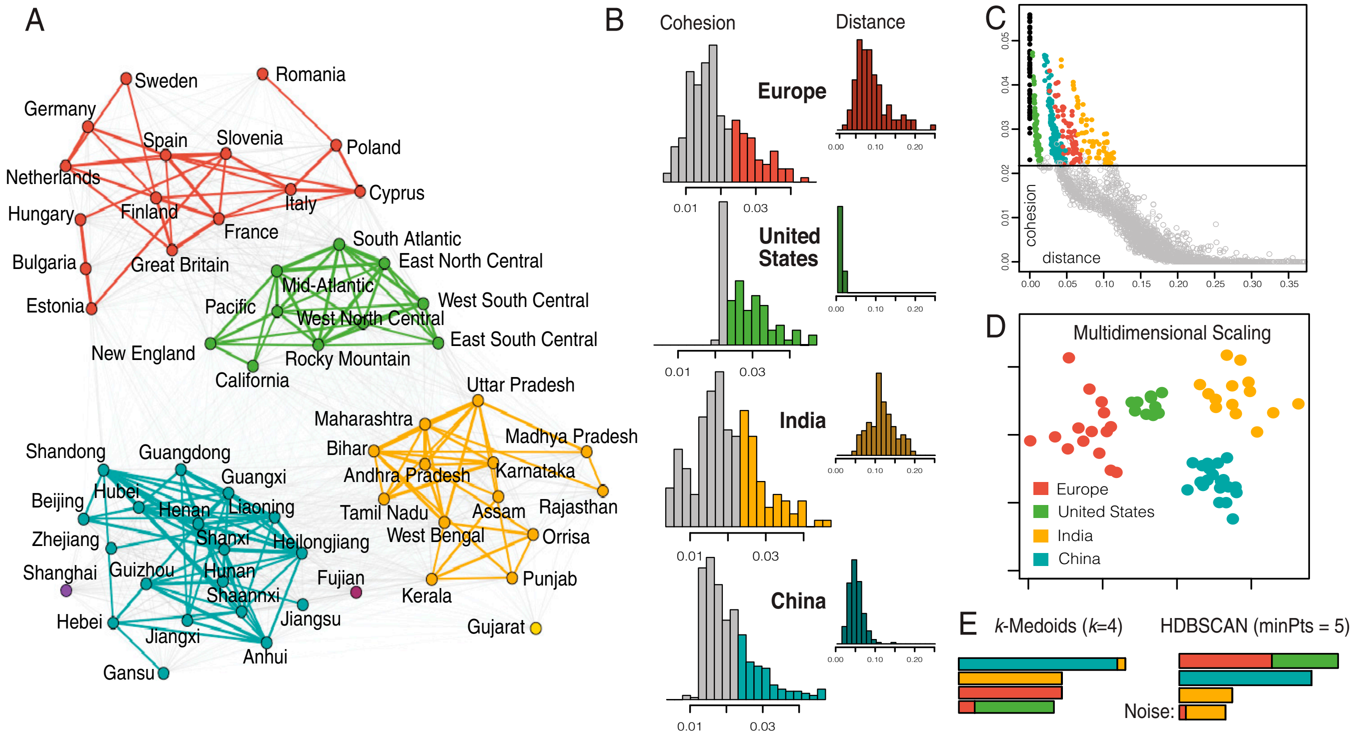

Fig. 7. In *A*, we display the community structure obtained from the cultural fixation index values from ref. 42 for regions within the United States, China, India, and the European Union. In *B*, we display the distribution of within-group cohesions and distances; note that distances are brought to comparable levels of cohesion. In *C* is a plot of cohesion against distance along with the threshold indicated by a horizontal line; pairs of points at distances greater than 0.36 all have cohesion equal to zero. In *D*, we provide a two-dimensional embedding via multidimensional scaling. In *E*, we indicate region membership for the clusters obtained from *k*-medoids (when *k* = 4) and HDBSCAN.

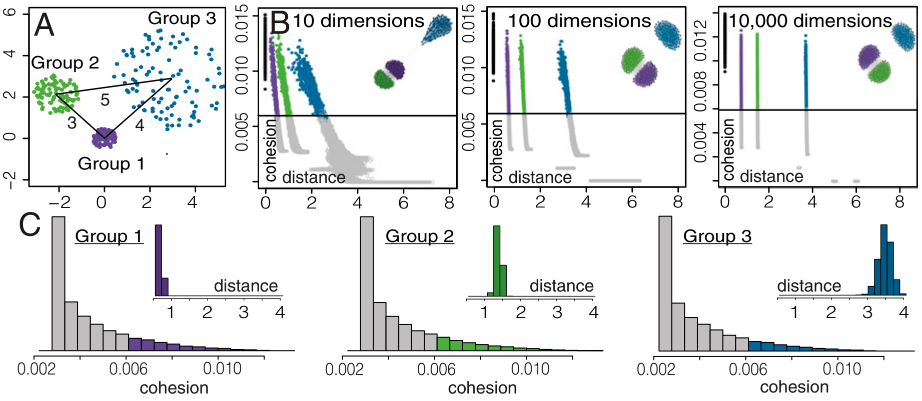

Fig. 8. We consider points sampled uniformly at random from balls of varying radii over increasing dimensions. In *A*, we consider the two-dimensional case to illustrate features of the generated data. In *B*, we display the associated plots of cohesion against distance together with small visuals of the corresponding cluster graphs. Although the variability of within-cluster distances decreases dramatically, with increasing dimension, the variability of cohesion is maintained and is comparable across clusters regardless of density. In *B* and *C*, values below the threshold are indicated in gray.

applied to the cultural distance data in Fig. 7, results in particularly strong connections appearing only at distances below 0.15, a quantity suggested by Wright as a heuristic cutoff above which there is great differentiation (see ref. 51). For further examples of connections between existing local–global heuristics and the given threshold, see *SI Appendix*, Figs. S12 and S13.

Discussion and Conclusions

The partitioning of local depths provides an integrated and holistic approach to depth, embedding, clustering, and local methods for general data that reflects relative positioning. In particular, cohesion provides a means to account for varying local density within data (see Theorems 1 to 3) which may be valuable in a variety of settings and applications. Notably, the resulting algorithm requires no extraneous inputs (beyond a collection of interpoint distance comparisons), optimization criteria, or distributional assumptions; this may be valuable for clarity, transparency, interpretability, and communication.

As mentioned earlier, we only leverage distance comparisons among triples of points, which can enable implementation when one has more confidence in relative comparisons as opposed to exact (numeric) dissimilarities. Since the magnitude of distances is indirectly employed, this permits the consideration of position in relatively high-dimensional data and that which is not inherently Euclidean.

With PaLD, each point is treated equally from a computational standpoint. Thus, in cases in which one can confidently identify extensive noise, some pre- or postprocessing may be warranted. In such instances, one may also choose to alter the manner in which the opposing point is selected to emphasize certain local interactions over others.

When a partitioning of the data into clusters is desired, the results obtained from PaLD can complement those provided by other perspectives. Recall, when further partitioning of the graph is of interest, community detection methods for networks could be applied directly to G_S^* (or G_S); see *PaLD and Existing Methods*. The manner in which cohesion accounts for varying local density might be useful in applications beyond those considered here (e.g., hotspot detection, classification, and smoothing). Since no additional parameters, distributional assumptions, or cost functions are employed, the approach can provide additional context for the results obtained from other methods.

Contribution to local depth only uses a collection of within-triplet distance comparisons, and hence, the inherent computational complexity is not influenced by any underlying Euclidean dimension or lack thereof. Note that for the cognate data set considered in Fig. 5, computation of the distances, local depths, and cohesion matrix along with the displaying of the community network required ∼0.80 s. A naive implementation of the algorithm to exactly compute the cohesion matrix is $O(n^3)$; see *SI Appendix*, Fig. S14 for pseudocode. PaLD in its entirety is presently applicable for datasets of size $n = 20{,}000$ (i.e., when calculation and ready storage of a Euclidean distance matrix is typically feasible). Note that the method is entirely deterministic, and one does not need to search a parameter space nor select initial values. In addition, the threshold in Eq. **4** provides for sparsification (in the form of the graph G_S^*, *SI Appendix*, Table S7); this may be valuable for memory considerations. Methods which exploit parallelizability, the redundancy of the collection of ordinal dissimilarity relationships, or the probabilistic nature of local depths may in the future be able to improve algorithmic complexity. Implementation of ideas related to nearest neighbor descent (52) may also be of value when considering particularly large datasets.

In closing, it is crucial to note that there is value in approaches which allow scientists from varying backgrounds to obtain informative, concrete, and interpretable results that are well suited for communication to the general public in a straightforward and actionable manner. The concept of data communities proposed here is derived from, and aligns with, a shared human social perspective.

Data Availability. All data are provided in GitHub at https://github.com/moorekatherine/pald-communities.

ACKNOWLEDGMENTS. We acknowledge the research assistance of G. Ballard, J. Guo, Y. Huang, R. Langefeld, and L. D'Agostino McGowan. We also thank R. W. R. Darling for valuable discussions, J. Healy for suggestions regarding integrating the PaLD approach into density-based algorithms, and M. Muthukrishna for providing the cultural distance data for Fig. 7. K.S.B. acknowledges financial support from a Baker Family Faculty Fellowship. K.S.B. and K.E.M. would like to personally thank students in a variety of recent classes within which PaLD has been covered.

1. R. Chiong, *Nature-Inspired Algorithms for Optimisation* (Springer, 2008), vol. 193.
2. A. Neme, S. Hernández, "Algorithms inspired in social phenomena" in *Nature-Inspired Algorithms for Optimization*, R. Chiong, Ed. (Springer, 2009), pp. 369–387.
3. P. D. Hoff, A. E. Raftery, M. S. Handcock, Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97**, 1090–1098 (2002).
4. Y. Zuo, R. Serfling, General notions of statistical depth function. *Ann. Stat.* **28**, 461–482 (2000).

5. M. Kleindessner, U. von Luxburg, Lens depth function and k-relative neighborhood graph: Versatile tools for ordinal data analysis. *J. Mach. Learn. Res.* **18**, 1889–1940 (2017).
6. M. M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, "LOF: Identifying density-based local outliers" in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, W. Chen, J. F. Naughton, P. A. Bernstein, Eds. (ACM, 2000), pp. 93–104.
7. R. J. G. B. Campello, P. Kröger, J. Sander, A. Zimek, Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **10**, 1343 (2020).
8. B. S. Everitt, S. Landau, M. Leese, D. Stahl, *Cluster Analysis* (John Wiley & Sons, Ltd, ed. 5, 2011).
9. R. Domingues, M. Filippone, P. Michiardi, J. Zouaoui, A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognit.* **74**, 406–421 (2018).
10. S. L. Feld, The focused organization of social ties. *Am. J. Sociol.* **86**, 1015–1035 (1981).
11. J. S. Coleman, Social capital in the creation of human capital. *Am. J. Sociol.* **94**, S95–S120 (1988).
12. M. S. Granovetter, The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380 (1973).
13. R. S. Burt, *Structural Holes: The Social Structure of Competition* (HUP, Cambridge, MA, 2009).
14. L. C. Freeman, The sociological concept of "group": An empirical test of two models. *Am. J. Sociol.* **98**, 152–166 (1992).
15. J. Moody, D. R. White, Structural cohesion and embeddedness: A hierarchical concept of social groups. *Am. Sociol. Rev.* **68**, 1 (2003).
16. C. Stadtfeld, K. Takács, A. Vörös, The emergence and stability of groups in social networks. *Soc. Networks* **60**, 129–145 (2020).
17. K. D. Bailey, Sociological classification and cluster analysis. *Qual. Quant.* **17**, 251–268 (1983).
18. J. R. S. Fonseca, Clustering in the field of social sciences: That is your choice. *Int. J. Soc. Res. Methodol.* **16**, 403–428 (2013).
19. C. Agostinelli, M. Romanazzi, Local depth. *J. Stat. Plan. Inference* **141**, 817–830 (2011).
20. D. Paindaveine, G. Van Bever, From depth to local depth: A focus on centrality. *J. Am. Stat. Assoc.* **108**, 1105–1119 (2013).
21. F. Bloch, M. O. Jackson, P. Tebaldi, Centrality measures in networks. https://dx.doi.org/10.2139/ssrn.2749124 (June 1, 2019).
22. M. Newman, *Networks* (Oxford University Press, 2018).
23. J. Snow, *On the Mode of Communication of Cholera* (John Churchill, 1855).
24. A. B. Lawson, Hotspot detection and clustering: Ways and means. *Environ. Ecol. Stat.* **17**, 231–245 (2010).
25. A. Ukkonen, "Crowdsourced correlation clustering with relative distance comparisons" in *2017 IEEE International Conference on Data Mining (ICDM)*, V. Raghavan, S. Aluru, G. Karypis, L. Miele, X. Wu, Eds. (IEEE Computer Society, 2017).
26. F. Ros, S. Guillaume, Munec: A mutual neighbor-based clustering algorithm. *Inf. Sci.* **486**, 148–170 (2019).
27. A. Gisbrecht, B. Hammer, Data visualization by nonlinear dimensionality reduction. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **5**, 51–73 (2015).
28. S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).
29. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* [Preprint] (2020). https://arxiv.org.arXiv:1802.03426 (Accessed 30 December 2021).
30. L. Wasserman, Topological data analysis. *Annu. Rev. Stat. Appl.* **5**, 501–532 (2018).
31. U. von Luxburg, A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
32. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *JSTAT* **2008**, P10008 (2008).
33. M. Newman, Communities, modules and large-scale structure in networks. *Nat. Phys.* **8**, 25–31 (2012).
34. I. Jolliffe, *Principal Component Analysis* (Wiley, 2002).
35. I. Borg, P. J. F. Groenen, *Modern Multidimensional Scaling* (Springer, 2005).
36. A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation. *ACM TKDD* **1**, 1–30 (2007).
37. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *JMLR* **12**, 2825–2830 (2011).
38. I. Dyen, J. B. Kruskal, P. Black, An Indoeuropean classification: A lexicostatistical experiment. *Trans. Am. Philos. Soc.* **82**, iii-132 (1992).
39. R. D. Gray, Q. D. Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439 (2003).
40. T. M. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. *Softw. Pract. Exper.* **21**, 1129–1164 (1991).
41. M. Love, R. Irizarray, tissuesGeneExpression, Version 1.0. https://github.com/genomicsclass/tissuesGeneExpression. Accessed 8 November 2019.
42. M. Muthukrishna *et al.*, Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychol. Sci.* **31**, 678–701 (2020).
43. R. Inglehart *et al.*, *World Values Survey: All Rounds-Country-Pooled Datafile 1981-2014* (JD Systems Institute, Madrid, 2014).
44. A. V. Bell, P. J. Richerson, R. McElreath, Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 17671–17674 (2009).
45. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton University Press, 1994).
46. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster fAnalysis* (JWS, 1990).
47. D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality" in *American Mathematical Society Conference Math Challenges of the 21st Century* (AMS, 2000).
48. A. N. Gorban, I. Y. Tyukin, Blessing of dimensionality: Mathematical foundations of the statistical physics of data. *Philos. Trans.- Royal Soc., Math. Phys. Eng. Sci.* **376**, 20170237 (2018).
49. O. Haggstrom, *Finite Markov Chains and Algorithmic Applications* (Cambridge University Press, 2002).
50. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, NY, 2009).
51. S. Wright, Evolution and the *Genetics* of *Populations: A Treatise* in *Four Volumes*: Vol. 4: *Variability Within* and *Among Natural Populations* (University of Chicago Press, 1978).
52. W. Dong, M. Charikar, K. Li, "Efficient K-nearest neighbor graph construction for generic similarity measures" in *Proceedings of the 20th International Conference on World Wide Web*, S. Srinivasan *et al.*, Eds. (ACM, 2011), pp. 577–586.
53. T. Caliński, J. Harabasz, A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* **3**, 1–27 (1974).
54. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).