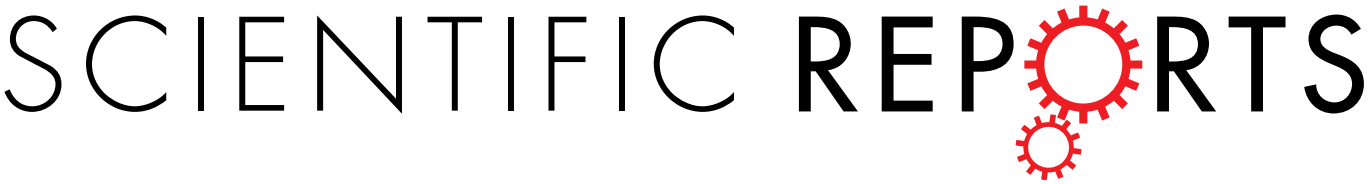

OPEN

Cluster-based network proximities for arbitrary nodal subsets

Kenneth S. Berenhaut[1], Peter S. Barr[2], Alyssa M. Kogel[1,3] & Ryan L. Melvin[1,4]

Received: 20 September 2017
Accepted: 7 August 2018
Published online: 25 September 2018

The concept of a *cluster* or *community* in a network context has been of considerable interest in a variety of settings in recent years. In this paper, employing random walks and geodesic distance, we introduce a unified measure of cluster-based proximity between nodes, relative to a given subset of interest. The inherent simplicity and informativeness of the approach could make it of value to researchers in a variety of scientific fields. Applicability is demonstrated via application to clustering for a number of existent data sets (including multipartite networks). We view community detection (i.e. when the full set of network nodes is considered) as simply the limiting instance of clustering (for arbitrary subsets). This perspective should add to the dialogue on what constitutes a cluster or community within a network. In regards to health-relevant attributes in social networks, identification of clusters of individuals with similar attributes can support targeting of collective interventions. The method performs well in comparisons with other approaches, based on comparative measures such as NMI and ARI.

There has been heightened interest recently regarding clustering of individuals in social networks based on characteristics such as tobacco use[1], alcohol consumption[2], level of happiness[3], emotion[4], divorce[5], cultural preferences[6,7], gun violence[8], and general health behaviors and attitudes[9–15] (see Fig. 1). However, there is little notion of what definitively constitutes a tightly or diffusely knit cluster in such instances. The requirement that individuals comprising a cluster be linked via path-wise attachment (through nodes of similar characteristic) may not be appropriate, particularly in cases where there may be missing data regarding links or nodal attributes. Here we provide a notion of proximity of nodes restricted to a subset for a network, which is then well-suited for analysis via extant clustering procedures. The method can be applied with informativeness through all levels of subset size, from only a few nodes in a large network through consideration of the limiting case of all nodes (commonly referred to as *community detection* or *graph partitioning*; see for instance Porter *et al.*[16], Newman[17], Schaeffer[18] and Fortunato[19,20]). The work here has applications in scientific fields where networks with nodal attributes arise including biology, ecology, neuroscience, physics, computer science, sociology, psychology, chemistry, and economics. One side benefit of the approach is that, applied in community detection, it is parsimonious and simple (see Eq. 1). To the best of our knowledge this is the first work specifically providing a measure of proximity between nodes that adequately reflects cluster membership for restriction to arbitrary nodal subsets on arbitrary networks (including non-spatial networks; see Related Work). This should add to the dialogue on what constitutes a community within a network. As mentioned in[21], in regards to health-relevant attributes in social networks, identification of cliques or clusters of individuals with similar attributes can support targeting of collective interventions.

The remainder of the paper proceeds as follows. We first introduce the concept of community-relative distance (see Community-relative Distance), and then turn to discussion of applications in the context of related work (see Related Work and Applications and Discussion). The paper ends with some technical computational considerations (see Materials and Methods).

Community-Relative Distance

Consider a network represented as a graph, $G=(V, E)$, with a set of vertices or nodes, V, and a set of edges, E (see Fig. 2 for an example of a 25-node, 40 edge graph). We assume the graph is connected and undirected, and the edges are unweighted (although it is not difficult to extend the work to weighted edges). Suppose some subset of nodes, S, is selected. These nodes could represent, for instance, infected individuals in a social network (or individuals with specific attributes, such as obesity or other health behaviors), suspected terrorists in a communication network, crimes on a spatial city street network, genes and conditions in gene expression networks, disease-related genes, proteins, or metabolites in an interaction network, etc., or simply nodes of high degree in a larger network.

[1]Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, NC, 27109, USA. [2]Department of Computer Science, Wake Forest University, Winston-Salem, NC, 27109, USA. [3]Present address: Department of Statistics, West Virginia University, Morgantown, WV, 26506, USA. [4]Department of Physics, Wake Forest University, Winston-Salem, NC, 27109, USA. Correspondence and requests for materials should be addressed to K.S.B. (email: berenhks@wfu.edu)

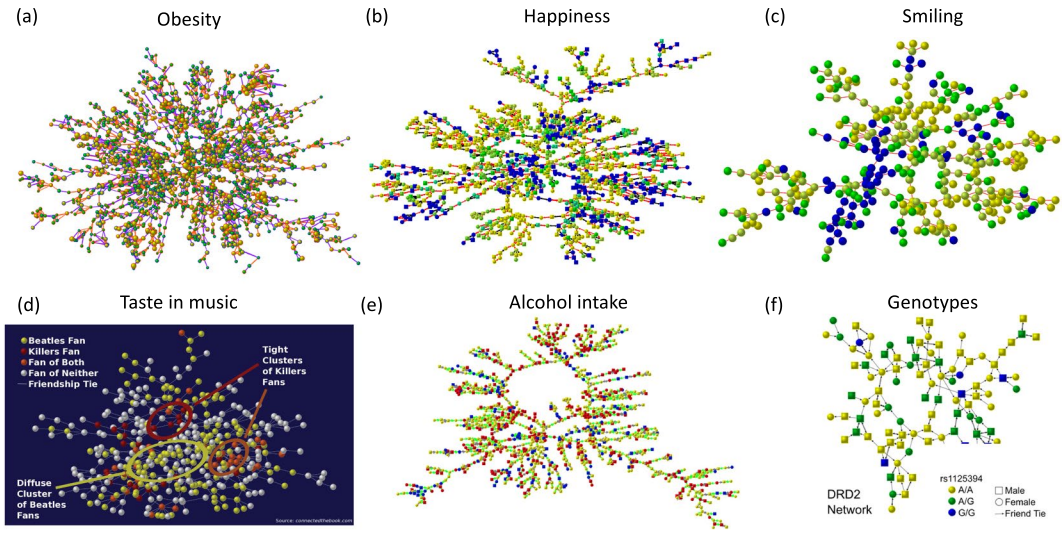

Figure 1. Examples of social networks with noted *clustering* of nodes of interest. The figures have been reproduced by permission of the authors of the respective manuscripts; for further details see the references as indicated. (**a**) A network of individuals in 2000 from the Framingham Heart Study (FHS) Social Network[9]. Connections arise from friendship, marital and familial ties. Yellow nodes indicate individuals with body mass index greater than or equal to 30, and nodes are colored green otherwise. The size of each node is proportional to the individual's body-mass index. The authors of[9] note that clusters of obese and non-obese individuals are visible in the network. (**b**) A network of individuals in 1996 from the FHS Social Network[3]. Colors indicate mean happiness of egos and all directly connected alters, on a spectrum from blue (unhappy) to yellow (happy). Happiness is measured via the Center for Epidemiological Studies depression scale. The authors of[3] note that "clusters of happy and unhappy people are visible in the network". (**c**) A social network of individuals in 2007 ascertained using Facebook[21]. Ties indicate the connected individuals were tagged in a photo together. Yellow nodes reflect individuals who are smiling in profile photographs and surrounded by others who are smiling. Similarly blue nodes reflect individuals who are frowning, surrounded by others who are frowning, and green indicate a mix of smiling and non-smiling friends. The graph suggests clustering of both blue and yellow nodes. In addition those who do not smile appear to be more scattered towards the peripherally in the network. (**d**) A social network of individuals in 2007 whose social ties were ascertained via Facebook[6]. The interior color of the nodes indicates the individual's taste in music. The graph suggests clustering (both diffuse and closely-knit) based on musical tastes within the network. (**e**) A network of individuals in 2000 from the FHS Social Network[2]. Node color denotes the alcohol intake of the subject, with red indicating an abstainer and blue indicating heavy intake (yellow nodes indicate moderate intake). As noted by the authors of[2], "the graph suggests clustering in abstention and heavy alcohol consumption behavior". (**f**) A network of individuals from the National Longitudinal Study of Adolescent Health (Add Health) Social Network[76], started in 1994[77]. Node color indicates genotypes for DRD2 (which has been associated with alcoholism). The graph suggests clustering of genotypes.

The underlying idea, here (see the example in Fig. 2), is that if two nodes, i and j, are part of a closely-knit community, then an individual resting at node i and perturbed off the node should encounter first (via randomly walking) a subset node "close" to node j. Specifically, for any two nodes, i and j in S, consider a random walk on G departing from node i. We define the *distance from i to j* (relative to the members of S), as the expected shortest-path (or geodesic) distance to j of the first node in S that the random walk encounters. We denote the resulting $|S| \times |S|$ matrix of distances as $\boldsymbol{D}$.

Now, define the *distance between i and j* (again, relative to the members of S) to be the smaller of the two associated distances (i to j and j to i), with the intuition that connections can be asymmetrically initiated (see Remark 1, below). Note that the reflexive distance between node i and itself is taken to be zero. In what follows, for convenience, we will refer to the resulting symmetric $|S| \times |S|$ matrix of distances as $\boldsymbol{D}^*$ (given for the example in Fig. 2a in Fig. S1), and the individual entries as *community-relative distances*. The reader is referred to Lovász[22] and Aldous and Fill[23] for discussion of random walks on graphs, and Pons and Latapy[24], Zhou and Lipowsky[25], and Zhou[26] for some discussion in the context of community detection (see also Related Work, below). For a survey on distance measures on graphs, see[27], and the references therein; for discussion of kernel-based measures, see for instance[28]. In general, one could replace the shortest-path distances used here with another context-dependent measure (including $\boldsymbol{D}$, in an iterative fashion).

Remark 1. Note that the distances from i to j and j to i may be quite different. For the network in Fig. 2(a), consider a random walk departing from node 4. The expected distance to node 9 of the first node in S that the walk encounters is 0.83 (reflecting likely encounters with nodes proximate to 9, such as 8 and 9, itself). On the other hand, for a random walk departing from node 9, the corresponding expected distance to node 4 is 1.52 (reflecting potential encounters with nodes distant to 4 such as 13 and 20).

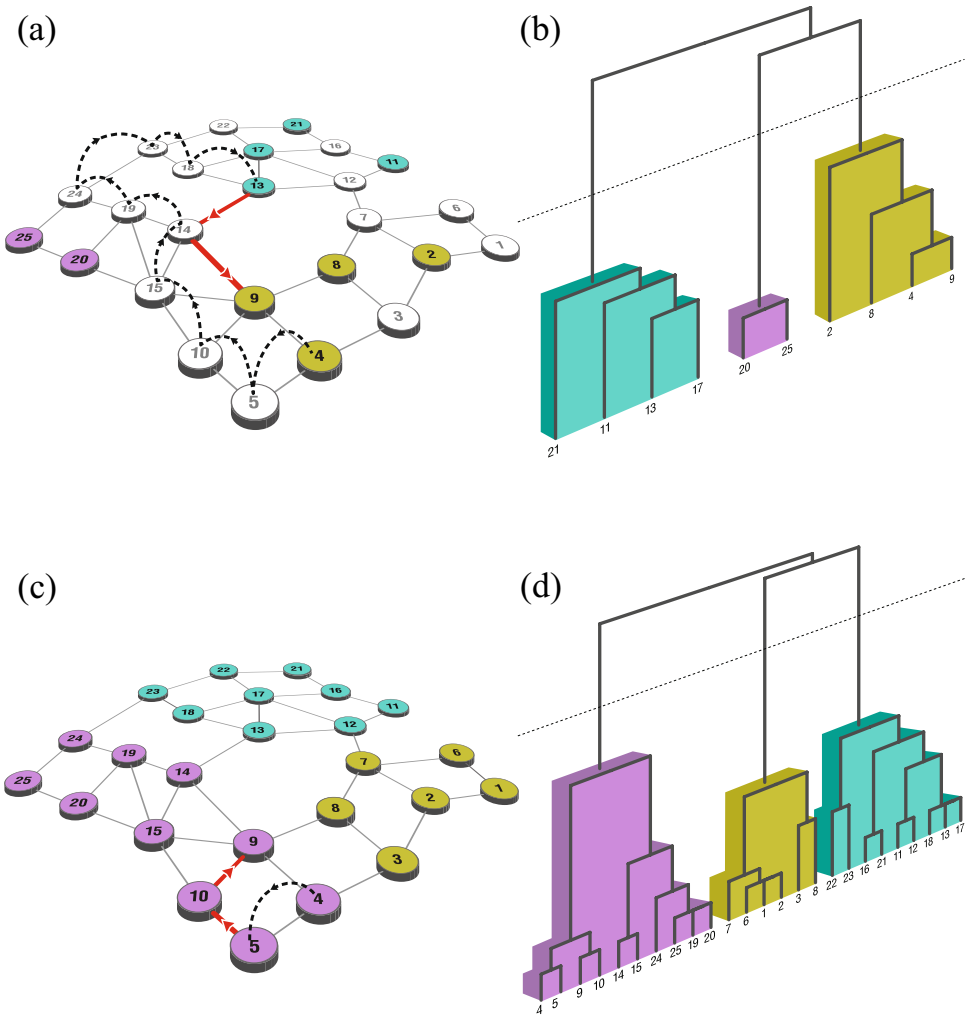

Figure 2. (**a**,**b**) A subset of ten selected nodes within a larger network of 25 nodes. A sample path for a random walk departing from node 4, and eventually entering the set of selected nodes at node 13 is indicated with dashed lines. The solid red line indicates a shortest path to the "target" node, node 9. A resulting dendrogram (via average-linkage clustering) is given in (**b**). A separation into three distinct clusters can be seen in the dendrogram. (**c**,**d**) The 25 node network with S comprised of all 25 nodes. A one-step sample path for a random walk departing from node 4, entering the set S at node 5 is indicated with dashed lines. The solid red line indicates a shortest path of length two to node 9. A resulting dendrogram is given in (**d**). A separation into three clusters (communities) can again be seen in the dendrogram.

Example 1. Figure 3a gives a dynamic perspective, which illustrates the connection between community-relative distances as suggested above and cluster membership. Consider the 165-node, 15×11 grid graph, G, with nodal subset S consisting of the 25 nodes in the 5×5 sub-grid in black, as well as the two nodes A and B shaded grey. For a random walk departing from Node A, the expected distance to B of the first node in S encountered is 2.55, as initial entry into S is likely to occur at a black node distant to B. Now, consider S augmented by the node at the position labeled 1; the expected distance now shrinks slightly to 2.53. The expected distances for the 28-node subsets obtained by augmenting with nodes 2 through 7 (in turn, in place of 1), are indicated adjacent to the corresponding node. Note that as the additional node is moved from positions 1 through 7, nodes A and B are in a sense drawn closer together, as the distance between them shrinks from 2.55 (with no added node) to 2.31 (for an added node at position 7). In a more general sense, two nodes at fixed position in a network will be drawn closer in community-relative distance when there are other proximate network nodes in the subset of interest. The scenario reflects a *cooperation* pattern between nodes in a community, and the leveraging of distance-based information from proximate nodes, particularly through weak ties[29]. □

Example 2. Figure 3b provides a simple example of an 8-node graph, with community-relative distances for pairs at shortest-path distance one indicated adjacent to the corresponding edge. Note that the tie between node 3 and node 5 is one of greater community-relative distance, suggesting separation, while the ties in the clique consisting of nodes 5–8 correspond to smaller community-relative distance. □

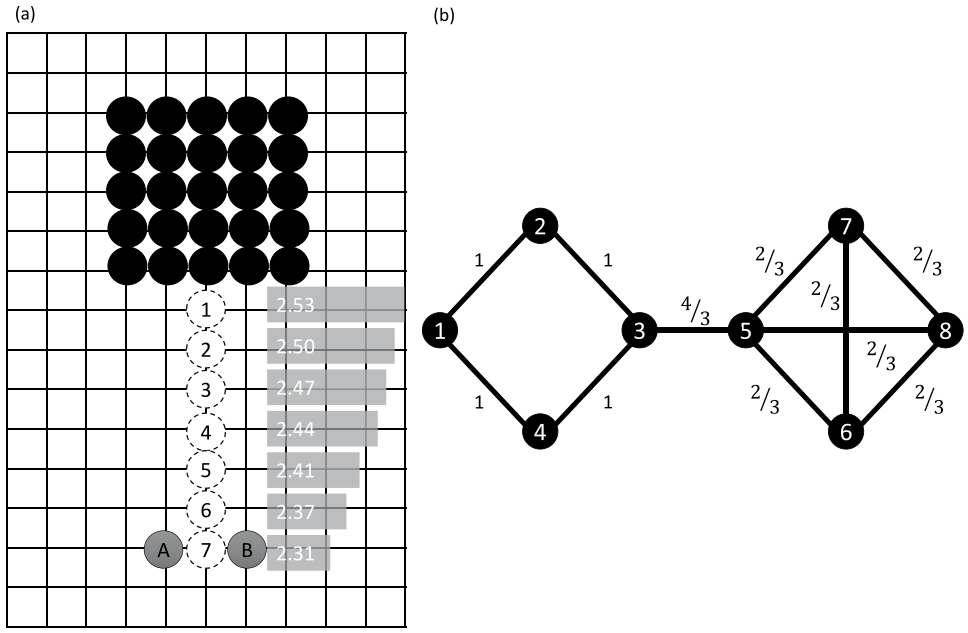

Figure 3. (**a**) A 15×11 grid with set S consisting of the 5×5 sub-grid at the top (in black) as well as the two nodes A and B. The community-relative distance between A and B is 2.55. The updated distance when a single node is added at one of the locations denoted 1 through 7, is given to the right of the respective location (see Example 1). Note that as the extra node approaches A and B, the two nodes become more proximate, in a sense shrinking the space between A and B, as they become part of a stronger community. (**b**) A simple 8-node graph. Community-relative distances for nodes at shortest-path distance one are indicated adjacent to the corresponding edge. Note that the tie between node 3 and node 5 is one of greater community-relative distance, suggesting separation, while the ties in the clique consisting of nodes 5–8 correspond to smaller community-relative distance (see Example 2).

Community-relative distances could have potential use in missing link or attribute prediction, wherein small distances between unconnected nodes could suggest potential edges, and a large cluster of mixed attribute nodes could suggest missing attributes.

Employing $\boldsymbol{D}^*$, it is possible to cluster the elements of S via standard extant procedures. Unless specified otherwise, the results in what follows here arise from employment of average-linkage hierarchical clustering on $\boldsymbol{D}^*$, i.e. sequentially combining two clusters with the lowest average distance between pairs (see for instance Lazega *et al.*[30] and Newman[31] for discussion of hierarchical clustering). As mentioned in Related work below (see also Figs 3b, S2 and S3 and Applications and Discussion), community-relative distances reflect separation between clusters, and provide robust results under different clustering procedures (see Fig. S3). It can be worthwhile to look at the resulting dendrograms for overall clustering patterns and, if desired, natural locations to "cut" S into clusters (see Fig. 2b,d). There are several methods available for searching for appropriate dendrogram cut-points (see for instance[32]). For comparison purposes, below, to estimate reasonable stopping conditions, we employ an average silhouette width criterion (ASW; see[33]) as well as the variance ratio criterion (VR) of Caliński and Harabasz[34] (see Related Work as well as Figs S4 and S5); similar results are obtained using other common extant methods. Also available are non-hierarchical methods such as partitioning around medoids (see[35]).

Importantly, note that in the case where $S = V$, i.e. all (say n) nodes are selected, the community-relative distance from node i to node j, described above, reduces to simply the average of the shortest-path distances from the direct neighbors of i to j. In fact, computation reduces simply to

$$\boldsymbol{D} = \boldsymbol{P}\boldsymbol{D}_s - \boldsymbol{I}, \tag{1}$$

where $\boldsymbol{P}$ is the transition matrix for a random walk on the graph G, $\boldsymbol{D}_s$ is the matrix of shortest path distances and $\boldsymbol{I}$ is the identity matrix (of size n). For further details and discussion of computational complexity in the general case, see Materials and Methods, below.

Related Work

Closest to the work presented here, specifically in the limiting case of community detection, is the popular *Walktrap* method of Pons and Latapy[24]. Therein, random walks are also employed to obtain distances which can then be used in agglomerative hierarchical procedures. In particular, therein, the distance, $r_{i,j}$ between nodes i and j is defined for fixed $t \in \{1, 2, \ldots\}$ via

$$r_{i,j}(t) = \left|\left|\boldsymbol{\Delta}^{-1/2}\boldsymbol{P}^t_{i,\cdot} - \boldsymbol{\Delta}^{-1/2}\boldsymbol{P}^t_{j,\cdot}\right|\right|, \tag{2}$$

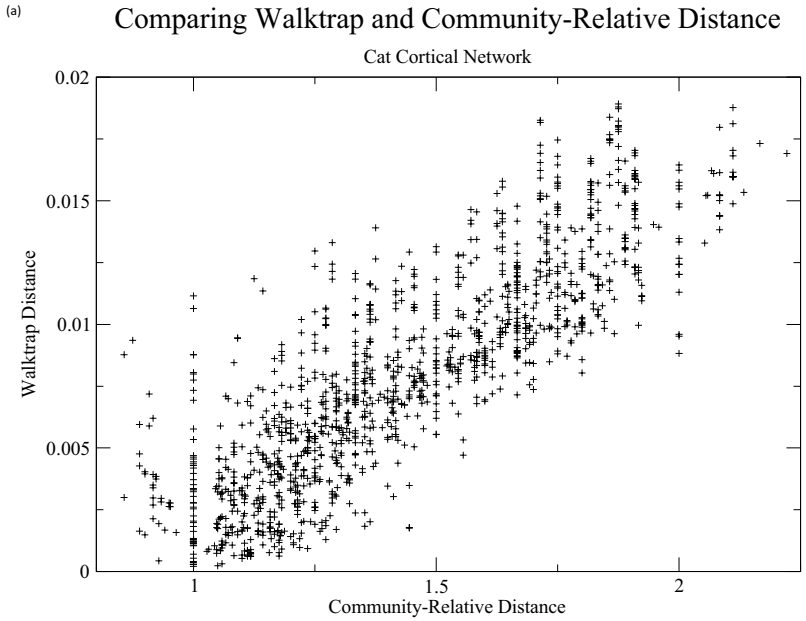

Figure 4. A plot of Walktrap (t = 4) distances against community-relative distances for the cat cortical network.

where $\boldsymbol{\Delta}$ is a diagonal matrix with diagonal entries $\Delta_{i,i} = d(i)$, $d(i)$ is the degree of v_i, $\boldsymbol{P}^t_{l,\cdot}$ is the column probability vector $(P^t_{l,k})_{1\leq k\leq n}$, $\boldsymbol{P} = [P_{i,j}]$ is the transition matrix for a random walk on the graph G, and $|\cdot|$ indicates the Euclidean norm on $\mathbb{R}^n$. A plot of these distances against community-relative distances for a cat cortical network (see[36] and Applications and Discussion, below) is given in Fig. 4. Note that the ordering of distances is quite different in the two cases. In terms of community detection, community-relative distance does have advantages: (i) there is no need to choose an appropriate parameter t. The Walktrap method can be sensitive to values of t, as well as the choice of agglomerative method (compare Figs S2 and S3). (ii) Community-relative distances are particularly simple and parsimonious (See Eq. 1), while computational times are similar for the two methods, (iii) units of resulting distances are easily interpretable in terms of shortest path distance and (iv) most importantly, there is no immediate counterpart to clustering restricted to subsets in the case of the Walktrap algorithm.

Figure 5a contains adjusted Rand index (ARI; see[37]), and normalized mutual information (NMI; see[38]) values for agglomerative clustering (employing average-linkage and a VR stopping condition) for some common networks possessing reasonable ground truths, via a range of common distance measures; for discussion of Jaccard and cosine similarity measures, see for instance[39] and the references therein. Note that community-relative distance performs comparably or considerably better for the six common networks considered. For similar results employing ASW, see Fig. S4. The networks are discussed further in Applications and Discussion, below.

Figure 5b contains ARI and NMI values for the six network data sets, employing nine methods built into the *igraph* package in R (see[40]), alongside those for community-relative distance using both ASW and VR stopping conditions. Again community-relative distance performs comparably or considerably better for the networks considered. Plots and dendrograms under community-relative distance are provided in Applications and Discussion, below; for plots of associated ASW and VR values, see Fig. S5. For general discussion regarding comparing clusterings see for instance[41]. For other work related to community detection and random walks see[25] and[26] and the references therein.

In terms of restriction to nodal subsets, there has been considerable work recently in the special case of types within bipartite networks (see[42–51]). For discussion of community-relative distance in this context see Applications and Discussion, below. It is important to note that, contrasted with methods specific to bipartite networks, the perspective proposed here imposes no assumptions on the edge structure of the network considered, nor the sets under consideration for clustering.

For some recent work on attributes in the context of clustering, see[52]. Although different in scope, it is worth noting connected work on clustering in spatial networks (see for instance[53]). Community-relative distance is applicable for arbitrary (potentially non-spatial) networks, and may be of some potential future use in existing algorithms for spatial networks, in place of often considered geodesic distance. In addition, there has been important recent work employing stochastic complementation[54] in the context of restriction to subsets of network nodes (see[55] and [[28], Section 10.4.5]).

Applications and Discussion

In this section, we consider community-relative distance applied to several data sets, first in the context of proper nodal subsets, S, of interest, and finally in the context of community detection.

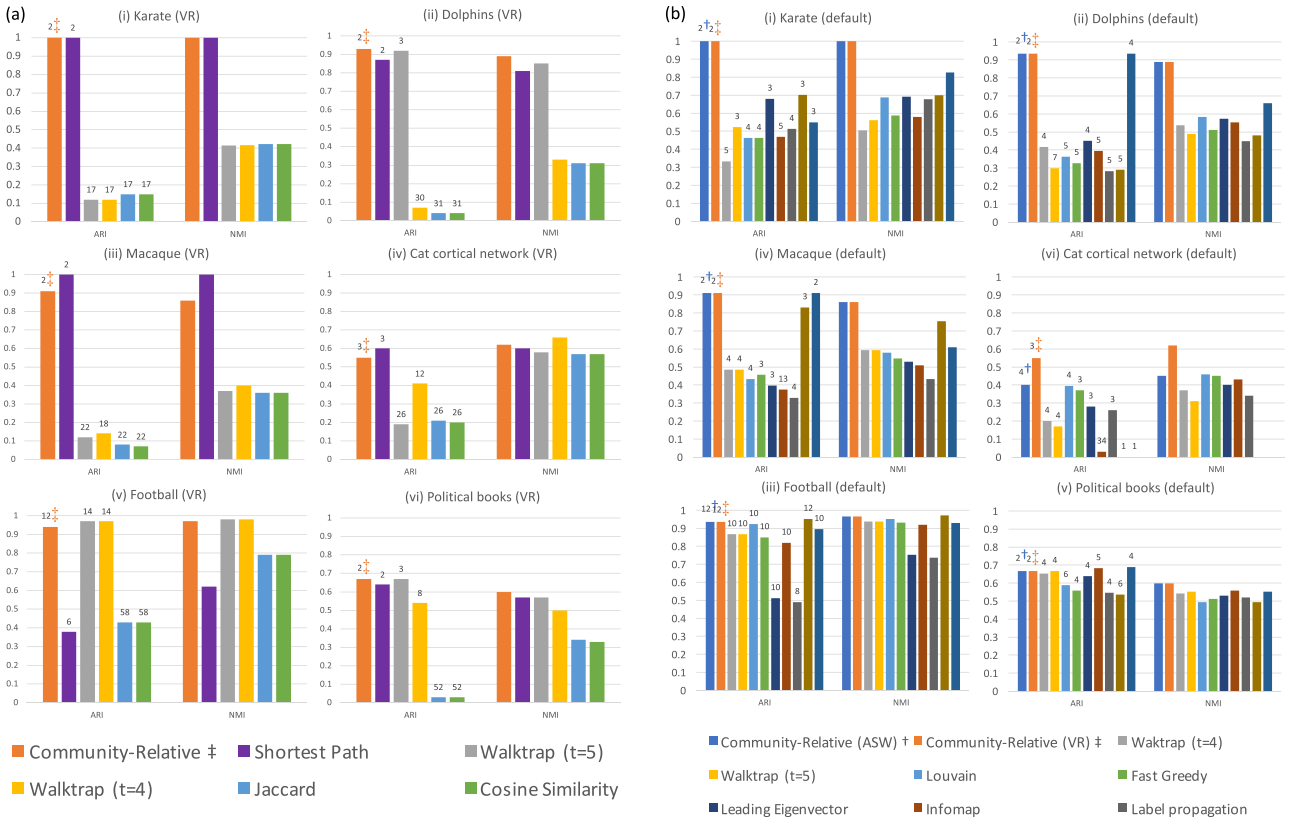

Figure 5. (**a**) ARI and NMI values for agglomerative clustering (employing average-linkage and a VR stopping condition) for some common networks possessing reasonable ground truths, via a range of common distance measures. The networks are discussed further in Applications and Discussion, below. For discussion of Jaccard and cosine similarity measures, see for instance[39] and the references therein. (**b**) ARI and NMI values for the six network data sets, employing nine methods built into the *igraph* package in R (see[40]), alongside those for community-relative distance using both ASW and VR stopping conditions.

Nodal subsets. In Fig. 6a–f, we consider a macaque cortical network[56]. Employing community-relative distances and average-linkage clustering on the subset consisting of the cortical areas within the visual cortex, we obtain a fairly clear partition into two clusters as indicated in Fig. 6a,b. A histogram displaying community-relative distances is given in Fig. 6c (see also Fig. S13). For comparison, there are only three distinct shortest path distances: 1, 2, and 3; a tabulation of these is given in Fig. 6d (see also Fig. S14). If agglomerative clustering were to be implemented given the shortest path distances, final results could depend heavily on the choices when dealing with tied distances (see Fig. 6e,f). In Fig. S6 we provide a two-clustering for each of the two factions which arose for the karate club at a large state university studied by Zachary[57]. The corresponding $\boldsymbol{D}^*$ matrices are given in Figs S7 and S8, respectively.

Since it is possible to consider any subset S contained in V, it is feasible to consider nodes of a particular type in a multipartite network. Davis *et al.*[58] studied a group of 18 women and their observed participation in social events. Here we obtain the dendrogram in Fig. 6g,h. The suggested structural split into clusters (via ASW or VR) matches well with those in the meta-analyses of 21 studies as presented in[59], Fig. 7; the match is exact with two of the methods considered therein. Similarly, consider the bipartite network of US Supreme court justice decisions for the 2000–2001 term[60], depicted in Fig. S10. Here edges are drawn from each of the nine justices to any of 24 important cases for which they voted in the minority (two of the 26 cases from the original data had unanimous decisions). It is possible to consider the justices and cases, separately, by taking S as the set of justices, or the set of cases, respectively. Clusterings of the justices into 4 groups, and cases into 7 groups match exactly those as suggested in[60], Fig. 1; see Fig. S10. For other discussion of analyzing community structures in two-mode (bipartite) networks, see for instance[61] and the references, therein. As mentioned earlier, contrasted with extant methods specific to bipartite networks, the perspective proposed here imposes no assumptions on the edge structure of the network considered, nor the sets under consideration for clustering.

For an additional example, in Fig. 7 we consider three disease subsets from the human disease network[62] consisting of disorders and the disease genes whose mutations are associated with the disorders. Histograms for community-relative distances for cancer, neurological and skeletal diseases are given in Fig. 7b–d, respectively. Note the distinct differences in the distributions of community-relative distances for the three disease node subsets. Cancer nodes are more closely positioned within the network; whereas, neurological and skeletal disease nodes are more diffusely positioned. Analyses informed by community-relative distance may aide in uncovering key cellular pathway components that lead to disease. A network plot and dendrograms for the three disorder classes are given in Figs S11 and S12.

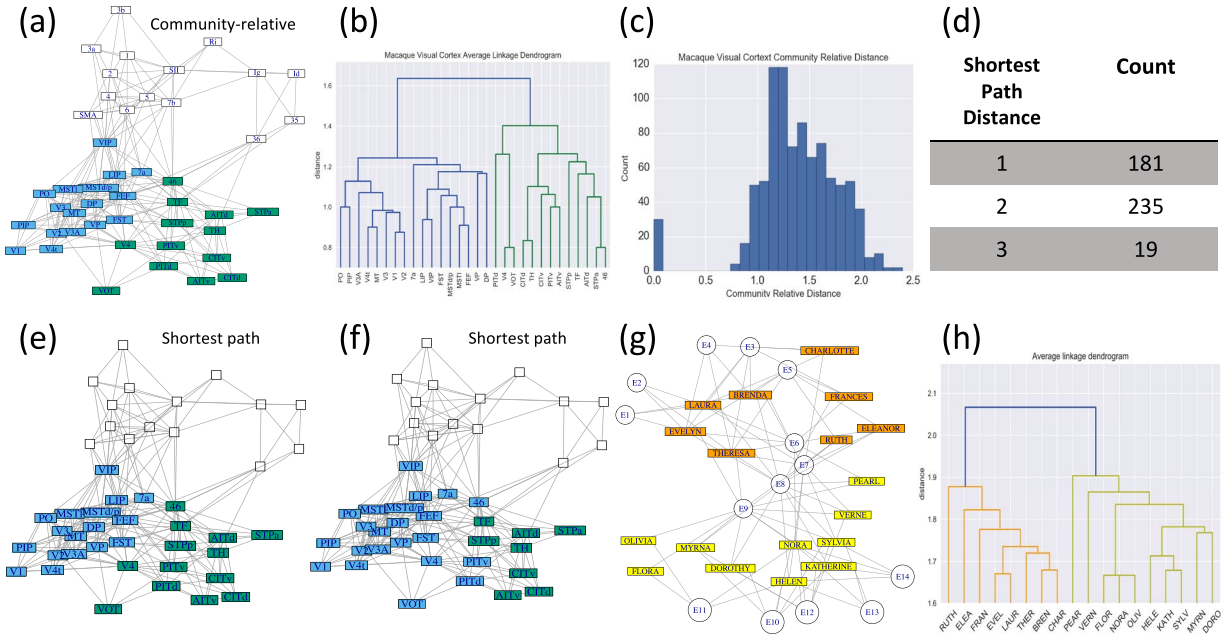

Figure 6. (**a**,**b**) An application of community-relative distance to the 30-node visual cortex subset within the 45-node cortical pathways network of the macaque monkey (see[56]). Note that we employ the force-directed layout algorithm of Fruchterman and Reingold[78], throughout for network plots. (**c**,**d**) A histogram of community-relative distances is given in (**c**). The 435 distinct pairs of subset nodes are comprised of 181 at (shortest-path) distance one, 235 at distance two and 19 at distance three (see (**d**)). (**e**,**f**) A cut into two clusters using average linkage hierarchical clustering and shortest-path distances for the 30-node visual cortex, for two permutations of the vertex order. Note the sensitivity to vertex order. (**g**,**h**) Two-clustering (via community relative distances and average-linkage clustering) for the nodes representing the 18 women in the bipartite network of Davis *et al.*[58].

Community detection. Figure 8a–l contains plots and dendrograms for the networks considered earlier from a numerical perspective in Related Work; for full numeric comparisons with other methods, see Fig. 5. For the case of the karate network (Fig. 8a,b), we obtain a two-clustering which captures the factions suggested in[57] (ARI and NMI values of 1). As noted therein, Individual 9 was "a structural part"[57] of the group assigned to in Fig. 8a; however, following fission of the original club, this individual did join the other group (due to some personal motivation). The corresponding full $\boldsymbol{D}^*$ matrix of community relative distances is given in Fig. S9.

In the case of the network produced by Lusseau from following a pod of dolphins in Doubtfull Sound, off the coast of New Zealand[63] (Fig. 8c,d), we find results which nearly match the factions encountered in[64] (2 nodes misclassified; an ARI value of 0.93; and NMI value of 0.89). In Fig. 8e,f, we consider the macaque brain network considered earlier. Here, we obtain a clear separation into two clusters which reflects membership in either the visual or sensorimotor cortices (1 node misclassified; ARI value of 0.91; and NMI value of 0.86). As a further example of community detection on a highly connected graph, we consider the cat cortical network discussed by Scannell *et al.* (see Fig. 8g,h)[36]. We find that the results obtained via community-relative distance is in strong agreement with the standard classification into four major thalamocortical systems[36] (5 nodes misclassified; ARI value of 0.55; NMI value of 0.62).

Finally, for the case of community detection, we provide two additional examples. The first is a network with (American) collegiate football teams as nodes, and edges representing games played against one another[65]. Community-relative distance and average-linkage clustering (see Fig. 8i,j) quite clearly split these teams into the underlying conferences with accuracy (4 nodes misclassified – all independent teams; NMI value of 0.97; ARI value of 0.94). Some novel characteristics of the dendrogram may be noted, including the fact that many teams which later joined the Atlantic Coast Conference (ACC) are situated close to the ACC teams in the dendrogram. We also obtain similarly appropriate results with a network of politically themed books with edges connecting books commonly purchased together on amazon.com[66]. Here community-relative distance and average linkage clustering split the books quite well into groups of political affiliation, as identified by Newman[67] (4 conservative or liberal nodes misclassified; ARI value of 0.67; and NMI value of 0.60).

Remark 2. Note for the latter two examples in the section, results can be improved, if we restrict the set S to only nodes of interest. In the case of the football network, if we exclude consideration of independent teams (without conference membership), we obtain a perfect eleven-clustering into conferences (ARI and NMI values of 1). Similarly for the political blogs network, restricting S to the set of non-neutral books leads to only 3 books being misclassified. □

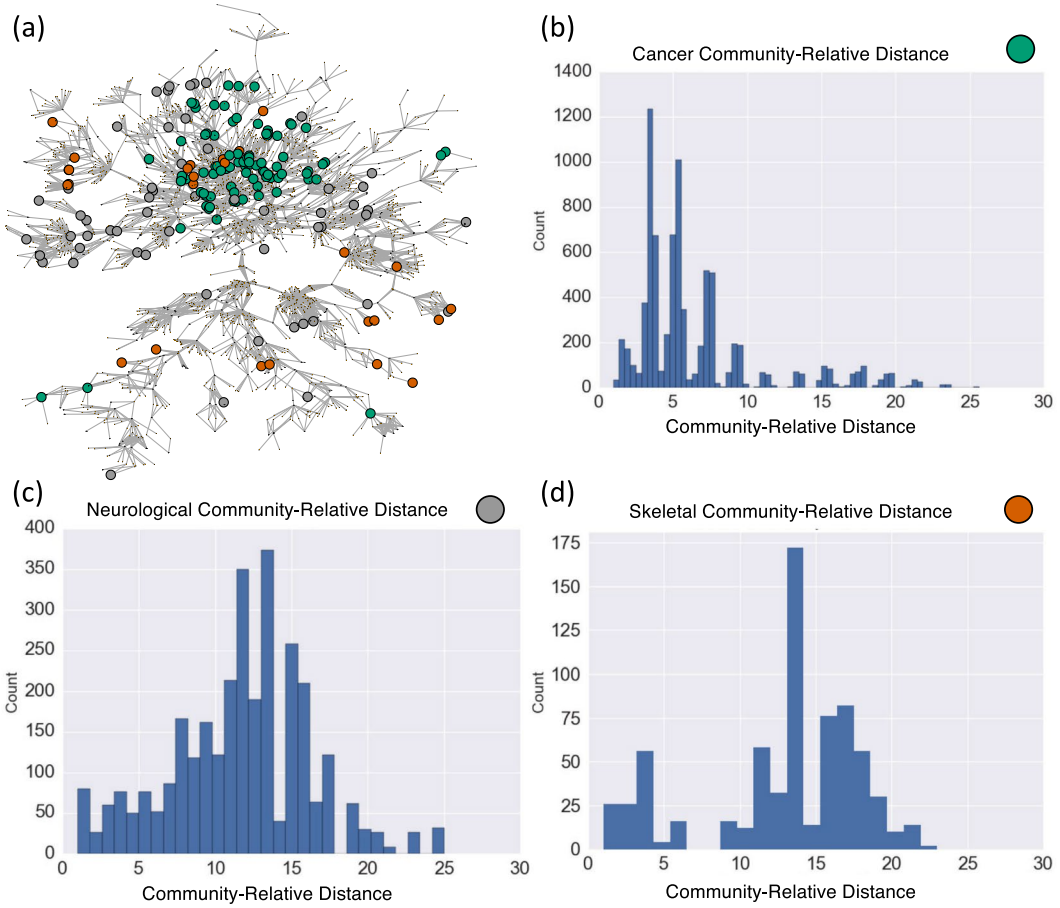

Figure 7. (**a**) The human disease network[62]. Here nodes corresponding to cancer, neurological and skeletal diseases are highlighted in green, grey and brown, respectively. (**b**–**d**) Histograms for the corresponding community-relative distances. For additional considerations for the human disease network, see Figs S11 and S12.

Materials and Methods

In this section we address computing of the (distance) entries in the matrix $D = [d_{i,j}]$. Code in the R programming language is available upon request.

Suppose $G = (V, E)$ and $S \subseteq V$ are fixed. Let $\boldsymbol{A} = [A_{i,j}]$ be the adjacency matrix for G, i.e. $A_{i,j} = 1$ if $(v_i, v_j) \in E$ and zero otherwise, $\boldsymbol{I}$ be the $n \times n$ identity matrix, and $\boldsymbol{\Delta}$ be a diagonal matrix with diagonal entries $\Delta_{i,i} = d(i)$, where $d(i)$ is the degree of v_i.

For a general subset $S = \{s_1, s_2, \ldots, s_m\} \subseteq V$, the matrix $\boldsymbol{D}$ can be obtained in the following manner. Define the matrix $\boldsymbol{L} = [L_{i,j}]$ via

$$L_{i,j} = \begin{cases} 1 & \text{if } i = j \\ -1/d(i) & \text{if } v_i \notin S \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Note that $\boldsymbol{L}$ is similar to the random-walk normalized Laplacian matrix, $\boldsymbol{L}^* = \boldsymbol{I} - \boldsymbol{\Delta}^{-1}\boldsymbol{A}$ except that if $v_i \in S$, then the i-th row of $\boldsymbol{L}^*$ is replaced with $e_i = (0, 0, \ldots, 0, 1, 0 \ldots, 0)$, i.e. the i-th row of the $n \times n$ identity matrix. Now, set

$$\hat{\boldsymbol{D}} = \boldsymbol{P}\boldsymbol{L}^{-1}\widetilde{\boldsymbol{D}}_S, \tag{4}$$

where $\widetilde{\boldsymbol{D}}_S$ is similar to the matrix of shortest path distances $\boldsymbol{D}_s$, except that if $v_i \notin S$, then the i-th row of $\boldsymbol{D}_s$ is replaced with $0 = (0, 0, \ldots, 0 \ldots, 0)$, i.e. a null n-vector of zeros. The (i, j)-entry in $\hat{\boldsymbol{D}}$ is then the community-relative distance from node i to node j relative to the set S.

As suggested in Eq. (4), the process of computing community-relative distances requires (i) all pairs shortest-path distances between nodes in S, (ii) a solution to $\boldsymbol{L}\boldsymbol{X} = \widetilde{\boldsymbol{D}}_S$, and (iii) computation of the product $\hat{\boldsymbol{D}} = \boldsymbol{P}\boldsymbol{X}$. Note that for (i), the full matrix of shortest-path distances (or an approximation, see for instance[68–70]) are often available, as these arise in standard preliminary network analyses (and elsewhere), even for relatively large networks. When this is not the case, some savings may be possible since only intra-set distances for S are

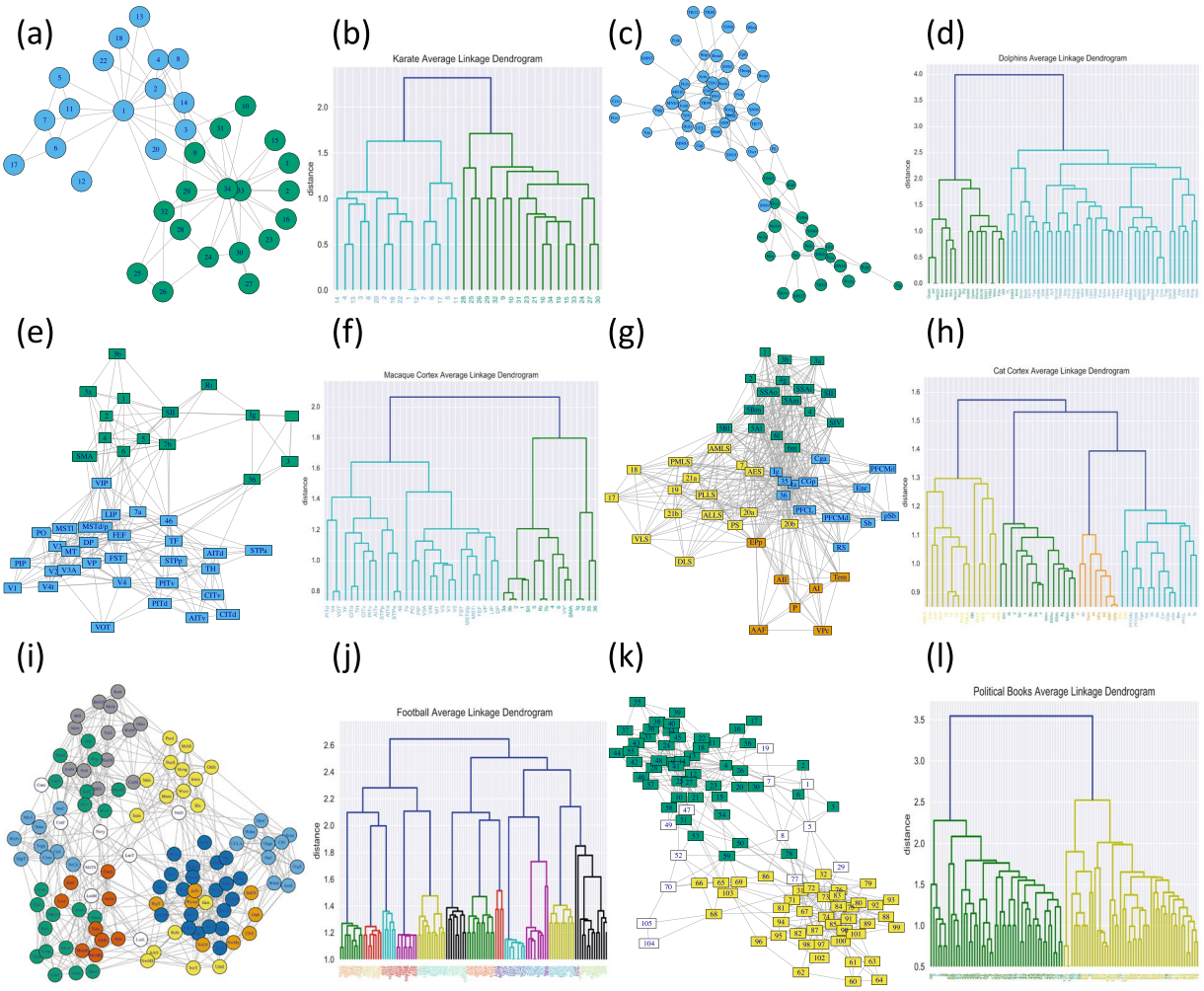

Figure 8. Graph plots and dendrograms (employing community-relative distance and average-linkage hierarchical clustering) for the six full networks considered in Fig. 5. (a,b) A two-clustering for the karate network of Zachary[57]. (**c**,**d**) A two-clustering clustering for the dolphin social network[63]. (**e**,**f**) A two-clustering for a 45-node cortical pathways network of the macaque monkey[56]. (**g**,**h**) A four-clustering for the cat cortical network[36]. (**i**,**j**) A twelve-clustering for the nodes representing the 2000–2001 NCAA football teams in the network of Girvan and Newman[65]. (**k**,**l**) A two-clustering for the political books network[66]. The nodes are colored to reflect apparent political affiliation (white for neutral, green for conservative and yellow for liberal), as suggested in[67]. Black is used in the dendrogram labels, in the neutral case, in place of white.

required. For (ii), $\boldsymbol{L}$ can be viewed as the normalized random walk Laplacian for a directed variant of the graph G, wherein outgoing edges from nodes in S have been removed. Here, recently developed Laplacian solvers (see[71]) may be employed and computation can then be sub-quadratic in n (at least for sparse graphs). As $|S|$ increases, the matrix $\boldsymbol{L}$ becomes increasingly sparse, and in the extreme case where $S = V$, we have that $\boldsymbol{L}$ is simply the identity matrix $\boldsymbol{I}$. Since, within each column of $\boldsymbol{X}$, one needs only solve for $|S^c|$ entries, computations can be reduced to

$$\widetilde{O}((|E_{cc}|^{3/4}|S^c| + |E_{cc}||S^c|^{2/3})|S|), \tag{5}$$

where E_{cc} denotes the set of within-S^c edges and the $\widetilde{O}$ notation suppresses polylogarithmic factors[7]. For the multiplication in (iii), note that to obtain the $|S| \times |S|$ matrix of within-S community-relative distances, we may consider the sparse multiplication of a $|S| \times n$ matrix $\widetilde{\boldsymbol{P}}$ and an $n \times |S|$ matrix $\widetilde{\boldsymbol{X}}$, where $\widetilde{\boldsymbol{P}}$ consists of the $|S|$ rows of $\boldsymbol{P}$ corresponding to the elements in S, and $\widetilde{\boldsymbol{X}}$ consists of the $|S|$ columns of $\boldsymbol{X}$ corresponding to the elements in S. Note that $\widetilde{\boldsymbol{P}}$ contains $|E_{S.}|$ non-zero entries, where $E_{S.}$ is the set of edges outgoing from S, and hence the number of operations is of order

$$O(|E_{S.}||S|). \tag{6}$$

As mentioned earlier, community-relative distances provide expanded separation between clusters. We have employed average-linkage hierarchichal clustering, here, which has complexity $O(|S|^2)$ (see[72,73]), in an effort to show that even naive clustering procedures can work well. One may chose to employ $\boldsymbol{D}$ in other proximity-based

methods, as appropriate in applications. For a discussion of exact and approximation methods, with savings in both time and space complexity, see[74,75].

In the case $S = V$, as mentioned earlier, the matrix $\boldsymbol{D}$ has a simple form given via

$$\boldsymbol{D} = \boldsymbol{P}\boldsymbol{D}_s - \boldsymbol{I}. \tag{7}$$

Code Availability

The computations here were performed using the R programming language; a documented package which employs optimized routines in C++, is available upon request.

References

1. Christakis, N. A. & Fowler, J. H. The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine* **358**, 2249–2258, https://doi.org/10.1056/NEJMsa0706154 (2008).
2. Rosenquist, J. N., Murabito, J., Fowler, J. H. & Christakis, N. A. The Spread of Alcohol Consumption Behavior in a Large Social Network. *Annals of Internal Medicine* **152**, 426–433 (2010).
3. Fowler, J. H. & Christakis, N. A. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ* **337**, a2338–a2338, https://doi.org/10.1136/bmj.a2338 (2008).
4. Hill, A. L., Rand, D. G., Nowak, M. A. & Christakis, N. A. Emotions as infectious diseases in a large social network: the SISa model. *Proceedings of the Royal Society B: Biological Sciences* **277**, 3827–3835, https://doi.org/10.1098/rspb.2010.1217 (2010).
5. McDermott, R., Fowler, J. H. & Christakis, N. A. Breaking Up Is Hard to Do, Unless Everyone Else Is Doing It Too: Social Network Effects on Divorce in a Longitudinal Sample. *Social Forces* **92**, 491–519, https://doi.org/10.1093/sf/sot096 (2013).
6. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. & Christakis, N. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks* **30**, 330–342, https://linkinghub.elsevier.com/retrieve/pii/S0378873308000385 (2008).
7. Lewis, K., Gonzalez, M. & Kaufman, J. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences* **109**, 68–72 (2012).
8. Green, B., Horel, T. & Papachristos, A. V. Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014. *JAMA Internal Medicine* **177**, 326, https://doi.org/10.1001/jamainternmed.2016.8245 (2017).
9. Christakis, N. A. & Fowler, J. H. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine* **357**, 370–379, https://doi.org/10.1056/NEJMsa066082 (2007).
10. de la Haye, K., Robins, G., Mohr, P. & Wilson, C. Obesity-related behaviors in adolescent friendship networks. *Social Networks* **32**, 161–167, https://doi.org/10.1016/j.socnet.2009.09.001, https://linkinghub.elsevier.com/retrieve/pii/S0378873309000495 (2010).
11. de la Haye, K., Robins, G., Mohr, P. & Wilson, C. How physical activity shapes, and is shaped by, adolescent friendships. *Social Science & Medicine* **73**, 719–728, https://doi.org/10.1016/j.socscimed.2011.06.023, https://linkinghub.elsevier.com/retrieve/pii/S0277953611003790 (2011).
12. Shakya, H. B., Christakis, N. A. & Fowler, J. H. Social network predictors of latrine ownership. *Social Science and Medicine* **125**, 129–138, https://doi.org/10.1016/j.socscimed.2014.03.009 (2015).
13. Shakya, H. B. *et al*. Intimate partner violence norms cluster within households: an observational social network study in rural Honduras. *BMC public health* **16**, 233, https://www.biomedcentral.com/1471-2458/16/233 (2016).
14. Latkin, C. *et al*. Relationships between social norms, social network characteristics, and HIV risk behaviors in Thailand and the United States. *Health Psychology* **28**, 323–329, https://doi.org/10.1037/a0014707 (2009).
15. Hruschka, D. J., Brewis, A. A., Wutich, A. & Morin, B. Shared Norms and Their Explanation for the Social Clustering of Obesity. *American Journal of Public Health* **101**, S295–S300, https://doi.org/10.2105/AJPH.2010.300053 (2011).
16. Porter, M. A., Onnela, J.-P. & Mucha, P. J. Communities in networks. *Notices of the AMS* **56**, 1082–1097 (2009).
17. Newman, M. E. Communities, modules and large-scale structure in networks. *Nature Physics* **8**, 25–31 (2012).
18. Schaeffer, S. E. Graph clustering. *Computer Science Review* **1**, 27–64 (2007).
19. Fortunato, S. Community detection in graphs. *Physics reports* **486**, 75–174 (2010).
20. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics Reports* **659**, 1–44, https://doi.org/10.1016/j.physrep.2016.09.002 (2016).
21. Christakis, N. A. & Fowler, J. H. Social network visualization in epidemiology. *Norsk Epidemiologi* **19**, 5–16 (2009).
22. Lovász, L. *et al*. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty* **2**, 353–398 (1996).
23. Aldous, D. & Fill, J. Reversible Markov Chains and Random Walks on Graphs, 2014, http://www.stat.berkeley.edu/~aldous/RWG/book.html.
24. Pons, P. & Latapy, M. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, 284–293 (Springer, 2005).
25. Zhou, H. & Lipowsky, R. Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *Computational Science-ICCS 2004*, 1062–1069 (Springer, 2004).
26. Zhou, H. Distance, dissimilarity index, and network community structure. *Physical Review E* **67**, 061901, https://doi.org/10.1103/PhysRevE.67.061901 (2003).
27. Yen, L., Saerens, M., Mantrach, A. & Shimbo, M. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 785–793 (ACM, 2008).
28. Fouss, F., Saerens, M. & Shimbo, M. *Algorithms and models for network data and link analysis* (Cambridge University Press, 2016).
29. Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology* **78**, 1360–1380, https://doi.org/10.1086/225469 (1973).
30. Stanley, W. & Faust, K. *Social network analysis: Methods and applications*. Cambridge University Press (1994).
31. Newman, M. *Networks: an introduction* (OUP Oxford, 2010).
32. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* **24**, 719–720 (2008).
33. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).
34. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* **3**, 1–27 (1974).
35. Kaufman, L. & Rousseeuw, P. J. *Partitioning Around Medoids (Program PAM)*, 68–125, https://doi.org/10.1002/9780470316801.ch2 (John Wiley & Sons, Inc., 2008).
36. Scannell, J., Burns, G., Hilgetag, C., O'Neil, M. & Young, M. P. The connectional organization of the cortico-thalamic system of the cat. *Cerebral Cortex* **9**, 277–299 (1999).
37. Hubert, L. & Arabie, P. Comparing partitions. *Journal of classification* **2**, 193–218 (1985).
38. Fred, A. L. N. & Jain, A. K. Robust data clustering. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003 Proceedings* **2**, II–128–II–133, https://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1211462 (2003).

39. Hamers, L. *et al*. Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Information Processing & Management* **25**, 315–318 (1989).
40. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* Complex Systems, 1695, https://igraph.org (2006).
41. Wagner, S. & Wagner, D. *Comparing clusterings: an overview. Tech. Rep.* (Karlsruhe: Universität Karlsruhe, Fakultät für Informatik, 2007).
42. Sun, H.-l. *et al*. A fast community detection method in bipartite networks by distance dynamics. *Phys. A: Stat. Mech. its Appl*. **496**, 108–120, https://doi.org/10.1016/j.physa.2017.12.099, http://linkinghub.elsevier.com/retrieve/pii/S0378437117313481 (2018).
43. Han, X. *et al*. Emergence of communities and diversity in social networks. *Proc. Natl. Acad. Sci.* **114**, 2887–2891, https://doi.org/10.1073/pnas.1608164114 (2017).
44. Li, K. & Pang, Y. A unified community detection algorithm in complex network. *Neurocomputing* **130**, 36–43, http://linkinghub.elsevier.com/retrieve/pii/S0925231213007479, https://doi.org/10.1016/j.neucom.2012.11.054 (2014).
45. Liu, J.-G., Hou, L., Pan, X., Guo, Q. & Zhou, T. Stability of similarity measurements for bipartite networks. *Sci. Reports* **6**, 18653, http://www.nature.com/articles/srep18653, https://doi.org/10.1038/srep18653 1512.01432 (2016).
46. Wang, X. & Qin, X. Asymmetric intimacy and algorithm for detecting communities in bipartite networks. *Phys. A: Stat. Mech. its Appl.* **462**, 569–578, http://linkinghub.elsevier.com/retrieve/pii/S0378437116303715, https://doi.org/10.1016/j.physa.2016.06.096 (2016).
47. Xu, Y., Chen, L., Li, B. & Liu, W. Density-based modularity for evaluating community structure in bipartite networks. *Inf. Sci.* **317**, 278–294, http://linkinghub.elsevier.com/retrieve/pii/S0020025515003412, https://doi.org/10.1016/j.ins.2015.04.049 (2015).
48. Larremore, D. B., Clauset, A. & Jacobs, A. Z. Efficiently inferring community structure in bipartite networks. *Phys. Rev. E* **90**, 012805, https://link.aps.org/doi/10.1103/PhysRevE.90.012805, https://doi.org/10.1103/PhysRevE.90.012805 1403.2933 (2014).
49. Cui, Y. & Wang, X. Uncovering overlapping community structures by the key bi-community and intimate degree in bipartite networks. *Phys. A: Stat. Mech. its Appl.* **407**, 7–14, http://linkinghub.elsevier.com/retrieve/pii/S037843711400288X, https://doi.org/10.1016/j.physa.2014.03.077 (2014).
50. Guimerà, R., Sales-Pardo, M. & Amaral, L. A. N. Module identification in bipartite and directed networks. *Phys. Rev. E* **76**, 036102, https://doi.org/10.1103/Phys-RevE.76.036102 0701151 (2007).
51. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74**, 036104, https://doi.org/10.1103/Phys-RevE.74.036104 0605087 (2006).
52. Newman, M. E. & Clauset, A. Structure and inference in annotated networks. *Nat. Commun.* **7**, 1–11, https://doi.org/10.1038/ncomms11863 1507.04001 (2016).
53. Okabe, A. & Sugihara, K. *Spatial analysis along networks: statistical and computational methods* (John Wiley & Sons, 2012).
54. Meyer, C. D. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM review* **31**, 240–272 (1989).
55. Yen, L., Saerens, M. & Fouss, F. A link analysis extension of correspondence analysis for mining relational databases. *IEEE Transactions on Knowledge and Data Engineering* **23**, 481–495 (2011).
56. Négyessy, L., Nepusz, T., Kocsis, L. & Bazsó, F. Prediction of the main cortical areas and connections involved in the tactile function of the visual cortex by network analysis. *European Journal of Neuroscience* **23**, 1919–1930, Data Accessed: 2016-07-1, https://github.com/igraph/igraphdata (2006).
57. Zachary, W. W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 452–473, Data Accessed: 2016-07-1, https://vlado.fmf.uni-lj.si/pub/networks/data./WaFa/default.htm (1977).
58. Davis, A., Gardner, B. B., Gardner, M. R. & Warner, W. L. *Deep South: A Sociological Anthropological Study of Caste and Class*, Data Accessed: 2016-07-1, https://networkdata.ics.uci.edu/netdata/html/davis.html (University of Chicago Press, 1941).
59. Freeman, L. C. *Finding social groups: A meta-analysis of the southern women data*. (The National Academies Press, Washington, DC, 2003).
60. Doreian, P., Batagelj, V. & Ferligoj, A. Generalized blockmodeling of two-mode network data. *Social Networks* **26**, 29–53 (2004).
61. Barber, M. J. Modularity and community detection in bipartite networks. *Physical Review E* **76**, 066102 (2007).
62. Goh, K.-I. *et al*. The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685–8690, http://www.pnas.org/cgi/content/abstract/104/21/8685, https://doi.org/10.1073/pnas.0701361104 (2007).
63. Lusseau, D. *et al*. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology* **54**, 396–405 (2003).
64. Lusseau, D. & Newman, M. E. J. Identifying the role that animals play in their social networks. *Proceedings of the Royal Society B: Biological Sciences* **271**, S477–S481, https://doi.org/10.1098/rsbl.2004.0225, 0112110v1 (2004).
65. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**, 7821–7826, https://doi.org/10.1073/pnas.122653799, 0112110v1 (2002).
66. Krebs, V. Books about US politics, https://networkdata.ics.uci.edu/data.php and https://www.orgnet.com (2004).
67. Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**, 8577–8582 (2006).
68. Akiba, T., Iwata, Y. & Yoshida, Y. Fast exact shortest-path distance queries on large networks by pruned landmark labeling. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 349–360 (ACM, 2013).
69. Gubichev, A., Bedathur, S., Seufert, S. & Weikum, G. Fast and accurate estimation of shortest paths in large graphs. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 499–508 (ACM, 2010).
70. Roditty, L. & Zwick, U. Dynamic approximate all-pairs shortest paths in undirected graphs. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, 499–508 (IEEE, 2004).
71. Cohen, M. B. *et al*. Faster algorithms for computing the stationary distribution, simulating random walks, and more. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, 583–592 (IEEE, 2016).
72. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview, ii. Wiley Interdiscip. *Rev. Data Min. Knowl. Discov.* **7** (2017).
73. Murtagh, F. Complexities of hierarchic clustering algorithms: state of the art. *Comput. Stat. Q.* **1**, 101–113 (1984).
74. Eppstein, D. Fast hierarchical clustering and other applications of dynamic closest pairs. *J. Exp. Algorithmics (JEA)* **5**, 1 (2000).
75. Cardinal, J. & Eppstein, D. Lazy algorithms for dynamic closest pair with arbitary distance measures. In ALENEX/ANALC, 112–119 (2004).
76. Harris, K. M. *et al*. The national longitudinal study of adolescent health: Research design. *Available at* https://www.cpc.unc.edu/projects/addhealth/design (2009).
77. Fowler, J. H., Settle, J. E. & Christakis, N. A. Correlated genotypes in friendship networks. *Proceedings of the National Academy of Sciences* **108**, 1993–1997, https://doi.org/10.1073/pnas.1011687108 (2011).
78. Fruchterman, T. M. & Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience* **21**, 1129–1164, https://doi.org/10.1002/spe.4380211102 (1991).

Acknowledgements

This work was partially supported by National Institutes of Health grant T32-GM095440, supporting R.L.M.

Author Contributions

All authors performed the research. K.S.B. and R.L.M. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-32172-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.