

On comparing clusterings: an element-centric framework unifies overlaps and hierarchy

Alexander J. Gates^{a,b,c,1}, Ian B. Wood^{a,b}, William P. Hetrick^d, Yong-Yeol Ahn^{a,b,c,1}

^aDepartment of Informatics, Indiana University. Bloomington, IN.

^bCenter for Complex Networks and Systems Research, Indiana University. Bloomington, IN.

^cProgram in Cognitive Science, Indiana University. Bloomington, IN.

^dDepartment of Psychological and Brain Sciences, Indiana University. Bloomington, IN.

¹To whom correspondence should be addressed.

E-mails: ajgates@indiana.edu and yyahn@indiana.edu

Abstract

Clustering is one of the most universal approaches for understanding complex data. A pivotal aspect of clustering analysis is quantitatively comparing clusterings; clustering comparison is the basis for tasks such as clustering evaluation, consensus clustering, and tracking the temporal evolution of clusters. For example, the extrinsic evaluation of clustering methods requires comparing the uncovered clusterings to planted clusterings or known metadata. Yet, as we demonstrate, existing clustering comparison measures have critical biases which undermine their usefulness, and no measure accommodates both overlapping and hierarchical clusterings. Here we unify the comparison of disjoint, overlapping, and hierarchically structured clusterings by proposing a new element-centric framework: elements are compared based on the relationships induced by the cluster structure, as opposed to the traditional cluster-centric philosophy. We demonstrate that, in contrast to standard clustering similarity measures, our framework does not suffer from critical biases and naturally provides unique insights into how the clusterings differ. We illustrate the strengths of our framework by revealing new insights into the organization of clusters in two applications: the improved classification of schizophrenia based on the overlapping and hierarchical community structure of fMRI brain networks, and the disentanglement of various social homophily factors in Facebook social networks. The universality of clustering suggests far-reaching impact of our framework throughout all areas of science.

Introduction

Clustering is one of the most basic and ubiquitous methods to analyze data. Classically, clustering is viewed as separating data elements into disjoint clusters of comparable sizes [1, 2]. However, complications to this simplistic picture are becoming more prevalent, particularly given the rise of network science and nuanced clustering methods that reveal heterogeneous cluster size distributions [3, 4], overlaps [5, 6, 7, 8], and hierarchical structure [9, 10, 11, 12] (see Figure 1a). These generalizations present new challenges for clustering comparison [13, 3] and render current methods susceptible to critical biases [14, 15, 3, 16, 17]. In addition to the consistent grouping of elements into clusters, similarity measures must account for many other aspects of clusterings, such as the number of clusters, the size distribution of those clusters, multiple element memberships when clusters overlap, and scaling relations between levels of hierarchical clusterings.

Despite the increasing prevalence of irregular cluster features, the effect of such structure on clustering similarity has received little attention. Here we illustrate that all of the most popular clustering similarity measures are vulnerable to critical biases, calling into question the appropriateness of their general usage. We also argue that these biases are maintained or exacerbated by extensions to accommodate overlapping or hierarchical clusterings [18, 19, 20], suggesting that none of the existing frameworks for clustering similarity are adequate for comparing overlapping and hierarchically structured clusterings.

Element-centric clustering comparisons

Here we propose a new *element-centric* framework that not only addresses the common biases, but also naturally incorporates overlaps and hierarchy. In our approach, elements are compared based on the relationships induced by the cluster structure, in contrast to the traditional *cluster-centric* philosophy. As we will see, this change in perspective resolves many of the aforementioned difficulties.

Our approach captures cluster-induced relationships between the elements through the *cluster affiliation graph*, which is a bipartite graph where one vertex set corresponds to the original elements and the other corresponds to the clusters (see Figure 1b, Methods and Supplemental Information, SI, section S3). It naturally incorporates overlaps with multiple edges, and hierarchy with weighted edges. The cluster affiliation graph is then projected onto the element vertices to produce the *cluster-induced element graph*, which is a weighted, directed graph that summarizes the inter-element relationships induced by common cluster memberships [21] (see Figure 1c and Methods).

The traditional notion of pair-wise co-occurrence in a cluster is captured by the presence of an edge in the cluster-induced element graph. However, the focus on element *pairs* misses high-order relations (triplets, quadruplets, etc.) which are useful for characterizing cluster structure [22]. Such high-order co-occurrences can be captured through the presence of paths in the cluster-induced element graph. The weight of the path accounts for the relative importance of elements in the presence of overlapping and hierarchical cluster structures. Here,

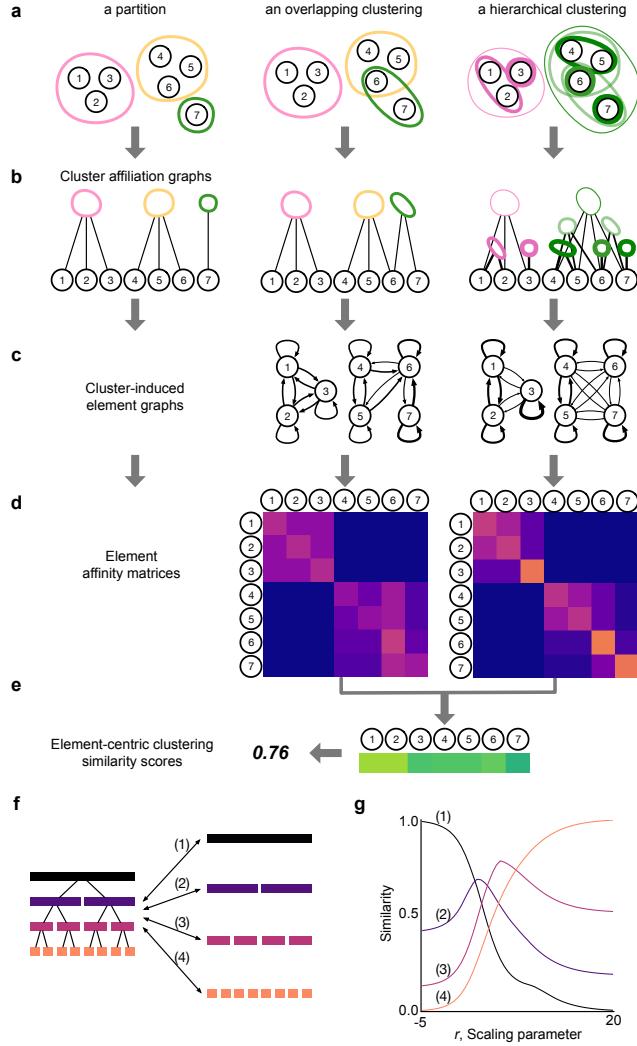


Figure 1: The element-centric perspective naturally incorporates overlaps and hierarchy. **a**, Three examples of clusterings: a partition, a clustering with overlap, and a clustering with both overlapping and hierarchical structure. **b**, Cluster affiliation graphs derived from the overlapping and hierarchical clusterings. **c**, Cluster-induced element graphs found by projecting the cluster affiliation graphs in **b** to the element vertices. **d**, The element-affinity matrices found as the personalized pagerank equilibrium distribution. **e**, The corrected L1 metric distance between each affinity distribution in **d** gives an element-wise similarity between clusterings, the average element-wise similarity provides the final clustering similarity score. **f**, A binary hierarchical clustering is compared to each of its individual levels. **g**, The hierarchical scaling parameter for element-centric similarity acts as a “zooming lens”, refocusing the similarity to different levels (1-4) of the hierarchical comparison in **f**.

we incorporate every possible path between elements obtaining the equilibrium distribution for a personalized diffusion process on the graph (often called “personalized pagerank” or “random walk with restart”) [23, 24, 25]. A similarity score is calculated for each element as the corrected L1 metric distance between these discrete probability distributions; the final similarity score between the two clusterings is the average of the element-wise scores (Figure 1d and Methods). As illustrated in Figure 1, our element-centric framework unifies disjoint, overlapping, and hierarchical clustering comparison in a single framework.

Beyond naturally accommodating generalized clusterings, our element-centric similarity can provide detailed insights into how two clusterings differ because the similarity is calculated at the level of individual elements. Simply examining individual element-wise scores reveal how consistently each element is grouped across clusterings. The rank-distribution of element-wise scores reflects the elements’ relative contributions to the total similarity: a flat distribution suggests the clusterings differ equally across all elements while a skewed distribution suggests the clusterings are distinguished by a subset of elements (see SI, section S3.4). Additionally, the measure can be averaged over the pair-wise comparisons within a set of clusterings. The element-wise *agreement* is revealed by the average of these element-wise scores over comparisons between uncovered clusterings and a reference clustering (SI, section S3.6). The element-wise scores can also be averaged over all pair-wise comparisons within the set of uncovered clusterings, revealing the *frustrated* elements that cannot be consistently clustered.

Our element-centric framework is flexible and allows several choices to accommodate alternative interpretations. For example, our choice of hierarchical weighting function and the scaling parameter, r , reflects a continuum in the hierarchy (Figure 1g): lower r emphasizes higher levels and reflects a divisive hierarchy, in which lower levels of the dendrogram are treated as refinements of the higher levels, while larger r puts emphasis on lower levels and reflects an agglomerative hierarchy, in which higher levels of the dendrogram are seen as a coarsening of the lower level cluster structure. Other interpretations of hierarchy can be implemented by changing the specific hierarchical weighting function. Moreover, our choice of L1 comparisons between personalized pagerank distributions, which is based on a principled extension of element co-occurrence, can be replaced by another measure of graph similarity or probability metric with an alternative intuition of the trade-offs associated with clustering similarity. In fact, several common clustering similarity measures can be recovered by adapting other choices of graph similarity; for example, choosing the graph-edit distance between the two graphs induced from disjoint partitions reduces our measure to the Rand index.

Results

Bias in clustering comparisons

Our element-centric similarity measure is the only clustering similarity method to follow our common-sense expectations and avoid critical biases when comparing generalized clusterings.

We demonstrate such biases by constructing a parametrized family of synthetic clusterings and observing the behavior of clustering similarity measures (listed in Figure 2d and SI, section S4).

In the first example, the consistent grouping of elements is tested by comparing a clustering with equally sized clusters against itself after a fraction of element memberships have been shuffled between clusters (Figure 2a). Intuition suggests that as the randomization increases, the similarity between the original clustering and the shuffled clustering should decrease from the maximum value (1.0 in all cases) to some non-zero value, reflecting the fact that the number and sizes of clusters are still identical. However, two measures reach zero, ignoring the similarity of the cluster size sequences. The overlapping normalized mutual information (ONMI) [19] is particularly conservative, reporting no similarity at just over 50% randomization; ONMI’s surprising behavior highlights the difficulty of accommodating overlaps in a traditional similarity framework.

The second example explores the bias favoring skewed cluster size sequences. Starting from an initial clustering with regularly sized clusters, we generate new, shuffled clusterings through a preferential attachment shuffling scheme (Figure 2b and SI, section S4). Intuition suggests that as the entropy of the cluster size sequence decreases (reflecting an increase in the cluster size heterogeneity), the two clusterings should become less similar. However, four similarity measures increase as the entropy of the cluster size sequence decreases.

Finally, we investigate a scenario where the number and sizes of clusters in two clusterings diverge (Figure 2c). This extreme case captures the bias of information theoretic measures towards comparisons with many clusters. Normalized mutual information (NMI) reports larger similarity if we simply increase the number of clusters in one of the clusterings.

These three examples suggest that the most common measures are subject to critical biases which render them inappropriate for comparing generalized clusterings—only our element-centric similarity measure displays the intuitive behavior in all examples. An extended analysis, additional examples, and additional measures (such as the variation of information, VI) are given in the SI, section S4.

Element-centric comparisons reveal insights into how K-means clusterings differ

Beyond serving as a global measure of clustering similarity, our element-centric similarity also provides detailed insights into how clusterings differ, in contrast to other measures. Consider an illustrative example from K-means clustering shown in Figure 3a; 19 clusters were randomly placed in a square with a randomly selected arrangement (Gaussian blob, anisotropic blob, circle, or spiral) and size (see SI, section S5.2). K-means has difficulty when the predefined clusters overlap or when circularly arranged [26]. This difficulty can be explicitly quantified by calculating the average element-wise similarity between the predefined clustering and 100 uncovered clusterings (Figure 3b). We then calculate the frustration by ..; the average of all pair-wise comparisons between the 100 uncovered clusterings reveal data points that are consistently grouped into similar clusters or are assigned to drastically

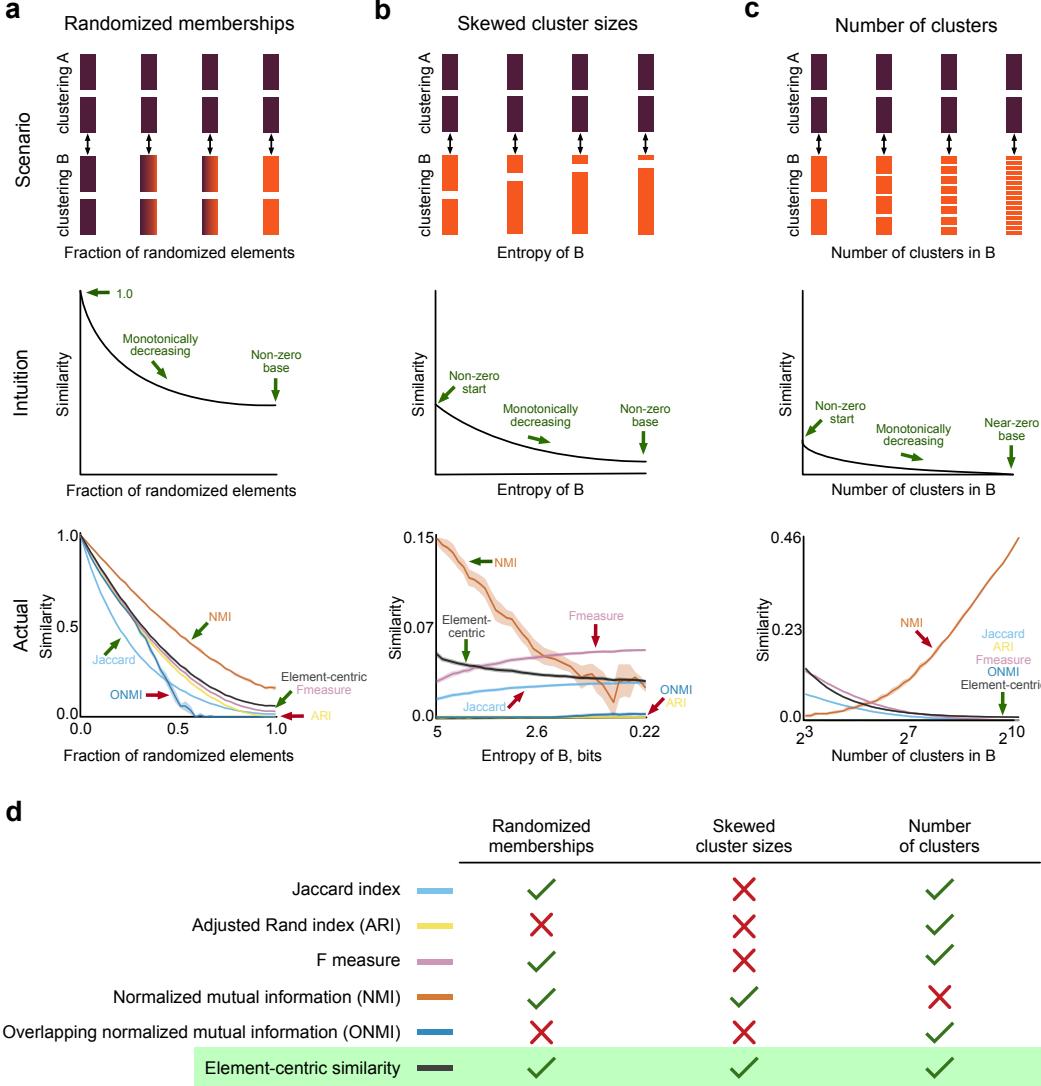


Figure 2: Element-centric similarity behaves intuitively in three clustering similarity scenarios while common clustering similarity measures exhibit counter-intuitive behaviors. 1,024 elements are assigned to clusters according to the following scenarios **a-c** and compared using the Jaccard index, adjusted Rand index, the F measure, normalized mutual information (NMI), overlapping normalized mutual information (ONMI), and our element-centric similarity. All results are averaged over 100 runs and error bars denote one standard deviation.

- a**, A clustering with 32 non-overlapping and equal-sized clusters is compared to a randomized version of itself where elements are shuffled.
- b**, A clustering with 32 non-overlapping and equal-sized clusters is compared against clusterings with increasing cluster size skewness.
- c**, A clustering with 8 non-overlapping and equal-sized clusters is compared against a clustering with n non-overlapping, equal-sized clusters and randomized element memberships for different values of n .
- d**, Only our element-centric similarity measure follows the intuitive behavior in all three scenarios.

different clusters (Figure 3c).

We also present a real-world example of handwriting recognition [27] (Figure 3d and SI, section S5.3). The same procedure reveals that some clusters of digits are correctly and consistently identified (“0”), while the error mostly results from incorrect grouping of other digit clusters (“9”, “8”, and “1”; Figure 3e). Element-wise clustering frustration shows that there are some digits which cannot be consistently classified (“3” and “8”, Figure 3f), while some errors are regularly made (“1” and “9”). The extreme examples of these two types of error are shown in Figure 3g.

The convolution of meta-data in social networks

We now use our framework to explore the community structure of college friendship networks on Facebook. Previous research has suggested that friendship networks at major universities are organized into clusters which reflect the graduation year, dormitory, or student major [28, 29]. However, the details of the organizing principles underlying this similarity are unknown. Here we demonstrate and visualize how multiple attributes interact and contribute to community structure. We first derive clusterings in binary friendship networks using the Louvain method (see SI section S5.4) and compare these to the aforementioned self-declared user attribute clusterings. Element-wise similarity reveals that school year closely captures the modular structure for most of the network, particularly for the students in early years, while students’ major gradually takes over the cohort-based connections (Figure 3h,i red arrows). This result, which has only become straight-forward through our framework, supports the intuition that network structure results from the convolution of multiple attributes [30].

Element-centric comparisons of overlapping and hierarchical clustering in brain networks

Finally, to illustrate the utility of our element-centric similarity measure, we demonstrate its ability to capture meaningful differences in clustering structure by classifying schizophrenic individuals based on the overlapping and hierarchical community structure of resting-state fMRI brain networks. There are several known distinctive and interpretable properties of resting-state fMRI brain networks in schizophrenia, but their classification utility is limited, with accuracies between 75% – 80% [31, 32, 33]. Network communities, in particular, are hypothesized to capture functionally integrated modules in the brain that reflect key properties of schizophrenia [31]. Here we demonstrate that employing our measure to compare communities derived from functional brain networks can improve the classification accuracy significantly. We extract communities with overlapping and hierarchical structure using OSLOM community detection [34] from the functional brain networks of 48 subjects (29 healthy controls and 19 individuals diagnosed with schizophrenia) analyzed in a previous study [32] (see SI, section S5.1 for details). The similarity between each pair of the subjects’ hierarchical and overlapping clusterings was found using our element-centric similarity measure, producing a 48×48 similarity matrix (Figure 4a). This similarity matrix was then

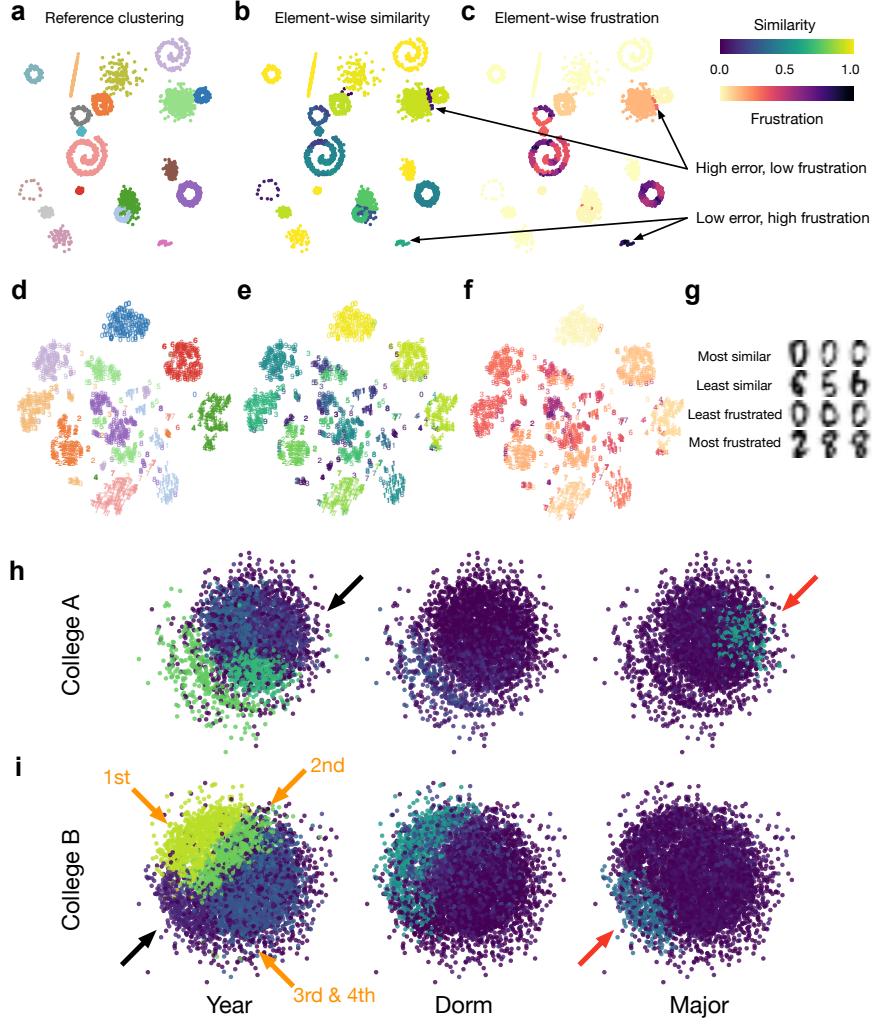


Figure 3: Element-wise clustering similarity reveals insights into how clusterings differ. **a-c**, A K-means clustering example. **a**, The planted clustering. **b**, The average element-wise similarity between the planted clustering and 100 K-means clusterings. **c**, The average element-wise similarity between 100 K-means clusterings. **d-g**, A handwriting classification example. **d**, The labeled handwritten digit data projected using t-SNE dimensionality reduction for visualization. **e**, The average element-wise similarity between the labels and 100 K-means clusterings. **f**, The average element-wise similarity between 100 K-means clusterings. **g**, Exemplar digits that are consistently grouped as in the ground-truth clustering, consistently clustered differently from the ground-truth clustering, least frustrated, and most frustrated. **h,i**, Facebook friendship networks for **h** College A and **i** College B. The element-wise similarity between user affiliation to school year, dorm, and major compared to Newman's modularity optimized by the Louvain method demonstrates that social networks can be organized by a convolution of different attributes (black vs red arrows). The similarity to school year attenuates with student's status (1st year - 4th year, orange arrows).

used in conjunction with a weighted k-nearest neighbors classifier to perform a binary classification of subjects as either schizophrenic or healthy controls (SI, section S5.1). Evaluated by a nested 10-fold cross-validation procedure, our approach achieves an average accuracy of 84%, outperforming other measures (ONMI) and state-of-the-art results (Figure 4b). Note that, classification based on individual levels from the hierarchy does not perform as well as the method using the full hierarchy.

Our element-centric clustering similarity measure also provides insights into which brain regions are consistently clustered within groups. To find such group differences, we consider the element-centric similarity between all healthy controls, and the element-centric similarity between all schizophrenic patients. As seen in Figure 4c, the difference between the means of these two groups highlights several regions which are consistently clustered into similar functional modules in the healthy controls or schizophrenic patients. In particular, regions of interest (ROIs) located in the Fusiform gyrus (Brodmann Area 37) were consistently clustered in the healthy controls but displayed great variability in cluster structure for the schizophrenic patients (verified with a Bayesian difference of means test, see SI section S5.1). This result is corroborated by the fact that the Fusiform gyrus has previously been associated with abnormal activation in schizophrenia during semantic tasks [35, 36].

Summary and discussion

In summary, we present an element-centric framework that intuitively unifies the comparison of disjoint, overlapping, and hierarchically structured clusterings. We have presented that our element-centric similarity does not suffer from the common counter-intuitive biases of existing measures, and that it also provides insights into how clusterings differ at the level of individual elements.

Our framework suggests straight-forward extensions to more complex scenarios, such as soft or fuzzy clusterings, hierarchical clusterings specified by dendograms with merge distance information, and hyper-graph similarity. The framework also provides a measure of pair-wise similarity between elements, akin to the nodal association matrix of Bassett *et al.* [37], and an element-wise clustering similarity which summarizes the difference in relationships induced by overlapping and hierarchically structured clusterings from the perspective of individual elements. Both of these objects hold promise for use in clustering ensemble methods [38, 39].

As clustering methods advance to uncover more nuanced and accurate organizational structure of complex systems, so too should clustering similarity measures facilitate meaningful comparisons of these organizations. The element-centric framework proposed here provides an intuitive quantification of clustering similarity that holds great promise for uncovering the relationships amongst all types of clusters, such as network communities, ontogenies, and dendograms. The ubiquity of clustering in all areas of science suggests the extensive potential impact of our framework.

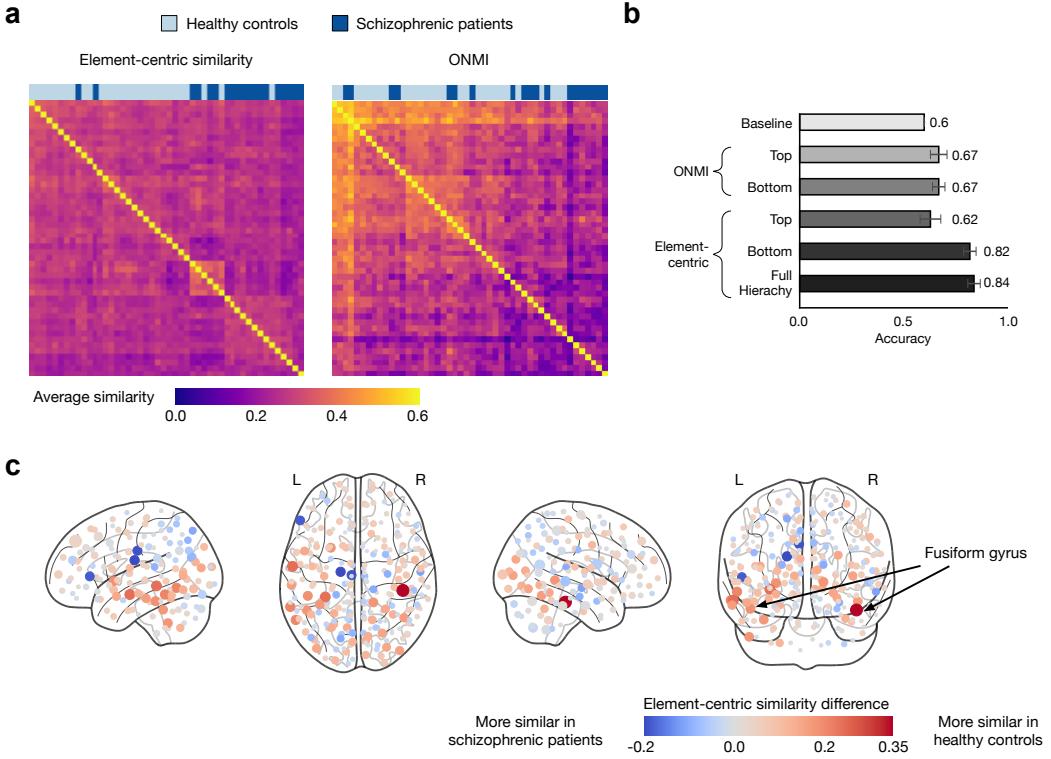


Figure 4: Our element-centric similarity better differentiates the overlapping and hierarchical community structure of functional brain networks in healthy and schizophrenic individuals. **a**, Hierarchical clustering of average pair-wise element-centric similarity using the entire OSLOM hierarchy closely reflects the true classification of participants as healthy (light blue) or schizophrenic (dark blue), while hierarchical clustering of the average pair-wise similarity using ONMI on the bottom level of the OSLOM hierarchy fails to uncover patient classification. **b**, Classification accuracy using different clustering similarity measures averaged over 100 instances of 10-fold cross-validation, error bars denote one standard deviation. **c**, The difference in element-centric similarity for each brain region when comparing amongst the healthy controls minus the similarity when comparing amongst the schizophrenic individuals; ROIs within the Fusiform gyrus are more consistently clustered in the healthy controls than in the schizophrenic individuals.

Methods

Graph representation of clusterings

Given a clustering \mathcal{C} of N elements $E = \{v_1, \dots, v_N\}$ into $K_{\mathcal{C}}$ clusters, the cluster affiliation graph is an undirected bipartite graph where one vertex set corresponds to the elements, the other corresponds to the clusters, and a weighted edge exists between a cluster and each of its elements. For hierarchically structured clusterings, each cluster $c_{\beta} \in \mathcal{C}$ is assigned a hierarchical level $l_{\beta} \in [0, 1]$ by rescaling the hierarchy's acyclic graph (dendrogram) according to the maximum path length from the roots [40]. The weight of the cluster affiliation edge is given by the hierarchy weighting function $h(l_{\beta})$:

$$h(l_{\beta}) = e^{rl_{\beta}} \quad (1)$$

where r is a scaling parameter that determines the relative importance of membership at different levels of the hierarchy. The cluster-induced element graph is formed by projecting the cluster affiliation graph (with $N \times K_{\mathcal{C}}$ bipartite adjacency matrix \mathbb{A}) onto the element vertices resulting in a directed graph with the edge w_{ij} between elements v_i and v_j having weight:

$$w_{ij} = \sum_{\gamma} \frac{a_{i\gamma} a_{j\gamma}}{\sum_{\kappa} a_{i\kappa} \sum_m a_{m\gamma}} \quad (2)$$

Personalized PageRank affinity

Given an cluster-induced element graph with weighted adjacency matrix \mathbf{W} , the personalized PageRank (PPR) affinity from element v_i to all elements v_j is found as the stationary distribution of a diffusion process with stochastic matrix $\boldsymbol{\Pi}_i$ and restart probability $1.0 - \alpha$ to v_i :

$$\boldsymbol{\Pi}_i = (1.0 - \alpha)\mathbf{v}_i + \alpha\mathbf{W} \quad (3)$$

The value of α controls the influence of overlapping clusters and hierarchical clusters with shared lineages; here we use $\alpha = 0.90$. The above equation can be exactly solved for partitions—the affinity value for each co-clustered element pair is inversely proportional to the cluster size, and 0 otherwise. For clusterings with overlaps or hierarchy, several algorithms are available to quickly approximate the PPR affinity [41]. See the SI, section S3 for further comments about implementation.

Element-centric similarity

The element-wise similarity of an element v_i in two clusterings \mathcal{A} and \mathcal{B} is found by comparing the stationary probability distributions $\mathbf{p}^{\mathcal{A}}$ and $\mathbf{p}^{\mathcal{B}}$ induced by the PPR processes on the two cluster-induced element graphs. Here, we use the normalized L1 metric for probability distributions corrected to account for the PPR process:

$$S_i(\mathcal{A}, \mathcal{B}) = 1.0 - L1_{\alpha}(\mathbf{p}_i^{\mathcal{A}}, \mathbf{p}_i^{\mathcal{B}}) = 1.0 - \frac{1}{2\alpha} \sum_{j=1}^N |p_j^{\mathcal{A}} - p_j^{\mathcal{B}}| \quad (4)$$

The final element-centric similarity score $S(\mathcal{A}, \mathcal{B})$ of two clusterings \mathcal{A}, \mathcal{B} is the average of the element-wise similarities:

$$S(\mathcal{A}, \mathcal{B}) = \frac{1}{N} \sum_{i=1}^N S_i(\mathcal{A}, \mathcal{B}). \quad (5)$$

Datasets

Details on the synthetic clusterings, fMRI brain networks, K-means point, handwriting, and Facebook social network datasets can be found in SI Text.

Acknowledgements

We would like to thank Dae-Jin Kim for assistance with the interpretation of the schizophrenia classifications, and Hu Cheng for assistance processing the fMRI timeseries. We thank Randall D. Beer, Luis M. Rocha, Filippo Radicchi, Sune Lehmann, Olaf Sporns, Alessio Cardillo, and Artemy Kolchensiky for helpful discussions. Supported in part by National Institute for Mental Health Grant 2R01MH074983 to WPH. YYA thanks Microsoft Research for support through a Microsoft Research Faculty Fellowship.

Author Contributions

AJG and YYA developed the method, AJG and IBW performed the analysis, HB contributed the fMRI data, AJG, IBW, HB, and YYA participated in interpreting results, AJG and YYA wrote the manuscript. All authors reviewed and edited the manuscript.

References

- [1] Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31**, 264–323 (1999).
- [2] Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).
- [3] He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**, 1263–1284 (2009).
- [4] Leskovec, J., Lang, K. J., Dasgupta, A. & Mahoney, M. W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* **6**, 29–123 (2009).
- [5] Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).

- [6] Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
- [7] Gopalan, P. K. & Blei, D. M. Efficient discovery of overlapping communities in massive networks. *PNAS* **110**, 14534–14539 (2013).
- [8] Yang, J. & Leskovec, J. Structure and overlaps of ground-truth communities in networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* **5**, 26 (2014).
- [9] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
- [10] Sales-Pardo, M., Guimera, R., Moreira, A. A. & Amaral, L. A. N. Extracting the hierarchical organization of complex systems. *PNAS* **104**, 15224–15229 (2007).
- [11] Delvenne, J.-C. C., Yaliraki, S. N. & Barahona, M. Stability of graph communities across time scales. *PNAS* **107**, 12755–12760 (2010).
- [12] Zhang, P. & Moore, C. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *PNAS* **111**, 18144–18149 (2014).
- [13] Meila, M. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* **98**, 873–895 (2007).
- [14] White, A. P. & Liu, W. Z. Technical note: Bias in information-based measures in decision tree induction. *Machine Learning* **15**, 321–329 (1994).
- [15] Fitzner, D., Leibbrandt, R. & Powers, D. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems* **19**, 361–394 (2009).
- [16] Zhang, P. Evaluating accuracy of community detection using the relative normalized mutual information. *Journal of Statistical Mechanics: Theory and Experiment* **2015**, P11006 (2015).
- [17] Gates, A. J. & Ahn, Y.-Y. The impact of random models on clustering similarity. *arxiv* **1701.06508**, 1–31 (2017).
- [18] Collins, L. M. & Dent, C. W. Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research* **23**, 231–242 (1988).
- [19] Lancichinetti, A., Fortunato, S. & Kertsz, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11**, 033015 (2009).
- [20] Perotti, J. I., Tessone, C. J. & Caldarelli, G. Hierarchical mutual information for the comparison of hierarchical community structures in complex networks. *Physical Review E* **92**, 062825 (2015).

- [21] Zhou, T., Ren, J., Medo, M. & Zhang, Y.-C. Bipartite network projection and personal recommendation. *Physical Review E* **76**, 046115 (2007).
- [22] Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193–218 (1985).
- [23] Haveliwala, T. H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering* **15**, 784–796 (2003).
- [24] Tong, H., Faloutsos, C. & Pan, J. Y. Fast random walk with restart and its applications. In *ICDM '06. Sixth International Conference on Data Mining*, 613–622 (IEEE Computer Society, 2006).
- [25] Kloumann, I. M., Ugander, J. & Kleinberg, J. Block models and personalized pagerank. *PNAS* **114**, 33–38 (2017).
- [26] Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* **31**, 651–666 (2010).
- [27] Alimoglu, F. & Alpaydin, E. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium* (1996).
- [28] Traud, A. L., Kelsic, E. D., Mucha, P. J. & Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review* **53**, 526–543 (2011).
- [29] Traud, A. L., Mucha, P. J. & Porter, M. A. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications* **391**, 4165–4180 (2012).
- [30] Peel, L., Larremore, D. B. & Clauset, A. The ground truth about metadata and community detection in networks. *arxiv* **1608.05878**, 1–24 (2016).
- [31] Alexander-Bloch, A. *et al.* The discovery of population differences in network community structure: new methods and applications to brain functional networks in schizophrenia. *Neuroimage* **59**, 3889–3900 (2012).
- [32] Cheng, H. *et al.* Nodal centrality of functional network in the differentiation of schizophrenia. *Schizophrenia Research* **168**, 345–352 (2015).
- [33] Fornito, A., Zalesky, A., Pantelis, C. & Bullmore, E. T. Schizophrenia, neuroimaging and connectomics. *Neuroimage* **62**, 2296–2314 (2012).
- [34] Lancichinetti, A., Radicchi, F., Ramasco, J. J. & Fortunato, S. Finding Statistically Significant Communities in Networks. *PLoS ONE* **6**, e18961 (2011).

- [35] Kircher, T. T. *et al.* Differential activation of temporal cortex during sentence completion in schizophrenic patients with and without formal thought disorder. *Schizophrenia research* **50**, 27–40 (2001).
- [36] Kuperberg, G. R., Deckersbach, T., Holt, D. J., Goff, D. & West, W. C. Increased temporal and prefrontal activity in response to semantic associations in schizophrenia. *Archives of General Psychiatry* **64**, 138–151 (2007).
- [37] Bassett, D. S. *et al.* Robust detection of dynamic community structure in networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **23**, 013142 (2013).
- [38] Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* **52**, 91–118 (2003).
- [39] Lancichinetti, A. & Fortunato, S. Consensus clustering in complex networks. *Scientific Reports* **2**, 00336 (2012).
- [40] Czgel, D. & Palla, G. Random walk hierarchy measure: what is more hierarchical, a chain, a tree or a star? *Scientific Reports* **5**, 17994 (2015).
- [41] Lofgren, P. A., Banerjee, S., Goel, A. & Seshadhri, C. Fast-ppr: scaling personalized pagerank estimation for large graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1436–1445 (ACM, 2014).
- [42] Meil, M. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, 173–187 (Springer, 2003).
- [43] Albatineh, A. N., Niewiadomska-Bugaj, M. & Mihalko, D. On similarity indices and correction for chance agreement. *Journal of Classification* **23**, 301–313 (2006).
- [44] Meila, M. & Heckerman, D. An experimental comparison of model-based clustering methods. *Machine Learning* **42**, 9–29 (2001).
- [45] Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods* **9**, 471–472 (2012).
- [46] Lancichinetti, A. & Fortunato, S. Community detection algorithms: a comparative analysis. *Physical Review E* **80**, 056117 (2009).
- [47] Danon, L., Diaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* **2005**, P09008–P09008 (2005).
- [48] Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846 (1971).

- [49] Fowlkes, E. B. & Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* **78**, 553–569 (1983).
- [50] Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1073–1080 (ACM, 2009).
- [51] Steinley, D., Brusco, M. J. & Hubert, L. The variance of the adjusted rand index. *Psychological Methods* (2016).
- [52] Ben-Hur, A., Elisseeff, A. & Guyon, I. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, 6–17 (2002).
- [53] Jaccard, P. The distribution of the flora in the alpine zone. *New Phytologist* **11**, 37–50 (1912).
- [54] Levandowsky, M. & Winter, D. Distance between sets. *Nature* **234**, 34–35 (1971).
- [55] Lancichinetti, A., Fortunato, S. & Kertesz, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11**, 033015 (2009).
- [56] Esquivel, A. V. & Rosvall, M. Comparing network covers using mutual information. *arXiv preprint arXiv:1202.0425* (2012).
- [57] Xie, J., Kelley, S. & Szymanski, B. K. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (csur)* **45**, 43 (2013).
- [58] Hric, D., Darst, R. K. & Fortunato, S. Community detection in networks: Structural communities versus ground truth. *Physical Review E* **90**, 062805 (2014).
- [59] Campello, R. J. A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters* **28**, 833–841 (2007).
- [60] Campello, R. J. Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters* **31**, 966–975 (2010).
- [61] Waterman, M. S. & Smith, T. F. On the similarity of dendograms. *Journal of Theoretical Biology* **73**, 789–800 (1978).
- [62] Morlini, I. & Zani, S. An overall index for comparing hierarchical clusterings. In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, 29–36 (Springer, 2012).
- [63] Yang, J. & Leskovec, J. Community-Affiliation Graph Model for Overlapping Network Community Detection. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*, 1170–1175 (IEEE, 2012).

- [64] Corominas-Murtra, B., Goñi, J., Solé, R. V. & Rodríguez-Caso, C. On the origins of hierarchy in complex networks. *PNAS* **110**, 13316–13321 (2013).
- [65] Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P. & Van Dooren, P. A Measure of similarity between graph vertices: applications to synonym extraction and web searching. *SIAM Review* **46**, 647–666 (2004).
- [66] Amelio, A. & Pizzuti, C. Is normalized mutual information a fair measure for comparing community detection methods? In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 1584–1585 (ACM, 2015).
- [67] First, M. B. *Structured clinical interview for DSM-IV-TR Axis I disorders: Patient edition* (Biometrics Research Department, Columbia University, 2005).
- [68] Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Groupwise whole-brain parcellation from resting-state fmri data for network node identification. *Neuroimage* **82**, 403–415 (2013).
- [69] Serrano, M. , Bogu, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *PNAS* **106**, 6483–6488 (2009).
- [70] Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning. *NY Springer* (2001).
- [71] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 111–147 (1974).
- [72] Rao, R. B., Fung, G. & Rosales, R. On the dangers of cross-validation. an experimental evaluation. In *SDM*, 588–596 (SIAM, 2008).
- [73] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- [74] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 85 (2008).
- [75] Newman, M. E. J. Modularity and community structure in networks. *PNAS* **103**, 8573–8574 (2006).
- [76] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

Supplemental Information

On comparing clusterings: an element-centric framework unifies overlaps and hierarchy

Alexander J. Gates, Ian B. Wood, William P. Hetrick, Yong-Yeol Ahn

Contents

S1 Clusterings	19
S2 Existing measures of clustering similarity	19
S2.1 Rand Index	19
S2.2 Adjusted Rand index (ARI)	21
S2.3 Jaccard index	21
S2.4 F measure	22
S2.5 Normalized mutual information (NMI)	22
S2.6 Overlapping NMI	23
S2.7 Variation of Information VI	24
S2.8 Extensions for overlaps or hierarchy	24
S3 Detailed methods	25
S3.1 Cluster induced relationships	25
S3.2 Cluster-induced element graph	26
S3.3 Generalizing element co-occurrence	26
S3.4 Element-centric similarity	28
S3.5 Element-centric similarity for strict partitions	28
S3.6 Average agreement and frustration	29
S4 Comparing clustering similarity measures: extend discussion	29
S5 Clustering similarity applications	31
S5.1 Functional brain networks	31
S5.1.1 Dataset	31
S5.1.2 Overlapping and hierarchically structured clusterings	33
S5.1.3 Classification	33
S5.2 Point clusters	34
S5.3 Handwriting digits	34
S5.4 Facebook friendship networks	35

Contents

S1 Clusterings

Throughout this work, we are focused on the grouping of elements (i.e. data points or vertices) into clusters (the groups). The set of clusters is called a *clustering*. Specifically, given a set of N distinct elements $V = \{v_1, \dots, v_N\}$, a clustering is a set $\mathcal{C} = \{C_1, \dots, C_{K_C}\}$ of K_C non-empty subsets of V such that every element v_i in V is in at least one cluster C_β : $\forall v_i \in V \exists C_\beta \text{ s.t. } v_i \in C_\beta$.

We consider three classes of clusterings. A *partition* is a clustering in which all elements are members of one, and only one, cluster. An *overlapping* clustering allows elements to be members of multiple clusters. *Hierarchical* clusterings capture the nested organization of clusters at different scales and are accompanied by a directed acyclic graph (or dendrogram) showing the hierarchical relationships between clusters.

The rest of this paper focuses on the similarity of two clusterings over the same set of N labeled elements, $\mathcal{A} = \{A_1, \dots, A_{K_A}\}$ (with K_A clusters of sizes a_i) and $\mathcal{B} = \{B_1, \dots, B_{K_B}\}$ (with K_B clusters of sizes b_i).

S2 Existing measures of clustering similarity

The clustering similarity measures can be roughly categorized into three classes [42]. The first class counts the pairs of elements co-assigned to the same cluster in both clusterings; Albatineh et al. [43] provides a list of 22 such clustering similarity measures based on pair counting. The second class identifies clusters which constitute a best match between the two clusterings [44]; examples include the maximum matching statistic [44], and the maximum matching ratio [45]. The third class captures the amount of information which exists about the cluster membership of a randomly selected element; examples include the mutual information and its normalized variants [46, 47], as well as the variation of information [42]. Here, we focus on five of the most prominent measures from the clustering literature: the adjusted Rand index, the Jaccard index, the F measure, normalized mutual information (NMI), and overlapping normalized mutual information (ONMI).

S2.1 Rand Index

The Rand index [48] counts the number of element pairs which are either members of the same cluster, or members of different clusters in both clusterings. The most common formulation of the Rand index focuses on the following four sets of the $\binom{N}{2}$ element pairs: N_{11} the number of element pairs which are grouped in the same cluster in both clusterings, N_{10} the number of element pairs which are grouped in the same cluster by \mathcal{A} but in different clusters by \mathcal{B} , N_{01} the number of element pairs which are grouped in the same cluster by \mathcal{B} but in different clusters by \mathcal{A} , and N_{00} the number of element pairs which are grouped in different clusters

\mathcal{A}/\mathcal{B}	B_1	B_2	\dots	B_{K_B}	Sums
A_1	n_{11}	n_{12}	\dots	n_{1K_B}	a_1
A_2	n_{21}	n_{22}	\dots	n_{2K_B}	a_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
A_{K_A}	$n_{K_A 1}$	$n_{K_A 2}$	\dots	$n_{K_A K_B}$	a_{K_A}
Sums	b_1	b_2	\dots	b_{K_B}	$\sum_{ij} n_{ij} = N$

Table S1: The contingency table \mathcal{T} for two clusterings $\mathcal{A} = \{A_1, \dots, A_{K_A}\}$ and $\mathcal{B} = \{B_1, \dots, B_{K_B}\}$ of N elements, where $n_{ij} = |A_i \cap B_j|$ are the number of elements in both cluster $A_i \in \mathcal{A}$ and cluster $B_j \in \mathcal{B}$.

by both \mathcal{A} and \mathcal{B} . Intuitively, N_{11} and N_{00} are indicators of the agreement between the two clusterings, while N_{10} and N_{01} reflect the disagreement between the clusterings.

The aforementioned pair counts are identified from the contingency table \mathcal{T} between two clusterings, shown in Table S1, by the following set of equations:

$$\begin{aligned}
 N_{11} &= \sum_{k,m=1}^{K_A, K_B} \binom{n_{km}}{2} = \frac{1}{2} \left(\sum_{k,m=1}^{K_A, K_B} n_{km}^2 - N \right) \\
 N_{10} &= \sum_{k=1}^{K_A} \binom{a_k}{2} - N_{11} = \frac{1}{2} \left(\sum_{k=1}^{K_A} a_k^2 - \sum_{k,m=1}^{K_A, K_B} n_{km}^2 \right) \\
 N_{01} &= \sum_{m=1}^{K_B} \binom{b_m}{2} - N_{11} = \frac{1}{2} \left(\sum_{m=1}^{K_B} b_m^2 - \sum_{k,m=1}^{K_A, K_B} n_{km}^2 \right) \\
 N_{00} &= \binom{N}{2} - N_{11} - N_{10} - N_{01}.
 \end{aligned} \tag{S1}$$

The Rand index between clusterings \mathcal{A} and \mathcal{B} , $RI(\mathcal{A}, \mathcal{B})$ is then given by the function:

$$\begin{aligned}
 RI(\mathcal{A}, \mathcal{B}) &= \frac{N_{11} + N_{00}}{\binom{N}{2}} \\
 &= \frac{2 \sum_{k,m=1}^{K_A, K_B} \binom{n_{km}}{2} - \sum_{k=1}^{K_A} \binom{a_k}{2} - \sum_{m=1}^{K_B} \binom{b_m}{2} + \binom{N}{2}}{\binom{N}{2}}.
 \end{aligned} \tag{S2}$$

It lies between 0 and 1, where 1 indicates the clusterings are identical and 0 occurs for clusters which do not share a single pair of elements (this only happens when one clustering is the full set of elements and the other clustering groups each element into its own singleton cluster). As the number of elements being clustered becomes large, the measure becomes dominated by the number of pairs which were classified into different clusters (N_{00}), resulting in decreased sensitivity to co-occurring element pairs [49].

S2.2 Adjusted Rand index (ARI)

A popular extension of the Rand index, called the adjusted Rand index (ARI), considers the average of the measure in the context of a random ensemble of clusterings [50, 22, 43, 17]. Such a correction for chance uses the expected similarity of all pair-wise comparisons between clusterings specified by a random null model to establish a baseline; the resulting similarity values have a new interpretation that facilitates comparisons within a set of clusterings. Specifically, once corrected for chance, a similarity value of 1 still corresponds to identical clusterings, but a value of 0 now corresponds to the expected value amongst random clusterings. Positive values of corrected similarity better reflect an intuitive comparison of clusterings, although they are still slightly biased [51]. However, the correction process also introduces negative values for similarity that occur when two clusterings are less similar than would be expected by chance.

The commonly used adjusted Rand index (ARI) of Hubert and Arabie [22] calculates the expectation of the Rand index under the assumption that random clusterings are drawn from the permutation model. In the permutation model the number and size of clusters within a clustering are fixed; all random clusterings are generated by shuffling the elements between the fixed clusters. The expectation of the Rand index with respect to the permutation model follows from the fact that the entries in Table S1 follow a generalized hypergeometric distribution. Taking $Q^{\mathcal{A}} = \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2}$ and $Q^{\mathcal{B}} = \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2}$, the expectation $\mathbb{E}_{perm}[RI(\mathcal{A}, \mathcal{B})]$ of the Rand index with respect to the permutation model for the cluster size sequences of clusterings \mathcal{A} and \mathcal{B} is given by:

$$\mathbb{E}_{perm}[RI(\mathcal{A}, \mathcal{B})] = \frac{2Q^{\mathcal{A}}Q^{\mathcal{B}} - \binom{N}{2}(Q^{\mathcal{A}} + Q^{\mathcal{B}}) + \binom{N}{2}^2}{\binom{N}{2}^2} \quad (S3)$$

(see Fowlkes and Mallows [49], Hubert and Arabie [22], or Albatineh and Niewiadomska-Bugaj [43] for the full derivation). Finally, the ARI between clusterings \mathcal{A} and \mathcal{B} is given by:

$$ARI(\mathcal{A}, \mathcal{B}) = \frac{R(\mathcal{A}, \mathcal{B}) - \mathbb{E}_{perm}[RI(\mathcal{A}, \mathcal{B})]}{1 - \mathbb{E}_{perm}[RI(\mathcal{A}, \mathcal{B})]} \quad (S4)$$

$$= \frac{\binom{N}{2} \sum_{k,m=1}^{K_{\mathcal{A}}K_{\mathcal{B}}} \binom{n_{km}}{2} - \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2} \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2}}{\frac{1}{2} \binom{N}{2} \left[\sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2} + \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2} \right] - \sum_{k=1}^{K_{\mathcal{A}}} \binom{a_k}{2} \sum_{m=1}^{K_{\mathcal{B}}} \binom{b_m}{2}}. \quad (S5)$$

S2.3 Jaccard index

Another popular clustering similarity measure which utilizes pair-wise co-occurrence between the elements is the Jaccard index or Jaccard similarity coefficient [52]. Originally proposed to compare regional floras [53], the Jaccard index is a similarity measure for finite sets. It is defined as the number of element pairs which are grouped in the same cluster in both clusterings divided by the number of element pairs which are grouped in the cluster in at

least one of the clusterings. Thus, it ignores the number of element pairs that are grouped into different clusters by both clusterings. One minus the Jaccard index is a metric on the collection of finite sets [54]. Using the above notation from the contingency table Table S1, the Jaccard index between clusterings \mathcal{A} and \mathcal{B} takes the form:

$$J(\mathcal{A}, \mathcal{B}) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}} \quad (\text{S6})$$

S2.4 F measure

The F measure has a long history of use in clustering validation, natural language processing, information retrieval, and machine learning. It is based off of two asymmetric measures (sometimes called Dice's asymmetric coefficients), that count the proportion of element pairs which are correctly co-assigned to the same cluster in one of the clusterings using the other clustering as a baseline. When one of these clusterings is considered to be a ground-truth clustering, these asymmetric measures are known as *precision* and *recall*. The F measure is the harmonic mean of the precision and recall. Specifically, the F measure between clusterings \mathcal{A} and \mathcal{B} is given by:

$$F(\mathcal{A}, \mathcal{B}) = \frac{2N_{11}}{2N_{11} + N_{10} + N_{01}} \quad (\text{S7})$$

The F measure F and Jaccard index J are related by $J = F/(2 - F)$.

S2.5 Normalized mutual information (NMI)

Another family of approaches for finding the similarity of two cluster coverings is based on the amount of information in each covering and the amount of information one covering contains about the other. These quantities can also be calculated from the contingency Table S1. The entropy H of a clustering \mathcal{A} is given by

$$H(\mathcal{A}) = - \sum_{k=1}^{K_{\mathcal{A}}} \frac{a_k}{N} \log \frac{a_k}{N} \quad (\text{S8})$$

and, using the entries n_{km} from the contingency table S1, the joint entropy between two clusterings \mathcal{A} and \mathcal{B} is

$$H(\mathcal{A}, \mathcal{B}) = - \sum_{k,m=1}^{K_{\mathcal{A}}, K_{\mathcal{B}}} K_{\mathcal{B}} \frac{n_{km}}{N} \log \frac{n_{km}}{N} \quad (\text{S9})$$

Thus, the mutual information between two clusterings is given by:

$$\begin{aligned} MI(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B}) \\ &= \sum_{k,m=1}^{K_{\mathcal{A}}, K_{\mathcal{B}}} \frac{n_{km}}{N} \log \frac{n_{km}N}{a_k b_m}. \end{aligned} \quad (\text{S10})$$

The mutual information can be interpreted as an inverse measure of independence between the clusterings, or a measure of the amount of information each clustering has about the other. As it can vary in the range $[0, \min\{H(\mathcal{A}), H(\mathcal{B})\}]$, to facilitate comparisons, it is desirable to normalize it to the range $[0, 1]$. There are at least six proposals in the literature for this upper bound, each with different advantages and drawbacks;

$$\begin{aligned} \min\{H(\mathcal{A}), H(\mathcal{B})\} &\leq \sqrt{H(\mathcal{A})H(\mathcal{B})} \leq \frac{H(\mathcal{A}) + H(\mathcal{B})}{2} \\ &\leq \max\{H(\mathcal{A}), H(\mathcal{B})\} \leq \max\{\log K_{\mathcal{A}}, \log K_{\mathcal{B}}\} \leq \log N. \end{aligned} \quad (\text{S11})$$

The resulting measures are all known as normalized mutual information (NMI). Here, we always use the average of the two clustering entropies $\frac{1}{2}(H(\mathcal{A}) + H(\mathcal{B}))$. Although some results have been shown to depend on the normalization term used for NMI, Figure S2 demonstrates that NMI behaves un-intuitively regardless of the normalization term.

S2.6 Overlapping NMI

The NMI has been modified to account for clusterings with overlapping clusters [55]. Consider a clustering \mathcal{A} with $K_{\mathcal{A}}$ possibly overlapping clusters $A_1, \dots, A_{K_{\mathcal{A}}}$. For each cluster A_k , we can consider a binary random variable X_k which represents the probability that a randomly selected node is a member of that cluster with distribution

$$P(X_k = 1) = \frac{a_k}{N}, \quad P(X_k = 0) = 1 - \frac{a_k}{N} \quad (\text{S12})$$

The same holds for a second clustering \mathcal{B} with $K_{\mathcal{B}}$ possibly overlapping clusters $B_1, \dots, B_{K_{\mathcal{B}}}$ and random variables Y_m . We can then define the joint probability distribution $P(X_k, Y_m)$:

$$\begin{aligned} P(X_k = 1, Y_m = 1) &= \frac{n_{km}}{N}, \quad P(X_k = 0, Y_m = 0) = 1 - \frac{n_{km}}{N} \\ P(X_k = 1, Y_m = 0) &= \frac{a_k - n_{km}}{N}, \quad P(X_k = 0, Y_m = 1) = \frac{b_m - n_{km}}{N} \end{aligned} \quad (\text{S13})$$

Given a particular cluster $A_k \in \mathcal{A}$, the amount of information it has about another cluster $B_m \in \mathcal{B}$ is found by the conditional entropy

$$H(X_k|Y_m) = H(X_k, Y_m) - H(Y_m). \quad (\text{S14})$$

In the case of overlapping clusters, there are many possible candidates for the best match between two clusters. The best match is the one with the minimal conditional entropy. Thus, the conditional entropy of X_k with respect to all of the clusters in \mathcal{B} is

$$H(X_k|\mathbf{Y}) = \min_{m \in \{1, \dots, M\}} H(X_k|Y_m). \quad (\text{S15})$$

However, in minimizing the entropy it may be the case that the optimal B_m^* is the complement of A_k , thus we have to add the following constraint to insure the above minimization identifies the best matching cluster:

$$h[P(1, 1)] + h[P(0, 0)] > h[P(0, 1)] + h[P(1, 0)]. \quad (\text{S16})$$

This entropy is normalized by the entropy of X_k and averaged over all X_k to give the normalized conditional entropy of \mathbf{X} with respect to \mathbf{Y}

$$H(\mathbf{X}|\mathbf{Y})_{\text{norm}} = \frac{1}{K} \sum_{k=1}^K \frac{H(X_k|\mathbf{Y})}{H(X_k)}. \quad (\text{S17})$$

Finally, the overlapping normalized mutual information ONMI is given by

$$\text{ONMI}(\mathcal{A}, \mathcal{B}) = 1 - \frac{1}{2}[H(\mathbf{X}|\mathbf{Y})_{\text{norm}} + H(\mathbf{Y}|\mathbf{X})_{\text{norm}}]. \quad (\text{S18})$$

It is interesting to note that when \mathcal{A} and \mathcal{B} are partitions, the $NMI(\mathcal{A}, \mathcal{B})$ and $\text{ONMI}(\mathcal{A}, \mathcal{B})$ do not necessarily agree. Although there have been several attempts to reformulate ONMI so that it agrees with NMI, the above formulation is pervasive in the literature [56, 57, 58].

S2.7 Variation of Information VI

Another popular clustering comparison measure based on information theory is the Variation of Information (VI). Unlike the similarity measures discussed above, the VI is a metric on the lattice of partitions [42]. Thus, it is measure of distance between clusterings instead of the similarity of the clusterings; it attains its minimum at 0 when the clusterings are identical, and attains positive values for clusterings which differ. Using the entropy and mutual information between clusterings defined in Section S2.5, the VI is given by:

$$\begin{aligned} \text{VI}(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - 2MI(\mathcal{A}, \mathcal{B}) \\ &= 2H(\mathcal{A}, \mathcal{B}) - H(\mathcal{A}) - H(\mathcal{B}). \end{aligned} \quad (\text{S19})$$

S2.8 Extensions for overlaps or hierarchy

There have been several extensions of these common measures to clusterings with either overlapping or hierarchical structures (but not both). One such measure is the Omega index, an extension of the adjusted Rand index for coverings with overlapping clusters [18]. The Omega index treats every set of cluster memberships as an independent grouping and counts all co-grouped pairs of elements in both clusterings, and all element pairs which are not grouped together in both clusterings. For partitions, the Omega index is equivalent to the ARI. There is also another measure derived for fuzzy overlapping clusterings introduced by Campello as an extension of the Rand index [59, 60].

The similarity of hierarchical cluster structures has received considerably less attention in the network literature. Most measures of hierarchical clustering similarity focus only on dendrograms; for example, the pair-wise edit distance [61], or the cluster similarity between successive cuts of the dendrograms [49, 62]. The one similarity measure for hierarchically structured clusterings which is closest to the framework proposed here is the normalized hierarchical mutual information [20]. Normalized hierarchical mutual information is an extension of NMI in which successive levels of the hierarchy reduce the overall uncertainty of the element memberships, hence, it is built around divisive hierarchies.

S3 Detailed methods

S3.1 Cluster induced relationships

Formally, cluster induced relationships are represented via the cluster affiliation graph [63]. A cluster affiliation graph is constructed for a clustering \mathcal{C} of labeled elements $\mathcal{V} = \{v_1, \dots, v_N\}$ as a bipartite graph $\mathcal{CTAG}(\mathcal{V} \cup \mathcal{C}, \mathcal{R})$ where one vertex set corresponds to the original elements \mathcal{V} and the other vertex set corresponds to the clusters \mathcal{C} . An undirected edge $a_{i\beta} \in \mathcal{R} \subset \mathcal{V} \times \mathcal{C}$ is placed between element $v_i \in \mathcal{V}$ and cluster $c_\beta \in \mathcal{C}$ if $v_i \in c_\beta$, i.e. the element is a member of the cluster. Notice that an element's membership in several overlapping clusters directly translates into several edges in the cluster affiliation graph.

The cluster affiliation graph also accommodates hierarchical cluster organization. Hierarchical cluster structure captures organization at different scales and is typically represented by a directed acyclic graph or a dendrogram, a tree-like structure in which more closely related elements have common ancestors lower in the tree than compared to more distantly related elements [64]. Both directed acyclic graphs and dendrograms have nodes representing the clusters of the clustering, and directed edges between two nodes whenever the target cluster is a decedent of the source cluster. Clusters which are not decadents of any other cluster are called root clusters, while clusters which have no descendants are known as leaves. Following Czegel & Palla [40], every cluster c_β in \mathcal{C} is given a rescaled hierarchical level l_β according to the following process. The hierarchical level of all roots is 0.0, while the hierarchical level of all leaves is 1.0. If a cluster is not a root or a leaf, then we find the path of maximum length between a root and a leaf which passes through the cluster. The hierarchical level of the cluster is then the cluster's position in the path relative to the root (maximum distance from the root) divided by the total path length.

Given a clustering with a rescaled hierarchical structure (such that every cluster has an hierarchical level, see above), the edge weights of the cluster affiliation graph are given by a function of the cluster's hierarchical level. Specifically, if an element $v_i \in \mathcal{V}$ is a member of a hierarchical cluster $c_\beta \in \mathcal{C}$ which occurs at level l_β in the acyclic graph or dendrogram capturing the hierarchical organization of \mathcal{C} , the appropriate edge $a_{i\beta}$ in the cluster affiliation graph has weight $a_{i\beta} = h(l_\beta)$ given by the hierarchy weight function $h : [0, 1] \rightarrow \mathbb{R}^+$. The function h reflects an important decision of hierarchical clustering similarity in general: one has to decide if the similarity of hierarchies should be more strongly focused on the coarser relationships (those at the top of the dendrogram) or the finer relationships (those at the bottom of the dendrogram). This distinction is related to the fact that hierarchies can be constructed in a divisive manner (a top-down approach in which clusters are successively subdivided into finer-grained structures) or an agglomerative manner (a bottom-up approach in which clusters are successively combined into coarser-grained structures). The shape of h will determine what trade-offs are made in terms of hierarchical similarity: a constant function flattens the hierarchy into an overlapping clustering with one level, a monotonically decreasing h will favor relationships induced by higher levels of the dendrogram, while monotonically increasing h will favor relationships induced by lower levels of the dendrogram over those at higher levels. A choice of h that is not monotonically increasing or decreasing

would suggest there are some resolutions which are more important than others but those resolutions are scattered throughout the dendrogram.

Here, we adapt the hierarchical weighting function:

$$h(l_\beta) = e^{rl_\beta} \quad (\text{S20})$$

where r is a constant that determines the relative importance of membership at different levels of the hierarchy. For $r < 0$, similarities between higher levels of the dendrogram are favored over lower levels, while for $r > 0$ similarities between lower levels are more important than higher levels. While the decision of an appropriate value of r depends on the specific application, we take the approach that similar hierarchical clusters should respect the finest graining of the network, and cluster memberships are further enhanced as one ascends the dendrogram. In general, we have found that the exact value of r for which the lowest level of the dendrogram is considered the most important will depend on the height of the hierarchy (length of the maximum directed path in the acyclic graph). For all comparisons between hierarchical clusterings conducted in this work, we use a value of $r = 8$, but further investigation into the sensitivity of the measure on r will be needed.

S3.2 Cluster-induced element graph

The cluster affiliation graph contains all of the same information as the original clustering structure, we now begin to summarize attributes of that structure which contribute to clustering similarity. To extract a coherent set of relationships between the elements induced by the clustering, the bipartite cluster affiliation graph is projected to its element vertices to form the cluster-induced element graph. Specifically, the cluster-induced element graph is a weighted, directed network where the edge w_{ij} captures the aggregated influence between elements v_i and v_j , normalized by the total weight incident to element v_i :

$$w_{ij} = \sum_\gamma \frac{a_{i\gamma}a_{j\gamma}}{\sum_\kappa a_{i\kappa} \sum_m a_{m\gamma}}. \quad (\text{S21})$$

Note that self-loops occur throughout the element interaction graph.

S3.3 Generalizing element co-occurrence

Next we extend the concept of element co-occurrence to the cluster-induced element graph. As discussed in section S2, many existing clustering similarity measures focus on the pair-wise co-occurrence of elements in clusters. In the cluster-induced element graph, the co-occurrence of two elements in at least one cluster is captured by the presence of an edge. The weight of this edge reflects the relative influence between the elements aggregated over all clusters in which the elements co-occur.

The focus on element pairs misses high-order relations which are induced by the cluster structure and are beneficial for differentiating cluster structure [22]. The cluster-induced element graph captures such high-order occurrences through the presence of paths. Thus,

all co-occurrences between 3 elements are captured by paths of length 2, while the co-occurrences between 4 elements are captured by paths of length 3, etc. The weight of the path accounts for the relative importance of neighboring elements in the presence of overlapping and hierarchical cluster structures. Note that in our generalization, singleton clusters are naturally accommodated by the presence of self-loops in the cluster-induced element graph, and hence paths which contain multiple passes through the singleton element.

The information contained in all possible paths through a graph can be integrated using a diffusion process on the graph. However, from the perspective of each element, all paths through the cluster-induced element graph are not created equal. Instead, we want to favor those paths which explore the local neighborhood around each element. Thus, for our element-centric similarity measure, we utilize the stationary distribution of a personalized diffusion process as a useful proximity measure that integrates both local and global graph structure around an element.

Specifically, given a cluster-induced element graph with weighted adjacency matrix \mathbf{W} , the Personalized PageRank affinity between element v_i and all elements v_j is found as the stationary distribution of a diffusion process on the element interaction graph with restart probability $1.0 - \alpha$ to e_i .

$$\Pi_i = (1.0 - \alpha)\mathbf{v}_i + \alpha\mathbf{W} \quad (\text{S22})$$

The value of α controls the influence of longer paths in the element interaction graph which relates to the relative importance of overlapping clusters and hierarchical clusters with shared lineages; here we use $\alpha = 0.9$. The complete matrix of pair-wise personalized pagerank affinities provides a relative measure of the similarity between two elements under the relationships induced by a clustering. One potential use of this matrix, not explored here, is to average the affinity matrices over several clusterings. The resulting object should function in a similar manor as the nodal-affinity matrices of Bassett et. al. [37] and can become the subject of further consensus clustering routines [39].

In general, for large data sets and clusterings with many overlapping and hierarchical clusters, the calculation of personalized pagerank can be a computationally expensive process; a different matrix must be inverted for every element with a resulting complexity of $\mathcal{O}(N^4)$. However, there are some computational simplifications that can be made. First, the personalized pagerank affinity of strict partitions can be analytically solved (see Section S3.5). Second, when several elements share exactly the same cluster memberships, their resulting personalized pagerank affinity vectors are related by simple permutations; therefore, the personalized pagerank affinity vector need only be calculated once for each common cluster membership set. Third, due to the utility of personalized pagerank for recommendation systems, there have been many algorithms for the approximation of personalized pagerank [41]. Because the worst case computational complexity of element-centric similarity will only occur for highly overlapping and deeply hierarchical clusterings, objects which were previously incomparable using traditional clustering similarity methods, we do not consider the computational complexity as a drawback of our method.

It is also useful to note that our choice of the personalized pagerank equilibrium distribution can be motivated in terms of the graph similarity between the two cluster-induced

element graphs [65]. While there are many different methods to assess the similarity of graphs, several other common choices have meaningful interpretations for clustering similarity. For example, the graph-edit distance between the cluster-induced element graphs derived from partitions results in the Rand index.

S3.4 Element-centric similarity

A convenient aspect of the element-centric similarity is that one can recover element-specific values of similarity under the different clusterings. For our element-centric similarity, we use the L1 metric for probability distributions, corrected to account for the personalized pagerank process:

$$S_i(\mathcal{A}, \mathcal{B}) = L_\alpha(\mathbf{p}_i^{\mathcal{A}}, \mathbf{p}_i^{\mathcal{B}}) = 1.0 - \frac{1}{2\alpha} \sum_{j=1}^N |p_j^{\mathcal{A}} - p_j^{\mathcal{B}}| \quad (\text{S23})$$

This correction accommodates the fact that personalized diffusion will always have at least $1 - \alpha$ probability centered at each vertex, so the largest difference between two personalized pagerank vectors is 2α and not 2.

The element-wise similarity scores provide insight into how the clusterings differ. The ranked-distribution of element-wise scores reflects the differences in cluster structure. A flat distribution occurs when all elements have the same similarity score. This suggests that all elements saw an roughly equal change in cluster structure. A highly skewed distribution occurs when some elements have much higher or lower similarity than the rest of the elements. This suggests that the two clusterings had their agreements and disagreements concentrated on this small set of elements.

The final element-centric similarity $S(\mathcal{A}, \mathcal{B})$ of two clusterings \mathcal{A}, \mathcal{B} is the average of the element-wise similarities.

$$S(\mathcal{A}, \mathcal{B}) = \frac{1}{N} \sum_{i=1}^N S_i(\mathcal{A}, \mathcal{B}) \quad (\text{S24})$$

S3.5 Element-centric similarity for strict partitions

When the clustering is a strict partition (a clustering without overlapping memberships or hierarchical structure), the calculation of the personalized pagerank matrix $\mathbf{\Pi}$ for the clustering can be analytically solved. First, note that in the absence of overlap and hierarchical structure, the element interaction graph of a clustering is clique graph where each connected component corresponds to a single cluster from the clustering and all edge weights are 1.0. For each element e_i , there is one cluster c_β of size $|c_\beta|$ to which e_i belongs, and the resulting peronsalized pagerank matrix has entries:

$$\mathbf{\Pi}_{ij} = (\alpha/|c_\beta| + (1 - \delta_{ij})(1 - \alpha)) \delta_{j \in c_\beta} \quad (\text{S25})$$

where δ is the Dirac delta function. Thus, the similarity of strict partitions has low computational overhead and can actually become much faster than traditional clustering similarity methods when many comparisons are made at once.

S3.6 Average agreement and frustration

Beyond a similarity measure between two clusterings, our element-centric similarity measure reveals how an arbitrary set of clusterings groups the elements. The *average agreement* between a reference clustering and a set of clusterings measures the regular grouping of elements with respect to a reference clustering. Specifically, given a clustering \mathcal{G} and a set of clusterings $\mathbf{R} = \{\mathcal{R}_1, \dots, \mathcal{R}_T\}$, the element-wise average agreement for element v_i is evaluated as:

$$\frac{1}{T} \sum_{j=1}^T S_i(\mathcal{G}, \mathcal{R}_j). \quad (\text{S26})$$

The *frustration* within a set of clusterings reflects the consistency with which elements are grouped by the clusterings. For the set of clusterings $\mathbf{R} = \{\mathcal{R}_1, \dots, \mathcal{R}_T\}$, the element-wise frustration for element v_i is given by:

$$\frac{1}{\binom{T}{2}} \sum_{j=2}^T \sum_{k=1}^{j-1} S_i(\mathcal{R}_k, \mathcal{R}_j). \quad (\text{S27})$$

S4 Comparing clustering similarity measures: extend discussion

In order to evaluate the behavior of clustering similarity measures, we introduced three comparison scenarios in which one clustering was held constant and the second clustering was randomly generated according to different constraints. These three scenarios specifically focused on the trade-offs made by clustering similarity measures when incorporating strong discrepancies in three aspects of the cluster structure: the consistent grouping of elements into clusters, the size distribution of the clusters, and the number of clusters. Here, we provide an extended analysis of the behaviors seen in the original three scenarios shown in Fig. 1 and introduce one additional comparison scenario.

For all of the scenarios, we consider a baseline clustering, named clustering \mathcal{A} , which contains 1,024 elements clustered into 32 clusters of equal size with no overlap. All results are averaged over 100 instantiations and reported with an error of one standard deviation.

In the first scenario, we compare clustering \mathcal{A} to a second clustering generated by randomly shuffling the membership of a fraction p of the elements in clustering \mathcal{A} , leaving the number and size sequence of the clusters unchanged. As expected, all clustering similarity measures decrease as p increases. The Jaccard index, F measure, NMI, and our element-centric similarity measure remain at a non-zero value reflecting the fact that even after all of the element memberships have been fully shuffled, there will be a fraction of the elements which are co-assigned to the same cluster in both clusterings—a property of all clusterings with a similar number and distribution of clusters. However, the adjusted Rand index and ONMI eventually reach a base value of 0. This is particularly noticeable in the case of ONMI which assess 0 similarity between the clusterings with only $p \approx 0.5$, losing the ability

to discern the similarity of clusterings with more randomization. The 0 base value for the adjusted Rand index reflects the underlying philosophy of the measure: Random clusterings should have a similarity of 0, regardless of the number of clusterings or cluster size sequence [22, 17].

In the second scenario, we explore the effect of a skewed cluster size sequence. For this case, we compare clustering \mathcal{A} to a second clustering \mathcal{B} generated using a preferential attachment model of element assignment. Specifically, using a clustering with 32 equal sized clusters (and randomized element memberships compared to clustering \mathcal{A}) as the seed, at each step of our algorithm, a random element is uniformly chosen for reassignment to a new cluster based on the current sizes of those clusters. A move is rejected if it resulted in an empty cluster. The shuffling procedure is run for a total of 10^4 steps and the subsequent samples from all 100 trials are grouped into 40 bins according to their clustering entropy. There are now three distinct types of behaviors exhibited by the clustering similarity measures. The NMI and our element-centric similarity measure exhibit the intuitive behavior and decrease as the clustering entropy decreases. The ONMI and ARI maintain a 0 similarity for all comparisons regardless of the clustering entropy. Note, the larger variation in NMI, ONMI, and ARI seen for small basin entropy results from the presence of singleton and binary clusters which contribute to statistical fluctuations in element memberships. Finally, the F measure and Jaccard index increase as the entropy decreases: They cannot account for the differences in the cluster size distribution. This increase is a consequence of their formulation in terms of the correctly co-assigned element pairs while disregarding the incorrectly co-assigned element pairs.

In the third scenario, we explore the effect of the number of clusters. Here, we compare clustering \mathcal{A} against a second clustering \mathcal{B} generated by randomly assigning the elements to c regularly sized clusters, where c is the control parameter for the scenario. Hence, one clustering remains the same size, while the other has c regularly sized clusters. Again, we see two distinctly different behaviors of the clustering similarity measures: the Jaccard index, F measure, ONMI, ARI and our element-centric similarity measure all follow our intuition and decrease with increasing c , while NMI increases with increasing c . The increasing behavior for NMI can be attributed to the information-theoretic bias towards comparisons with more clusters [14, 66, 16, 17]. This bias makes NMI a particularly troubling measure for hierarchical clusterings where we expect the number of clusters to vary over several orders of magnitude.

Finally, we introduce one additional scenario, depicted in Figure S1, in which both clustering \mathcal{A} and clustering \mathcal{B} are generated by randomly assigning the elements to c regularly sized clusters. This scenario prominently demonstrates the trade-offs with which a clustering similarity measure must contend. Namely, in this scenario, the effect of randomized element memberships is opposed by the increasing similarity of the number of clusters and cluster size sequences. This trade-off is clearly illustrated by the behavior of our element-centric similarity measure; the initial decrease, resulting from the increasingly random element memberships, is eventually overcome by the relative similarity in the number and sizes of the clusters. Eventually, the similarity reaches the expected value of 1 when there are

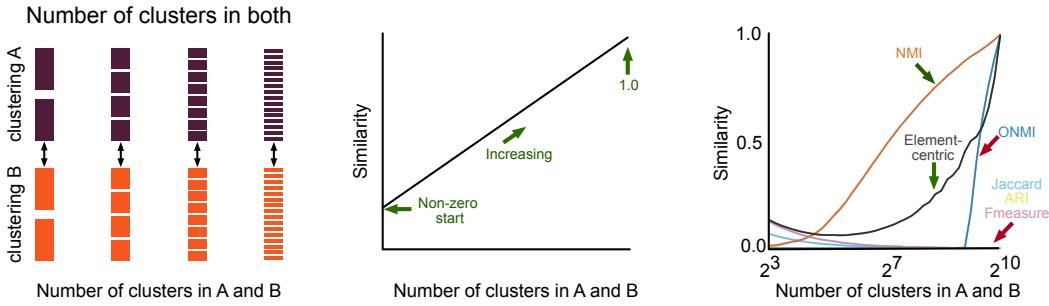


Figure S1: A fourth scenario demonstrates the trade-offs between element randomization, cluster size sequence, and the number of clusters. Two clusterings with random element memberships into $2^3 < c < 2^{10}$ non-overlapping and equal-sized clusters for different values of c .

2^{10} clusters—every element is placed within a singleton cluster in both clusterings and the randomization of element memberships has no effect. In contrast, NMI always increases as the number of clusters increases suggesting the aforementioned bias towards clusterings with more clusters is always stronger than the effect of element randomization. Once again, the extreme behavior of ONMI can be seen when the measure jumps to a similarity of 1 at 2^{10} clusters. The decreasing behavior for the Jaccard index and F measure results from their scaling behavior—when the number of clusters is large relative to the number of elements, there are very few elements co-occurring in each cluster.

S5 Clustering similarity applications

S5.1 Functional brain networks

S5.1.1 Dataset

The dataset used here was originally collected for Cheng et al. [32]; please refer to that work for specific details of the data acquisition and pre-processing, here we only offer a brief overview.

Data was acquired from 19 individuals diagnosed with schizophrenia (6 female, mean age 33.1 ± 10.9 years) and 29 healthy controls (15 female, mean age 28.1 ± 8.4 years). Diagnosis of schizophrenia was based on the Structured Clinical Interview for the DSM-IV Axis I Disorders (SCID-I) [67] and medical chart review. All subjects were scanned on a Siemens TIM Trio 3 T MRI scanner using a 32-channel head coil. The high anatomical scan had a resolution of 1 mm^3 . A total of 200 volumes of resting state fMRI data were acquired with EPI sequences for 8 min and 20s. During the resting state fMRI scan, the subjects were at rest with eyes closed and instructed not to think of anything in particular. All functional data were motion corrected in FSL.

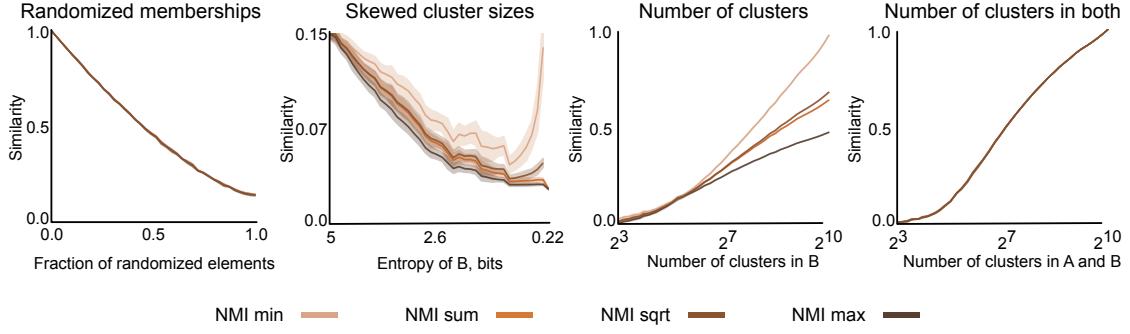


Figure S2: NMI's bias towards the number of clusters is independent of normalization term. The three scenarios from the main text, and one additional scenario described in Figure S1 for different normalization terms of NMI: the minimum of cluster entropies (min), the average of the cluster entropies (sum), the geometric mean of the cluster entropies (sqrt), and the maximum of the cluster entropies (max). See Section S2.5 for the measure details.

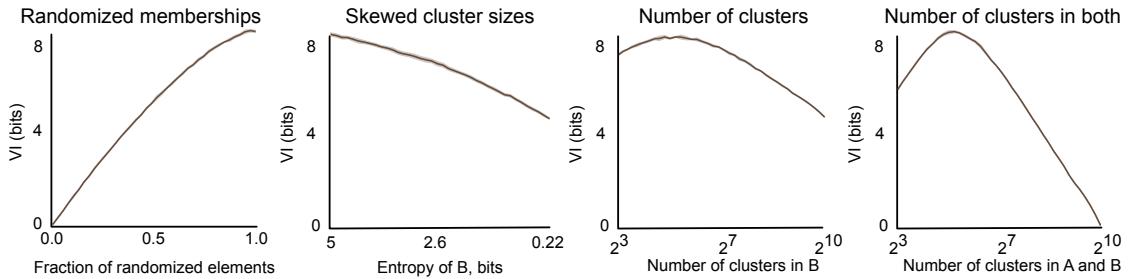


Figure S3: The VI displays counter-intuitive behavior for skewed cluster sequences and differing number of clusters. The three scenarios from the main text, and one additional scenario described in Figure S1. Since the VI is a metric, the intuitive behavior differs from the similarity measures discussed in the main text; one would now expect the measure to increase in **a-c** and decrease in **d**. See Section S2.7 for the measure details.

In conjunction with the anatomical image, the functional images were parcellated using a parcellation scheme proposed by Shen et al. [68]. This parcellation divides the cerebral cortex into 278 regions of interest (ROIs), and was derived from resting state functional data of the healthy subjects by maximizing functional homogeneity within each ROI. After regressing out head motion, the time signal was band-pass filtered between 0.01 – 0.10 Hz and the time courses were extracted from the 278 brain ROIs as the average over voxels.

The functional network was computed from the wavelet coherence between all pair-wise combinations of ROIs, giving rise to a square symmetric matrix (278×278). The resulting functional connectivity matrix has only positive edges. In order to identify a backbone network structure, the multiscale network backbone [69] was extracted using an alpha of $\alpha = 0.2$. Technically, the multiscale backbone is a directed network, however, since our original graph was undirected, we convert the multiscale backbone back into an undirected network. The network was not corrected to insure a single connected component.

S5.1.2 Overlapping and hierarchically structured clusterings

Overlapping and hierarchically structured clusterings were derived using Order Statistics Local Optimization Method (OSLOM) network community detection [34] with the following parameters: weighted, undirected edges, $p = 0.1$, 100 runs for the detection at the bottom of the hierarchy and 1000 runs for the detection at the top of the hierarchy. All singlet communities were kept. Due to the variability in clustering structure between runs of the algorithm, 10 clusterings were extracted for each patient.

The subject similarity matrix was then constructed as follows. The similarity of each diagonal entry is 1.0. Each off-diagonal entry in the (48×48) subject similarity matrix is the average element-centric similarity of all comparisons $10 \times 10 = 100$ between the 10 OSLOM communities uncovered for each subject. For all comparisons, we set $\alpha = 0.9$ and $r = 8.0$. Our choice of the scaling parameter, $r = 8.0$, was grounded in the explorations of synthetic binary hierarchies of equivalent height. The dis-similarity matrix is one minus the similarity matrix. Four additional matrices were found by using the community structure found by slicing each OSLOM community dendrogram and retaining only the bottom or top communities and performing all pair-wise comparisons with either the element-centric similarity or ONMI similarity measures.

S5.1.3 Classification

Given a dis-similarity matrix, a distance weighted k-Nearest Neighbors (kNN) classifier was trained using nested and stratified 10-fold validation [70]. Specifically, the data was randomly split into 10 groups such that the proportions of each class were kept relatively equal in each group. Each group in turn was then used as the testing set, while the other 9 groups formed the training set. For each training set, we first find the best k for the kNN classifier using a grid search for k between 1 and 15 and another stratified 10-fold validation. The classifier was then retrained on the entire training set for the specified k . Finally, the accuracy of the trained classifier was found on the testing set. In the paper, we report the average accuracy

identified in 100 random initializations of the nested 10-fold validation technique [71, 72].

S5.2 Point clusters

5,000 points were randomly formed into clusters in an algorithm akin to the process for constructing benchmark graphs [46]. Cluster sizes were randomly drawn from a powerlaw distribution with a minimum cluster size of 10, a maximum cluster size of 1000, and an exponent of 1.0. The center of those clusters was uniformly selected from points in a 40×40 box. The standard deviation (or spread) of each cluster was also drawn from a powerlaw distribution with a minimum of 0.2, a maximum of 2.0, and an exponent of 1.0. Next, the type of each cluster was uniformly selected from four options. The first option is the 2-D Gaussian blob with mean given by the cluster center and standard deviation given by the cluster standard deviation. The second option is the 2-D Anisotropic blob with a mean given by the cluster center, standard deviation given by the cluster standard deviation, and transformation given by the rotational matrix:

$$\begin{bmatrix} a \cos(\theta) & -a \sin(\theta) \\ b \sin(\theta) & b \cos(\theta) \end{bmatrix}, \quad (\text{S28})$$

where a, b randomly drawn from the unit interval and θ was randomly drawn from the range $[0, \pi]$. The third option is the circle centered at the cluster center with radius given by the cluster standard deviation; the points were uniformly spread along the circle and Gaussian noise with mean 0 and standard deviation 0.2 was added to all points. The forth option is the spiral with points uniformly spread in the range $[0, 10]$, converted to circular coordinates by $(x, y) \rightarrow (\sigma\sqrt{x} \cos(x), \sigma\sqrt{y} \cos(y))$, where σ is the cluster standard deviation, randomly rotated by the rotation matrix of equation (S28) with $a = b = 1$ and θ randomly drawn from the range $[0, \pi]$, and Gaussian noise with mean 0 and standard deviation 0.2 was added to all points.

The sci-kit learn [73] implementation of K -means clustering was initialized with $K = 19$ clusters and random initial centroids. The identification method was then run from 100 random centroid initializations. Clustering agreement was calculated by comparing all 100 uncovered clusterings with the ground-truth clustering using the element-wise similarity vector was found for each comparison and then averaged over the uncovered clusterings. Clustering frustration was calculated from all pair-wise comparisons between the 100 uncovered clusterings using the element-wise similarity vector was found for each comparison and then averaged over each comparison.

S5.3 Handwriting digits

The digits data set is bundled with the sci-kit learn source code and consists of 1797 images of 88 gray level pixels for handwritten digits. The reference clustering contains 10 clusters corresponding to the true digit. The data set was originally assembled by Alimoglu

and Alpaydin [27]. To provide a visualization, the data was projected to 2-d using the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction method [74] initialized from the pca decomposition.

The sci-kit learn [73] implementation of K -means clustering was initialized with $K = 10$ clusters and random initial centroids. The identification method was then run from 100 random centroid initializations. Clustering agreement was calculated by comparing all 100 uncovered clusterings with the ground-truth clustering using the element-wise similarity vector was found for each comparison and then averaged over the uncovered clusterings. Clustering frustration was calculated from all pair-wise comparisons between the 100 uncovered clusterings using the element-wise similarity vector was found for each comparison and then averaged over each comparison.

S5.4 Facebook friendship networks

The Facebook friendship networks analyzed here were originally released as part of the the Facebook 100 data set [28, 29]. This data set contains a snapshot of all friendships at each of 100 schools in the fall of 2005. Additionally, the data includes several categorical variables volunteered by the users on their individual pages: gender, class year, high school, major, and dormitory residence. Here, we only analyze the networks corresponding to two schools: the Oberlin (College A) and Rochester networks (College B). For each school we took the largest connected component. The extracted clusterings were uncovered using the Louvain method, a optimization scheme that identifies clusters with high Newman’s modularity [75, 76]. The categorical data for year, dorm and major were used to create three non-overlapping clusterings. Every student with missing categorical data was placed into an individual singleton cluster.

Comparisons between the categorical clusterings and the extracted clustering were made using our element-centric similarity measure. For both schools, there is a high similarity to year, confirming previous results [28, 29]. The element-wise similarity scores indicate that this similarity is strongest for the first-year students, and fails to capture the clustering structure of other students (original text, Fig. 4 h,i black arrows). Those regions with low similarity to year actually have higher similarity to major.