The metric backbone preserves community structure and is a primary transmission subgraph in contact networks

Rion Brattig Correia[1,2,6], Alain Barrat[3,4] and Luis M. Rocha[1,5,6,*]
[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal
[2]CAPES Foundation, Ministry of Education of Brazil, Brasília DF, Brazil
[3]Aix Marseille Univ, Université de Toulon, CNRS, CPT,
Turing Center for Living Systems, Marseille, France
[4]Tokyo Tech World Research Hub Initiative (WRHI),
Institute of Innovative Research, Tokyo Institute of Technology, Japan
[5]Department of Systems Science and Industrial Engineering,
Binghamton University, Binghamton NY, USA
[6]Center for Social and Biomedical Complexity, Indiana University, Bloomington IN, USA
[*]rocha@binghamton.edu

Abstract

The structure of social networks strongly affects how different phenomena spread in human society, from the transmission of information to the propagation of contagious diseases. It is well-known that heterogeneous connectivity strongly favors spread, but a precise characterization of the redundancy present in social networks and its effect on the robustness of transmission is still lacking. This gap is addressed by the recently introduced metric backbone, a subgraph that is sufficient to compute all shortest paths of weighted graphs. We show that the metric backbones of nine contact networks obtained from proximity sensors in a variety of social contexts are generally very small, 49% of the original graph for one and ranging from about 6% to 20% for the others. This reflects a surprising amount of redundancy and reveals that shortest paths on these networks are very robust to random attacks and failures. We show that the metric backbone preserves the community structures of the full contact networks and is a primary subgraph in epidemic transmission. The metric backbone, thus, is a principled graph reduction technique that does not affect shortest path connectivity for any nodes in a network. It is an important subgraph to characterize epidemic spread, the robustness of social networks, and any communication dynamics that depend on complex network shortest paths.

keywords: complex networks, network backbone, community structure, epidemic spreading, contact networks.

1 Introduction

As the current COVID-19 pandemic illustrates, our social lives and overall public health depend heavily on interactions that scale from the molecular networks of minute pathogens to the complex socio-technical networks of our transportation, health, economy, ecology, and governance systems. Many insights into the organization of such systems come from advances in network science based on the study of patterns of connectivity (network structure), and one of the fields in which network science has led to the most concrete advances is the epidemiology of infectious diseases [1]. In disease propagation, the structure of the contact network plays a crucial role, and networks with heterogeneous connectivity strongly favor spread [2]. Much remains to be understood, however, about how the structure of complex networks affects their dynamics and robustness [3, 4, 1]. For instance, the collection of social behavior data from social media and mobile devices allows us to map the structure of social interaction to understand health and disease in an unprecedented

manner [5, 6]. Yet, much remains to be understood about the dynamics of transmission on these networks for us to be able to predict and control specific biomedical phenomena, such as epidemics, affecting society.

Weighted graphs, where every edge is enriched with a positive real number, are often used to capture distance associations between linked nodes within a set of node variables. This type of network is very useful to infer social dynamics in real-world settings and often used in epidemiology to model disease spread and evaluate disease containment policies at various scales [7, 8, 9, 1]. Here we address the link between structure and dynamics by focusing on important *patterns of redundancy* in social networks modeled as weighted graphs.

Redundancy is considered a primary factor in the evolvability of complex systems [10], and recent characterizations of the phenomenon have shown it greatly contributes to their dynamics, controlability, and robustness [11, 12, 13]. A full understanding of the interplay between network structure and dynamics requires a study of multivariate dynamics [12] and its redundancy [14]. Often, however, we do not possess enough time-resolved data or computational power to precisely characterize the multivariate dynamics of large networks. In these cases, network structure is very useful for understanding the dynamics of spread and communication phenomena, which can be inferred from the shortest paths and community structure among the variables. We know complex networks generally contain substantial connectivity redundancy whereby all shortest paths can be computed from the smaller subgraph of the recently introduced *metric backbone* (known in general for any measure of path length as the distance backbone) [13].

Weighted graphs inferred from real-world data can be large and dense, and thus, cumbersome to manipulate [15]. To overcome this problem, several methods have been proposed to extract network backbones by removing weak, non-significant, or redundant edges. Methods that remove edges based on a distance (or strength) threshold tend to disrupt global network connectivity by altering the distribution of shortest paths and possibly creating disconnected components (islands) [16]. Indeed, even very weak edges may be very important for computing shortest paths [17] and maintaining multiscale organization (including community structure) in complex networks [18]. Other methods to extract backbones are based on network properties (e.g., degree, betweenness, and effective resistance) [19, 20] or comparison with an expected connectivity distribution (i.e., a null model) [18, 15, 21, 22, 23, 24]. All methods, ultimately, remove edges (and potentially nodes) based on thresholding edge weights (retaining only the edges with a proximity weight larger than a given value) or comparing to a null-model distribution. Thus, in either case there is an arbitrary parameter that tunes the removal of edges (and nodes).

In contrast, the metric backbone is a parameter-free, principled method to obtain a subgraph that fully preserves the shortest paths of the original graph[1][13]. It is typically a subgraph much smaller than the original network across domains ranging from topical spaces of large document corpora [25, 26, 27] to the brain connectome and functional networks [13, 28, 29, 30]. For instance, the metric backbone of a knowledge graph of more than 3 million concepts extracted from Wikipedia and used for automated fact-checking [31] contains only 2% of the original edges but is sufficient to compute all shortest-paths of the original graph [13].

The small relative size of the metric backbone reveals that network shortest-path robustness to attacks and failures likely stems from surprisingly vast amounts of redundancy [13]. Importantly, other backbone methods end up removing edges that are not redundant for shortest paths. Even the *disparity filter backbone* [18], which has been proposed to preserve the multiscale structure of complex networks, alters the distribution of shortest paths, overall connectivity, and can remove nodes [13]. For instance, the disparity filter backbone of an air traffic network of over 1000 U.S. airports is composed of 24% of the original edges, but it alters the shortest path distribution and removes 23% of all nodes [18][2]. In contrast, the metric backbone of the same airport traffic network is much smaller, composed of 16% of the original edges, and keeps all node variables with the same connectivity and shortest path length distribution [13]. Moreover, the metric backbone is parameter free (see section 2) while the disparity filter backbone depends on a significance level parameter in comparison to a null model distribution to remove edges.

[1] No shortest path is affected by the metric backbone. In particular, even the distance between nodes directly linked by a removed edge is not affected as the shortest path between these nodes is then an indirect path via other nodes (edges not on the backbone break the triangle inequality; see [13] for details).

[2] For $\alpha = 0.2$ the disparity filter backbone of the airport network is composed of these proportions of edges and nodes, where α is the p-value threshold for the null model-derived normalized-edge-weight distribution. For $\alpha = 0.05$, the disparity filter backbone is composed of 17% of the original edges but also removes 34% of the original nodes (see [18] for details).

Here we present the metric backbones of nine networks obtained by measuring the contacts between pairs of individuals (using wearable sensors) in a variety of social settings [32, 33, 34]. Such networks are relevant in behavioral studies [35] and as input to data-driven numerical simulations of epidemic spread [36, 9, 34]. We already know that the metric backbone preserves all shortest paths in these networks and removes edges truly redundant for this purpose [13]. Here we show that in contact networks from various social contexts, the metric backbones: (a) are much smaller than the original networks (which have a lot of redundancy), (b) preserve community structure, and (c) are primary subgraphs for simple transmission dynamics.

2 The metric backbone of contact networks

Social contact networks are built by recording when individuals meet. The data we consider (see section 5) describe close proximity events in populations with a finite temporal resolution, i.e., in successive time windows of approximately 20 seconds [37, 32, 33, 34]. These data can be represented as graphs, $R(X)$, in which each node $x_i \in X$ represents a person in a population X, and an edge exists between two nodes if the corresponding individuals have been in close contact at least once. The number of time windows in which individuals x_i and x_j were in close proximity is denoted by an entry r_{ij} in the graph adjacency matrix. The diagonal entries of this matrix, r_{ii}, denote the total number of time windows in which individual x_i was in a close social interaction with any other individual in the population X: $r_{ii} = \sum_{x_j \in X: j \neq i} r_{ij}$. A normalized transformation of this data, used to quantify the strength of association between individuals, can be obtained via the Jaccard measure [38, 39]:

$$p_{ij} = \frac{r_{ij}}{r_{ii} + r_{jj} - r_{ij}}, \quad \forall x_i, x_j \in X, \tag{1}$$

where $p_{ij} \in [0, 1]$ denotes a *proximity* of two individuals, with $p_{ij} = 0$ for individuals x_i and x_j have no contact, and $p_{ij} = 1$ when they are both active in the same time windows and always in contact; naturally, $p_{ii} = 1$. The *proximity graph*, $P(X)$, can be produced by measures other than Jaccard's as long as the proximity strength is symmetrical and proportional to the intensity of the interaction [27]. Two additional forms of normalizing interactions are described in SM, section S1.

Because computing shortest paths requires a measure of length, rather than proximity, for the edges we use *distance graphs* $D(X)$ obtained via the nonlinear map φ[3]:

$$d_{ij} = \varphi(p_{ij}) = \frac{1}{p_{ij}} - 1 = \frac{r_{ii} + r_{jj} - 2r_{ij}}{r_{ij}}, \quad \forall x_i, x_j \in X, \tag{2}$$

where the resulting distance weights are symmetrical and inversely proportional to the intensity of social interaction, with $d_{ij} = +\infty$ for individuals x_i and x_j who have no contact, $d_{ij} = 0$ when they are simultaneously active and always in contact, and $d_{ii} = 0$.

The *metric backbone* [13] of a distance graph $D(X)$ is defined as its invariant subgraph $B(X)$ in the computation of the graph's *metric closure* $D^T(X)$. The metric closure is the graph obtained after computing the shortest paths between all pairs of nodes of the distance graph and replacing the original distance edges d_{ij} with the *length of the shortest path* between x_i and x_j. Length is computed by summing the edge weights in the (shortest) path via δ indirect (non-repeating) nodes, where δ is no larger than the diameter of the graph[4], thus, $d^T_{ij} = \ell_{ij} = d_{ik_1} + d_{k_1k_2} + \ldots + d_{k_\delta j}$. The edge weights of the metric backbone graph are given

[3] Other maps are possible, but without loss of generality we use the simplest nonlinear isomorphism possible between proximity and distance graphs [27].

[4] The metric closure is one of infinite possible distance closures that are isomorphic to transitive closures in generalized probabilistic or fuzzy metric spaces. In this case, the topological space is defined by the algebraic structure $([0, +\infty], \min, +)$, where $([0, +\infty], \min)$ and $([0, +\infty], +)$ are monoids that enforce shortest paths with lengths computed by summing edge weights. There are many other ways to compute length besides summing edges [27] that lead to many other meaningful distance backbones [13]. However, the metric closure, or the *All Pairs Shortest Paths* (APSP) problem [40] is the most common in network science and typically computed via Dijkstra's algorithm [41].

by:

$$b_{ij} = \begin{cases} d_{ij}, & \text{if } d_{ij} = d^T_{ij} \\ +\infty, & \text{if } d_{ij} > d^T_{ij} \end{cases}, \forall x_i, x_j \in X, \tag{3}$$

where $b_{ij} = +\infty$ means that there is no direct edge between x_i and x_j in the distance backbone graph.

The edge weights of $D(X)$ that do not change after computation of the metric closure $D^T(X)$ are called *metric* because they obey the *triangle inequality*:

$$d_{ij} \leq d^T_{ik} + d^T_{kj}, \forall x_i, x_j, x_k \in X. \tag{4}$$

The edge weights that become smaller with the metric closure break the triangle inequality in $D(X)$ and are not included on the backbone. When an edge d_{ij} of $D(X)$ breaks the triangle inequality, it means that the *length* of at least one indirect path between x_i and x_j is shorter than the direct distance: $\ell_{ij} = d^T_{ij} < d_{ij}$. These are known as *semi-metric edges* [25] and do not contribute to any shortest path [27]. Thus, metric edges alone define the backbone and are sufficient to compute the closure, as shown in Simas et al. [13].

The *proportion of semi-metric edges* is, therefore, the proportion of edges in graph $D(X)$ that are not necessary to compute shortest-paths. This measure of *edge redundancy* is given by:

$$\sigma(D) = \frac{|\{d_{ij} : d_{ij} > d^T_{ij}\}|}{|\{d_{ij}\}|}, \forall x_i, x_j \in X : i > j. \tag{5}$$

Similarly, the *proportion of metric edges* in graph $D(X)$ is the *relative size of its metric backbone* $B(X)$:

$$\tau(D) = \frac{|\{d_{ij} : d_{ij} = d^T_{ij}\}|}{|\{d_{ij}\}|} = \frac{|\{b_{ij}\}|}{|\{d_{ij}\}|}, \forall x_i, x_j \in X : i > j. \tag{6}$$

It follows that $\tau = 1 - \sigma$. Because distance graphs are symmetric ($d_{ij} = d_{ji}$) and edges are nondirected, in formulae 5 and 6 we count each edge only once and do not tally reflexive edges, d_{ii}. This means we tally only the lower diagonal entries of the adjacency matrix: $d_{ij} : i > j$.

The metric closure, computation of all pairs shortest paths (APSP), induces a topological *distortion* [27] of the original graph obtained from the multivariate associations observed in the social contact data, whereby semi-metric edges are made to conform to the triangle inequality (eq. (4)). However, only the semi-metric edges get distorted; the metric edges and the metric backbone they compose remain invariant in the APSP. Therefore, σ also denotes the proportion of edges topologically distorted by the metric closure, whereas τ denotes the proportion of edges topologically invariant under the metric closure. A measure of *semi-metric edge distortion* between nodes x_i and x_j is then obtained via the ratio of the direct distance over the shortest indirect path length.:

$$s_{ij} = \frac{d_{ij}}{d^T_{ij}}, \forall x_i, x_j \in X : i \neq j. \tag{7}$$

If an edge d_{ij} is metric, $s_{ij} = 1$, meaning there is no distortion. If an edge d_{ij} is semi-metric, $s_{ij} > 1$, meaning there is distortion; the larger the value over 1, the more the edge breaks the triangle inequality, and thus, the more distorted it is in the metric closure. While semi-metric edges are redundant for shortest paths and have null betweenness centrality, their distortion (the distribution of s) varies widely. This in turn affects the robustness of shortest paths to edge removals [13] (see fig. 4 and section 4).

3 Results

We have computed the metric backbone for nine different contact networks from a variety of social contexts, from an elementary school in Utah (USA) to an art exhibit in Dublin (Ireland). As shown in table 1, these networks vary widely in the number of nodes and edges, yet the metric backbone typically comprises a small proportion of edges.

Network (D)	Location	Social context	$\lvert X\rvert$	$\lvert d_{i>j}\rvert$	$\tau(D)$	$\sigma(D)$
Fr-Ho [42]	Lyon, France	Hospital	75	1,139	19.05	80.95
It-SC [43]	Turin, Italy	Scientific Conference	113	2,196	14.03	85.97
Ir-Ex [43]	Dublin, Ireland	Exhibit	159±63	645±468	48.44±9.27.	51.56±9.27.
Fr-Wo [44]	Paris, France	Workplace	232	4,274	17.43	82.57
Fr-PS [45]	Lyon, France	Primary School	242	8,317	9.5	90.5
Fr-HS [46]	Marseille, France	High School	327	5,818	10.36	89.64
US-ES [34]	Utah, USA	Elementary School	339	16,546	6.82	93.18
US-MS [34]	Utah, USA	Middle School	591	56,867	6.19	93.81
US-HS [33]	USA	High School	788	118,291	7.84	92.16

Table 1: Distance graphs (D) of the social contact networks analyzed, with respective references. $|X|$: number of nodes; ($|d_{i>j}|$): number of finite distance edges; $\tau(D)$ relative size of the metric backbone; $\sigma(D)$: edge redundancy. Values of τ, and σ are shown as percentages (%). For the Exhibit networks, values of $|X|$, ($|d_{i>j}|$), τ and σ denote the mean $\pm$ standard deviation of the networks computed for each of the 69 days for which data were gathered. See section 5.1 for additional details and a description of the networks.

Our metric backbone analysis is exemplified by the contact network collected by the SocioPatterns collaboration [37] over a period of 4 days in 2013 in a High School (Fr-HS) in Marseilles, France [46]. A similar analysis of all other networks is provided in Supplemental Materials (SM). The Fr-HS contact network was built from 188,508 contact records among $|X| = 327$ students enrolled in distinct *préparatoires* classes, which are designed for college-bound students in the last two years of their high school studies. This student body had been largely separated from other high school students (i.e., different building and lunch), forming an almost closed population with few contacts with the outside world, at least during school days. Nine classes were organized in four different specializations: "MP" focus on mathematics and physics (3 classes), "PC" on physics and chemistry (2 classes), "PSI" on engineering (1 class), and "BIO" on biology (3 classes). The total student participation in the study was 86.3% [46].

The distance graph $D(X)$ and its metric backbone $B(X)$ have been obtained for the Fr-HS data via eqs. (2) and (3) in section 2 and are depicted in fig. 1, with other processing details described in section 5.1.

3.1 The metric backbone preserves community structure

As shown in table 1, the Fr-HS metric backbone comprises only $\tau \approx 10\%$ of the edges of the original network. Interestingly, in addition to preserving all shortest paths by design [13], the metric backbone also preserves and highlights the community structure of the original social network. This is clearly seen in fig. 1A-C when we compare its center (B) and right (C) panels. The metric backbone depicted in fig. 1B the nodes are placed according to the computation of the `ForceAtlas2` algorithm [47] (in `Gephi`[48]) on the original distance graph (shown in fig. 1A). In contrast, in fig. 1C the nodes are placed after recomputing the same algorithm using only the backbone subgraph. Interestingly, we observe that the community structure remains largely unaltered with respect to the fig. 1B panel.

Similar results have been observed for all other networks in table 1 (see section S2), with the interesting exception of the American high school (US-HS) contact network. In this case, the original distance graph is devoid of any obvious social structure. After edge removal, however, four distinct communities that likely correspond to the four grades typical of American high school education can clearly be seen (see fig. 1D-F). In this larger network ($|X| = 788$), removing the $\sigma \approx 92\%$ of edges that are redundant reveals a clear community structure in the metric backbone graph. Moreover, while the metric backbone comprises only $\tau \approx 8\%$ of the edges, it also preserves all of the original shortest paths.

It is useful to quantify how well the metric backbone preserves or highlights the social organization of contact networks beyond visual inspection. While visual representation of networks is an important part of social network analysis, especially for small- or mid-size networks, visual representations and inspections become increasingly difficult as the number of nodes and edges grows. Indeed, it is well known that the task of clustering of nodes into communities is extremely challenging [49].

We have developed and used a set of measures to compare the community structure of each distance graph using its backbone and same-size threshold and random subgraphs (see section 5.2 for details). Because all measures support the same conclusions, in the main text we present the results for the *bidirectional similarity*

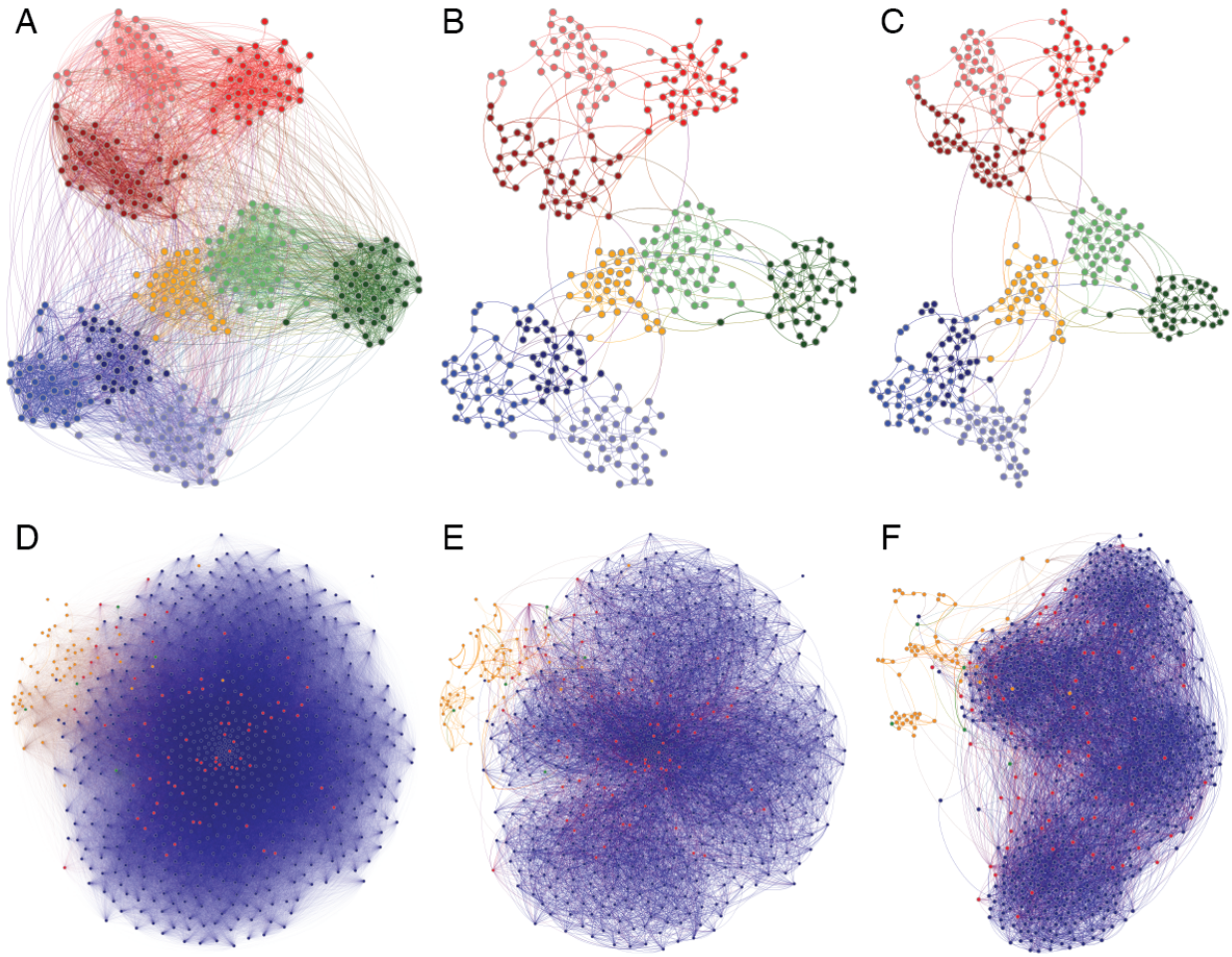

Figure 1: (**A-C**) French high school (Fr-HS) contact network of $|X| = 327$ students. Colors represent the four student specializations: "MP" in blue, "PC" in green, "PSI" in orange, and "BIO" in red; lighter or darker colors separate the distinct classes within each specialization. (**D-F**) American high school (US-HS) contact network of $|X| = 788$ students, staff, and teachers. Colors represent students in blue, teachers in red, and staff in orange. Other (unspecified) individuals are shown in green. No student class metadata is available. (**A, D**) Distance graphs $D(X)$ with node layout computed by `ForceAtlas2` algorithm [47]. (**B, E**) Metric backbones $B(X)$ rendered with the same node layout as computed for $D(X)$. (**C, F**) $B(X)$ with recomputed node layouts for backbone subgraphs. Plotted with `Gephi` [48].

measure, y_{AB}, as defined by eq. (8) (section 5.2), focusing on the Fr-HS social contact network, as shown in table 2. Results for all measures and all other networks are shown in section S2.

The Fr-HS dataset includes $m = 9$ classes in the metadata (see fig. 1A-C), while the `Louvain` community detection algorithm [50] identifies $m = 10$ communities in the original distance graph, showing the additional very small community splitting from one of the MP classes, as shown in fig. 2A. In other words, the community structure of the social contact data represented by the distance graph, as captured by `Louvain`, is very similar to the metalabel communities. Indeed, $y_{AB} = 0.88$ measures a high amount of bidirectional similarity between the two module partitions[5]. This confirms that the social contacts of students tend to occur more frequently within their assigned classes and a few small groups of students interact more with students from other classes/modules, though almost always within the same specializations, as shown in fig. 2A.

We do not have access to a "gold standard," such as an independent survey of the observed cohorts, that shows the true community structure of the contact networks we analyse. We only have access to either the metalabels provided in the data (e.g., the classes each student of the Fr-HS network is enrolled in), or the

[5]The larger the value of $y_{AB} \in [0, 1]$, the more similar the community structure between the two graphs defined on the same set of nodes, see section 5.2.

Table 2: Community structure analysis of the SocioPatterns Fr-HS contact network. Rows show values for number of communities, m, and bidirectional modularity similarity, y_{AB} (eq. (8) in section 5.2). Columns show values for original distance graph, $D(X)$, metric backbone subgraph, $B(X)$, and same-size threshold and random subgraphs. Mean and standard deviation shown for 100 random subgraphs.

		$D(X)$	$B(X)$	Threshold	Random
m	Metalabels	9	-	-	-
	`Louvain`	10	17	27	47.19±3.79
y_{AB}	Metalabels	0.88	0.68	0.55	0.38±0.02
	$D(X)$	-	0.74	0.61	0.37±0.02

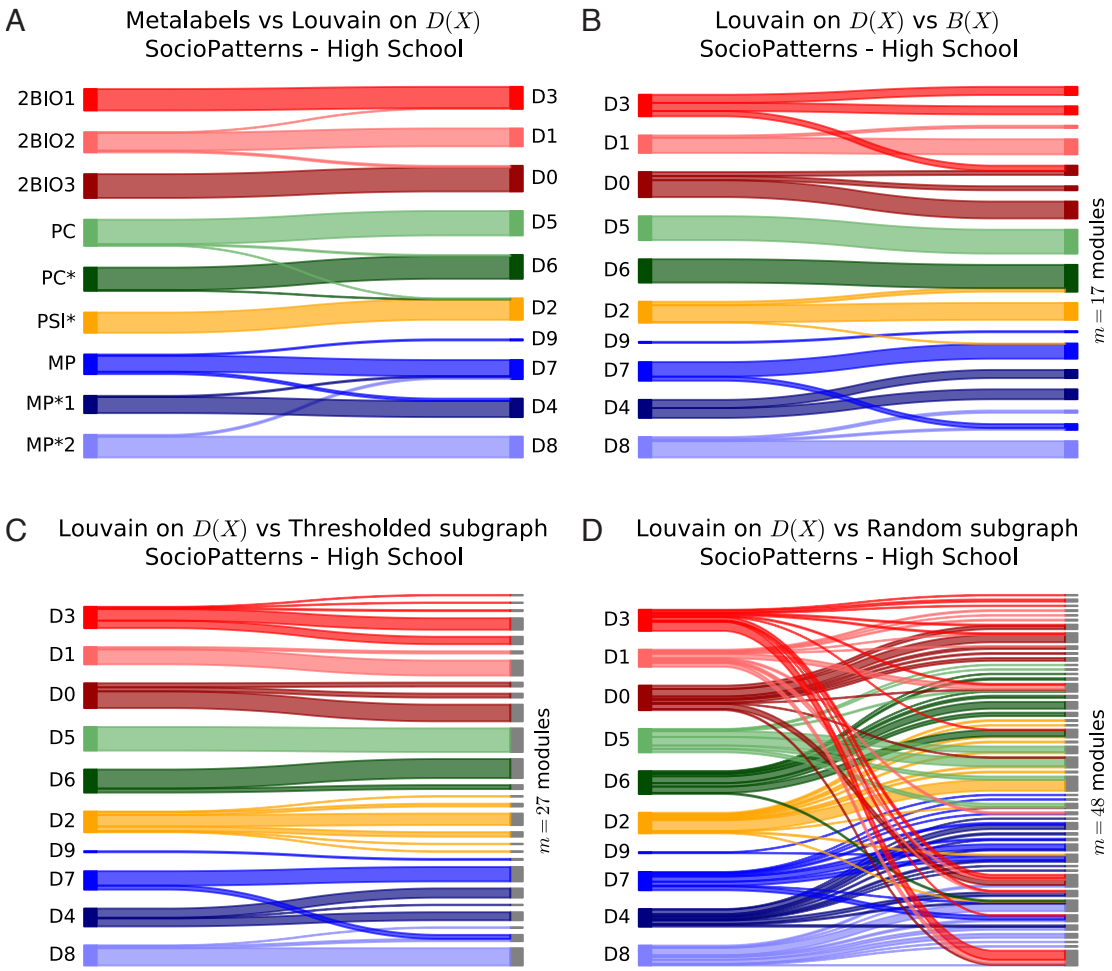

Figure 2: Modularity comparison for the French high school (Fr-HS) dataset using Sankey plots. (**A**) Comparison of original metalabels—student attribution to their classes—with the modules obtained with the distance graph, $D(X)$. (**B**) Comparison of the modules obtained with the distance graph, $D(X)$, and the metric backbone, $B(X)$. (**C**) Comparison of modules obtained with the distance graph, $D(X)$, and the threshold subgraph. (**D**) Comparison of modules obtained with the distance graph, $D(X)$, and a random subgraph realization. Metalabel module colors assigned to student specialization and classes as in fig. 1A-C. Modules computed using the Louvain algorithm [50].

social organization uncovered by community detection algorithms, such as the `Louvain` algorithm applied to the original network. It is important to note that neither is guaranteed to perfectly characterize the true social organization. While the metalabels denote organizational roles (e.g., classes and professions), individuals may contact members of other (metalabel) groups more than others within their own group due to personal relationships outside of their organizational roles. Additionally, the original network may contain

spurious contact edges, especially if the definition of contact from the sensors' properties do not correspond to close proximity, and if contacts with very short duration can be registered. The social organization of the original network as captured by community detection algorithms may thus be confounded by many contact edges that do not in reality denote a strong social link. Indeed, the US-HS network discussed above is one such case, as shown in fig. 1D-F. Therefore, to provide evidence that the metric backbone preserves a correct picture of the organization of contact networks, we compare its community structure with same-size subgraphs obtained by thresholding or randomly removing edges[6].

In the case of the Fr-HS network, it is clear from table 2 that the metric backbone preserves the community structure of the original distance graph as well as that of its metalabel partition better than the same-size threshold and random subgraphs. The measure of bidirectional modularity similarity of the metric backbone ($y_{AB} = 0.74$), with only about 10% of the edges, shows that it preserves the `Louvain`-community structure of the original distance graph much better than the same-size threshold subgraph ($y_{AB} = 0.61$) and better than a sample of 100 same-size random subgraphs ($y_{AB} = 0.37 \pm .02$). The difference can also be visualized as in fig. 2, where it is clear that the metric backbone breaks into fewer modules ($m = 17$, B) than do the threshold ($m = 27$, C) and random ($m = 47.19 \pm 3.79$, D) subgraphs. In the case of the threshold graph this is likely due to deletion of (weak) bridges within and between modules that are nonetheless required to preserve shortest paths [13]. It is also clear that random deletion not only breaks community structure more, it also mixes up the social organization, with students from different specializations being grouped together in small modules more than as observed with the metric backbone or threshold subgraph. While there is some variation across different social networks, as detailed in SM, the metric backbone typically preserves the community structure of the original distance graph better than same-size threshold and random backbones for the various similarity measures defined in section 5.2.

A similar behavior occurs with the metalabel partition, which is also better captured by the metric backbone ($y_{AB} = 0.68$) than the same-size threshold subgraph ($y_{AB} = 0.55$) and the sample of 100 same-size random subgraphs ($y_{AB} = 0.38 \pm .02$). However, as mentioned above, it is not clear that metalabel partitions capture the true social organization, especially in several of the networks analyzed in SM where metalabels are not always expected to form social contact modules in the observed contexts (e.g., teachers in classroom settings or in-patients in a hospital). Still, in school settings like that of the Fr-HS network, it is clear that threshold and random subgraphs break the metalabel partition into many more `Louvain`-communities than does the metric backbone, and the random subgraphs mix the metalabel partition much more. Interestingly, with the French Primary School network (Fr-PS) we observe that the `Louvain`-communities of the metric backbone can be closer to the metalabel partition than the `Louvain`-communities of the original distance graph (see section S2.1), which suggests that the metric backbone can in some cases filter out noisy or redundant contact data, as was also observed in the US-HS network (fig. 1D-F).

3.2 Epidemic spreading on the metric backbone

Networks are the supporting structure of various types of dynamical processes, ranging from the synchronization of oscillators to the spread of information or infectious diseases [2]. How these processes are altered when they occur on a subgraph, with respect to how they would unfold on the whole network, is a natural question when devising a filtering strategy that retains only a subgraph of the original network [51, 20]. We thus explore this issue in the metric backbone.

As many processes are based on diffusion phenomena, we focus on simple and paradigmatic contagion processes typically used to simulate the spread of infectious diseases in a population. Moreover, because our focus is on how a spreading process occurring on a whole network differs from that on the metric backbone (i.e., we are not interested in recovering dynamics) we consider the simplest such model, the *Susceptible-Infected* (SI) model. In this model, each node can be in only one of two states: S (susceptible) or I (infected

[6] We do not compare community structure to other backbone methods, such as the disparity filter backbone, [18] because they can remove nodes (see section 1) which would result in a different universe X to partition into communities. Additionally, distance backbones such as the metric backbone are parameter-free, so we compare its community structure to threshold and random graphs obtained by removing exactly the same number of edges as removed in forming the metric backbone ($\sigma(D).|d_{i>j}|$) in a parameter-free manner. A comparison and discussion of the advantages of distance backbones in regard to alternative backbone subgraphs is already available [13].

and infectious). A susceptible node, x_j, in contact with an infectious one, x_i, along an edge of (proximity) weight p_{ij} can become infectious at rate $\beta.p_{ij}$ (i.e., with probability $\beta.p_{ij}.dt$ in each time interval dt)[7]. Once infected, nodes remain in that state.

Each simulation starts with an infectious seed node chosen at random and ends when no contagion event is possible (either because all nodes are in state I or because the remaining S nodes have no link with the infectious nodes). In order to characterise the spread, we do not focus on the final size (number of nodes reached by the spread, which is always close to the total number of nodes as they do not recover) but rather on the velocity of the spread [52]. More specifically, we compute the time needed by the process to reach either half of the nodes ($t_{1/2}$) or all the nodes (t_1) [52, 53, 54, 55].

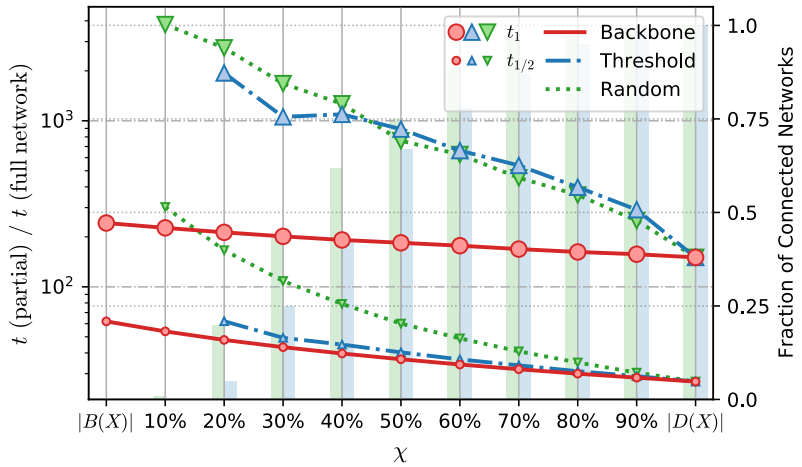

Figure 3: Time to infection on the metric backbone, threshold, and random subgraphs of the French High School (Fr-HS) network. Horizontal axis shows the number of edges in subgraphs analyzed, ranging from the backbone (leftmost, $|B(X)|$) to the original network (rightmost, $|D(X)|$), as a proportion χ of edges not in the respective subgraph. Left vertical axis shows the average time (number of time steps per node in the SI model evolution) to the desired level of infection. t_1 denotes the time for all nodes in the network to be infected; $t_{1/2}$ denotes the time for half of the population to be infected. The green and blue bars, quantified against the right vertical axis, denote the fraction of networks in threshold and random baseline ensembles that are connected for a given χ. Non-connected networks are discarded to compute the spreading times. For the simulations shown, the spreading parameter was set as $\beta = 0.9/p_{max}$ where p_{max} is the largest weight of the original proximity network $P(X)$.

In addition to comparing spreading times on the original network with those on the metric backbone we also, as in the section 3.1 analysis, consider two baselines: subgraphs obtained by thresholding the weights and by randomly removing edges[8]. For a fair comparison, we need to contrast the various subgraphs using the same number of edges. However, in many cases the threshold and random subgraphs are composed of several disconnected components though they have the same number of edges as the metric backbone because only the metric backbone guarantees that connectivity and shortest path distribution are preserved [13]. Therefore, we have devised the following procedure: Starting from the whole network, remove a fraction of the semi-metric edges edges randomly to yield a subgraph composed of the metric backbone plus a percentage χ of semi-metric edges (similar to the procedure in Presigny et al. [51]).

For this study, we have performed $n_r = 10$ runs of the SI model starting from a random seed and computed $t_{1/2}$ and t_1 averaged over these runs and over 100 realizations of the set of randomly selected semi-metric edges that are not removed. We then built threshold and random subgraphs with exactly the same number of edges as the metric backbone. We performed exactly the same procedure in each case to obtain a series of intermediate subgraphs between the original graph and the final threshold or random subgraphs of the same size as the metric backbone. We randomly removed a fraction $1-\chi$ of edges that did

[7]The infection probability uses the proximity graph, given by eq. (1) in section 2, rather than the isomorphic distance graph $D(X)$ used to define the metric backbone. This probability can naturally be defined in terms of distance weights as $\beta.(1/d_{ij} - 1).dt$

[8]As already noted in section 3.1, we do not compare to other backbone methods because they can remove nodes in addition to edges or introduce parameters that would complicate the transmission study pursued. A comparison and discussion of the advantages of distance backbones in regard to alternative backbone subgraphs is available [13].

not belong to the threshold or to the random subgraph.

Figure 3 depicts $t_{1/2}$ and t_1 as a function of χ for the various cases: The rightmost point corresponds to the whole network, $D(X)$, and the leftmost to the metric backbone, $B(X)$, shown in red; corresponding same-size threshold and random subgraphs shown in blue and green, respectively. In all cases, as edges are removed from the original network, the time to infection or half-infection increases as fewer paths are available to transmit the disease. However, it is quite clear that the infection on the metric backbone propagates faster and remains much closer to the infection times observed for the full original network than do the infection times observed on the baseline subgraphs. This is especially striking for the time to full infection τ_1, but also observed for time to half infection $\tau_{1/2}$.

Moreover, as χ was decreased we frequently obtained disconnected threshold and random subgraphs, for which t_1 is by definition infinite as not all nodes could be reached, and $t_{1/2}$ is finite only if the seed node existed in a connected component that contained at least half of the nodes. The fraction of connected networks for both baselines is indicated by blue and green bars in fig. 3. It is clear that this fraction becomes rather small well before we reach the size of the metric backbone ($\chi = 0$). Indeed, by $\chi = 20\%$, only 5% and 20% of connected networks exist in the threshold and random baseline ensembles, respectively. Naturally, the values of t_1 and $t_{1/2}$ shown are only computed for connected networks, thus simulations for subgraphs of the same size as the metric backbone ($|B(X)|$) were not possible for both threshold and random baseline ensembles; in the case of the threshold baseline, even subgraphs with $|B(X)| + \chi = 10\%$ were not connected. This means that we cannot find threshold and random subgraphs of the same size as the metric backbone that can sustain propagation to the whole (or even half) of the original network. In contrast, time to full (or half) infection is always finite in the case of the backbone—and any intermediate case between it and the full network as χ varies—as it preserves the connectivity (and shortest-path distribution) of the original network.

Altogether the SI simulations show that the metric backbone is a primary subgraph for simple transmission dynamics. Furthermore, considering that the threshold subgraph is built by removing the weaker edges (low proximity or large distance), it follows that there are weak links in the metric backbone that are removed from the network in the threshold subgraph but which are essential for transmission. These weak edges are potentially good candidates for interventions aiming at reducing transmission while minimizing reduction of social contact.

4 Discussion

It is only in weighted graphs—where weights discriminate and characterize the degree of association between nodes as proximity or its isomorphic distance—that the concept of metric backbone is meaningful because the backbone of non-weighted (regular) graphs is the graph itself. In weighted graphs the metric backbone (and its generalized distance backbone [13]) is a very useful construct because, unlike other graph reduction techniques, the backbone graph $B(X)$ is guaranteed to preserve all bridges, connectivity, and distribution of shortest paths of the original graph $D(X)$, and is defined in a non-parametric manner from simple algebraic principles. Thus, the metric backbone of a distance graph is unique, typically small, and provides a principled graph reduction technique that preserves shortest paths.

We have analyzed the metric backbones of nine contact networks collected in a variety of social settings and recorded based on the interactions between pairs of individuals via wearable sensors. It was already known that the metric backbone, by design, preserves network connectivity and all shortest paths after removal of all redundant edges for that purpose [13]. Here we have shown that the metric backbones of social contact networks are also much smaller than their associated original networks (large redundancy in shortest-path calculation), as previously observed in biological, technological, and knowledge networks [13]. Built-in redundancy is a hallmark of complex systems, allowing them natural protection against perturbations [56, 57]. In the social networks analyzed here, the small backbones reveal this in that their distribution of shortest paths is very robust to random removal of edges. Specifically, the proportion of semi-metric edges (σ) ranges from 52 to 94%, with eight of the nine networks observing $\sigma > 80\%$ redundancy. This means that all shortest paths can be computed in these eight networks with fewer than $\tau = 20\%$ of the edges. Indeed, five networks have metric backbones composed of $\tau \leq 10\%$ of their original edges (see table 1).

The only network that has a fairly large metric backbone ($\sigma = 48 \pm 9\%$) pertains to a context where we do not expect a typical social organization: an exhibit in the Dublin Science Gallery (Ir-Ex). In a museum setting, most visitors pass by alone or in small disconnected groups, thus a large fraction of (weak) edges are required to preserve shortest-path connectivity. Indeed, as shown in fig. 4, the distribution of semi-metric distortion reveals a social organization of almost random encounters. The values of semi-metric distortion are quite low ($\tilde{s}_{ij} \approx 2$) with a small variation range (see also fig. S18). This means that the distance graph is very close to being entirely metric, as previously observed in random fluctuation phenomena that occurs in an physical (Euclidean) space [13]. In such cases, the backbone is expected to comprise 50% of the graph, but the (near 50%) semi-metric edges are almost all very close to metric (low distortion).

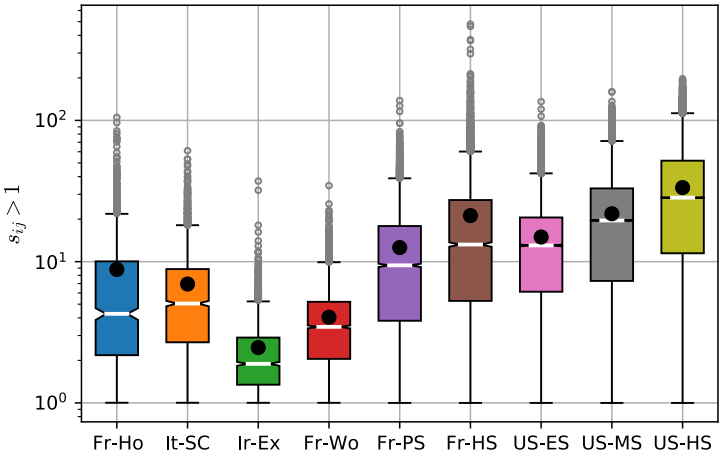

Figure 4: Distribution of semi-metric edge distortion for contact networks. Box plots of the distributions of semi-metric distortion for semi-metric edges ($s_{ij} > 1$) for all contact networks in table 1. White bars and black circles denote the median and mean values, respectively. Actual distributions are shown in SM.

Because shortest paths reveal the strength of social connectedness and the likelihood of transmission phenomena on these networks, we have focused on analyzing how both are captured by the metric backbone. Our results show that in comparison to same-size threshold and random subgraphs, the metric backbone best preserves community structure (section 3.1) and is a primary subgraph for simple transmission dynamics (section 3.2). Additionally, the metric backbone clearly helps in the identification and visualization of the social community structure of the network, especially when the original graph is too dense to be visually inspected (e.g., fig. 1D-F).

Thresholding is often used to decrease the number of edges in densely connected networks. The common assumption is that edges with small proximity (large distance) weights are spurious or simply noise. However, as we show here in all networks studied, the set of weak edges includes those that do in fact contribute to the computation of shortest paths and preserve connectivity and are, thus, important to keep communities connected and sustain spreading phenomena. Thresholding a network can isolate communities, creating disconnected components where dynamical processes, such as epidemic spread, cannot be sustained. Indeed, our epidemic spread simulations on the nine social networks (section 3.2) show that threshold or random subgraphs much larger than the metric backbone frequently lead to disconnected graphs, but the metric backbone preserves connectivity and shortest paths. The importance of weak edges has long been explored in social networks [17]. Our analysis, thus, suggests that weak edges on the metric backbone are potentially good candidates for interventions that aim at reducing transmission while minimizing reduction of social contact in communities.

It is worth also considering the role of semi-metric edges, those not on the backbone (section 3.2), as their properties show that the distance backbone methodology is not exclusively a network reduction technique. While all (semi-metric) edges with $s_{ij} > 1$ have null betweeness centrality, the distribution of semi-metric edge distortion (eq. (7)) typically varies widely, often spanning many scales [13]. Indeed, the distribution of s_{ij} is informative about the robustness of shortest-paths on a given network. If there are many semi-metric edges with low values of s, then the removal of metric edges from the backbone is likely to have a small

effect on the distribution of shortest-paths because alternative paths with low distortion are likely to exist. In contrast, if there are high values of s, removing some edges on the backbone is likely to result in a large disruption of the shortest-path distribution. Thus, while the size of the backbone captures how robust the distribution of shortest paths is to a random attack on the whole network, the distribution of s captures how robust the backbone itself is to attack.

This knowledge is relevant when devising intervention strategies to disrupt spreading phenomena on social networks. If the backbone is robust, with low central tendency of s and a more homogeneous distribution, then removing edges from it (hindering specific social connections) will have little impact on shortest paths; this is the case of the Exhibit (Ir-Ex) and Workspace (Fr-Wo) networks as shown in fig. 4 (see also figs. S15 and S18). If, on the other hand, the distribution of s is heterogeneous with a large central tendency and variation, then targeting specific edges on the backbone would result in a large disruption to shortest paths. The latter is the case of all school contact networks we have analyzed, especially the French (Fr-HS) and US (US-HS) high school networks as shown in fig. 4 (see also SM for s distributions of all school networks). For instance, the median semi-metric distortion of the US-HS network is $\tilde{s}_{ij} \approx 30$. This means that 50% of the (semi-metric) edges not on the backbone break the triangle inequality by a factor of at least 30. In other words, they refer to social distances that are at least 30 times shorter via indirect paths on the backbone. Thus, deleting edges from this backbone is likely to result in a very large impact on the distribution of shortest paths. Because the backbone is typically very small (for the US-HS $\tau(D) = 7.84\%$), such intervention strategies are easier to design on it than on the whole network, which reveals the importance of targeting the backbone in containing epidemic spread.

Here we have studied social contact networks statically and left considering attack interventions to the network or backbone for future work. We also intend to include other distance backbones in upcoming studies. Indeed, the metric backbone generalizes to any measure of path length [13], not just the summation of distance edges as in the standard shortest paths of the metric backbone. Thus, other distance backbones based on ultra-metric, euclidean, and diffusion distances are likely to be relevant for both understanding the organization of and transmission phenomena on social networks. Taken together, our results indicate that the parameter-free metric backbone in particular and distance backbones in general can be widely applicable to network science problems, especially when shortest paths and distance or proximity measures are relevant in explaining properties of the system.

5 Methods

5.1 Datasets and network computation

The datasets range in date from 2009 to 2015 and were collected from a number of different social environments, such an elementary school [34], a primary school [45], two high schools [46, 33], a hospital ward [42], a workplace [44], a scientific conference [43], and a museum exhibit [43]. The data were collected using proximity sensors worn by individuals. Most of the resulting datasets provide a contact record data file, where each row consists of the anonymized IDs of individuals and the time that they met someone, with a time resolution of approximately 20 seconds. Some datasets also contain additional metadata. For instance, in the primary school, class and type (i.e., student or teacher) are provided for each contact ID. Details for each dataset and their analysis are provided in SM, section S2.

5.2 Module similarity between Backbone and Original Network

Several measures are available to quantify how much the metric backbone is able to preserve the social organization captured by the original network and to compare the value with the social organization captured by null model subgraphs obtained by removing the same number of edges. Given two distance graphs $A(X)$ and $B(X)$ defined on the same set of nodes X from contact or proximity data as described in section 2,

we consider the partitions of X to be the community structure of each graph. Specifically, graph $A(X)$ is partitioned into a set of m_A nonempty modules A_i, and graph $B(X)$ is partitioned into a set of m_B nonempty modules B_j[9]

The module partitions we consider are either the metalabels of the original networks or the communities found by the `Louvain` algorithm [50] of a target distance graph. The goal is then to compare how similar the module partition of a distance graph is to the partition of its backbone—or threshold or random subgraph of the same size.

We consider various measures for computing similarity between modularity partitions of the same set of nodes, which is an old problem in classification [58]. A straightforward way to obtain a value that captures *bidirectional similarity* of modularity between two distance graphs is to use a normalized measure based on the Jaccard proximity [38] between all pairs of modules with one from each distance graph:

$$y_{AB} = \frac{\sum_{i}^{m_A}\sum_{j}^{m_B} P(A_i, B_j)}{\sqrt{m_A \cdot m_B}} \quad \text{where} \quad P(A_i, B_j) = \frac{|A_i \cap B_j|}{|A_i \cup B_j|} \,. \tag{8}$$

This measures how similar partitions $\{A_1 \cdots A_{m_A}\}$ and $\{B_1 \cdots B_{m_B}\}$ of X are to each other and varies in the unit interval ($y_{AB} \in [0, 1]$). The larger the value, the more similar the partitions.

We also consider a *directional similarity* based on the Jaccard proximity to compare each module in a partition to its most similar counterpart in the other partition ($J_{A\to B}$) and vice-versa ($J_{B\to A}$) as follows:

$$J_{A\to B} = \frac{\sum_{i}^{m_A} \max_{j}^{m_B} P(A_i, B_j)}{m_A} \quad , \quad J_{B\to A} = \frac{\sum_{j}^{m_B} \max_{i}^{m_A} P(A_i, B_j)}{m_B} \,. \tag{9}$$

These dual measures also vary in the unit interval ($J_{A\to B}, J_{B\to A} \in [0, 1]$). The larger the value of $J_{A\to B}$ the more each module in partition $\{A_1 \cdots A_{m_A}\}$ has a similar counterpart module in partition $\{B_1 \cdots B_{m_B}\}$, and vice versa for $J_{B\to A}$.

A measure of *modularity dispersion*, intuitively, yields a comparison opposite to those provided in eqs. (8) and (9):

$$h_{A\to B} = \frac{\sum_{k}^{m_A} H(\mathcal{A}_i)}{m_A \cdot \log_2(m_B)} \qquad h_{B\to A} = \frac{\sum_{l}^{m_B} H(\mathcal{B}_j)}{m_B \cdot \log_2(m_A)} \quad . \tag{10}$$

$H(\mathcal{A}_i)$ and $H(\mathcal{B}_j)$ denote the Shannon entropy of probability distributions associated with the dispersion of each module of one partition into all modules of the other partition:

$$\mathcal{A}_i = \big\{S(A_i, B_1), \ldots, S(A_i, B_{m_B})\big\} \quad \text{and} \quad \mathcal{B}_j = \big\{S(B_j, A_1), \ldots, S(B_j, A_{m_A})\big\} \quad .$$

$S(A_i, B_j)$ denotes the subsethood [59, 60] of module A_i in module B_j, and $S(B_j, A_i)$ denotes the subsethood of module B_j in module A_i:

$$S(A_i, B_j) = \frac{|A_i \cap B_j|}{|A_i|} \quad \text{and} \quad S(B_j, A_i) = \frac{|A_i \cap B_j|}{|B_j|} \quad .$$

Typically $S(A_i, B_j) \neq S(B_j, A_i)$. Thus, $h_{A\to B} \in [0, 1]$ captures how much the modules of partition $\{A_1 \cdots A_{m_A}\}$ are on average dispersed into the modules of partition $\{B_1 \cdots B_{m_B}\}$, and $h_{B\to A} \in [0, 1]$ the other way around. The smaller the values are, the more the community structure of one network is preserved in the other.

For comparison, in addition to the measures in formulae 8 to 10, we also compute the *CluSim* [61, 62] measure of modularity similarity, which considers module "nestedness," and the *Adjusted Rand Index* [58]. All computed measures are shown in the respective dataset details in SM section S2.

[9] We consider strict partitions such that $\cup_{i=1}^{m_A} A_i \equiv X$, $A_k \cap A_l \equiv \emptyset \quad \forall k \neq l$, and $A_i \neq \emptyset \quad \forall i \in \{1 \ldots m_A\}$. Naturally, $i = 1, \ldots, m_A$ and $j = 1, \ldots, m_B$, with $m_A, m_B \in \mathbb{N}$.

Data Availability Statement. Computations presented in the analysis were performed in Python. Source code is made available on Github at https://github.com/rionbr/socialbackbone. The `distanceclosure` Python package used to compute the metric backbones can be found at https://github.com/rionbr/distanceclosure. The SocioPatterns datasets analyzed can be found in http://www.sociopatterns.org/datasets. The Toth et al. [34] and Salathé et al. [33] datasets can be found in their respective publications.

Author Contributions. All authors designed and conducted the research, and contributed to the final manuscript. RBC developed the Python code, performed all backbone computations and visualizations, and drafted the manuscript. AB designed and performed the epidemic spread simulations. LMR developed the metric backbone methodology and the new measures of community similarity employed. He also took a leading role in writing and editing the final manuscript.

Acknowledgments. We are thankful to Deborah Rocha for very thorough line editing. This work was partially funded by National Institutes of Health, National Library of Medicine Program, grant 01LM011945-01 (LMR and RBC), a Fulbright Commission fellowship (LMR), the NSF-NRT grant 1735095 "Interdisciplinary Training in Complex Networks and Systems" (LMR), a CAPES Foundation fellowship, grant 18668127 (RBC), the Fundação para a Ciência e a Tecnologia, grant PTDC-MEC-AND-30221-2017 (RBC), and the Agence Nationale de la Recherche (ANR) project DATAREDUX ANR-19-CE46-0008 (AB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Pastor-Satorras R, Castellano C, Van Mieghem P, and Vespignani A. Epidemic processes in complex networks. Reviews of modern physics. 2015 Aug; 87:925. DOI: 10.1103/RevModPhys.87.925
2. Barrat A, Barthelemy M, and Vespignani A. Dynamical processes on complex networks. Cambridge University Press, 2008
3. Vespignani A. Twenty years of network science. Nature. 2018; 558:528–9. DOI: 10.1038/d41586-018-05444-y
4. Newman ME. Complex systems: A survey. American Journal of Physics. 2011; 79:800–10. DOI: 10.1119/1.3590372
5. Correia RB, Wood IB, Bollen J, and Rocha LM. Mining social media data for biomedical signals and health-related behavior. Annual Review of Biomedical Data Science. 2020; 3(1):433–58. DOI: 10.1146/annurev-biodatasci-030320-040844
6. Pescosolido BA et al. The social symbiome framework: Linking genes-to-global cultures in public health using network science. Handbook of Applied System Science. 2016 :25–48. DOI: 10.4324/9781315748771
7. Colizza V, Barrat A, Barthelemy M, Valleron AJ, and Vespignani A. Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. PLOS Medicine. 2007; 4(1):e13. DOI: 10.1371/journal.pmed.0040013
8. Colizza V and Vespignani A. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. J. Theor. Biol. 2008; 251:450–67
9. Gemmetto V, Barrat A, and Cattuto C. Mitigation of infectious disease at school: targeted class closure vs school closure. BMC infectious diseases. 2014; 14:695
10. Conrad M. The geometry of evolution. BioSystems. 1990; 24:61–81
11. Marques-Pita M and Rocha LM. Canalization and control in automata networks: body segmentation in Drosophila melanogaster. PloS one. 2013; 8:e55946
12. Gates AJ and Rocha LM. Control of complex networks requires both structure and dynamics. Scientific reports. 2016; 6:24456. DOI: doi.org/10.1038/srep24456
13. Simas T, Correia RB, and Rocha LM. The distance backbone of complex networks. Journal of Complex Networks. 2021; 9(6):cnab021. DOI: 10.1093/comnet/cnab021

14. Gates AJ, Correia RB, Wang X, and Rocha LM. The effective graph reveals redundancy, canalization, and control pathways in biochemical regulation and signaling. Proceedings of the National Academy of Sciences. 2021; 118:e2022598118. DOI: 10.1073/pnas.2022598118

15. Radicchi F, Ramasco JJ, and Fortunato S. Information filtering in complex weighted networks. Phys. Rev. E. 2011 Apr; 83(4):046101. DOI: 10.1103/PhysRevE.83.046101

16. Bellingeri M, Bevacqua D, Scotognella F, Alfieri R, and Cassi D. A comparative analysis of link removal strategies in real complex weighted networks. Scientific Reports. 2020; 10:3911. DOI: 10.1038/s41598-020-60298-7

17. Granovetter MS. The Strength of Weak Ties. Social Networks. 1977 :347–67. DOI: 10.1016/B978-0-12-442450-0.50025-0

18. Serrano MÁ, Boguñá M, and Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proceedings of the National Academy of Sciences. 2009; 106:6483–8. DOI: 10.1073/pnas.0808904106

19. Zhang RJ, Stanley HE, and Ye FY. Extracting h-Backbone as a Core Structure in Weighted Networks. Scientific Reports. 2018; 8:14356. DOI: 10.1038/s41598-018-32430-1

20. Mercier AM, Scarpino SV, and Moore C. Effective Resistance for Pandemics: Mobility Network Sparsification for High-Fidelity Epidemic Simulation. 2021. arXiv: 2111.02449v1

21. Tumminello M, Micciché S, Lillo F, Piilo J, and Mantegna RN. Statistically Validated Networks in Bipartite Complex Systems. PLOS ONE. 2011 Mar; 6:0017994. DOI: 10.1371/journal.pone.0017994

22. Zhan XX, Hanjalic A, and Wang H. Information diffusion backbones in temporal networks. Scientific Reports. 2019; 9:6798. DOI: 10.1038/s41598-019-43029-5

23. Marcaccioli R and Livan G. A Pólya urn approach to information filtering in complex networks. Nature Communications. 2019; 10:745. DOI: 10.1038/s41467-019-08667-3

24. Kobayashi T, Takaguchi T, and Barrat A. The structured backbone of temporal social ties. Nature Communications. 2019; 10:220. DOI: 10.1038/s41467-018-08160-3

25. Rocha LM. Soft Computing Agents: A New Perspective for Dynamic Information Systems. Ed. by Loia V. IOS Press, 2002. Chap. Semi-metric Behavior in Document Networks and its Application to Recommendation Systems:137–63

26. Rocha LM, Simas T, Rechtsteiner A, Giacomo MD, and Luce R. MyLibrary@LANL: Proximity and Semi-metric Networks for a Collaborative and Recommender Web Service. *2005 IEEE/WIC/ACM International Conference on Web Intelligente (WI'05)*. IEEE Press. IEEE Press, 2005 :565–71

27. Simas T and Rocha LM. Distance closures on complex networks. Network Science. 2015 Jun; 3(2):227–68. DOI: 10.1017/nws.2015.11

28. Simas T and Suckling J. Commentary: Semi-Metric Topology of the Human Connectome: Sensitivity and Specificity to Autism and Major Depressive Disorder. Frontiers in Neuroscience. 2016 Aug; 10:353. DOI: 10.3389/fnins.2016.00353

29. Peeters S, Simas T, Suckling J, Gronenschild E, Patel A, Habets P, Os J van, and Marcelis M. Semi-metric analysis of the functional brain network: Relationship with familial risk for psychotic disorder. NeuroImage: Clinical. 2015; 9:607–16. DOI: 10.1016/j.nicl.2015.10.003

30. Kalavri V, Simas T, and Logothetis D. The Shortest Path is Not Always a Straight Line: Leveraging Semi-metricity in Graph Analysis. Proceedings of the VLDB Endowment. 2016 May; 9:672–83. DOI: 10.14778/2947618.2947623

31. Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F, and Flammini A. Computational Fact Checking from Knowledge Networks. PLoS ONE. 2015; 10:e0128193. DOI: 10.1371/journal.pone.0128193

32. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, and Vespignani A. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. PLoS ONE. 2010 Jul; 5:e11596. DOI: 10.1371/journal.pone.0011596

33. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, and Jones JH. A high-resolution human contact network for infectious disease transmission. Proceedings of the National Academy of Sciences. 2010; 107:22020–5. DOI: 10.1073/pnas.1009094108

34. Toth DJA, Leecaster M, Pettey WBP, Gundlapalli AV, Gao H, Rainey JJ, Uzicanin A, and Samore MH. The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. Journal of The Royal Society Interface. 2015; 12. DOI: 10.1098/rsif.2015.0279

35. Stehlé J, Charbonnier F, Picard T, Cattuto C, and Barrat A. Gender homophily from spatial behavior in a primary school: A sociometric study. Social Networks. 2013; 35:604–13. DOI: 10.1016/j.socnet.2013.08.003

36. Stehlé J et al. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. BMC Medicine. 2011; 9:87. DOI: 10.1186/1741-7015-9-87

37. SocioPatterns: data-driven social dynamics and human activity. http://www.sociopatterns.org. Accessed on Sept 10. 2018

38. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bulletin de la Société Vaudoise des Sciences Naturelles. 1901; 37:241–72

39. Grefenstette G. Explorations in Automatic Thesaurus Discovery. Kluwer Academic, 1994

40. Zwick U. All Pairs Shortest Paths using Bridging Sets Retangular Matrix Multiplication. Journal of the ACM. 2002 May; 49:289–317

41. Dijkstra E. A note on two problems in connexion with graphs. Numerische mathematik. 1959; 1:269–71. DOI: 10.1007/BF01386390

42. Vanhems P, Barrat A, Cattuto C, Pinton JF, Khanafer N, Régis C, Kim Ba, Comte B, and Voirin N. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. PLOS ONE. 2013 Sep; 8:e73970. DOI: 10.1371/journal.pone.0073970

43. Isella L, Stehlé J, Barrat A, Cattuto C, Pinton JF, and Van den Broeck W. What's in a crowd? Analysis of face-to-face behavioral networks. Journal of Theoretical Biology. 2011 Feb; 271:166–80. DOI: 10.1016/j.jtbi.2010.11.033

44. Génois M and Barrat A. Can co-location be used as a proxy for face-to-face contacts? EPJ Data Science. 2018; 7:11. DOI: 10.1140/epjds/s13688-018-0140-1

45. Stehlé J et al. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. PLoS ONE. 2011 Aug; 6. Ed. by Viboud C:e23176. DOI: 10.1371/journal.pone.0023176

46. Mastrandrea R, Fournet J, and Barrat A. Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys. PLoS ONE. 2015 Sep; 10:0136497. DOI: 10.1371/journal.pone.0136497

47. Jacomy M, Venturini T, Heymann S, and Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLOS ONE. 2014 Jun; 9:1–12. DOI: 10.1371/journal.pone.0098679

48. Bastian M, Heymann S, and Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. 2009. Available from: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

49. Fortunato S. Community detection in graphs. Physics Reports. 2010; 486:75–174. DOI: 10.1016/j.physrep.2009.11.002

50. Blondel VD, Guillaume JL, Lambiotte R, and Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008 Oct; 10. DOI: 10.1088/1742-5468/2008/10/P10008

51. Presigny C, Holme P, and Barrat A. Building surrogate temporal network data from observed backbones. Phys. Rev. E. 2021 May; 103(5):052304. DOI: 10.1103/PhysRevE.103.052304

52. Barthélemy M, Barrat A, Pastor-Satorras R, and Vespignani A. Velocity and Hierarchical Spread of Epidemic Outbreaks in Scale-Free Networks. Physical Review Letters. 2004 Apr; 92(17):178701. DOI: 10.1103/PhysRevLett.92.178701

53. Karsai M, Kivelä M, Pan RK, Kaski K, Kertész J, Barabási AL, and Saramäki J. Small But Slow World: How Network Topology and Burstiness Slow Down Spreading. Physical Review E. 2011; 83:025102. DOI: 10.1103/physreve.83.025102

54. Starnini M, Machens A, Cattuto C, Barrat A, and Pastor-Satorras R. Immunization strategies for epidemic processes in time-varying contact networks. Journal of Theoretical Biology. 2013; 337:89–100. DOI: 10.1016/j.jtbi.2013.07.004

55. Delvenne J, Lambiotte R, and Rocha LEC. Diffusion on Networked Systems Is a Question of Time Or Structure. Nature Communications. 2015; 6:7366. DOI: 10.1038/ncomms8366

56. Tononi G, Sporns O, and Edelman GM. Measures of degeneracy and redundancy in biological networks. Proceedings of the National Academy of Sciences. 1999 Mar; 96:3257–62. DOI: 10.1073/pnas.96.6.3257

57. Ulitsky I and Shamira R. Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. Mol Syst Biol. 2007; 3:104. DOI: 10.1038/msb4100144

58. Hubert L and Arabie P. Comparing partitions. Journal of classification. 1985; 2:193–218. DOI: 10.1007/BF01908075

59. Klir G and Yuan B. Fuzzy sets and fuzzy logic. Vol. 4. Prentice hall New Jersey, 1995

60. Kosko B. Neural Networks and Fuzzy Systems. Englewood Cliffs, NJ, Prentice Hall. 1991

61. Gates AJ, Wood IB, Hetrick WP, and Ahn YY. Element-centric clustering comparison unifies overlaps and hierarchy. Scientific Reports. 2019 Jun; 9:8574. DOI: 10.1038/s41598-019-44892-y

62. Gates AJ and Ahn YY. CluSim: a python package for calculating clustering similarity. The Journal of Open Sourrce Software. 2019 Jan; 4:1264. DOI: 10.21105/joss.01264